Can MiRBase Provide Positive Data for Machine Learning for the Detection of MiRNA Hairpins?

Müşerref Duygu Saçar[1], Hamid Hamzeiy[1], Jens Allmer[1,*]

[1]Molecular Biology and Genetics, Izmir Institute of Technology, Gulbahce, Urla, Izmir, Turkey

Summary

Experimental detection and validation of miRNAs is a tedious, time-consuming, and expensive process. Computational methods for miRNA gene detection are being developed so that the number of candidates that need experimental validation can be reduced to a manageable amount. Computational methods involve homology-based and *ab inito* algorithms. Both approaches are dependent on positive and negative training examples. Positive examples are usually derived from miRBase, the main resource for experimentally validated miRNAs. We encountered some problems with miRBase which we would like to report here. Some problems, among others, we encountered are that folds presented in miRBase are not always the fold with the minimum free energy; some entries do not seem to conform to expectations of miRNAs, and some external accession numbers are not valid. In addition, we compared the prediction accuracy for the same negative dataset when the positive data came from miRBase or miRTarBase and found that the latter led to more precise prediction models. We suggest that miRBase should introduce some automated facilities for ensuring data quality to overcome these problems.

1 Introduction

Existing efforts to differentiate miRNA genes have led to the detection of thousands of miRNAs in various species, but countless remain undiscovered [1]. These attempts are predominantly based on experimental methods such as directional cloning of endogenous small RNAs, which are time consuming, expensive, and work intensive [2]. Difficulties for such experimental approaches are that miRNAs may only be expressed in specific cell types, at low levels or only in response to changing environmental stimuli. In order to overcome this problem, several computational methods have been designed and applied to the computational detection of miRNA hairpins [3].

Computational methods for miRNA gene detection can be grouped into two main categories; homology-based methods and *ab initio* approaches and both have specific advantages and drawbacks [4]. Machine learning algorithms are different from rule-based miRNA detection algorithms since the rules to decide whether a given sequence is a miRNA, are not manually created, instead these rules are fit, trained, or learned from examples [5]. Usually, machine learning-based methods start with the learning process of sequence, structure, and/or thermodynamic characteristics of miRNAs based on some curated parameters. After, the trained classifier can decide whether an unknown sequence is a miRNA, based on the information gained from training with positive and negative data sets. Normally, the parameters are a set of numerical features defining a candidate miRNA, such as minimum free energy of folding or terminal loop size. If the parameters describing the hairpin are properly chosen, and if the training data is informative, an unknown sequence could be classified into either being a miRNA precursor or not at high accuracy.

* To whom correspondence should be addressed. Email: jens@allmer.de

doi:10.2390/biecoll-jib-2013-215

However, there are two main obstacles for machine learning of miRNA genes. First one is the imbalance of positive and negative examples. Since the exact number of real miRNAs in any genome is unknown, it is supposed that there are few miRNA precursors within the millions of hairpins that can be found in a eukaryotic genome [6]. Also, the number of positive examples should be significantly smaller than that of negative examples. For instance, one of the commonly used negative dataset for miRNA detection algorithms consists of approximately 9000 pseudo hairpins while the number of human miRNAs that can be obtained from miRBase [7] (http://www.mirbase.org/) is less than 1600 [2]. The imbalance problem between positive and negative datasets can significantly reduce the performance of current machine learning approaches [6]. Although several negative datasets have been proposed, none of them provides a ground truth as it is currently not possible to experimentally confirm whether the selected negative examples are truly negative. The missing of ground truth data is the other obstacle.

MiRBase [7] is the main repository for miRNA sequences and associated information. Although there are other competing databases, to the best of our knowledge all miRNA hairpin detection algorithms have derived their positive datasets from miRBase. While developing our positive and negative datasets for *ab initio* miRNA hairpin detection, we came across some problems with positive data located in miRBase and further noticed some inconsistencies that we found worth to give some additional attention.

