Comparison of four *Ab Initio* MicroRNA Prediction Tools

Müşerref Duygu Saçar[1] and Jens Allmer[1]
[1]*Molecular Biology and Genetics, Izmir Institute of Technology,Urla, Izmir, Turkey*
duygusacar@gmail.com, jens@allmer.de

Keywords: miRNA, *ab initio*, miRNA prediction, comparison

Abstract: MicroRNAs are small RNA sequences of 18-24 nucleotides in length, which serve as templates to drive post transcriptional gene silencing. The canonical microRNA pathway starts with transcription from DNA and is followed by processing by the Microprocessor complex, yielding a hairpin structure. This is then exported into the cytosol where it is processed by Dicer and next incorporated into the RNA induced silencing complex. All of these biogenesis steps add to the overall specificity of miRNA production and effect. Unfortunately, experimental detection of miRNAs is cumbersome and therefore computational tools are necessary. Homology-based miRNA prediction tools are limited by fast miRNA evolution and by the fact that they are template driven. *Ab initio* miRNA prediction methods have been proposed but they have not been analyzed competitively so that their relative performance is largely unknown. Here we implement the features proposed in four miRNA *ab initio* studies and evaluate them on two data sets. Using the features described in Bentwich 2008 leads to the highest accuracy but still does not provide enough confidence into the results to warrant experimental validation of all predictions in a larger genome like the human genome.

1 INTRODUCTION

MicroRNAs (miRNAs) are a group of small noncoding RNAs, discovered in the early 90s by Ambrost and colleagues (Lee et al., 1993), which convey posttranscriptional regulation. In most cases miRNAs lead to down regulation of their target mRNAs but translational activation has also been observed (Ørom et al., 2008). It has been estimated that 60% of all human genes are regulated by miRNAs (Friedman et al., 2009). Another estimate is that there are more than 1000 miRNAs in the human genome (Berezikov et al., 2005). MiRNAs can come from introns (Morlando et al., 2008), coding regions (Rodriguez et al., 2004), or intergenic miRNA gene clusters (Altuvia et al., 2005). It has been suggested that a miRNA may regulate hundreds of targets (Enright et al., 2003). MiRNAs can, therefore, form complex regulatory networks. Not surprisingly, miRNAs are implicated in diseases such as cardiovascular disease (Elton et al., 2011) and cancer (Suzuki and Miyazono, 2011). The biogenesis of miRNAs follows largely the canonical pathway. Initially, DNA is transcribed into RNA by either RNA polymerase II (Lee et al., 2004) or III (Borchert et al., 2006) and then the microprocessor complex (Han et al., 2006) cleaves hairpin structures from the transcript. These hairpins are exported into the cytosol by Exportin 5 (Lund et al., 2004; Zeng and Cullen, 2004; Okada et al., 2009) where they are cleaved by Dicer (Cifuentes et al., 2010) and then loaded onto the RNA induced silencing complex

Despite the great effort that has been put into the elucidation of the miRNA pathway, not much is known which would facilitate computational modeling that is based on clear processing facts instead of data mining approaches. Two approaches are available for miRNA prediction, one based on homology and the other free of any references named *ab initio*.

While homology-based methods seem straight forward, they only retrieve results similar to already known miRNAs and rarely allow the detection of novel miRNAs (Bentwich et al., 2005). Furthermore, miRNA evolution progresses at a high rate (Liang and Li, 2009; Lu et al., 2008), which limits the applicability of homology-based methods.

Ab initio prediction methods (Ng and Mishra, 2007; Lai et al., 2003; Bentwich, 2008; Ding et al., 2010; Jiang et al., 2007; Pfeffer et al., 2005; Xue et al., 2005; Yousef et al., 2006; Grundhoff, 2011; Burgt et al., 2012; Cakir and Allmer, 2010; Ritchie et al., 2012) try to extract parameters which describe

hairpin structures, an element which is deemed important in the miRNA genesis process, and uses these features for machine learning to detect miRNAs. Although many of the *ab inito* algorithms that have been proposed report their accuracy, they cannot be easily compared as they are run on different data sets. Furthermore, most of the algorithms cannot be obtained. Therefore, we implemented all features described in Ding et al., Jiang et al, Ng and Mishra, and Bentwich (Ng and Mishra, 2007; Bentwich, 2008; Ding et al., 2010; Jiang et al., 2007) and compared them on the same data set to investigate relative algorithm performance. Ding et al. performed best on the data sets which we prepared but the maximum achieved accuracy of 0.996 would produce too many false positives so that experimental validation of all predictions would not be cost and time effective. Furthermore, this high accuracy seems to be an outlier and in all other cases that we tested Bentwich 2008 outperforms Ding et al. 2010 with a maximum accuracy of 0.986 that is closely reproduced among data sets. Therefore, we advise the use of the features described in Bentwich 2008 when attempting *ab initio* hairpin prediction.

