Computational and Bioinformatics Methods for miRNA Gene Prediction

Jens Allmer

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

Original Article

The original publication is available at http://link.springer.com/protocol/10.1007/978-1-62703-748-8_9.

Key Words

miRNA, secondary structure prediction, homology-based prediction, *ab initio* prediction, miRNA prediction accuracy, multiple sequence alignment-based prediction

Running Head

Bioinformatics Methods for miRNA Gene Prediction

Summary

MicroRNAs (miRNA) have attracted ever increasing interest in recent years. Since experimental approaches for determining miRNAs are non-trivial in their application, computational methods for the prediction of miRNAs have gained popularity. Such methods can be grouped into two broad categories 1) performing *ab initio* predictions of miRNAs from primary sequence alone, and 2) additionally employing phylogenetic conservation. Most methods acknowledge the importance of hairpin or stem loop structures and employ various methods for the prediction of RNA secondary structure. Machine learning has been employed in both categories with classification being the predominant method. In most cases, positive and negative examples are necessary for performing classification. Since it is currently elusive to experimentally determine all possible miRNAs for an organism, true negative examples are hard to come by and therefore the accuracy assessment of algorithms is hampered. In this chapter, first RNA secondary structure prediction is introduced since it provides a basis for miRNA prediction. This is followed by an assessment of homology and then *ab initio* miRNA prediction methods.

1. Introduction

Non coding RNAs (ncRNA) represent a large portion of the transcriptome and have recently received much attention ***(1)*** although the term ncRNA may not have been chosen well since many so called ncRNAs also lead to mRNAs ***(2)***. These ncRNAs have been grouped into families ***(3, 4)*** one of which, microRNAs, is the focus of this book. MiRNAs can originate from any part of a genome ***(5)*** and can lead to silencing of transcripts originating from anywhere in the genome. MiRNA genes' presence or their effects have been shown in many species and even viruses make use of miRNAs to regulate host and virus encoded genes ***(6)***.

There are at least two computational challenges: 1) the prediction of miRNAs in a genome and 2) the mapping of the miRNAs to likely targets. This chapter focuses on the prediction of miRNAs within a genome. Computational miRNA gene prediction can be grouped into several approaches. Generally, either homology modeling or machine learning is applied to extract likely miRNAs from a genome. Although homology modeling can glean information from already successfully established miRNAs from a related organism's genome, it is also limited since completely novel miRNAs cannot be determined in this way. Furthermore, miRNAs evolve quickly and very close homology is thus needed for successful miRNA gene prediction ***(7)***. Another approach, machine learning, is hampered in a similar manner but assumes that the examples for learning are derived from the organism in question. In the following, first miRNA gene prediction will be further explored followed by a brief discussion of RNA secondary structure prediction, a process vital to all approaches in miRNA gene prediction. Then homology-based miRNA gene prediction and *ab initio* gene prediction will be discussed.

2. miRNA Gene Prediction

Identification of miRNA genes is computationally challenging since a genome can be divided into millions of putative miRNAs of appropriate sequence length (e.g.: 80--200 nucleotides for pre-MiRNAs). Folding all these sequences *in silico* increases the complexity and may only be practical for small genomes. Furthermore, many hairpin structures can be found in the predictions and will thus lead to an abundance of putative miRNAs many of which may represent false positive results. An inherent problem to the experimental validation of miRNAs occurs because their expression may only happen in response to specific signals or at certain developmental stages ***(8)***. See chapters 13 and 14 in this volume or Bentwich 2005 for more

details on miRNA gene validation ***(9)***. In order to decrease the number of false positive results many filtering strategies have been developed and will be discussed later in this chapter. Since both, homology guided detection algorithms and *ab initio* miRNA gene prediction algorithms rely on the prediction of RNA secondary structure, a number of such tools shall be introduced first before the two miRNA gene prediction approaches are discussed in more detail.

2.1. RNA Secondary Structure Prediction

Prediction of RNA secondary structure is integral to many algorithms trying to find hairpins, also known as stem-loop structures and pre-miRNAs, which may give rise to miRNAs. In general, the prediction of secondary structure is much easier for shorter sequences, which means that the longer the sequence becomes the more difficult the prediction which is further reflected in exponentially increasing algorithm run time. Therefore, most algorithms which use secondary structure prediction resort to merely predicting the hairpin structure which is always contained in a sequence of less than 500 nucleotides which can successfully be folded in a short time. There are a number of algorithms which can be used for RNA secondary structure prediction (Table 1). The table is sorted by usage statistics not by successfulness of the algorithm. A recent paper has shown, however, that in the realm of predicting the secondary structure of short nucleotide sequences RNAfold seems to be most successful ***(10)***.

For both methods, the homology-based prediction of miRNA genes and their *ab initio* prediction, RNA structure prediction is vital. One feature of miRNAs is the stem loop structure which seems to be important for processing of the pre-miRNA into a mature miRNA with Drosha and Dicer. The homology-based prediction of miRNA genes is inherently simpler than their *ab initio* prediction and shall thus be discussed first.

2.2. Homology-Based miRNA Gene Prediction

In contrast to *ab initio* gene prediction where miRNA genes need to be found without additional knowledge, homology-based mapping methods can build on available and experimentally validated miRNAs and find similar structures and sequences in related species.

All software that enable mapping of a known miRNAs to homologous genomes take sequence similarity as well as RNA secondary structure into account (Table 2). The assumption is that a mature miRNA derives from a hairpin structure formed by folding its pre-miRNA. The approach taken by one of the most recent developments, MapMi ***(21)***, first scans the miRNA sequences against the target genome and then creates two potential pre-miRNAs from it. The ViennaRNA package ***(13)*** is used to fold the extracted RNA sequences. Finally, the results are scored, ranked and displayed. Both a web service with rich display facilities and a downloadable, local, version are freely available for this program which as the authors report achieves 92% sensitivity at 98% specificity.