In this study, we report that miRBase includes entries which are not fitting the criteria for a sequence to be labelled as a true miRNA. Moreover, by performing classification we show that comparing to miRNAs proven with strong experimental evidence from miRTarBase [8] database the sensitivity, specificity and classification accuracy using the miRBase derived positive dataset is lower. In addition to that we assessed whether the structures presented on miRBase have actually been predicted by RNAFold [9] as published, and found that this was not true for the structures we analysed. Finally, a database that calls itself the official miRNA registry, should offer some kind of support lest errors are encountered. Unfortunately, reporting problems with mapping of miRBase stored Ensembl [10] accession numbers to the Ensembl genome browser for human [11] via email has not been replied to within one year. We suggest that all sequences submitted to miRBase should be folded by an automatic process to overcome the problem that folds which do not have minimal free energy are presented on miRBase, misleading users into believing that the presented fold is the best. Another automatic process should occasionally validate the provided accession numbers and finally an automatic process should check all submissions for whether they adhere to certain criteria describing a hairpin. If they fail these checks the data may be send back to the submitter for checking, manually validated by the miRBase team or put into a separate section of miRBase for not validated data.

2 Materials and Methods

2.1 Dataset

Machine learning for miRNA hairpin detection depends at least on positive examples but in general additionally needs negative examples [3]. For negative examples, the widely used pseudo hairpin sequences by Ng and Mishra [2] were chosen since in our previous work, we have shown that this negative dataset provides better distinction between miRNA and non-miRNA sequences than other datasets [4]. From the negative examples we randomly selected 180 pseudo hairpins and used this as negative data for all analyses (Supplementary File: pseudoHairpins.txt; http://bioinformatics.iyte.edu.tr/index.php?n=Data.MiRBase).

Since miRTarBase [8] has only little data for human we used all 180 available examples for the positive dataset from miRTarBase. MiRTarBase claims that this data has strong experimental evidence showing miRNA-target interactions (Supplementary File: miRTarBasePosEx.txt).

MiRBase contains about 1600 human miRNAs, but in an attempt to keep the comparison fair, we randomly extracted 180 sequences (Supplementary File: miRBasePosEx.txt).

This leads to two positive datasets which both consist of 180 examples one from miRBase and one from miRTarBase which will be compared in (Section 3.2). Both positive datasets were used in conjunction with the same negative dataset in order to ensure fair comparison.

Since the data available in miRBase stems from many organisms most of which have no competing databases available, we chose to focus on the available human data described above. Although a generalization of these results is therefore not necessarily possible, we offer possible solutions which would aid solving a general problem.

2.2 Parameters

We previously assessed 12 *ab initio* miRNA detection algorithms [2], [6], [12–21] and determined the parameters that were used to describe a miRNA hairpin. More than 200 different parameters have been described and more than 100 have been used in machine learning for miRNA hairpin detection. In this study, we used the 10 most frequently used features in the 12 *ab initio* miRNA hairpin detection studies. The selected parameters are: hairpin loop length (hll), base pairing propensity (bpp), hairpin minimum free energy (hpmfe), dinucleotide shuffling, p-value of hpmfe, and the frequencies of the following triplet structures U(((, U(.(, C(.(, A..., G(((.

Selected features are of major importance for the accuracy and generalization of the model established through machine learning. In this particular case, we perform a relative comparison and therefore the only requirement is to use the same features for all models. The reason why we used only the 10 most frequently used features is motivated by the fact that many algorithms are using them and not due to their discrimination ability which is unknown. Also, implementing all 200 features is beyond the scope of this study and therefore the number of features needed to be restricted. Slightly increasing or decreasing the amount of features does not influence the outcome of the relative comparison (data not shown) and therefore an arbitrary number of 10 features was selected for this study.

2.3 Classification

Orange Canvas [22] was used to perform SVM, Naive Bayes, and Logistic Regression classifications with 10 fold cross-validation. All the classifiers are used with their default settings.

3 Results and Discussion

3.1 Folds Available in MiRBase

It has been shown that for effective miRNA processing by Drosha [23] and Dicer [24] the terminal loop region is very essential and the mutations in this region affect mature miRNA production [25]. Figure 1 shows a miRBase entry for a human miRNA which is supposed to have a mature miRNA located partially within the terminal loop, something not supported by the current understanding of miRNAs.

