On the Computational Prediction of miRNA Promoters

Charalampos Michail, Aigli Korfiati, Konstantinos Theofilatos, Spiros Likothanassis, Seferina Mavroudi

On the computational prediction of miRNA promoters

Charalampos Michail[1], Aigli Korfiati[1,2], Konstantinos Theofilatos[2], Spiros Likothanassis[1], Seferina Mavroudi[3]

[1] Department of Computer Engineering and Informatics, University of Patras, Patra, Greece
{cmichail,korfiati,likothan}@ceid.upatras.gr
[2]InSyBio Ltd, London, UK
k.theofilatos@insybio.com
[3] Department of Social Work, Technological Institute of Western Greece, Patra, Greece
mavroudi@teiwest.gr

Abstract. MicroRNAs transcription regulation is an open topic in molecular biology and the identification of the promoters of microRNAs would give us relevant insights on cellular regulatory mechanisms. In the present study, we introduce a new computational methodology for the prediction of microRNA promoters, which is based on the hybrid combination of an adaptive genetic algorithm with a nu-Support Vector Regression (nu-SVR) classifier. This methodology uses genetic algorithms to locate the optimal features set and to optimize the parameters of the nu-SVR classifier. The main advantage of the proposed solution is that it systematically studies and calculates a vast number of features that can be used for promoters prediction including frequency-based properties, regulatory elements and epigenetic features. The proposed method also handles efficiently the issues of over-fitting, feature selection, convergence and class imbalance. Experimental results give accuracy over 87% in the miRNA promoter prediction.

Keywords: miRNA promoters, classification, computational prediction, feature selection, transcription start sites

1 Introduction

One of the current trends in molecular biology is studying the various types of short and long non-coding RNAs (ncRNAs) [1]. MicroRNAs (miRNAs) are the most thoroughly characterized subclass of short RNAs in the recent literature [2]. miRNAs are short (21-23 nt) and single stranded endogenous RNA molecules. They regulate protein coding genes by binding to the 3' untranslated regions (3' UTRs) of their target mRNAs. This binding event causes translational repression of the target gene or stimulates rapid degradation of the target transcript [3]. miRNAs are involved in diverse biological processes, including development, differentiation, apoptosis, cell proliferation, and disease [3]. A growing number of studies indicate that miRNAs play crucial roles in human disease development, progression, prognosis, diagnosis

and evaluation of treatment response [4] and miRNAs have been linked to cancer, neurodegenerative and cardiovascular diseases.

Many algorithms are able to predict miRNA genes and their targets, but their transcription regulation is still under investigation [5]. It is generally believed that intragenic (intronic, exonic) miRNAs (located in introns or exons of protein coding genes) are co-transcribed with their host genes [6], but literature has indicated that intragenic miRNA genes may be transcribed by their own promoter [7, 8]. Intergenic miRNAs (located between protein coding genes) are independent transcription units, with their own transcriptional regulatory elements [9]. For intergenic miRNAs the distances between transcription initiation sites (TSSs) and miRNA-coding regions dramatically vary, ranging from a few hundred bases to 30-kb upstream and the nature of the primary transcript of intergenic miRNAs and promoter organization are largely unknown.

Transcription initiation is a key step in the regulation of gene expression. During this process, transcription factors bind promoter region of a gene in a sequence-specific manner and recruit the RNA polymerase to form an active initiation complex around the transcription start site (TSS) [10]. The promoter is commonly referred to as the region upstream of a gene that contains the information permitting the proper activation or repression of the gene that it controls [11]. The promoter region is divided into three parts:

- the core-promoter is 100 bp long, surrounds the TSS and contains binding sites for RNA polymerase II (Pol II) and general transcription factors;
- the proximal promoter is several hundred base pairs long upstream the core promoter and contains several regulatory elements;
- the distal promoter is up to thousands of base pairs long upstream of the TSS and contains additional regulatory elements called enhancers and silencers.

As it contains primary information to control gene transcription, it is a fundamental step to identify the core-promoter in study of gene expression patterns and constructing gene transcription networks.