Although mapping by homology is a straightforward approach, it can only reproduce results and cannot find new miRNA genes. Since many miRNAs are species specific these will always be missed by this method and therefore other strategies need to be used in tandem. Additionally, miRNA genes evolve very rapidly which further limits the applicability of homology-based methods ***(37, 38)***.

A recent study by Keshavan and colleagues pointed out that it is important to make sanity checks when constructing a computational pipeline for miRNA gene detection since in their case the temperature at which Ciona intestinalis operates is only 18 degrees Celsius while most folding

programs default to 37 degrees Celsius ***(39)***. They were able to confirm about half of their predictions by either microarray analysis or by the fact that the predicted hairpins were already in other databases.

The two aforementioned studies are just a small selection of the large amount of available studies but the following section aims to briefly summarize common approaches among different studies.

2.2.1. Methods Used in Homology-Based miRNA Detection

There are many ways to detect and filter hairpin structures and miRNAs. The list below is separated into two sections, the first one showing methods for hairpin/ miRNA detection and the second one listing methods used to filter/remove false positive identifications. The methods below are a non-comprehensive list and some methods may not be used synergistically while others can be combined. In general any algorithm used for homology detection of hairpins or miRNAs uses a combination of some of the methods in the list but no algorithm has been proposed that integrates most of the detection and filtering methods below.

Methods for initially detecting miRNAs

- Difference in evolutionary conservation
    - coding arm, non-coding arm, seed region
    - loop, stem flanking regions
    - effect on secondary structure
- Scanning for hairpin structures conserved in closely related species
    - Sliding window (70--110), folding sequences for each window

    - Level of expected similarity can be adjusted
- Windows with high sequence conservation (sometimes higher than for coding sequences) flanked by windows with high sequence variation
- Homology of the miRNA targets among genomes

Since studies have shown that excessive number of conserved hairpin structures can be found ***(40)***, additional criteria for their filtering need to be established, some of which are listed below.

Methods for filtering detected hairpins

- Varying level of sequence conservation within stem structure
- Using general properties of hairpin structures that can be learned from examples
- Repetitively detected structures are generally discarded
- Minimum free energy
- Length of stem loop structure
- If matching to certain annotation of a genome (e. g.: coding sequence) the detections may be discarded
- Base composition
- MiRNA gene clustering
- Upstream and downstream conserved regions surrounding miRNA genes
- Sequence entropy
- Identity with a multiple sequence alignment
- Position of mature sequence within hairpin structure

- Maximal internal loop and bulge sizes
- EST sequences can confirm that sequences are transcribed
- Text mining

2.2.2. Accuracy of Homology-Based miRNA Prediction

MirScan ***(27)*** has been applied to *Caenorhabditis elegans* and the predictions were validated experimentally setting the sensitivity to 0.50 at a specificity of 0.70 ***(27)***. The same study has also shown that many miRNAs are present at high levels, between 1000 and 50000 molecules per cell. Another study which also validated the predictions experimentally, studied the conservation among ten primate species and found that sequences representing stem loops are conserved whereas flanking regions and loop region are highly variable ***(30)***. The sensitivity of the method was reported at 0.83 but the specificity was not given. It may be rather low due to their prediction of 976 putative miRNAs where 179 were confirmed in miRNA databases and only 16 out of 69 predicted miRNAs have been confirmed via Northern Blot analysis. A study using two Drosophila species had a similar sensitivity (0.75) to the other studies presented above, but no value for the specificity was presented. 24 new miRNAs were, however, confirmed by Northern blotting ***(40)***. Huang and colleagues presented MirFinder which on their training and test set achieved an accuracy of 99.6% and had an area under the receiver operator characteristic (ROC) curve of almost 1 ***(41)***. They compared their ROC curve to those from other studies but this may be at best be misleading due to the usage of completely different training and test datasets. Artzi and colleagues set their filter specificity to 95%; they estimated the sensitivity of their algorithm at 88% (85% - 94% on seven mammalian species) ***(23)***. The content of the miROrtho database has been constructed with a hairpin prediction accuracy of 95%, yielding a

sensitivity of 84% at 97% specificity ***(24)***. They then filtered these hairpins by homology with an independent accuracy of 91%, but they do not report the overall accuracy measures. MapMi reports a sensitivity of 92% at a specificity of 98%. Wang and colleagues did not explicitly report on the accuracy of their algorithm but were able to confirm 67% of their predicted miRNAs by Northern blotting ***(33)***.

That the reported accuracies have to be viewed critically can be seen in a study by Leung and colleagues who found quite different sensitivities for ProMirII and miR-abela then the ones reported in their respective publications ***(21)***. They also report that they were able to increase the positive prediction value by more than 15% at high sensitivity. Since all accuracy measures reported above are derived from different studies, using different datasets, they are not integrated into a table for easy comparison since that would be misleading. In fact, the measures reported above can hardly be compared and are most likely highly optimistic. A study independently comparing these measures objectively needs yet to be done. Experimental validation may seem to actually proof the existence and effect of a miRNA but the opposite is not true so that these approaches can only be used to confirm the existence but never to prove the absence of a miRNA.

Two examples of algorithms for homology-based miRNA gene prediction will be presented as anecdotes in the next section.

2.2.3. Selected Examples Performing Homology-Based miRNA Gene Detection

Due to the large number of available miRNA gene prediction algorithms only the most cited one, MiRscan ***(27)***, and ProMiR II ***(29)*** will be discussed in some more detail followed by a more general statement about prediction accuracy.