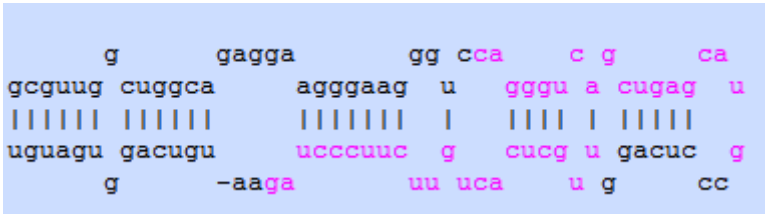

Figure 1: Predicted secondary structure of hsa-mir-1178 (MI0006271) as presented on miRBase, showing stem-loop structure. Pink bases indicate mature miRNA sequences.

Drosha requires a partner protein for performing RNA binding [26]. DGCR8 (Pasha) is the cofactor that interacts with Drosha and forms a functional complex known as Microprocessor complex. Two double strand RNA binding domains (dsRBDs) distinguish the single stranded RNA (ssRNA) segments flanking a stem of proper length [27]. DGCR8 anchors at the ssRNA-dsRNA (double stranded RNA) junction and leads Drosha to cleave ≈11 bp away from the junction [27]. Thus, flanking ends of a hairpin are very important for production [28], but as it can be seen in Figure 1 and Figure 2, miRBase entries do not always have flanking ends. In our opinion, there are two problems with the hsa-mir-1178 entry on miRBase, one the terminal loop is rather small and thus the miRNA may not be effectively processed, and two the miRNA is located within the terminal loop.

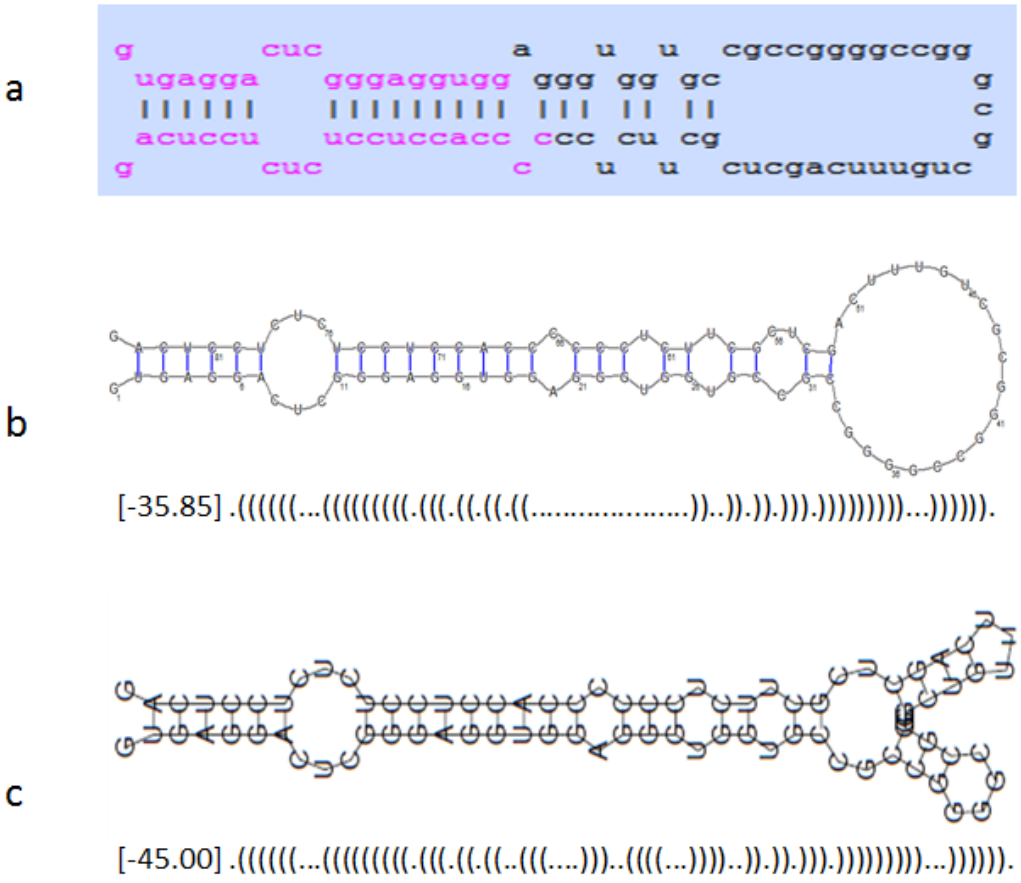

Figure 2: miRBase entry hsa-mir-1224 (MI0003764) showing stem-loop structure (a). Pink bases indicate mature miRNA sequences. Manually created dot-bracket structure and structure graph by RNAShapes (b). RNAFold prediction (c). Minimum free energy is given within brackets.