In the present study, a new computational methodology for the prediction of miRNA promoters is introduced and it is based on the hybrid combination of an adaptive genetic algorithm with a nu-SVR classifier. This methodology uses genetic algorithms to locate the optimal features set and optimize the parameters of the nu-SVR classifier. The main advantage of the proposed solution is that it systematically studies the different features that can be used for miRNA promoters prediction. The simple method and script provided can be used to calculate effectively most of the features that have been correlated with promoter attributes without the need for combining different tools. In terms of classification performance, the main advantages of the proposed method are that it handles efficiently the issues of over-fitting, feature selection, convergence and class imbalance. Experimental results give accuracy over 87% in the miRNA promoter prediction.

The rest of the article is organized as follows: section 2 describes the existing methods for the prediction of protein coding gene and miRNA promoters, section 3

analyzes the proposed methodology and the relevant datasets and features which were used, section 4 presents the experimental results and section 5 concludes the paper and discusses interesting future research directions.

2 Promoter prediction methods

The promoter of a gene is a significant region for its transcription initiation and thus, identifying miRNA promoters would give us insights on their regulatory mechanism. A common practice in miRNA promoters identification is to first apply a promoter prediction method to predict their promoters, and then to verify the predictions by wet lab experiments. Developing the promoter identification algorithm is a very challenging problem [12]. A number of computational methods for predicting promoters of protein-coding genes have been developed, however their performances are far from satisfactory, because our understanding of the transcription process is incomplete.

Literature indicates that there are features of the promoter regions that differentiate them from other parts in the genome. These features include TATA-box, GC-box, CAAT-box, and Inr [10]. Others features include the CpG islands close to the TSS, binding sites of typical transcription factors, chromatin modifications and statistical properties of the core and proximal promoter. The similarities between orthologous promoters and information from mRNA transcripts have also been used to identify promoters [11]. Some well-known promoter prediction programs are CoreBoost_HM [10], McPromoter [13] and EP3 [11].

Concerning miRNA promoter prediction, initial approaches trained classifiers on protein coding genes promoters and applied them to identify miRNA promoters [9], [12], [14,15]. These techniques provided the first indications of miRNA transcription start site (TSS) positions on a genome-wide scale. However, they were not built based on the promoters of microRNA genes and they exhibit high false-positive rates. Additionally, although miRNA promoters present several similarities with RNA Pol II promoters, this is mainly true for intergenic miRNAs, as too little is still known about intragenic miRNA promoters. For these reasons, a supervised method trained on protein-coding genes is not the optimal choice for identifying miRNA promoters [16].

Other studies for miRNA promoter prediction are based on experimental data, such as cap analysis of gene expression (CAGE) data, RNA Pol II data or histone modification data. CAGE tags were used to identify miRNA TSSs by considering its possibility to capture the 5' cap. MiRStart [15], PROmiRNA [16] and the method of Saini et al. [9] are representative examples. However, there exist uncapped pre-miRNAs that can't be captured by CAGE technology. The methods of Wang et al. [10], Zhou et al. [12] and Corcoran et al. [7] are based on RNA Pol II data. However, these studies were limited to small amount of miRNAs due to the insufficiency of Pol II data. Chromatin signature based methods use histone mark profiles, such as H3K4me3 [17,18] or nucleosome positioning patterns [8] in specific cell lines to annotate miRNA promoters de novo. A recent method [19] combined data of H3K4me3 and DNase I hypersensitive sites (DHSs) with conservation and sequence features to iden-

tify cell-specific TSSs. Although histone mark-based methods have good results, they have been designed for specific cell lines. Additionally, due to the nature of ChIP-seq experiments, chromatin-based methods represent a valuable strategy for detecting intergenic and host gene miRNA promoters, but they might lack the sensitivity required to identify intronic promoters [16].

3 Materials and Methods

3.1 Datasets

Exclusively experimentally validated miRNA TSSs have been used in order to construct a positive dataset of high quality. For each TSS, the [-1000, 1000] bp region around it has been used as the promoter region. Even though a rather smaller region, [-250, 50] or [-450, 50] bp around the TSS, is usually used in other methods, we have included 1000 bases downstream the TSS, since in [20] it is suggested that downstream elements also regulate transcription. The experimentally validated TSSs have been downloaded from the miRT [21] database. From the total of 670 TSSs, we have selected only the 306 that are related to the gene assembly hg19. With these TSSs we have then queried the UCSC DAS server [22] in order to extract the promoter sequences.

For the construction of the negative dataset, i.e. a set of sequences that do not contain miRNA promoters, a pool of 1224 sequences was formed: the four (two upstream and two downstream) non-overlapping consecutive segments immediately upstream and downstream of the positive dataset, as suggested in [23].

