The challenge for the identification of new miRNAs is that both miRNAs and their targets need to be expressed simultaneously [10]. This obstacle led to the development of computational approaches for miRNA prediction. There are two major methods used; homology based methods based on evolutionary relations and *ab initio* strategies mostly relying on machine learning (ML) [11]. Performance of the latter depends on many parameters like; quality of data sets, influence of features, and feature selection scheme as well as the choice of ML algorithm.

With the thousands of features proposed to describe a miRNA hairpin it is particularly important to perform effective feature selection. We implemented a distributed genetic algorithm to select a feature subset while preserving high classification accuracy. The outcome was a feature subset consisting of 21 features with 99% accuracy. This performance was achieved after 108 generations of the genetic algorithm using a random forest classifier.

In this study we created a workflow to predict miRNAs in 5 retro-transcribing virus genomes with human host (Human endogenous retrovirus K113, Hepatitis B virus (strain ayw), Human T lymphotropic virus 1, Human T lymphotropic virus 2, Human immunodeficiency virus 2, Human immunodeficiency virus 1), by using the information from known virus and human miRNAs. Our results indicate that these viruses might produce miRNA precursors which require further experimental validation.

2 Architecture/Implementation

Feature selection methods search the feature subset space consisting, in this case, of 2^N possible feature combinations [12] where N is the number of features. Exhaustive search of 2^N combinations is an NP-hard problem [13], so we used a genetic algorithm (GA), a heuristic embedded feature selection algorithm.

2.1 Genetic Algorithm

A genetic algorithm is an approach which mimics the evolution process for solving problems or modeling evolutionary systems within the computational environment [14]. It is commonly used to produce solutions for search and optimization problems. The intuition of the algorithm is based on 'survival of fittest individual' during a natural selection process [15]. We developed a distributed genetic algorithm for this feature selection process to enable processing in reasonable time. HTCondor [16] was used to distribute the workload of evaluation of feature subsets using the KNIME data analytics platform [12] workflow which was constructed to calculate classification accuracy of feature subsets. 500 feature subsets were generated randomly using a 2% fixed mutation probability. These were evolved over until a stop criterion was reached. During the evolution of generations, stop condition is determined in terms of improvement of the fittest individual. If the score of the best individual is not improved during five generation, the genetic algorithm is terminated. Therefore, after 104 generation, the process was terminated since there was no further improvement for the fittest individual of the population. Then, the fittest individual's features were used for creating machine learning models (Figure 1). The HTCondor implementation of the GA used in this study enables the use of unused cycles of our computer pools and office equipment. The use of a KNIME workflow as the fitness function enables the application of the GA for a wide range of optimization problems exemplified with feature selection for pre-miRNA detection in this study. The GA and all novel features will be presented elsewhere (Toprak et al., manuscript in preparation).

2.2 Machine Learning

2.2.1 Feature Sets

More than 800 features defining pre-miRNAs based on structural, sequential, probabilistic and thermodynamic characteristics were calculated for training data sets [17–19]. These data sets were used for selection of features via the genetic algorithm. All of the available features designed by us and found in the literature can be found and calculated with JLab miRNA feature calculator software available on our website (http://jlab.iyte.edu.tr/software/mirna) or other similar tools [20].

2.2.2 Positive Data Sets

Virus hairpin sequences from miRBase (308 hairpins), human miRNA hairpins that have experimental validation listed in miRTarBase [21] (Support Type (Weak) samples were filtered, 388 hairpins). This was done because a larger fraction of entries in miRTarBase seems to represent real pre-miRNAs as compared to miRBase [22].

2.2.3 Negative Data Set

Pseudo hairpins obtained from Ng et al. The data set includes approximately 8000 hairpins [23], after calculation of features missing values were removed leading to 3589 hairpins. Genomes of Retro-transcribing viruses with human as host were obtained from NCBI (http://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?taxid=35268&host=human). A 1:1 ratio was maintained between positive and negative data sets [18]. The negative data set was randomly sampled to match the number of positive samples.