2 MATERIALS AND METHODS

The four tools which we wanted to compare are not available as software so we implemented all features that were proposed in the papers in Java™ and used our code to calculate the values from the negative and positive data sets.

2.1 Features

The features that are used to discriminate between a true miRNA hairpin and a pseudo one are different among studies, but sometimes the feature is the same and just the naming differs. In the following we list the features that were used in the studies compared here and add the acronym that we gave to the feature in parentheses. Features are also summarized in Table 1 for ease of reference.

2.1.1 Features used in Ng and Mishra 2007

Sequence based features; 16 dinucleotide frequencies %NN (%AA, %AC, %AG, %AU, %CA, %CC, %CG, %CU, %GA, %GC, %GG, %GU, %UA, %UC, %UG, %UU) and 1 aggregate dinucleotide frequency %G+C (%G++%C).

Probability based features derived from dinucleotide shuffling (dns); adjusted base pairing propensity dP (dns_p(bpp), dns_p(bpp/hpl)), adjusted Minimum Free Energy of folding (MFE) dG (dns_p(hpmfe_rf), dns_p(hpmfe_rf/hpl)), MFE index 1 MFEI1 (hpmfe_rf_I1), adjusted base pair distance dD (dns_p(bpd), dns_p(bpd/hpl)), adjusted shannon entropy dQ (dns_p(Q), dns_p(Q/hpl)),MFE index 2 MFEI2 (hpmfe_rf/ns), degree of compactness dF (dc) and normalized variants of dP, dG, dQ, dD and dF (dns_z(bpp), dns_z(bpp/hpl), dns_z(hpmfe_rf), dns_z(hpmfe_rf/hpl), dns_z(Q), dns_z(bpd/hpl), dns_z(bpd), dns_z(Q/hpl)).

2.1.2 Features used in Jiang et al., 2007

Structural features; 32 triplet elements i.e. U(((', 'A((.', etc. [*U(((, *U..., *U.((, *U(.., *U(.(, *U..(, *U..., *U.(., *C(((, *C(.(, *C..(, *C((., *C.(., *C(.., *C.((, *C..., *A(((, *A..., *A(.., *A(.(, *A.(., *A..(, *A.((, *A((., *G(((, *G(.(, *G((., *G(.., *G.((, *G.(., *G..(, *G...)], mfe (hpmfe_rf/ns) and P-value (dns_p(hpmfe_rf)).

2.1.3 Features used in Bentwich, 2008

Structural features; hairpin length (hpl), loop length (hll), free energy per nucleotide (hpmfe_rf/hpl), matching base pairs (bpp) and maximal bulge size (mbs).

Sequence based features; abundance of any dinucleotide, AA, AT, etc. (#AA, #AC, #AG, #AU, #CA, #CC, #CG, #CU, #GA, #GC, #GG, #GU, #UA, #UC, #UG, #UU), regular internal repeat (dr), inverted internal repeat (ir), free energy (hpmfe_rf) and GC content (%GC).

2.1.4 Features used in Ding et al. 2010

Structural features; triplet elements A(((, A…, U(((, U(.(, U…, G(((, C(((, C(.(, [*A(((, *A…, *U(((, *U(.(, *U…, *G(((, *C(((, *C(.(].

Sequence based features; base pairing propensity dP (bpp), dP/n_loops (bpp/nl), Avg_bp_stem (bpp/sl), diversity (bpd), |A-U|/L (%AU), |G-C|/L (%GC), %(A-U)/n_loops (st(A-U)/ns), %(G-C)/n_loops (st(G-C)/ns).