Since we wanted to preserve a 1:1 ratio between the positive and the negative dataset, 306 sequences from the negative pool of 1224 are selected at random in every execution of the proposed method. This rate has been maintained because an imbalanced distribution could affect the performance of the classifier.

3.2 Features

Representative features are essential in order to train the proposed model for efficiently distinguishing miRNA promoters from a negative data set. Features from several broad categories are presented in the literature [23,24,25] and since the proposed method is able to handle large numbers of features and to extract the optimal subset of them, we have used them all as inputs. They include (i) frequency-based properties of the promoters such as k-mers, word commonality, skew, palindromes; (ii) regulatory elements such as CpG islands, repetitive elements; (iii) epigenetic features such as chromatin states. The features employed are summarized in Table 1. For the features calculation we have implemented a Python script, which is freely available at: https://github.com/bioinfoceid/miRNAPromoters. Some features are calculated over the whole examined sequences, while others on sliding windows. When sliding windows are employed, their size is 100 bp and their step is 50 bp, thus resulting in 39 windows for an examined sequence of 2000 bp.

Table 1. Number of features per category

Feature category	*Number of features*
K-mers	2*(4+16+64)=168
Observed/expected ratio di- and tri-nucleotides	16+64=80
Word commonality	39
AT- and CG-skews	2*39=78
Palindromes	39
CpG islands	2*39=78
Repetitive elements	39
Chromatin states	7*15=105
Total	626

Concerning the k-mers, we have calculated the frequencies of mono-, di-, tri- consecutive nucleotides (4, 16, 64 frequencies, respectively) in the upstream [-1000,-1] and downstream [+1,+1000] regions relative to TSSs separately. This generates 2*(4+16+64) = 168 features.

The observed/expected ratio of di-nucleotides is

$$\frac{Obs}{Exp} b_i b_j = \frac{\#(b_i b_j)}{\#b_i * \#b_j} * N$$

and of tri-nucleotides is

$$\frac{Obs}{Exp} b_i b_j b_k = \frac{\#(b_i b_j b_k)}{\#b_i * \#b_j * \#b_k} * N^2$$

where b_i, b_j, b_k are the nucleotides A,C,G,T and $\#b_i, \#b_j, \#b_k, \#(b_i b_j), \#(b_i b_j b_k)$, are numbers of mono-, di- and tri-nucleotides and N is the total number of nucleotides in the examined sequence. The calculation of these ratios yields 16 features for the di-nucleotides and 64 for the tri-nucleotides.

For the word commonality feature category we have downloaded 1000 random sequences of 1000 bp from the gene assembly hg19 and then counted the frequency of all possible hexamers. The frequency of each hexamer has been normalized so that the least common hexamer has the score 0 and the most common one has the score 1. Finally, we have calculated a score in each sliding window of the examined sequence by adding the score of all hexamers occurring in that window, resulting in 39 features.

The AT-skew:

$$ATskew = \frac{\#A - \#T}{\#A + \#T}$$

and CG-skew:

$$CGskew = \frac{\#C - \#G}{\#C + \#G}$$

where $\#A, \#T, \#C, \#G$ represent the number of A, T, C and G have been calculated over each of the 39 sliding windows, thus producing 78 features.

In each sliding window, we have calculated the number of nucleotides overlapping with any palindrome of length six or more, taking into account that a sequence is considered a palindrome if it is equal to its reverse complement. This generates 39 features.

For the CpG islands features category we have calculated the following two metrics in each of the sliding windows:

$$\mathrm{M1} = \frac{\#CG}{\#C + \#G} * window\_length$$

$$\mathrm{M2} = \frac{\#C + \#G}{\mathrm{window\_length}} * 100$$

This results in 2*39=78 features.

In each sliding window, we have calculated the number of nucleotides overlapping with any repetitive element, producing 39 features.

For each examined sequence we have calculated the percentage of the total number of positions overlapping with each of 15 different chromatin states in each of 7 cell types (GM12878, H1-hESC, HMEC, HSMM, HUVEC, NHEK, and NHLF) [26]. The coordinates of the states have been downloaded from the UCSC Genome Browser. This produces 7*15= 105 features.