Therefore, this entry seems unlikely since Dicer cuts dsRNA and does not processes single stranded RNA and since terminal overhangs are usually around two nucleotides of length. In

addition, Drosha usually cuts an RNA double strand 11 base-pairs away from a ssRNA-dsRNA junction. For this, flanking ends of the hairpin are needed which are not given in this entry. Furthermore, the location of the mature miRNA is not conforming to the Drosha cleavage pattern. Thus we cannot explain the model in Figure 1 in light of the canonical miRNA pathway.

The entry in Figure 2-a contains the mature sequences at 3' and 5' ends of the hairpin which indicates missing flanking ends. Furthermore, the terminal overhang produced by Drosha is not respected (although blunt cutting has been observed). Another problem about this entry is that, when we draw the structure graph for this hairpin sequence by using RNAfold [9], we obtained the structure in Figure 2-c with a minimum free energy (mfe) value of -45.00, but the structure shown in miRBase is very different (Figure 2-a). To compare mfe values, we created a dot bracket (the format used to represent, transfer, and share RNA secondary structures) representation from the miRBase structure manually and drew it by using RNAShapes [29] (Figure 2-b). RNAeval from Vienna package [30] was then used to calculate mfe values for both structures.

We believe that the fold presented in Figure 2-a should not be in miRBase as it would mislead users to believe that it is the best fold while it is significantly worse than the fold in Figure 2-c. It seems unlikely that a mature miRNA can be created from the structure in Figure 2-a following the canonical pathway. It may be a different small regulatory RNA and should be marked as such. We cannot assess whether the Drosha cleavage was conformant since the flanking ends are missing but the Dicer cleavage may have been produced according to the canonical pathway.

Figure 3-a shows an example which has neither terminal loop nor flanking ends and therefore is unlikely to be a true miRNA. Figure 3-b further confirms this assumption as the structure that is predicted by RNAFold is significantly different and displays something that does not seem to be processable by the canonical miRNA pathway.