Thermodynamics based features; ensemble free energy NEFE (efe), minimum free energy index 1-4 MFEI1 (hpmfe_rf_I1), MFEI2 (hpmfe_rf/ns), MFEI3 (hpmfe_rf/ns/hpl), MFEI4 (hpmfe_rf/hpl), dG (dG), Diff (dme), ensemble frequency Freq (efq), melting temperature Tm (Tm), enthalpy divided by length dH/L (dH/hpl), entropy divided by length dS/L (dS/hpl), Tm/L (Tm/hpl), p-value_MFE (dns_p(hpmfe_rf)), p-value_EFE (dns_p(efe)), z-

score_MFE (dns_z(hmpfe_rf)), z-score_EFE (dns_z(efe)).

Studies	Sequence-Based	Probability-Based	Structural	Thermodynamic-Based
Ng and Mishra 2007	16 dfs %NN and 1 aggregate df %G+C ratio	dP, dG, MFEI1, dD, dQ, MFEI2, dF, normalized variants of dP, dG, dQ, dD and dF		
	%AA-%UU, %G+C	dns_p(bpp, bpp/hpl, hpmfe_rf, hpmfe_rf/hpl, bpd, bpd/hpl, Q, Q/hpl), dns_z(bpp, bpp/hpl, hpmfe_rf, hpmfe_rf/hpl, bpd, bpd/hpl, Q, Q/hpl), hpmfe_rf_I1, hpmfe_rf/ns, dc		
Jiang et al. 2007			32 triplet elements, mfe, P-value	
			*A… - *U(((, (hpmfe_rf/ns), dns_p(hpmfe_rf)	
Bentwich 2008	Dinucleotide abundance, regular internal repeats, inverted internal repeats, mfe, GC content		hairpin length, loop length, free energy per nucleotide, matching base pairs, maximal bulge size	
	#AA - #UU, dr, ir, (hpmfe_rf), %GC		hpl, hll, (hpmfe_rf/hpl), bpp, mbs	
Ding et al. 2010	base pairing propensity (dP), Avg_bp_stem, diversity, |A-U|/L, |G-C|/L, %(A-U)/n_loops, %(G-C)/n_loops		triplet elements	NEFE, MFEI1, MFEI2, MFEI3, MFEI4, dG, Freq, Tm, dH/L, dS/L, Tm/L, p-value_MFE, p-value_EFE, z-score_MFE, z-score_EFE
	bpp, (bpp/nl), (bpp/sl), bpd, %AU, %GC, (st(A-U)/ns), (st(G-C)/ns)		*A(((, *A…, *U(((, *U(.(, *U…, *G(((, *C(((, *C(.(	efe, hpmfe_rf_I1, hpmfe_rf/ns, hpmfe_rf/ns/hpl, dG, dme, efq, Tm, dH/hpl, dS/hpl, Tm/hpl, dns_p(hpmfe_rf, efe), dns_z(hpmfe_rf, efe)

Table 1: Features that were proposed in the selected studies are presented in the first row of the respective study and the acronyms we chose for those features are presented in the following row.

2.2 Data Sets

2.2.1 Positive Data Set

MirBase is the *de facto* standard repository for miRNAs (Griffiths-Jones et al., 2008). It contains about 1500 entries for human counting both guide and passenger strands. We downloaded all human miRNAs as positive data. From the entries we removed the ones that contain more than one hairpin when folded by RNAFold (Hofacker, 2003) or RNAShapes (Steffen et al., 2006). If no proper link to Ensemble could be established, the entries were removed as well. From the remaining, about 1000, miRNA examples we created five random subsets containing 500 positive examples each.

2.2.2 Negative Data Sets

Negative data sets are especially difficult to establish for miRNAs, experimentally or computationally (Ding et al., 2010; Ritchie et al., 2012; Wu et al., 2011; Yousef et al., 2008). Since most machine learning approaches that have been proposed for *ab initio* miRNA prediction are built on two class classification we designed one data set which

consists of random sequences of the same length as the selected miRNAs in the positive data set. We consider this data set to be easy to solve. Another data set, the pseudo miRNA data set, was taken from the Ng and Mishra study and we consider it to be more difficult (Ng and Mishra, 2007).

2.3 Machine Learning

We created five combined data sets consisting of 60% training and 40% test data from the overall data set. These data sets were used to train and test SVM classification using default settings in Orange Canvas (http://orange.biolab.si/). For performance evaluation test on 'test data' was used. This approach was used since fivefold cross validation could not be used with multiple studies at the same time; but had to be repeated individually, thus leading to different data sets and therefore to a potentially unfair comparison.