3.3 Proposed Methodology

The proposed method is an embedded classification method that combines an adaptive GA with a nu-SVR classifier. It is inspired by EnsembleGASVR, a method suggested in [27] for classifying missense single nucleotide polymorphisms. In principle, SVR classifiers present high classification performance and low complexity. The nu parameter of a nu-SVR classifier allows for the tuning of the number of the support vectors in the resulting classification model. GAs are stochastic meta-heuristic optimization algorithms. One advantage of GAs is their ability to explore efficiently large search spaces and identify possible solutions, without getting trapped in local optima, while at the same time locating near-to-optimal solutions. In the proposed method, the adaptive GA is used to identify the best feature subset and to tune the nu-SVR parameters.

The produced hybrid algorithm mainly consists of the iterative application of the evaluation, selection, crossover and mutation steps in a population of candidate solutions (chromosomes) which are initially randomly generated. Binary encoding has been used to represent each chromosome. Specifically, a 680-bit string is used where 626 bits encode features and 54 bits encode parameters. The parameters are (i) classifier parameters *C* (20 bits) and ii) *nu* (10 bits); (iii) radial basis kernel bandwidth *gamma* (14 bits); and (iv) classification *threshold* (10 bits).

A rank-based roulette wheel selection method controls the selection of the best candidates in each GA generation. This selection mechanism is preferred compared

with the single roulette wheel selection to raise the selection pressure toward better solutions when all solutions of the population present similar fitness values. Elitism is used to force the best solution of each population to be selected at least once in the next generation. This selection mechanism has been suggested in [27].

The evaluation of each chromosome in the population is performed according to the following fitness function:

$$Fitness = a * Accuracy + b * GeometricMean - c * 10^2 * MSE - d * \frac{1}{626} Features - e * \frac{1}{408} * SupportVectors$$

where $Accuracy$ is the nu-SVR's accuracy, $GeometricMean$ is the geometric mean of sensitivity and specificity, MSE is the mean square of errors, $Features$ is the size of the selected features subset and $SupportVectors$ is the number of support vectors included in the trained nu-SVR model.

The ranges of the examined variables in the proposed fitness function are $Accuracy \in [0,1]$, $GeometricMean \in [0,1]$, $MSE \in [0,0.01]$, $Features \in [1,626]$, where 626 represents the maximum number of features that can be selected by this method, and $SupportVectors \in [1,408]$, where 408 represents the number of the training samples. $MSE, Features$ and $SupportVectors$ are multiplied by constants, as shown in the equation, to normalize their values in the range $[0,1]$. The constants a = 0.5, b = 0.5, c = 0.01, d = 0.005 and e = 0.001 are user-defined weights assigned without experimentation and selected so as to reflect the priorities of each goal. More specifically, the classification accuracy and the geometric mean are the most significant. Then the MSE of the classifier follows. The number of selected features follows next and the number of support vectors is the least significant objective. To avoid over-fitting problems, we did not attempt to optimize these values, as suggested in [27]. The weights of the goals have been set so as to achieve high classification performance and simultaneously generate a simple and effective model.

Then the differentiation operators, crossover and mutation are applied to the top-ranked candidate solutions to create a new population. The crossover operator applies 2-point crossover to obtain a new offspring from two parents. The crossover rate is constant and set to 0.9, in order to leave some part of the population to survive unchanged to the next generation. This property is essential when good solutions emerge in early stages of the algorithm, as proposed in [27].

The mutation operator is very important to avoid local optima and explore a wide area of the search space. In the first generations it is preferable to explore a wider search space (exploration), while in the last generations it is preferable to search locally near the most promising areas of the search space (exploitation). To balance the tradeoff between the exploration and exploitation, the proposed method uses an adaptive mutation probability starting with a high value, 0.2, and gradually decreasing. The mutation rate is computed according to the following equation:

$$P_m(n) = 0.2 - n * \frac{0.2 - \frac{1}{P\_S}}{MAX\_G}$$

where n is the current generation, P_S is the size of the population and MAX_G is the maximum generation specified by the termination criterion. The mean similarity of every chromosome with the best chromosome of the population is measured at every generation. If the mean similarity is greater than 90% then the mutation rate is increased by a factor of $\frac{0.2-\frac{1}{P\_S}}{MAX\_G}$ instead of being decreased to avoid stagnation, i.e. getting trapped to local optima

The size of the population is set to 80 chromosomes and the termination criterion is 250 generations.