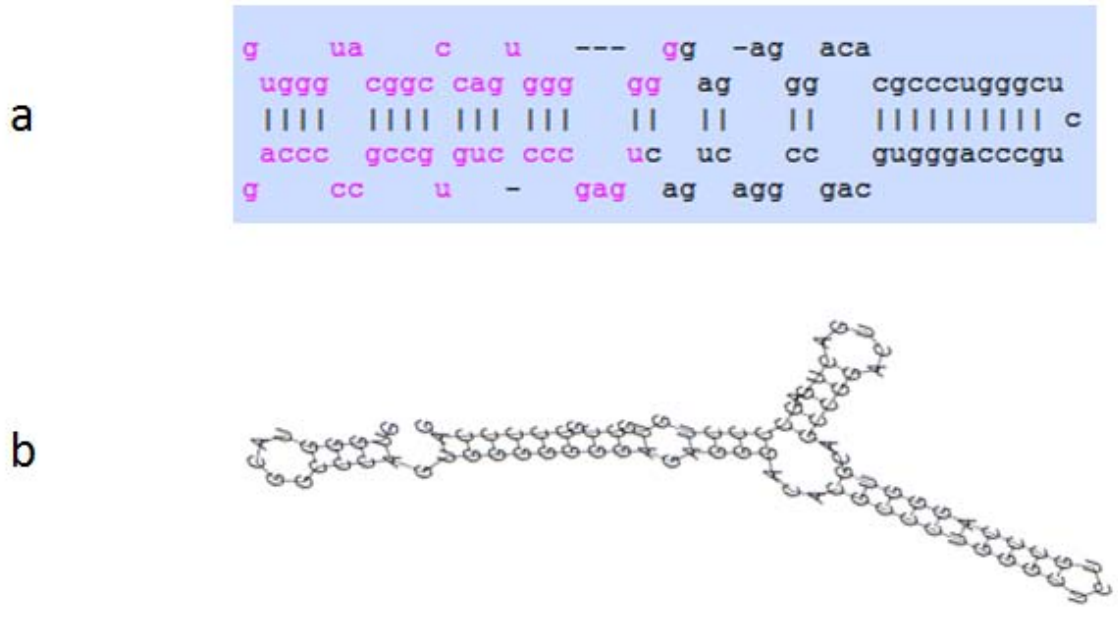

Figure 3: MiRBase entry for hsa-mir-1225 (MI0006311) showing the stem-loop structure as presented by miRBase (a). Pink bases indicate mature sequences. The hairpin structure has neither terminal loop nor flanking ends. Using RNAFold, a completely different structure is determined (b).

The three examples we give above a few of many we were able to pick out. Unfortunately, providing statistics is not possible since the folds stored in miRBase are not accessible in dot bracket notation making an automated comparison impossible. If we are right in assuming that the above examples and further examples that are not shown are not true microRNAs, then this should impact the classification accuracy for any miRNA hairpin detection method. We analyze this in the following section.

3.2 Classification Accuracy Comparison

In order to analyze whether entries in miRBase that are falsely annotated as miRNAs have an impact on classification accuracy, we randomly selected 180 entries as a positive dataset. A competing, equally sized, positive dataset was gleaned from miRTarBase. Both datasets were paired with a negative dataset consisting of 180 pseudo miRNAs from the Ng and Mishra dataset. Orange Canvas, a data mining software with graphical user interface, was used for training of three classifiers and for performance evaluation via 10 fold cross validation. The results of classifications for all three employed classifiers show that using the miRNAs having strong experimental evidence to interact with an mRNA as positive dataset (miRTarBase) provides a higher sensitivity, specificity, and classification accuracy (CA). Using positive examples derived from miRBase on the other hand leads to lower statistics (Table 1).

Table 1: Classification of miRNA hairpins and pseudo hairpins using Orange Canvas. SVM, Naïve Bayes, and Logistic Regression were used to compare performance of miRBase as positive data and miRTarBase as positive data.

	miRBase Entries			miRTarBase Entries		
Classifier	**Sensitivity**	**Specificity**	**CA**	**Sensitivity**	**Specificity**	**CA**
SVM	0.85	0.86	0.85	0.94	0.92	0.93
Naïve Bayes	0.90	0.82	0.86	0.93	0.90	0.91
Logistic Regression	0.91	0.92	0.92	0.93	0.94	0.94

The classifiers trained on the miRTarBase examples consistently outperform the classifiers based on miRBase examples. This is very obvious for support vector machine (SVM) and naïve Bayes but less so for logistic regression which seems to be able to somewhat compensate for false positive examples. The biggest gain is in the area of specificity which is in accordance with our assumption that some of the miRBase entries are falsely classified as miRNAs.

3.3 Other Observations

When we prepared a full miRNA dataset for human from all information downloadable from miRBase, we realized that 390 sequences were not properly mapped to the Ensembl human genome browser. Also, about 460 entries had only one entry for mature sequence which can be attributed to the assumption that only one mature sequence will be produced. This is however being disputed currently. About 40 entries seem to have more than one stem loop structure when predicted by RNAFold. Folds with more than 1 stem loop structure are probably rare in nature and thus should be signified as a possibly false positive miRNAs.

Another observation we made is that although some mature miRNAs are completely identical, they stem from different hairpins (Figure 4).