3 RESULTS AND DISCUSSION

In our opinion, the two datasets that were used (see Data Sets) are of different difficulty with the random dataset being easier to solve than the pseudo miRNA data set. This can also be deduced from the best results reported in Tables 2 and 3. The results for the random miRNAs in Table 2 lead to higher accuracy than the data in Table 3 which is achieved with pseudo miRNAs. For both tables the best and the average accuracy are provided along with the standard deviation, calculated from fivefold cross validation. Using the features described in the study by Bentwich 2008 leads to the highest accuracy without any standard deviation. Judging from these results it seems trivial to discriminate between true and false hairpins when using properly selected features.

Table 2: Accuracy measurements for human miRNAs (positive dataset) and random miRNAs (negative dataset).

Studies	Accuracy Values		
	Best	Average	Standard Deviation
Ng and Mishra 2007	0.919	0.894	0.183
Bentwich 2008	1.000	1.000	0
Ding et al. 2010	1.000	0.676	0.217
Jiang et al. 2007	0.954	0.952	0.003

We did not expect such a perfect result as achieved by Bentwich 2008 features in Table 2 for the pseudo miRNAs and Table 3 displays no such success. For the pseudo miRNA dataset, the features used in Ding et al. 2010 achieve the highest accuracy although with a high standard deviation over the cross validation.

Table 3: Accuracy measurements for human miRNAs (positive dataset) and pseudo miRNAs (negative dataset).

Studies	Accuracy Values		
	Best	Average	Standard Deviation
Ng and Mishra 2007	0.930	0.895	0.060
Bentwich 2008	0.986	0.983	0.002
Ding et al. 2010	0.996	0.599	0.198
Jiang et al. 2007	0.910	0.877	0.018

The best accuracy of 0.996, achieved by the features described in the study by Ding et al. 2010, when used for a million putative hairpins in human would lead to 4000 false positive identifications. Unfortunately, the number of putative hairpins in human is large and the accuracy calculated here does not fully reflect the true accuracy.

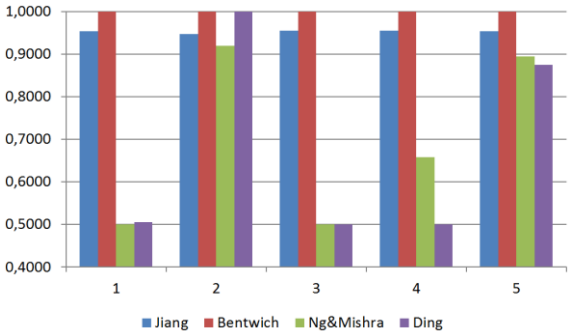

Figure 1: Accuracy measurements for human miRNAs (positive dataset) and random miRNAs (negative dataset). All cross validation results are shown individually.

This is due to the fact that it is not entirely known what differentiates a true from a false hairpin. A much higher false positive rate must therefore be expected for real data and thus the number of false positives may not allow costly experimental validation of all predicted miRNAs.

Figures 1 and 2 further support this point by showing that the accuracy strongly depends on the data set used for training and testing. This can be

deduced from the variation among the accuracies for the 5 data sets used in the fivefold cross validation.

For the random miRNAs, the variation among datasets is large for most studies. However, Bentwich 2008 always achieves perfect separation and Jiang et al. 2007 achieves a low variation and an overall good result (Figure 1).

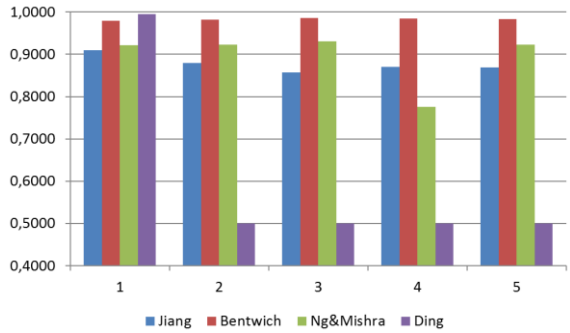

Figure 2: Accuracy measurements for human miRNAs (positive dataset) and pseudo miRNAs (negative dataset). All cross validation results are shown individually.

With the pseudo miRNAs the variation is even more important to be analysed. Although Ding et al. 2010 achieves the highest accuracy in one case, it fails in all other cases which shows a strong dependence on the training and test data set and a poor generalization for the features from that study (Figure 2). Bentwich 2008 does not have such generalization problems and outperforms all other studies on the remaining four data sets.