2.2.4 Model Generation

Two classifiers with various settings were used for model generation; Random Forest with 3 split criteria (Information Gain (IG), Information Gain Ratio (IGR), Gini Index (Gini)) and LIBSVM with 5 types and 4 kernels (C-SVC, nu-SVC, one-class SVM, epsilon-SVR, nu-SVR and linear, polynomial, radial basis function, sigmoid). 10-fold cross validation with stratified sampling was applied during training (Figure 1). All learning and prediction workflows were generated and applied in KNIME [24].

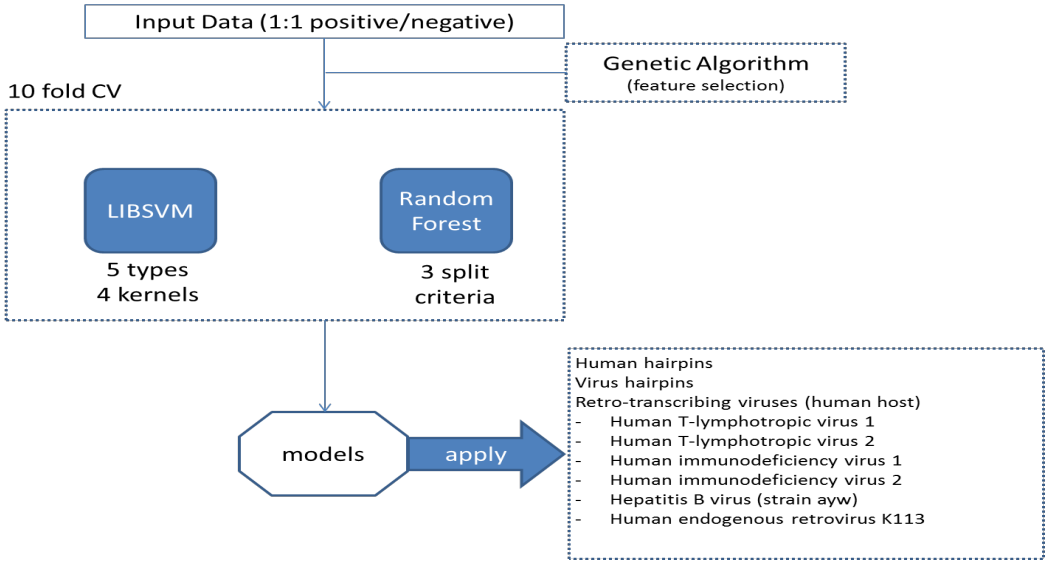

Figure 1: Workflow for the classification system used in this study.

2.2.5 Prediction Data Sets

Retro-transcribing viruses' genomes were split into overlapping fragments (500/250) and transcribed into RNA sequences (T => U as + strand, the complement as - strand). For all of these 500 nt long fragments secondary structure was calculated using RNAfold [25] and hairpins were extracted. After filtering the hairpins according to their length distribution (min: 36, max: 180) and removing the duplicate sequences, for 412 (+) strand and 420 (-) strand hairpins features are calculated.

3 Application

As a result of genetic algorithm, a feature subset consisting of 21 features remained after 104 generation. This feature set achieved 99% accuracy and was used in training classifiers (Figure 1).

The model was analyzed and optimized using multiple settings such as different kernels for support vector machines (SVM). Most models achieved high areas under the receiver operating characteristic (ROC) curve, but interestingly, most SVM models led to random decisions (Figure 2).