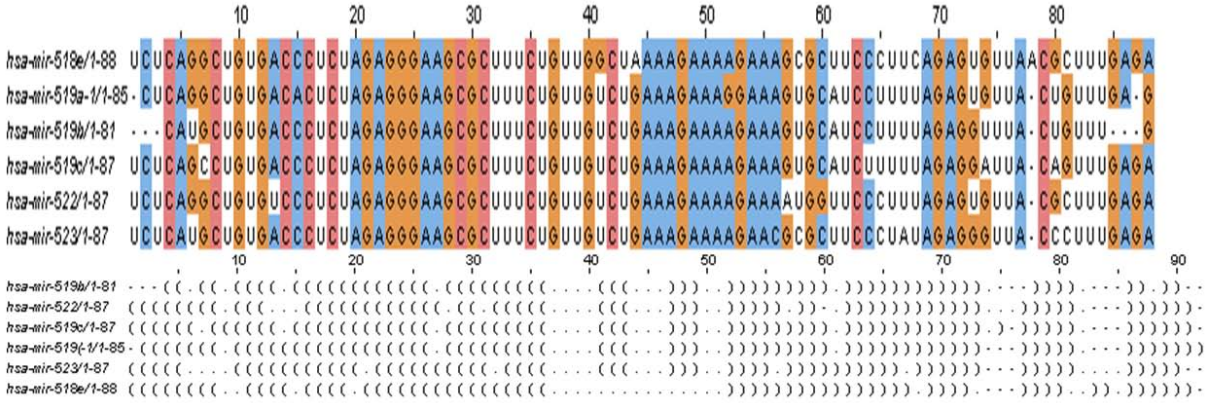

Figure 4: Multiple sequence alignment, performed by ClustalW [31] **and visualized by JalView** [32]**, of six miRBase entries which lead to the same mature miRNA. The structural alignment below shows that although the sequence alignment is highly conserved, the structures are still quite different with hsa-mir-518 being the most divergent.**

A comparison of the multiple sequence alignment of the sequences leads us to believe, however, that these may not really be different hairpins but that these may be due to sequencing errors or incomplete sequence submissions. It would be beneficial to add information about the genomic mapping of the miRNAs so that it can immediately be understood whether they are mapped to the same locus or not. This would remove any ambiguity. Although the sequence alignment is quite conserved, one of the structures is strikingly different which we cannot easily explain. It is interesting, that all models map to different Ensembl accessions which are actually retrievable. Here a mapping score or alignment of the hairpin to its location within the genome would be beneficial.

4 Conclusion

MiRBase claims experimental evidence for most entries and we do not want to dispute the fact that there may be experimental evidence. However, our concern is that entries annotated as miRNAs are not true miRNAs but likely represent other small regulatory RNA sequences which may lead to the same effect.

These falsely as miRNA annotated entries may impact prediction accuracy by as much as 8% as we were able to show above. In respect to the human genome and hundreds of millions putative hairpins, 8% is an error which is intolerable so we have to negate our initial question and we have to conclude that currently miRBase should not be used unfiltered for training of miRNA hairpin detection algorithms.

Other small errors in the database, like missing or wrong mapping to Ensemble, makes it difficult to extract surrounding sequences to retrieve the flanking ends and to calculate stability information about the miRNA hairpin within a larger sequence context.

Furthermore, in respect to currently available systems in biology, it is not easily understandable why not all data can be accessed somehow like for example the hairpin structures shown in miRBase as images. Here it would be essential to be able to retrieve the folds with their minimum free energies in some standard representation like the dot bracket representation. Moreover, it would be beneficial to enable grouping of sequences based on the experimental procedure they were confirmed by such as deep sequencing, cloning, or computational predictions in order to increase transparency.

MiRBase is the most comprehensive collection of miRNAs and it is essential in the field of miRNA research. The problems that we have shown can easily be remedied by implementing

a few automated processes which check all submissions. First of all, structures shouldn't be accepted as user submission, but be calculated automatically by miRBase which would ensure that the structure with the lowest minimum free energy will be displayed. MiRBase should determine a number of criteria which a miRNA hairpin must satisfy for example flanking ends, a minimum and maximum loop size, etc. Then an automated process could check whether a submission adheres to the assumptions and if not either bounce it back to the submitter or store it in a special area of miRBase for unsure submissions. Links and accession numbers to other database should periodically be checked for their validity and be updated if they are not valid any more. Since our previous emails have remained unanswered, we would recommend adding of a bug tracking system which would allow users to submit errors or issues to the miRBase developers. This would add a layer of transparency to the end-user who would be able to see how their requests are handled.

Acknowledgements

The study was in part supported by an award for outstanding young scientists from the Turkish Academy of Sciences (TÜBA, www.tuba.gov.tr).

References

[1] I. Bentwich, "Prediction and validation of microRNAs and their targets," *FEBS Lett.*, vol. 579, no. 26, pp. 5904–5910, Oct. 2005.