ProMiR II

ProMiR was first introduced in 2005 as an algorithm that simultaneously considers structure and sequences of pre-miRNAs ***(42)***. A machine learning approach was used with positive examples from known human miRNAs and negative examples extracted quasi-randomly from the human genome. Their hidden Markov model includes both sequence and structure and predicts for each element of the sequence whether it is part of a pre-miRNA or not. The predicted pre-miRNAs are then further evaluated in regards to their minimum free energy and their conservation among vertebrates.

ProMiR II extends ProMiR by adding knowledge about miRNA gene clustering, G/C ratio conservation, and entropy of candidate sequences ***(29)***. Another improvement of ProMiR II is that different criteria are now implemented in modules making the approach very extensible. In addition to that, several databases are integrated into the analysis without need for user intervention. ProMiR II is a web server available at: http://cbit.snu.ac.kr/~ProMiR2/introduction.html. No values for specificity and sensitivity are reported but the provided ROC curve seems to have an area under the curve somewhere between 80% and 95% which is similar to other algorithms (see Table 2).

MiRscan II

MiRscan was first introduced in 2003 to find miRNA genes conserved between two species ***(27)***. Initially, a screen for hairpin structures conserved between two genomes is performed; afterward the hairpin structures are evaluated in respect to their features. Among these features, which are used to discriminate between true and false miRNA genes, are stringent base pairing in the miRNA:mRNA target duplex seed region, less stringent base pairing in the remaining structure, sequence bias in the first 5 bases, loop symmetry, and bulges.

MiRscan II ***(28)*** extends MiRscan by including the genomic sequence upstream of the miRNA gene into the analysis algorithm. In addition to general conservation for the miRNA gene flanking regions, at about 200 bp a conserved motif was observed. These findings and orthology of host genes for intronic RNA were incorporated into the new program. The new version supersedes MiRScan and is the one referenced on the web server (http://genes.mit.edu/burgelab/MiRscanII/).

2.3. *Ab initio* miRNA Gene Prediction

Ab initio miRNA gene prediction needs no other information than the primary sequence in order to determine whether it is a true miRNA. Two possible modes of operation are possible with one using multiple sequences and the other based on single sequences.

2.3.1. Multiple Sequence Based miRNA Gene Prediction

RNAmicro is an SVM-based classifier that enables detection of hairpin structures in multiple sequence alignments ***(43)***. The approach tries to balance sensitivity and specificity unlike most other approaches in miRNA detection which try to minimize the number of false positives. In their initial tests they achieved a sensitivity of 91% at a specificity of 99% which as they point

out cannot be achieved in a real dataset due to limitations of RNAz which places an optimistic upper bound of 80% sensitivity at 99% specificity on experiments with real data.

2.3.2. Single Sequence Based miRNA Gene Prediction

One important field of research is the detection of novel miRNA genes. While there are experimental methods (see chapters 1-3 and 6 in this volume) to perform this task like forward genetic screens and identification in small RNA libraries ***(44)*** as well as deep sequencing methods ***(45)***; also refer to ***(46)*** for approaches to identify ncRNAs ***(47)***. These methods are either time consuming, inefficient, or expensive. Therefore, it is necessary to also develop computational methods to predict miRNA genes that can be used in tandem with experimental strategies. Some of the current approaches are listed in Table 3. In general a mature miRNA should be derived from the stem part of a short hairpin RNA (shRNA) which should form a large number of Watson-Crick pairs and few internal loops and bulges (cf. Figure 1). Other criteria are, for instance, that the mature miRNA is conserved in closely related species (see Homology-based miRNA gene prediction). Presence of Drosha and Dicer in the organism and accumulation of relevant product in deficient mutants is an experimental validation for a miRNA. Thermodynamic stability of hairpins and similarity to known miRNAs can also serve as supporting evidence when predicting new miRNAs. This can, for example, be done by defining features of known miRNAs and training a classifier such as a support vector machine (SVM). Clustering of miRNA genes in a genomic locus can further support the validity of miRNA genes ***(48)***.

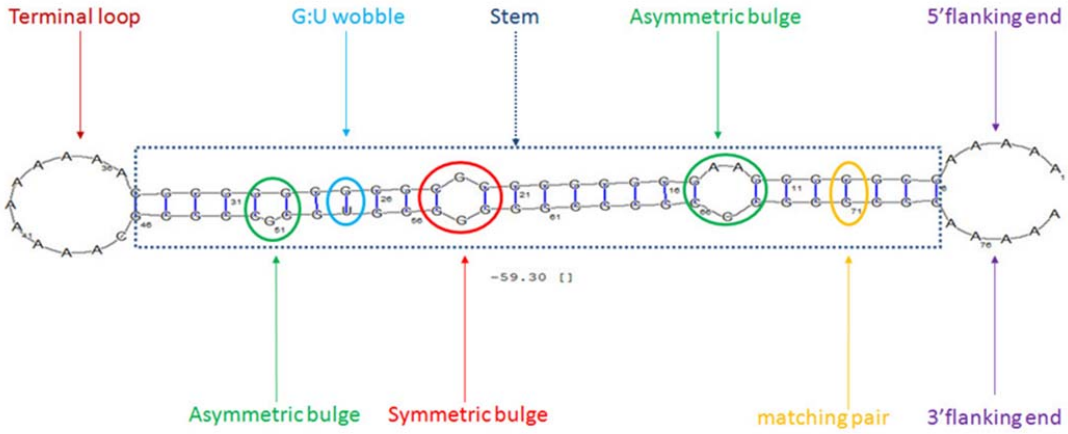

Figure 1*: The primary sequence for this hairpin was manually designed such that the selected elements were guaranteed to be present in one hairpin. The sequence was folded using RNAShapes.*

In a more systems driven approach the predicted mature miRNAs can be validated further by looking for targets in, for example 3'UTRs, and by evaluating the multiplicity of targets per miRNA and target sites per regulated mRNA (see chapters 12-14 in this volume). The non-comprehensive list of miRNA gene prediction programs and web servers in Table 3 contains algorithms using different strategies which are summarized in the following section for single sequence miRNA gene prediction.