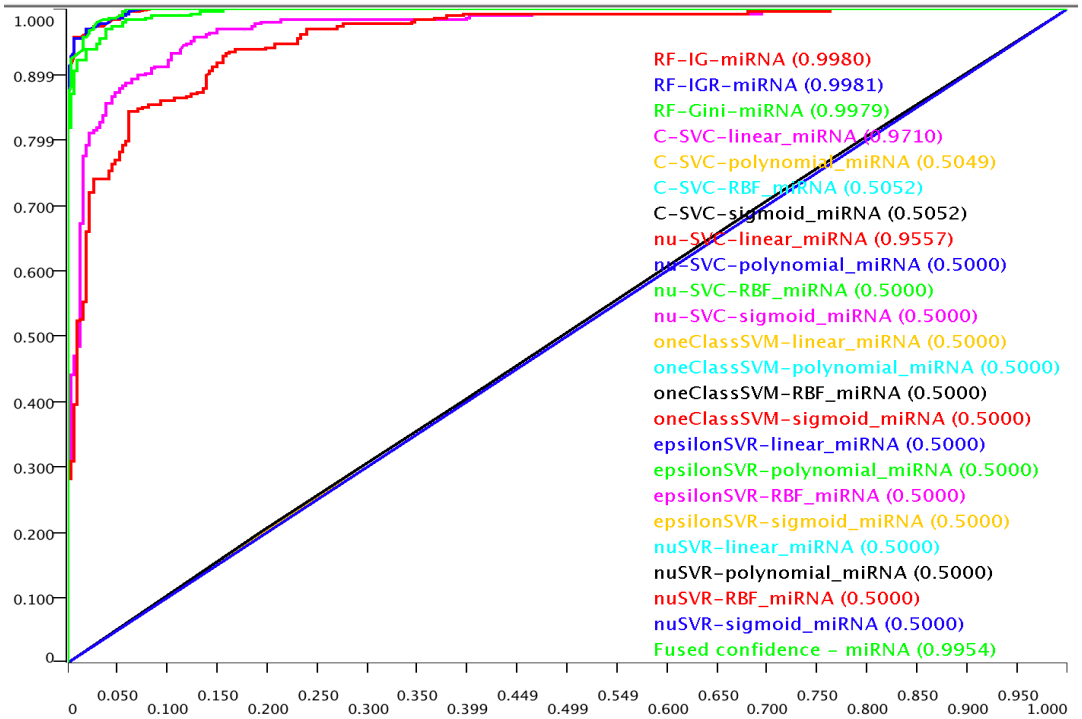

Figure 2: ROC curve graph from the case virus positive data set is used for model generation. False positive rate (x-axis) and true positive rate (y-axis) values are shown for different scenarios. IG: Information Gain; IGR: IG Ratio; Gini: Gini Index SVM types: C-SVC, nu-SVC, one-class SVM, epsilon-SVR, nu-SVR; Kernels: linear, polynomial, RBF: radial basis function, sigmoid.

Application of the trained models to the hairpins extracted from the virus genomes revealed a number of putative pre-miRNAs conforming to the human model at a prediction score cutoff of 0.95 (Table 1). Such pre-miRNAs may give the virus leverage to modulate its host gene expression and create a suitable environment for its replication. Figure 2 shows that the usage of SVM is highly dependent on optimization of its parameters whereas RF led to acceptable outcomes for all tested scenarios. Overall, RF achieved an accuracy better than 0.99 for all three tested split criteria.

Table 1: Number of predicted hairpins in 5 virus genomes by using models learned from known virus hairpins. RF: random forest, + and - indicate strands. Prediction score cut-off value is 0.95.