4 CONCLUSIONS

Although many algorithms have been proposed for *ab initio* miRNA gene prediction, they have not been compared for their relative performance. We compared four of twelve available *ab initio* algorithms in this study and found that Bentwich 2008 achieves the highest accuracy on the random data set and the second best accuracy on the pseudo hairpin data set but with a very low variation. Unfortunately, the achieved accuracy of 0.986 would lead to many false positives which would turn any attempt at experimental validation of all predicted miRNAs into a futile endeavor.

In the future, we plan to expand this assessment of available algorithms to all currently available ones. We believe it is necessary to establish the accuracy of existing algorithms not independently but transparently and comparable. To the best of our knowledge, this is the first independent assessment of multiple *ab initio* miRNA prediction methods.

As negative data sets are hard to come by, we will try to establish another set of negative data and further try one-class classification with the proposed parameters in follow-up studies.

Currently, we advise the use of the features used in the Bentwich 2008 study when trying *ab initio* prediction of hairpins.

ACKNOWLEDGEMENTS

This study was in part supported by an award received from the Turkish Academy of Sciences (http://www.tuba.gov.tr) for outstanding young scientists (TUBA GEBIP).

REFERENCES

Altuvia, Y., Landgraf, P., Lithwick, G., Elefant, N., Pfeffer, S., Aravin, A., Brownstein, M. J., Tuschl, T., and Margalit, H., 2005. Clustering and conservation patterns of human microRNAs. Nucleic acids research 33, 2697–706.

Bentwich, I., 2008. Identifying human microRNAs. RNA interference 320, 257–269.

Bentwich, I., Avniel, A., Karov, Y., Aharonov, R., Gilad, S., Barad, O., Barzilai, A., Einat, P., Einav, U., Meiri, E., et al., 2005. Identification of hundreds of conserved and nonconserved human microRNAs. Nature genetics 37, 766–70.

Berezikov, E., Guryev, V., van de Belt, J., Wienholds, E., Plasterk, R. H. A., and Cuppen, E., 2005. Phylogenetic shadowing and computational identification of human microRNA genes. Cell 120, 21–4.

Borchert, G. M., Lanier, W., and Davidson, B. L., 2006. RNA polymerase III transcribes human microRNAs. Nature Structural & Molecular Biology 13, 1097–1101.

Burgt, A. V. D., Fiers, M. W. J. E., Nap, J., and Van, R. C. H. J., 2012. In silico miRNA prediction in metazoan genomes : balancing between sensitivity and specificity. BMC Genomics 24, 1–24.

Cakir, M. V., and Allmer, J., 2010. Systematic computational analysis of potential RNAi regulation in Toxoplasma gondii. in Health Informatics and Bioinformatics, HIBIT), 2010 5th International Symposium on, Ankara, Turkey: IEEE), 31–38.

Cifuentes, D., Xue, H., Taylor, D. W., Patnode, H., Mishima, Y., Cheloufi, S., Ma, E., Mane, S., Hannon, G. J., Lawson, N. D., et al., 2010. A novel miRNA processing pathway independent of Dicer requires Argonaute2 catalytic activity. Science 328, 1694–1698.

Ding, J., Zhou, S., and Guan, J., 2010. MiRenSVM: towards better prediction of microRNA precursors using an ensemble SVM classifier with multi-loop features. BMC Bioinformatics 11 Suppl 1, S11.

Elton, T. S., Martin, M. M., Sansom, S. E., Belevych, A. E., Györke, S., and Terentyev, D., 2011. miRNAs got rhythm. Life Sciences 88, 373–383.

Enright, A. J., John, B., Gaul, U., Tuschl, T., Sander, C., and Marks, D. S., 2003. MicroRNA targets in Drosophila. Genome Biology 5, R1.

Friedman, R. C., Farh, K. K.-H., Burge, C. B., and Bartel, D. P., 2009. Most mammalian mRNAs are conserved targets of microRNAs. Genome Research 19, 92–105.

Griffiths-Jones, S., Saini, H. K., van Dongen, S., and Enright, A. J., 2008. miRBase: tools for microRNA genomics. Nucleic acids research 36, D154–8.

Grundhoff, A., 2011. Computational prediction of viral miRNAs. Methods in Molecular Biology, Clifton, N.J.) 721, 143–152.