Methods Used in Single Sequence Based miRNA Gene Prediction

- Proximity to known miRNA genes since miRNAs sometimes reside in clusters

- Varying level of sequence conservation within stem structure
- Using general properties of hairpin structures that can be learned from examples
- Minimum free energy threshold
- Length of stem loop structure threshold
- Base composition
- Local contiguous substructures paired with central sequence information of the substructure
- P-value derived from the predicted structure compared with randomized structures of the same sequence

Filtering Strategies

Obviously, it would be beneficial to include as much information as possible for discriminating false positive identification to increase prediction accuracy although the use of too many parameters can lead to over-training (see chapters 7 and 10 in this volume). For instance, sequence, structure and homology information can be used in tandem. Some of the information that can be used to distinguish true from false positive miRNA gene predictions are given below:

- miRNA genes are small noncoding genes (<150 nt)
    - miRNA length
        - varies between plants and animals
- originates from pre-miRNA (80--120 nt)
    - forms a characteristic hairpin structure
    - low free energy
    - sequence composition

        - G/C composition varies between plants and animals
- sequence conservation by homology
    - sequence
        - different for plants and animals
    - stem loop structure
        - varies between plants and animals
- Clustering of multiple miRNAs in a genome locus
- Each miRNA needs a target with sufficient complementarity
- Location of miRNA and target
    - Origin (intron, exon, intergenic)
    - Target (exon, 3'UTR)

Methods for filtering detected hairpins

Whether a computationally detected hairpin is truly interesting and whether it affords spending time and money on follow-up experimental research is not always clear. Some filtering can be performed to narrow down the number of putative miRNAs to an amount that suitable for budget and time constraints.

- Varying level of sequence conservation within stem structure (for homology-based predictions or post filtering for *ab initio* approaches)
- Using general properties of hairpin structures that can be learned from examples
- Repetitively detected structures are generally discarded
- Minimum free energy threshold filtering
- Length of stem loop structure threshold filtering

- If matching to certain annotation of a genome (e. g.: coding sequence) the detections may be discarded
- Base composition
- miRNA gene clustering
- upstream motif about 200 nucleotides before miRNA genes
- Text mining

Other information that could be included is, for example, the existence of a cap and a poly-A tail for pri-miRNAs that are often found in experimentally validated miRNAs.

Although the annotation of the genomic region has been used for filtering, it is clear that miRNAs can come from any region of a genome ***(5)*** and this filtering can thus only be used for reducing computational complexity and not for a biological valid reason.

Selected Examples Performing *ab initio* miRNA Gene Prediction

Due to the large number of available miRNA gene prediction algorithms only, two of them, miR-abela ***(3)*** and MiPred ***(53)*** will be discussed in some more detail followed by a more general statement about prediction accuracy.

MiR-abela

The approach for *ab initio* prediction of miRNAs by Sewer et al. assumes that miRNAs cluster and that they may be co-transcribed ***(3)***. Therefore, they restrict the search of novel miRNAs to areas having close proximity to already known miRNAs. For determining miRNAs, they first check for robust stem loop structures in the area around known miRNAs because they state that the structure is important for recognition and processing by Drosha and Dicer. For this, the

similarity to known stem loops is calculated using a support vector machine based on weighted sequence and structural features. Overall they describe 40 features for pre-miRNA determination with 16 features describing stem loop structures, 10 features for symmetrical regions of a stem loop, 11 features with relaxed symmetry constraints and 3 features in respect to mature miRNA sized portions of a hairpin.

When using their method to predict hairpins in the proximity of known hairpins from the Rfam database in human, mouse, and rat, they were able to achieve a sensitivity of about 89% for their hairpin detection in these species for their artificial negative examples they achieved a false negative discovery rate of 29%, a sensitivity of 71% with only 3% false positives.

MiPred

Ng and Mishra proposed an SVM based *ab initio* prediction method for finding miRNAs in 2007 ***(53)***. In this study, 29 features have been employed to describe a hairpin at the di-nucleotide, folding, thermodynamic, and topological levels. Ng and Mishra trained the classifier on human pre-miRNAs and later used the model to predict miRNAs for human with high sensitivity and specificity. When they used the same model to test the generalization ability of miPred, an average sensitivity of 88% at an average specificity of 98% on a variety of species was achieved. They also compared their method with other existing predictors and found that their method and RNAmicro ***(43)*** perform similar, both outperforming the other tools tested, by large. While RNAmicro employs multiple sequences for the prediction, miPred only uses a single sequence which makes these programs not directly comparable. Thus, according to the authors, miPred is the most successful quasi *ab initio* miRNA predictor for single sequences among the methods tested in their assessment.

Accuracy of miRNA Gene Prediction

It is hard to assess the specificity and sensitivity of algorithms in the absence of at least one fully annotated genome therefore this section does not compare the accuracy of existing algorithms. The reported values from different publications are listed but the reader should be aware of the fact that these values cannot be compared and may even be misleading (cf. Accuracy of Homology-Based miRNA Detection Section).

NovoMir, software for plant miRNA gene prediction, achieved a sensitivity of 80% at a specificity of 99% ***(54)***. MiRenSVM an algorithm combining three SVM achieved a sensitivity of 93% at a specificity of 97% ***(50)***.