[2] 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*, vol. 23, no. 11, pp. 1321–30, Jun. 2007.

[3] J. Allmer and M. Yousef, "Computational methods for ab initio detection of microRNAs.," *Frontiers in genetics*, vol. 3, p. 209, Jan. 2012.

[4] M. D. Saçar and J. Allmer, "Comparison of four Ab Initio MicroRNA Prediction Tools," in *4th International Conference on Bioinformatics Models, Methods and Algorithms*, 2013, p. (accepted).

[5] L. Li, J. Xu, D. Yang, X. Tan, and H. Wang, "Computational approaches for microRNA studies: a review.," *Mammalian genome : official journal of the International Mammalian Genome Society*, vol. 21, no. 1–2, pp. 1–12, Feb. 2010.

[6] J. Ding, S. Zhou, and J. Guan, "MiRenSVM: towards better prediction of microRNA precursors using an ensemble SVM classifier with multi-loop features.," *BMC Bioinformatics*, vol. 11 Suppl 1, no. Suppl 11, p. S11, Jan. 2010.

[7] S. Griffiths-Jones, "miRBase: microRNA sequences and annotation," *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al*, vol. Chapter 12, p. Unit 12.9.1–10, Mar. 2010.

[8] 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*, vol. 39, no. Database issue, pp. D163–9, Jan. 2011.

[9] I. Hofacker, W. Fontana, P. Stadler, L. Bonhoeffer, M. Tacker, and P. Schuster, "Fast Folding and Comparison of RNA Secondary Structures," *Monatshefte für Chemie*, vol. 125, no. 2, pp. 167–188, Feb. 1994.

[10] J. Stalker, B. Gibbins, P. Meidl, J. Smith, W. Spooner, H.-R. H. R. Hotz, and A. V. A. V. Cox, "The Ensembl Web site: mechanics of a genome browser.," *Genome Res.*, vol. 14, no. 5, pp. 951–5, May 2004.

[11] P. Flicek, I. Ahmed, M. R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, L. Gil, C. García-Girón, L. Gordon, T. Hourlier, S. Hunt, T. Juettemann, A. K. Kähäri, S. Keenan, M. Komorowska, E. Kulesha, I. Longden, T. Maurel, W. M. McLaren, M. Muffato, R. Nag, B. Overduin, M. Pignatelli, B. Pritchard, E. Pritchard, H. S. Riat, G. R. S. Ritchie, M. Ruffier, M. Schuster, D. Sheppard, D. Sobral, K. Taylor, A. Thormann, S. Trevanion, S. White, S. P. Wilder, B. L. Aken, E. Birney, F. Cunningham, I. Dunham, J. Harrow, J. Herrero, T. J. P. Hubbard, N. Johnson, R. Kinsella, A. Parker, G. Spudich, A. Yates, A. Zadissa, and S. M. J. Searle, "Ensembl 2013.," *Nucleic acids research*, Nov. 2012.

[12] E. C. Lai, P. Tomancak, R. W. Williams, and G. M. Rubin, "Computational identification of Drosophila microRNA genes," *Genome Biol*, vol. 4, no. 7, p. R42, 2003.

[13] I. Bentwich, "Identifying human microRNAs.," *Current Topics In Microbiology And Immunology*, vol. 320, pp. 257–69, Jan. 2008.

[14] P. Jiang, H. Wu, W. Wang, W. Ma, X. Sun, and Z. Lu, "MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features," *Nucleic Acids Res.*, vol. 35, no. Web Server issue, pp. W339–344, Jul. 2007.

[15] S. Pfeffer, A. Sewer, M. Lagos-Quintana, R. Sheridan, C. Sander, F. A. Grässer, L. F. van Dyk, C. K. Ho, S. Shuman, M. Chien, J. J. Russo, J. Ju, G. Randall, B. D. Lindenbach, C. M. Rice, V. Simon, D. D. Ho, M. Zavolan, and T. Tuschl, "Identification of microRNAs of the herpesvirus family," *Nat. Methods*, vol. 2, no. 4, pp. 269–276, Apr. 2005.

[16] C. Xue, F. Li, T. He, G.-P. Liu, Y. Li, and X. Zhang, "Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine," *BMC Bioinformatics*, vol. 6, p. 310, 2005.