Viral Genome	LibSVM(-)	LibSVM(+)	RF(-)	RF(+)	both(-)	both(+)
Human endogenous retrovirus K113	22	18	2	2	2	1
Hepatitis B virus (strain_ayw)	13	9	4	3	3	3
Human T lymphotropic virus 1	45	10	15	1	14	1
Human T lymphotropic virus 2	37	14	9	1	8	1
Human immunodeficiency virus 2	17	14	3	2	1	2
Human immunodeficiency virus 1	23	25	7	2	7	1

Out of three entries for HIV hairpins in miRBase, we only identified hiv1-mir-TAR with libSVM but it did not pass our prediction score cut-off value (0.95) for random forest (virus learning data model; libSVM score: 0.98, RF score: 0.94, human learning data model: libSVM score: 0.96, RF score: 0.88). Overall RF selected less pre-miRNAs (Table 1) which is likely due to its higher accuracy when compared to SVM in this study. Further filtering by requiring both LibSVM and RF to accept a pre-miRNA leads to a very small amount of putative pre-miRNAs (Table 1). While Table 1 considers pre-miRNA detected with a model based on known viral pre-miRNA examples, Table 2 contains those that pass a model trained on known human pre-miRNAs.

Table 2: Number of predicted hairpins in 5 virus genomes by using models learned from known human hairpins. RF: random forest, + and - indicate strands. Prediction score cut-off value is 0.95.

Viral Genome	LibSVM(-)	LibSVM(+)	RF(-)	RF(+)	both (-)	both(+)
Human endogenous retrovirus K113	4	5	2	4	1	3
Hepatitis B virus (strain_ayw)	3		2		2	
Human T lymphotropic virus 1	9		3		3	
Human T lymphotropic virus 2	14	1	4		2	
Human immunodeficiency virus 2	4	2	1	1		
Human immunodeficiency virus 1	4	11	1	1		1

By using the hairpins that passed both models in Table 1 and 2, we searched for possible human target genes of these viral miRNAs in the human genome. To achieve this, online miRNA target prediction tool psRNATarget (http://plantgrn.noble.org/psRNATarget/) was used (Table 3). As mature miRNA input, the viral hairpins that were divided into 30nt long fragments with 15nt overlaps were used. The human genome pool available in psRNATarget server was used for target search of these mature sequences. All of the analyzed miRNAs were used in a single target search. At the end, the targets were listed per organism (Table 3). Note that, one gene might be targeted by different miRNAs so they appear in the table more than once.

Table 3: Targets of predicted viral miRNAs in the human genome.