Xue and colleagues trained an SVM to distinguish between real and pseudo pre-miRNAs which achieved about 90% accuracy within human, from which the training data were derived, but interestingly also achieved high accuracies of up to 90% in other species ***(51)***. On human data they achieved a sensitivity of about 93% at a specificity of about 88%.

A study by Jiang and colleagues ***(52)*** which reused the same approach as Xue and colleagues ***(51),*** but added P-value and minimum free energy to the classification parameters and also used Random Forrest, a different classification algorithm, achieved a sensitivity of 95% at a specificity of 98%.

A recent study by Zeller and coworkers first extracted all shRNAs from the *Ciona intestinalis* genome filtered the results by structure/sequence conservation, homology to known microRNAs, and phylgenetic footprinting. For all 458 putative miRNAs predicted in this way a microarray

was designed ***(39)***. They were able to identify 100 of these using the microarray and 170 as homologous in the small RNA database for *C. intestinalis* ***(57)***.

Many algorithms for miRNA gene prediction are based on machine learning strategies. In general, these algorithms need a sufficient number of positive as well as negative examples. Although many miRNA genes seem to be unique in any organism, positive training examples can easily be found, whereas negative examples are hard to come by. They are also difficult to be established experimentally since an mRNA needs to be expressed in order to be affected by a miRNA which may only be possible in some specific developmental stadium. Some negative examples that were picked in studies like mRNA sequences ***(3)*** are dubious since to our current knowledge miRNAs can originate from any part of an mRNA. Therefore, one class classifiers which do not need negative examples may be of help in the future ***(58)***.

Without an encompassing knowledge of miRNA genesis only a systems approach can increase the accuracy of current methods. To the best of my knowledge, there is no existing systems approach that evaluates all initially introduced descriptors and discriminators for miRNA genes and further validates them with additional discriminating information such as transcription factor binding sites, expression assays using microarrays and many other more. Usage of several of these features in tandem is obligatory since when scanning the genome for putative miRNAs the number is enormous thus it needs to be strictly scrutinized.

3. Methods for Filtering of Predicted miRNAs

Rfam is a database grouping non coding RNAs, from over 200 complete genome sequences, into families aiming to facilitate the identification and classification of new non coding RNA

sequences ***(4, 59)***. This resource can help assessing whether a predicted miRNA actually fits to the miRNA family and thus aid in deciding whether it should be retained or removed from the predictions. Further databases such as UCbase ***(60)*** and others, can provide supporting information for confirmation of potential miRNAs.

Our recent assessments of miRBase, however, contradicts the above statement since we found many sequences which were labeled as miRNA but obviously must use a different mechanism since they do not fit to the current definition of miRNAs and their genesis or to the proposed processing pathway via Drosha, Dicer and RISC (Saçar, Hamzeiy, and Allmer, submitted).

4. Conclusion

Today's databases contain many miRNAs but at least one study suggests that these miRNAs may only represent abundant variants ***(40)*** another study found that the miRNAs, they were able to confirm experimentally, also turned out to be quite abundant ***(27)***, somewhat confirming the previous suggestion. Therefore, there is a large need for *ab initio* prediction of miRNAs in addition to homology detection. *Ab initio* prediction of genes has been discussed in this chapter but despite many approaches (Table 3) there is no user friendly software which would allow the *ab initio* prediction of miRNAs from sequences.

5. Outlook

Future miRNA gene prediction approaches should take a system approach and evaluate all parts of the system here, for instance, miRNA genesis and miRNA targeting at the same time. This can raise the confidence in individual predictions and reduce the number of false predictions ***(61, 62)***.

They could further include text mining ***(63)***, gene ontologies and networks ***(64)***, promoter sequences ***(65)***,

Integrative approaches like MMIA ***(62)***, which uses multiple miRNA target prediction algorithms in parallel, will also enhance prediction coverage and accuracy in the future.

Besides miRNAs, very similar structures adjacent to them, termed moRs, have been shown to induce gene silencing ***(57)*** which shows that we have not yet seen all biological regulatory options.

Strategies that make use of experimental data, such as deep sequencing data, for miRNA prediction ***(66)*** will in the future be more abundant and likely lead to detection of new miRNAs which do not closely resemble currently known miRNAs.

Other new findings, like spliced miRNAs ***(55)***, may be found in the future, further complicating the already complex prediction of miRNAs.

6. Acknowledgements

I would like to thank Müşerref Duygu Saçar for preparing Figure 1. This study was in part supported by an award received from the Turkish Academy of Sciences for outstanding young scientists (TUBA GEBIP, http://www.tuba.gov.tr).

7. References

1. Soldà, G., Makunin, I. V, Sezerman, O.U., et al. (2009) An Ariadne's thread to the identification and annotation of noncoding RNAs in eukaryotes. *Brief. Bioinform.* **10**, 475–89.

2. Dinger, M.E., Pang, K.C., Mercer, T.R., et al. (2008) Differentiating protein-coding and noncoding RNA: challenges and ambiguities. *PLoS Comput. Biol.* **4**, e1000176.

3. Sewer, A., Paul, N., Landgraf, P., et al. (2005) Identification of clustered microRNAs using an ab initio prediction method. *BMC Bioinformatics* **6**, 267.

4. Griffiths-Jones, S., Moxon, S., Marshall, M., et al. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* **33**, D121–124.

5. Rodriguez, A., Griffiths-Jones, S., Ashurst, J.L., et al. (2004) Identification of mammalian microRNA host genes and transcription units. *Genome Res.* **14**, 1902–1910.

6. Pfeffer, S., Zavolan, M., Grässer, F.A., et al. (2004) Identification of virus-encoded microRNAs. *Science* **304**, 734–6.