[17] A. Grundhoff, "Computational prediction of viral miRNAs," *Methods in Molecular Biology (Clifton, N.J.)*, vol. 721, pp. 143–152, 2011.

[18] A. van der Burgt, M. W. J. E. Fiers, J.-P. Nap, and R. C. H. J. van Ham, "In silico miRNA prediction in metazoan genomes: balancing between sensitivity and specificity.," *BMC genomics*, vol. 10, p. 204, Jan. 2009.

[19] M. V. Cakir and J. Allmer, “Systematic computational analysis of potential RNAi regulation in Toxoplasma gondii,” in *Health Informatics and Bioinformatics (HIBIT), 2010 5th International Symposium on*, 2010, pp. 31–38.

[20] W. Ritchie, D. Gao, and J. E. J. Rasko, “Defining and providing robust controls for microRNA prediction.,” *Bioinformatics (Oxford, England)*, vol. 28, no. 8, pp. 1058–61, Apr. 2012.

[21] Y. Xu, X. Zhou, and W. Zhang, “MicroRNA prediction with a novel ranking algorithm based on random walks.,” *Bioinformatics*, vol. 24, no. 13, pp. i50–8, Jul. 2008.

[22] T. Curk, J. Demsar, Q. Xu, G. Leban, U. Petrovic, I. Bratko, G. Shaulsky, and B. Zupan, “Microarray data mining with visual programming.,” *Bioinformatics*, vol. 21, no. 3, pp. 396–8, 2005.

[23] Y. Feng, X. Zhang, Q. Song, T. Li, and Y. Zeng, “Drosha processing controls the specificity and efficiency of global microRNA expression,” *Biochimica Et Biophysica Acta*, Jun. 2011.

[24] I. J. MacRae, K. Zhou, F. Li, A. Repic, A. N. Brooks, W. Z. Cande, P. D. Adams, and J. A. Doudna, “Structural basis for double-stranded RNA processing by Dicer.,” *Science (New York, N.Y.)*, vol. 311, no. 5758, pp. 195–8, Jan. 2006.

[25] X. Zhang, “The terminal loop region controls microRNA processing by Drosha and Dicer,” *Nucleic acids research*, vol. 38, no. 21, pp. 7689–7697, 2010.

[26] J. Han, J. S. Pedersen, S. C. Kwon, C. D. Belair, Y. Kim, K. Yeom, W. Yang, D. Haussler, R. Blelloch, and V. N. Kim, “Posttranscriptional Crossregulation between Drosha and DGCR8,” *Cell*, vol. 136, no. 1, pp. 75–84, 2009.

[27] J. Han, Y. Lee, K.-H. Yeom, J.-W. Nam, I. Heo, J.-K. Rhee, S. Y. Sohn, Y. Cho, B.-T. Zhang, and V. N. Kim, “Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex.,” *Cell*, vol. 125, no. 5, pp. 887–901, Jun. 2006.

[28] Y. Zeng and B. R. Cullen, “Efficient processing of primary microRNA hairpins by Drosha requires flanking nonstructured RNA sequences.,” *The Journal of biological chemistry*, vol. 280, no. 30, pp. 27595–603, Jul. 2005.

[29] P. Steffen, B. Voss, M. Rehmsmeier, J. Reeder, and R. Giegerich, “RNAshapes: an integrated RNA analysis package based on abstract shapes.,” *Bioinformatics (Oxford, England)*, vol. 22, no. 4, pp. 500–3, Feb. 2006.

[30] I. L. Hofacker, “Vienna RNA secondary structure server,” *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3429–3431, Jul. 2003.

[31] M. Goujon, H. McWilliam, W. Li, F. Valentin, S. Squizzato, J. Paern, and R. Lopez, “A new bioinformatics analysis tools framework at EMBL-EBI.,” *Nucleic acids research*, vol. 38, no. Web Server issue, pp. W695–9, Jul. 2010.

[32] A. M. Waterhouse, J. B. Procter, D. M. A. Martin, M. Clamp, and G. J. Barton, "Jalview Version 2--a multiple sequence alignment editor and analysis workbench.," *Bioinformatics (Oxford, England)*, vol. 25, no. 9, pp. 1189–91, May 2009.