4 Experimental results

In order to evaluate the performance of our method for predicting miRNA promoters against other sequences, we have performed 10 5-fold external cross validation experiments and then we took the average in order to better assess the performance. In each fold, the $\frac{2}{3}$of the samples were used to train the SVM model and $\frac{1}{3}$of the samples were used as validation samples to measure the performance and calculate the fitness values. Table 2 summarizes the results which have been achieved by the proposed method. It presents the average values of all 10 5-fold external cross validation experiments for the classification metrics: accuracy, specificity, sensitivity and geometric mean. The last column is the average value for the 10 5-fold cross validation experiments of the number of the selected features which are used as inputs in our method.

Table 2. Metrics

Accuracy	*Sensitivity*	*Specificity*	*Geometric mean*	*Number of features*
0.8786	0.8447	0.9126	0.8780	307

The proposed methodology achieves on average accuracy 87.86%, sensitivity 84.47%, specificity 91.26% and geometric mean 87.8%. The average of selected features is 307 out of the total 626. These results are comparable with those of the better performing methods in the literature. The recent method of Hua et al. [19] presents 84% sensitivity and 91.3% precision and the miRStart method [15] presents sensitivity of 90.36%, specificity of 90.05%, accuracy of 90.21% and precision of 90.08%. The accuracy, sensitivity, specificity, precision and Matthews Correlation Coefficient (MCC) of [25] are 92.00%, 91.56%, 92.15%, 79.74% and 0.80, respectively.

5 Conclusion

The proposed approach for the prediction of miRNA promoters is a computational methodology based on the hybrid combination of an adaptive genetic algorithm with a nu-SVR classifier. The adaptive genetic algorithm is responsible for locating the optimal features set and optimizing the parameters of the nu-SVR classifier. The main

advantage of the proposed solution is that it systematically studies a vast number of features that can be used for miRNA promoters prediction. They include (i) frequency-based properties of the promoters such as k-mers, word commonality, skew, palindromes; (ii) regulatory elements such as CpG islands, repetitive elements; (iii) epigenetic features such as chromatin states. The provided script can be used to calculate effectively most of the features that have been correlated with promoter attributes without the need for combining different tools. The proposed script is of general usage as it can be used to structurally, sequentially and epigenetically annotate candidate promoters not only for miRNAs but also for protein coding genes and other non-coding RNA categories. In terms of classification performance, the proposed method handles efficiently the issues of over-fitting, feature selection, convergence and class imbalance. Experimental results give accuracy over 87%, sensitivity over 84% and specificity over 91% in the miRNA promoter prediction.

Our future research plans involve a more extensive study on the calculated features in order to gain insight on miRNA promoters characteristics. Additionally, in order to better handle the class imbalance issue, we plan to employ the Synthetic Minority Over-sampling Technique (SMOTE) [28]. Finally, we plan to compare the proposed solution with other existing solutions in the same datasets in order to gain more fair comparative results and to better validate the performance of the proposed solution.

Acknowledgement

Insybio participates in NBG Business Seeds Program by NBG.

References

1. Stefani, G., & Slack, F. J. (2008). Small non-coding RNAs in animal development. Nature reviews Molecular cell biology, 9(3), 219-230.
2. Krol, J., Loedige, I., & Filipowicz, W. (2010). The widespread regulation of microRNA biogenesis, function and decay. Nature Reviews Genetics, 11(9), 597-610.
3. Bartel, D. P. (2009). MicroRNAs: target recognition and regulatory functions. Cell, 136(2), 215-233.
4. Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., ... & Liu, Y. (2009). miR2Disease: a manually curated database for microRNA deregulation in human disease. Nucleic acids research, 37(suppl 1), D98-D104.
5. Kleftogiannis, D., Korfiati, A., Theofilatos, K., Likothanassis, S., Tsakalidis, A., & Mavroudi, S. (2013). Where we stand, where we are moving: Surveying computational techniques for identifying miRNA genes and uncovering their regulatory role. Journal of biomedical informatics, 46(3), 563-573.
6. Rodriguez, A., Griffiths-Jones, S., Ashurst, J. L., & Bradley, A. (2004). Identification of mammalian microRNA host genes and transcription units. Genome research, 14(10a), 1902-1910.
7. Corcoran, D. L., Pandit, K. V., Gordon, B., Bhattacharjee, A., Kaminski, N., & Benos, P. V. (2009). Features of mammalian microRNA promoters emerge from polymerase II chromatin immunoprecipitation data. PloS one, 4(4), e5279.