7. Fahlgren, N., Jogdeo, S., Kasschau, K.D., et al. (2010) MicroRNA Gene Evolution in Arabidopsis lyrata and Arabidopsis thaliana. *The Plant Cell Online* **22**, 1074 –1089.

8. Aravin, A. and Tuschl, T. (2005) Identification and characterization of small RNAs involved in RNA silencing. *FEBS Lett.* **579**, 5830–40.

9. Bentwich, I. (2005) Prediction and validation of microRNAs and their targets. *FEBS Lett.* **579**, 5904–5910.

10. Janssen, S., Schudoma, C., Steger, G., et al. (2011) Lost in folding space? Comparing four variants of the thermodynamic model for RNA secondary structure prediction. *BMC Bioinformatics* **12**, 429.

11. Mathews, D.H. and Turner, D.H. (2002) Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. *J. Mol. Biol.* **317**, 191–203.

12. Juan, V. and Wilson, C. (1999) RNA secondary structure prediction based on free energy and phylogenetic analysis. *J. Mol. Biol.* **289**, 935–47.

13. Hofacker, I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res.* **31**, 3429–3431.

14. Krüger, J. and Rehmsmeier, M. (2006) RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic Acids Res.* **34**, W451–454.

15. Reuter, J.S. and Mathews, D.H. (2010) RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* **11**, 129.

16. Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31**, 3406–3415.

17. Shapiro, B.A. (1988) An algorithm for comparing multiple RNA secondary structures. *Comput. Appl. Biosci.* **4**, 387–393.

18. Aksay, C., Salari, R., Karakoc, E., et al. (2007) taveRNA: a web suite for RNA algorithms and applications. *Nucleic Acids Res.* **35**, W325–329.

19. Janssen, S. and Giegerich, R. (2010) Faster computation of exact RNA shape probabilities. *Bioinformatics* **26**, 632–639.

20. Markham, N.R. and Zuker, M. (2008) UNAFold: Software for Nucleic Acid Folding and Hybridization., In: Keith, J.M. (ed.) *Bioinformatics : Structure, Function and Applications*, pp. 3–31 Humana Press, Totowa, NJ.

21. Leung, W.-S., Lin, M.C.M., Cheung, D.W., et al. (2008) Filtering of false positive microRNA candidates by a clustering-based approach. *BMC Bioinformatics* **9 Suppl 12**, S3.

22. Dezulian, T., Remmert, M., Palatnik, J.F., et al. (2006) Identification of plant microRNA homologs. *Bioinformatics* **22**, 359–360.

23. Artzi, S., Kiezun, A., and Shomron, N. (2008) miRNAminer: A tool for homologous microRNA gene search. *BMC Bioinformatics* **9**, 39.

24. Gerlach, D., Kriventseva, E. V, Rahman, N., et al. (2009) miROrtho: computational survey of microRNA genes. *Nucleic Acids Res.* **37**, D111–117.

25. Maselli, V., Bernardo, D. Di, and Banfi, S. (2008) CoGemiR: A comparative genomics microRNA database. *BMC Genomics* **9**, 457.

26. Guerra-Assunção, J.A. and Enright, A.J. (2010) MapMi: automated mapping of microRNA loci. *BMC Bioinformatics* **11**, 133.

27. Lim, L.P., Lau, N.C., Weinstein, E.G., et al. (2003) The microRNAs of Caenorhabditis elegans. *Genes & Development* **17**, 991–1008.

28. Ohler, U., Yekta, S., Lim, L.P., et al. (2004) Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification. *RNA* **10**, 1309–1322.

29. Nam, J.-W., Kim, J., Kim, S.-K., et al. (2006) ProMiR II: a web server for the probabilistic prediction of clustered, nonclustered, conserved and nonconserved microRNAs. *Nucleic Acids Res.* **34**, W455–458.

30. Berezikov, E., Guryev, V., Belt, J. van de, et al. (2005) Phylogenetic shadowing and computational identification of human microRNA genes. *Cell* **120**, 21–4.

31. Huang, T.-H., Fan, B., Rothschild, M.F., et al. (2007) MiRFinder: an improved approach and software implementation for genome-wide fast microRNA precursor scans. *BMC Bioinformatics* **8**, 341.

32. Bonnet, E., Wuyts, J., Rouzé, P., et al. (2004) Detection of 91 potential conserved plant microRNAs in Arabidopsis thaliana and Oryza sativa identifies important target genes. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 11511–6.

33. Wang, X.-J., Reyes, J.L., Chua, N.-H., et al. (2004) Prediction and identification of Arabidopsis thaliana microRNAs and their mRNA targets. *Genome Biology* **5**, R65.

34. Lang, Q., Jin, C., Lai, L., et al. (2011) Tobacco microRNAs prediction and their expression infected with Cucumber mosaic virus and Potato virus X. *Mol. Biol. Rep.* **38**, 1523–31.

35. Gruber, A.R., Findeiß, S., Washietl, S., et al. (2010) Rnaz 2.0: Improved Noncoding Rna Detection. *Pacific Symposium on Biocomputing* **15**, 69–79.

36. Rivas, E., Klein, R.J., Jones, T.A., et al. (2001) Computational identification of noncoding RNAs in E. coli by comparative genomics. *Curr. Biol.* **11**, 1369–1373.

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

38. Lu, J., Shen, Y., Wu, Q., et al. (2008) The birth and death of microRNA genes in Drosophila. *Nature genetics* **40**, 351–5.

39. Keshavan, R., Virata, M., Keshavan, A., et al. (2010) Computational identification of Ciona intestinalis microRNAs. *Zoological Science* **27**, 162–170.

40. Lai, E.C., Tomancak, P., Williams, R.W., et al. (2003) Computational identification of Drosophila microRNA genes. *Genome Biol* **4**, R42.