Han, J., Lee, Y., Yeom, K.-H., Nam, J.-W., Heo, I., Rhee, J.-K., Sohn, S. Y., Cho, Y., Zhang, B.-T., and Kim, V. N., 2006. Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. Cell 125, 887–901.

Hofacker, I. L., 2003. Vienna RNA secondary structure server. Nucleic Acids Research 31, 3429–3431.

Jiang, P., Wu, H., Wang, W., Ma, W., Sun, X., and Lu, Z., 2007. MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. Nucleic Acids Research 35, W339–344.

Lai, E. C., Tomancak, P., Williams, R. W., and Rubin, G. M., 2003. Computational identification of Drosophila microRNA genes. Genome Biol 4, R42.

Lee, R. C., Feinbaum, R. L., and Ambrost, V., 1993. The C . elegans Heterochronic Gene lin-4 Encodes Small RNAs with Antisense Complementarity to & II-14. Cell 75, 843–854.

Lee, Y., Kim, M., Han, J., Yeom, K.-H., Lee, S., Baek, S. H., and Kim, V. N., 2004. MicroRNA genes are transcribed by RNA polymerase II. The EMBO Journal 23, 4051–4060.

Liang, H., and Li, W.-H., 2009. Lowly expressed human microRNA genes evolve rapidly. Molecular biology and evolution 26, 1195–8.

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

Lund, E., Güttinger, S., Calado, A., Dahlberg, J. E., and Kutay, U., 2004. Nuclear export of microRNA precursors. Science 303, 95–8.

Morlando, M., Ballarino, M., Gromak, N., Pagano, F., Bozzoni, I., and Proudfoot, N. J., 2008. Primary microRNA transcripts are processed co-transcriptionally. Nature Structural & Molecular Biology 15.

Ng, K. L. S., and Mishra, S. K., 2007. De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. Bioinformatics 23, 1321–30.

Okada, C., Yamashita, E., Lee, S. J., Shibata, S., Katahira, J., Nakagawa, A., Yoneda, Y., and Tsukihara, T., 2009. A high-resolution structure of the pre-microRNA nuclear export machinery. Science 326, 1275–1279.

Pfeffer, S., Sewer, A., Lagos-Quintana, M., Sheridan, R., Sander, C., Grässer, F. A., van Dyk, L. F., Ho, C. K., Shuman, S., Chien, M., et al., 2005. Identification of microRNAs of the herpesvirus family. Nature Methods 2, 269–276.

Ritchie, W., Gao, D., and Rasko, J. E. J., 2012. Defining and providing robust controls for microRNA prediction. Bioinformatics, Oxford, England) 28, 1058–61.

Rodriguez, A., Griffiths-Jones, S., Ashurst, J. L., and Bradley, A., 2004. Identification of mammalian microRNA host genes and transcription units. Genome Research 14, 1902–1910.

Steffen, P., Voss, B., Rehmsmeier, M., Reeder, J., and Giegerich, R., 2006. RNAshapes: an integrated RNA analysis package based on abstract shapes. Bioinformatics, Oxford, England) 22, 500–3.

Suzuki, H. I., and Miyazono, K., 2011. Emerging complexity of microRNA generation cascades. Journal of biochemistry 149, 15–25.

Wu, Y., Wei, B., Liu, H., Li, T., and Rayner, S., 2011. MiRPara: a SVM-based software tool for prediction of most probable microRNA coding regions in genome scale sequences. BMC Bioinformatics 12, 107.

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

Yousef, M., Jung, S., Showe, L. C., and Showe, M. K., 2008. Learning from positive examples when the negative class is undetermined--microRNA gene identification. Algorithms for molecular biology : AMB 3, 2.

Yousef, M., Nebozhyn, M., Shatkay, H., Kanterakis, S., Showe, L. C., and Showe, M. K., 2006. Combining multi-species genomic data for microRNA identification using a Naive Bayes classifier. Bioinformatics, Oxford, England) 22, 1325–34.

Zeng, Y., and Cullen, B. R., 2004. Structural requirements for pre-microRNA binding and nuclear export by Exportin 5. Nucleic acids research 32, 4776–85.

Ørom, U. A., Nielsen, F. C., and Lund, A. H., 2008. MicroRNA-10a binds the 5'UTR of ribosomal protein mRNAs and enhances their translation. Molecular cell 30, 460–71.