Source	Targets
Hepatitis B virus (strain ayw)	ACP1, STK38L, BEND2, ACP1, STK38L, BEND2, WWP2, NEO1, NEO1, WWP2, NEO1, NEO1, CLDN12, ZNF829, PPIL4, GYPA, GYPA, PHKA2, PHKA2, ABCA13, ZNF711, ACOX1, ACOX1, SH3RF1, ZNF677, BRPF3, FCRL5, FCRL5, GLI3, CCDC129, ERMN, MPZL1, MPZL1, RBM20, SLC24A3, IFIT5, POLR2M, GCOM1, GCOM1, MCHR2, DISC1, DISC1, DISC1, NSL1, NSL1, PRKCI, WWP2, ZNF704, GPR107, PBX1, PBX1, PSME3, PSME3, PSME3, PSME3, PDE9A, RBM5, RBM5, RBM5, RBM5, KCNK7, MARCKS, DDX51, NUFIP2, NUFIP2, SLC25A37, OBFC1, PHACTR2, ACP1, STK38L, BEND2, ACP1, STK38L, BEND2, WWP2, NEO1, NEO1, WWP2, NEO1, NEO1
Human T-lymphotropic virus 1	MAP1A, ZC3H4, IGF2, SPRED2, CACNB4, ARF3, NCOA1, NCOA1, CAMK2N2, FSTL1, KRAS, KRAS, VPS26B, COL18A1, ALG2, SSH2, CHRDL1, UPP2, SLC27A4, TUSC5, FOCAD, RACGAP1, RACGAP1, RACGAP1, RACGAP1, RACGAP1, RACGAP1, RACGAP1, RACGAP1, RACGAP1, RACGAP1, RACGAP1, RACGAP1, RACGAP1, FLRT3, ATG16L1, CTBS, DCLK3, ZFX, ZFX, PROX1, CTBS, DCLK3, ZFX, ZFX, PROX1, PNPLA8, PNPLA8, CASP3, CASP3, TNPO1, TNPO1, SERINC5, XIAP, CLEC2B, ARHGEF7, TSHZ2, SH2D4A, SH2D4A, GMFB, VAMP2, CTBS, DCLK3, ZFX, ZFX, PROX1, CTBS, DCLK3, ZFX, ZFX, PROX1, PNPLA8, PNPLA8, CASP3, CASP3, TNPO1, TNPO1, PARD6G, ABHD17C, SLAIN2, CDK6, SELL, CACNB4, SZT2, C11orf87, CXorf40A, IPPK, SC5D, KANK2, HECTD4, DDX17, DDX17, DCUN1D1, CAMKK2, CAMKK2, CAMKK2, CAMKK2, CAMKK2, CAMKK2, CAMKK2, CAMKK2, CAMKK2, CAMKK2, CAMKK2, CAMKK2, CAMKK2, CAMKK2, CAMKK2, CAMKK2, CAMKK2, STX17, CABLES1, CDH4, PGAP3, DDR1, ELAVL1, KCNMB1, MAF, TNFSF15, SCAMP1, VMA21, SLC30A10, TAF9B, FZR1, STK40, ARFIP1
Human T-lymphotropic virus 2	PTPDC1, PAX6, PAX6, PAX6, PAX6, BCL10, MEX3A, ARVCF, PPARA, NSL1, NSL1, SH3GLB1, BRIP1, BRIP1, TP53RK, CDH13, INO80D, DDHD1, DDHD1, EPB41, EPB41, PSD4, FREM2, CRTC3, PSD4, FREM2, CRTC3, ZNHIT6, GRIK3, ZNHIT6, GRIK3, NOMO3, NOMO1, ASB11, HTN3, BCAR1, BCAR1, ASB6, ESYT2, ASB6, HMBOX1, SYT13, PBX1, PBX1, PBX1, PSMA5, DDHD1, DDHD1, EPB41, EPB41, PSD4, FREM2, CRTC3, PSD4, FREM2, CRTC3, ZNHIT6, GRIK3, ZNHIT6, GRIK3, RHBDL3, DDHD1, EPB41, PSD4, ZNHIT6, GRIK3, MME, C21orf33, MYF6, DHX33, DHX33
Human endogenous retrovirus K113	METTL21A, NEK5, THEM4, CFLAR, CFLAR, CFLAR, CFLAR, NEK5, METTL21A, GATAD2B, IL24, IL24, LCOR, LCOR, NDRG1, CFLAR, CFLAR, NCS1, NEK5, METTL21A, CFLAR, CFLAR, OTUD4, ARL8B, NEK5, METTL21A, CFLAR, CFLAR, KCMF1, AP2B1, AP2B1, XG, MBLAC2, METTL21A, NEK5, CFLAR, CFLAR, THEM4, NEK5, CFLAR, CFLAR, GFRA1, CBLN2, SNX30, ATP6V0A2, ATP6V0A2, HBE1, GSAP, C10orf32, C10orf32, MLLT4, MLLT4, MLLT4, MLLT4, MLLT4, MLLT4, MLLT4, HHAT, ARRDC2, TMEM170B, DISC1, DISC1, DISC1, RALGPS2, CTDSPL2, CTDSPL2, CTDSPL2, CTDSPL2, CTDSPL2, AKAP12, STXBP3, ZNF845, SNED1, JPH2, GGTLC2, ZNF264, ZNF264, SLC9A7, GSE1, SLC7A10, SLC25A24, SLC25A24, SGMS1, CREB1, CREB1, CREB1, CREB1, CREB1, ZFP91, MIS18BP1, UNC5C, PLEKHG4B, QRFPR, FRMD8, ZNF518A, HNRNPK, HNRNPK, XKR7, EPHA10, WDR37, NCMAP, ZNF780A, CORO2B, CHML, ZNF592, SMCR8, SPEF1
Human immunodeficiency virus 1	RAP2B, KIR2DL5A, KIR2DL5B, KIR2DS5, RASSF8, OAZ1, ACTC1, UNC80, KLHDC3, FGFRL1, THAP1, THAP1, ADCY1, SATL1, AQP11, TMEM167A, GRIN2A, DLG1, ABHD5, SCN8A, ZFP91, SLC6A3, TESPA1, SEPT11, ZBTB44, CPNE8, RET, RET, MCUR1, FPR3
Human immunodeficiency virus 2	EFS, IGF1R, IGF1R, EPT1, MIS18A, POLDIP2, DPM2, ARHGAP26, RBM46, RBM46, PHLDA1, TOM1L2, SLC9A7, ZNF3, ZNF3, POLR2J, NUP50, NUP50, NUP50, NUP50, SENP8, SART3, PTBP3, LMLN, LMLN, CYP19A1, NXPH2