41. Huang, J.C., Morris, Q.D., and Frey, B.J. (2007) Bayesian inference of MicroRNA targets from sequence and expression data. *J. Comput. Biol.* **14**, 550–563.

42. Nam, J.-W., Shin, K.-R., Han, J., et al. (2005) Human microRNA prediction through a probabilistic co-learning model of sequence and structure. *Nucleic Acids Res.* **33**, 3570–3581.

43. Hertel, J. and Stadler, P.F. (2006) Hairpins in a Haystack: recognizing microRNA precursors in comparative genomics data. *Bioinformatics* **22**, 197–202.

44. Berezikov, E., Cuppen, E., and Plasterk, R.H.A. (2006) Approaches to microRNA discovery. *Nat. Genet.* **38 Suppl**, 2–7.

45. Hafner, M., Landthaler, M., Burger, L., et al. (2010) Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP. *Cell* **141**, 129–141.

46. Vogel, J. and Sharma, C.M. (2005) How to find small non-coding RNAs in bacteria. *Biol. Chem.* **386**, 1219–1238.

47. Hüttenhofer, A. and Vogel, J. (2006) Experimental approaches to identify non-coding RNAs. *Nucleic Acids Res.* **34**, 635–46.

48. Lau, N.C., Lim, L.P., Weinstein, E.G., et al. (2001) An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science* **294**, 858–62.

49. Bentwich, I. (2008) Identifying human microRNAs. *Curr. Top. Microbiol. Immunol.* **320**, 257–69.

50. 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.

51. Xue, C., Li, F., He, T., et al. (2005) Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics* **6**, 310.

52. Jiang, P., Wu, H., Wang, W., et al. (2007) MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Res.* **35**, W339–344.

53. 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.

54. Teune, J.-H. and Steger, G. (2010) NOVOMIR: De Novo Prediction of MicroRNA-Coding Regions in a Single Plant-Genome. *J. Nucleic Acids* **2010**,.

55. Thieme, C.J., Gramzow, L., Lobbes, D., et al. (2011) SplamiR--prediction of spliced miRNAs in plants. *Bioinformatics (Oxford, England)* **27**, 1215–1223.

56. Wu, Y., Wei, B., Liu, H., et al. (2011) MiRPara: a SVM-based software tool for prediction of most probable microRNA coding regions in genome scale sequences. *BMC Bioinformatics* **12**, 107.

57. Shi, W., Hendrix, D., Levine, M., et al. (2009) A distinct class of small RNAs arises from pre-miRNA-proximal regions in a simple chordate. *Nature Structural & Molecular Biology* **16**, 183–189.

58. Yousef, M., Jung, S., Showe, L.C., et al. (2008) Learning from positive examples when the negative class is undetermined--microRNA gene identification. *Algorithms Mol. Biol.* **3**, 2.

59. Gardner, P.P., Daub, J., Tate, J.G., et al. (2009) Rfam: updates to the RNA families database. *Nucleic Acids Res.* **37**, D136–140.

60. Taccioli, C., Fabbri, E., Visone, R., et al. (2009) UCbase & miRfunc: a database of ultraconserved sequences and microRNA function. *Nucleic Acids Res.* **37**, D41–48.

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

62. Nam, S., Li, M., Choi, K., et al. (2009) MicroRNA and mRNA integrated analysis (MMIA): a web tool for examining biological functions of microRNA expression. *Nucleic Acids Res.* **37**, W356–362.

63. Naeem, H., Küffner, R., Csaba, G., et al. (2010) miRSel: automated extraction of associations between microRNAs and genes from the biomedical literature. *BMC Bioinformatics* **11**, 135.

64. Backes, C., Meese, E., Lenhof, H., et al. (2010) A dictionary on microRNAs and their putative target pathways. *Nucleic Acids Res.* **38**, 4476–4486.

65. Long, Y.-S., Deng, G.-F., Sun, X.-S., et al. (2011) Identification of the transcriptional promoters in the proximal regions of human microRNA genes. *Mol. Biol. Rep.* **38**, 4153–7.

66. Hendrix, D., Levine, M., and Shi, W. (2010) miRTRAP, a computational method for the systematic identification of miRNAs from high throughput sequencing data. *Genome Biology* **11**, R39.

8. Tables

Table 1: Non comprehensive list of programs predicting the secondary structure from primary RNA sequence. The rows are sorted decreasingly by average citation count per year.

Name	**Summary**	**Systems**	**Availability**	**Reference**
Dynalign	Aligns two nucleotide sequences and predicts their common structure.	ANSI C++ code, Part of RNAstrucuture (MS Windows)	rna.chem.rochester.edu, Open Source	***(11)***
Unnamed	Predicts RNA secondary structure using covariational and free energy methods.	-	-	***(12)***
RNAfold	Predicts RNA secondary structure using minimum free energy.	Web service, local installation	http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi	***(13)***
RNAHybr	Finds the minimum free energy	Web service, limited local	http://bibiserv.techfak.uni-	***(14)***

id	hybridization of two RNAs	installation	bielefeld.de/rnahybrid/	
RNAStructure	Determines secondary structure using dynamic programming with free energy minimization	MS Windows, C++	rna.chem.rochester.edu	***(15)***
mfold	Determines secondary structure using dynamic programming with free energy minimization	Fortran, C, UNIX	www.ibc.wustl.edu/~zuker/rna/form1.cgi	***(16)***
RNADistance	Calculates the distance among structures based on string editing and base pair distance.	Local installation	http://www.tbi.univie.ac.at/~ivo/RNA/man/RNAdistance.html	***(17)***
ViennaRNA	Unified access to various RNA tools of	Web service, Software package	rna.tbi.univie.ac.at	***(13)***

	the Vienna package			
taveRNA	A package containing secondary structure prediction, RNA-RNA interaction, and a database pruning algorithm.	Web service	compbio.cs.sfu.ca/taverna	***(18)***
RNAShapes	Predicts secondary structure by evaluating promising shapes with Boltzman probabilities.	Web service, local installation	http://bibiserv.techfak.uni-bielefeld.de/rapidshapes/submission.html	***(19)***
UNAFold	Simulates folding, hybridization, and melting pathways for up to two sequences	Local installation	http://mfold.rna.albany.edu/	***(20)***

Table 2: Non-comprehensive selection of software that allows homology mapping of miRNAs to the source genome or to related species. The rows are sorted decreasingly by average citation count per year.