8. Ozsolak, F., Poling, L. L., Wang, Z., Liu, H., Liu, X. S., Roeder, R. G., ... & Fisher, D. E. (2008). Chromatin structure analyses identify miRNA promoters. Genes & development, 22(22), 3172-3183.
9. Saini, H. K., Griffiths-Jones, S., & Enright, A. J. (2007). Genomic analysis of human microRNA transcripts. Proceedings of the National Academy of Sciences, 104(45), 17719-17724.
10. Wang, X., Xuan, Z., Zhao, X., Li, Y., & Zhang, M. Q. (2009). High-resolution human core-promoter prediction with CoreBoost_HM. Genome research, 19(2), 266-275.
11. Abeel, T., Saeys, Y., Bonnet, E., Rouzé, P., & Van de Peer, Y. (2008). Generic eukaryotic core promoter prediction using structural features of DNA. Genome research, 18(2), 310-323.
12. Zhou, X., Ruan, J., Wang, G., & Zhang, W. (2007). Characterization and identification of microRNA core promoters in four model species. PLoS Comput Biol, 3(3), e37.
13. Ohler, U., Niemann, H., Liao, G. C., & Rubin, G. M. (2001). Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition. Bioinformatics, 17(suppl 1), S199-S206.
14. Monteys, A. M., Spengler, R. M., Wan, J., Tecedor, L., Lennox, K. A., Xing, Y., & Davidson, B. L. (2010). Structure and activity of putative intronic miRNA promoters. Rna, 16(3), 495-505.
15. Chien, C. H., Sun, Y. M., Chang, W. C., Chiang-Hsieh, P. Y., Lee, T. Y., Tsai, W. C., ... & Huang, H. D. (2011). Identifying transcriptional start sites of human microRNAs based on high-throughput sequencing data. Nucleic acids research, 39(21), 9345-9356.
16. Marsico, A., Huska, M. R., Lasserre, J., Hu, H., Vucicevic, D., Musahl, A., ... & Vingron, M. (2013). PROmiRNA: a new miRNA promoter recognition method uncovers the complex regulation of intronic miRNAs. Genome Biol, 14(8), R84.
17. Barski, A., Jothi, R., Cuddapah, S., Cui, K., Roh, T. Y., Schones, D. E., & Zhao, K. (2009). Chromatin poises miRNA-and protein-coding genes for expression. Genome research, 19(10), 1742-1751.
18. Marson, A., Levine, S. S., Cole, M. F., Frampton, G. M., Brambrink, T., Johnstone, S., ... & Calabrese, J. M. (2008). Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. Cell, 134(3), 521-533.
19. Hua, X., Chen, L., Wang, J., Li, J., & Wingender, E. (2016). Identifying cell-specific microRNA transcriptional start sites. Bioinformatics, btw171.
20. Maston, G. A., Evans, S. K., & Green, M. R. (2006). Transcriptional regulatory elements in the human genome. Annu. Rev. Genomics Hum. Genet., 7, 29-59.
21. Bhattacharyya, M., Das, M., & Bandyopadhyay, S. (2012). miRT: a database of validated transcription start sites of human microRNAs. Genomics, proteomics & bioinformatics, 10(5), 310-316.
22. Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D., & Kent, W. J. (2004). The UCSC Table Browser data retrieval tool. Nucleic acids research, 32(suppl 1), D493-D496.
23. Zhao, X., Xuan, Z., & Zhang, M. Q. (2007). Boosting with stumps for predicting transcription start sites. Genome biology, 8(2), R17.
24. Alam, T., Medvedeva, Y. A., Jia, H., Brown, J. B., Lipovich, L., & Bajic, V. B. (2014). Promoter analysis reveals globally differential regulation of human long non-coding RNA and protein-coding genes. PLoS One, 9(10), e109443.
25. Bhattacharyya, M., Feuerbach, L., Bhadra, T., Lengauer, T., & Bandyopadhyay, S. (2012). MicroRNA transcription start site prediction with multi-objective feature selection. Statistical applications in genetics and molecular biology, 11(1), 1-25.

26. Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., ... & Ku, M. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature, 473(7345), 43-49.
27. Rapakoulia, T., Theofilatos, K., Kleftogiannis, D., Likothanasis, S., Tsakalidis, A., & Mavroudi, S. (2014). EnsembleGASVR: a novel ensemble method for classifying missense single nucleotide polymorphisms. Bioinformatics, 30(16), 2324-2333.
28. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 321-357.