4 Discussion

MiRNAs and their actions in various cellular processes have attracted great interest. However, there are many obstacles to overcome to achieve a better understanding in the overall miRNA involved pathways. In recent years, potential interactions between host and parasite's miRNAs have been proposed which further increased the complexity of miRNA analyses.

We tested the influence of some of the elements in a classification system. Our previous work showed that selection of features has a high impact on accuracy of the classifier [17]. To obtain a high quality feature list we developed a genetic algorithm for feature selection. This is especially important since about a thousand features have been proposed for miRNA hairpin prediction. The genetic algorithm was able to decrease the number of features required to 21 at a very high accuracy and area under the ROC curve.

One of the factors affecting overall classification performance is the choice and settings of classifiers. By applying different criteria with two widely used classification methods libSVM and random forest to virus and human data, we obtained quite a big difference among classifier performances.

Our results indicate that the viruses examined in this study have the capacity to produce functional miRNAs. Most of the predictions we obtained seem to have a secondary structure that can be recognized by miRNA biogenesis proteins like Drosha or Dicer (Table 1, Table 2). Furthermore, there are reports claiming that hiv1-mir-TAR is processed by Dicer *in vitro* [4] Nevertheless there is no agreement for the possible HIV-1 generated miRNAs so more experimental analysis is required. The number of predicted pre-miRNAs from virus genomes in this study shows that with the approach used here it is possible to experimentally validate them.

Acknowledgements

The work was supported by the Scientific and Technological Research Council of Turkey [grant number 113E326] to JA.

References

[1] A. Kozomara and S. Griffiths-Jones. miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic acids research, 39(Database issue):D152-7, Jan. 2011.

[2] M. D. Saçar, C. Bağcı, and J. Allmer. Computational Prediction of MicroRNAs from Toxoplasma gondii Potentially Regulating the Hosts' Gene Expression. Genomics, proteomics & bioinformatics, 12(5):228–238, Oct. 2014.

[3] V. Scaria, M. Hariharan, S. Maiti, B. Pillai, S. K. Brahmachari, D. Bartel, et al. Host-virus interaction: a new role for microRNAs. Retrovirology, 3(1):68, 2006.

[4] R. L. Skalsky and B. R. Cullen. Viruses, microRNAs, and host interactions. Annual review of microbiology, 64:123–41, Jan. 2010.

[5] A. Grundhoff and C. S. Sullivan. Virus-encoded microRNAs. Virology, 411(2):325–43, Mar. 2011.

[6] G. S. França, M. D. Vibranovski, P. A. F. Galante, D. P. Bartel, M. R. Fabian, N. Sonenberg, et al. Host gene constraints and genomic context impact the expression and evolution of human microRNAs. Nature Communications, 7:11438, Apr. 2016.

[7] J. Lu, Y. Shen, Q. Wu, S. Kumar, B. He, S. Shi, R. W. Carthew, S. M. Wang, and C.-I. Wu. The birth and death of microRNA genes in Drosophila. Nature genetics, 40(3):351–5, Mar. 2008.