Name	**Summary**	**Clade**	**URI**	**Reference**
MicroHarvester	BLAST search for candidates filtered by structural features specific to plant miRNAs	Plant	http://www-ab.informatik.uni-tuebingen.de/brisbane/tb/index.php?view=microharvester2	***(22)***
miRNAminer	BLAST search for homologs with filtering by minimum free energy and alignment conservation	Animal	http://groups.csail.mit.edu/pag/mirnaminer/	***(23)***
miROrtho	Homology extended alignments of known miRBase families and putative miRNA families using SVM and orthology	Animal	http://cegg.unige.ch/mirortho	***(24)***

CoGemiR	Sequence similarity and secondary structure analysis similar to miRNAminer but with a larger number of species	Animal	http://cogemir.tigem.it/	***(25)***
MapMi	Maps miRNAs within species and across species using sequence homology and structure	Any	http://www.ebi.ac.uk/enright-srv/MapMi/	***(26)***
MiRscan	Trained on examples conserved between two closely related species derived from a fold-first find-homologs later strategy	Worms	http://genes.mit.edu/mirscan/	***(27)***
MiRscanII	Supersedes MiRscan, adds conservation of miRNA gene flanking regions and a conserved motif	Worms	http://genes.mit.edu/burgelab/MiRscanII/	***(28)***

ProMiR II	Integrative approach using several databases and criteria as well as several custom modules	Animal	http://cbit.snu.ac.kr/~ProMiR2/introduction.html	***(29)***
unnamed	Homologous miRNA genes among primates used to determine general characteristics of miRNA genes in vertebrates	Vertebrates	Not associated website allowing phylogenetic shadowing: http://eshadow.dcode.org/	***(30)***
MiRFinder	Based on pairwise genome searches for shRNA using SVM for filtering, introduces mutation model for hairpins	Any	http://www.bioinformatics.org/mirfinder/	***(31)***
Unnamed	Homology between Arabidopsis and Oryza; approach also takes target information into account	Plant	-	***(32)***
Unnamed	Homology between	Plant	-	***(33)***

	Arabidopsis and Oryza; approach also takes target information into account			
Unnamed	Exploit clustering of miRNAs to filter miRNA predictions	Mammals	-	***(21)***
Unnamed	GSS, EST versus known miRNAs and proteins with subsequent feature based filtering	Plant	-	***(34)***
RNAz	Detects thermodynamically stable and evolutionarily conserved ncRNA secondary structures in MSA	Any	http://rna.tbi.univie.ac.at/cgi-bin/RNAz.cgi	***(35)***
QRNA	Uses comparative genome sequence analysis to detect conserved ncRNA	Any	http://selab.janelia.org/software.html	***(36)***

	secondary structures			

Table 3: Non-comprehensive list of software that allows the *ab initio* prediction of miRNA genes. Rows are sorted by number of citations.

Program	**Summary**	**Clade**	**URL**	**Reference**
miRseeker	First homologous miRNA gene fishing then structure and nucleotide sequence divergence filtering	flies	Not functional: http://www.fruitfly.org/seq_tools/miRseeker.html	***(40)***
PalGrade	Hairpin structural and sequence characteristics model with subsequent experimental validation	human	-	***(49)***
Dynalign	Finds ncRNAs by optimizing total free energy between RNA sequences, alternative fast SVM	Any	http://rna.urmc.rochester.edu/dynalign.html	***(11)***

	classification			
MiRenSVM	Employs multiple targeted SVM to model different types of miRNAs	Any	-	***(50)***
MiR-abela	Assumes miRNA gene clustering and searches for new genes in proximity of known genes	Human, mouse, rat	http://www.mirz.unibas.ch/cgi/pred_miRNA_genes.cgi	***(3)***
Triplet-SVM	Forms structure sequence triplets from hairpins and classifies them using a SVM	Any	http://bioinfo.au.tsinghua.edu.cn/mirnasvm/	***(51)***
RNAmicro	First structure of shRNAs (RNAz) then SVM filtering of MSAs	Any	http://www.tbi.univie.ac.at/~jana/software/RNAmicro.html	***(43)***
miPred	Introduces a new machine learning approach, random forest, and improves upon Triplet-SVM	Any	http://www.bioinf.seu.edu.cn/miRNA/	***(52)***

miPred	Uses SVM classification without homology and defines 29 parameters to describe hairpin structures	Any	Not functional: http://web.bii.a-star.edu.sg/~stanley/Publications	***(53)***
NovoMir	Uses a series of filter steps and statistical models to determine pre-miRNAs in a plant genome.	Plant	www.biophys.uni-duesseldorf.de/~teune/Data/novomir-2010-10-10.tgz	***(54)***
SplamiR	Predicts miRNAs which derive from spliced transcripts.	Plant	www.uni-jena.de/SplamiR.html	***(55)***
MiRPara	Predicts miRNAs from high throughput sequencing data using a SVM.	Any	www.whiov.ac.cn/bioinformatics/mirpara	***(56)***