[8] E. Berezikov. Evolution of microRNA diversity and regulation in animals. Nature Reviews Genetics, 12(12):846–860, Nov. 2011.
[9] J. Meunier, F. Lemoine, M. Soumillon, A. Liechti, M. Weier, K. Guschanski, H. Hu, P. Khaitovich, and H. Kaessmann. Birth and expression evolution of mammalian microRNA genes. Genome Research, 23(1):34–45, Jan. 2013.
[10] M. D. Saçar and J. Allmer. Current Limitations for Computational Analysis of miRNAs in Cancer. Pakistan Journal of Clinical and Biomedical Research, 1(2):3–5, 2013.
[11] M. D. Saçar and J. Allmer. Machine Learning Methods for MicroRNA Gene Prediction. in miRNomics: MicroRNA Biology and Computational Analysis SE - 10, 1107:177–187, M. Yousef and J. Allmer, Eds. Humana Press, 2014, 177–187.
[12] M. DASH and H. LIU. Feature selection for classification. Intelligent Data Analysis, 1(1–4):131–156, 1997.
[13] E. Amaldi and V. Kann. On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. Theoretical Computer Science, 209(1–2):237–260, 1998.
[14] S. Forrest. Genetic algorithms: principles of natural selection applied to computation. Science (New York, N.Y.), 261(5123):872–8, Aug. 1993.
[15] D. Beasley, D. R. Bull, and R. R. Martin. An Overview of Genetic Algorithms : Part 1, Fundamentals. .
[16] M. J. Litzkow, M. Livny, and M. W. Mutka. Condor-a hunter of idle workstations. in *[1988] Proceedings. The 8th International Conference on Distributed*, 1988, :104–111.
[17] M. D. Saçar and J. Allmer. Comparison of Four Ab Initio MicroRNA Prediction Tools. in *Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms*, 2013, :190–195.
[18] M. D. Sacar and J. Allmer. Data mining for microrna gene prediction: On the impact of class imbalance and feature number for microrna gene prediction. in *2013 8th International Symposium on Health Informatics and Bioinformatics*, 2013, :1–6.
[19] M. Yousef, M. D. Saçar Demirci, W. Khalifa, and J. Allmer. Feature Selection Has a Large Impact on One-Class Classification Accuracy for MicroRNAs in Plants. Advances in Bioinformatics, 2016:1–6, 2016.
[20] C. A. Yones, G. Stegmayer, L. Kamenetzky, and D. H. Milone. miRNAfe: A comprehensive tool for feature extraction in microRNA prediction. Bio Systems, 138:1–5, 2015.
[21] S.-D. Hsu, F.-M. Lin, W.-Y. Wu, C. Liang, W.-C. Huang, W.-L. Chan, W.-T. Tsai, G.-Z. Chen, C.-J. Lee, C.-M. Chiu, C.-H. Chien, M.-C. Wu, C.-Y. Huang, A.-P. Tsou, and H.-D. Huang. miRTarBase: a database curates experimentally validated microRNA-target interactions. Nucleic acids research, 39(Database issue):D163-9, Jan. 2011.
[22] M. D. Saçar, H. Hamzeiy, and J. Allmer. Can MiRBase provide positive data for machine learning for the detection of MiRNA hairpins? Journal of integrative bioinformatics, 10(2):215, Jan. 2013.
[23] K. L. S. Ng and S. K. Mishra. De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. Bioinformatics, 23(11):1321–30, Jun. 2007.
[24] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, and B. Wiswedel. KNIME: The Konstanz Information Miner. in SIGKDD Explorations, 11(1):319–326, 2008, 319–326.
[25] I. L. Hofacker. Vienna RNA secondary structure server. Nucleic Acids Research, 31(13):3429–3431, Jul. 2003.