

# microRNA gene finding and target prediction - basic principles and challenges

## Abstract

Understanding regulative potential and biological significance of miRNAs relies on our ability to correctly identify their target genes. Many experimental and computational efforts have been made to identify miRNA targets. However, reliable microRNA target prediction is still unsolved computational challenge. In this review we will focus on the basic principles and assumptions on which are based current computational approaches for finding and target prediction of miRNAs as well as the challenges for their future development.

**Keywords:** miRNA biogenesis, RNA folding, target prediction, sequence alignment

Volume 1 Issue 4 - 2014

Roumen Dimitrov

Department of Physics, Sofia University of "St. Kliment Ochridski", Bulgaria

**Correspondence:** Roumen Dimitrov, Department of Physics, Sofia University of "St. Kliment Ochridski", Sofia, Bulgaria, Email roumen.dimitrov@gmail.com

**Received:** June 25, 2014 | **Published:** August 15, 2014

**Abbreviations:** miRNA, microRNA; ncRNAs, non-coding RNAs; pre-miRNA, miRNA long single-stranded precursor; FEM, free energy minimisation

## Introduction

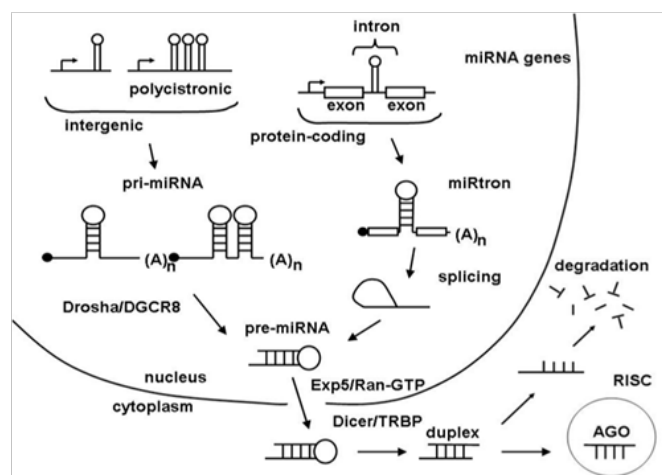
Encoded by eukaryotic nuclear DNA in plants, animals and some DNA based viruses miRNAs are a class of small ~21 nucleotides non-coding RNA (ncRNAs).<sup>1,2</sup> Why these short stretches of nucleotides are so important? Despite they small size it happens that miRNAs by binding to complementary sites on mRNAs gene transcripts are capable to induce cleavage or repression of productive translation for dozens or even hundreds of different mRNAs.<sup>3</sup> As a result individual miRNAs can impact multiple cellular pathways like - signaling pathways; cell differentiation, proliferation/growth, mobility, and apoptosis; brain function, subcellular compartmentalization and chromatin remodeling.<sup>4-7</sup> The ability to regulate almost every cell pathway makes miRNA very powerful factor in cells function and development. Understanding miRNAs will help in the discovery of treatments for diseases like cancer.<sup>8-10</sup>

Understanding regulative potential and biological significance of miRNAs relies on our ability to correctly identify their target genes. Many experimental and computational efforts have been made to identify miRNA targets.<sup>11-20</sup> However, reliable microRNA target prediction is still unsolved computational challenge.<sup>21</sup> In this review we will focus on the basic principles and assumptions on which are based current computational approaches for finding and target prediction of miRNAs as well as the challenges for their future development in time of incredible technological advances and fast growing number of sequenced genomes.

## miRNA biogenesis pathway

In animals miRNA gene is transcribed by RNA polymerase II in the cell nucleus in the form of long single-stranded precursor (pre-miRNA) transcripts that fold into very stable long hairpin structure. Starting from the cell nucleus and continuing on a few possible roads and processing steps toward the cytoplasm mature miRNA is excised from one of the stem arms of pre-miRNA hairpin. In the cytoplasm mature miRNA is loaded on a RISC protein complex and start to search for targets (Figure 1).<sup>22,23</sup> The instruction for target recognitions is coded in a mature miRNA nucleotide sequence. Based on sequence

homology, miRNAs can be grouped into sub-families, and many of them are evolutionarily conserved.<sup>24,25</sup> Some miRNAs have tissue-specific or developmental-stage-specific expression patterns and their abundance can vary greatly depending on cell type.<sup>26</sup>



**Figure 1** miRNA biogenesis pathway in animals. miRNAs are transcribed from RNA polymerase II as a long RNA primary transcript known as a pri-miRNA which is cleaved by Drosha to produce the ~70-bp long stem-loop structure of pre-miRNA. It is then exported into the cytoplasm by Exportin complex, where it is cleaved by Dicer endonuclease to produce ~21-bp miRNA duplex. In most cases the strand with the 5' terminus located at thermodynamically less-stable end of the duplex is selected as mature miRNA which is loaded onto the miRISC complex. Upon loading, the passenger strand of the miRNA duplex in most cases is degraded. The mature miRNA within miRISC serves as a guide for recognizing target mRNAs by base-pairing with the so called seed region (the first seven to eight nucleotides), which leads to a block in the translation of the mRNA target and/or its degradation.

In plants, miRNA is transcribed and processed by homologous to the animals proteins, but the entire process of miRNA biogenesis is undertaken within the plant nucleus.<sup>27</sup> The length of miRNA hairpin is more heterogeneous compare to animals, and can range from 70 to hundreds of nucleotides.<sup>28</sup>

A pre-miRNA is considered as an independent miRNA gene and seems to possess promoter and enhancer elements that are similar to those of protein-coding genes.<sup>29</sup> MiRNA genes are scattered all over

the genome but mostly they hide in intergenic regions or as part of introns of protein-coding and non-protein-coding genes in which case are excised by splicing machinery. MiRNA genes like to cluster together. Polycistronic transcripts with multiple miRNA-generating hairpins have been detected in animals but are rare in plants.<sup>30</sup> Animal genomes have small but abundant miRNA families, while plants have fewer but larger miRNA gene families. Although many miRNA genes are conserved across species, the same gene family varies in size and genomic organization in different species.<sup>31,32</sup> MiRNA community is dynamic – new genes and families are created and loss constantly.

There are a few major way miRNA genes can appear. First, the miRNAs genes in a given family have significant sequence homology to each other and almost identical mature miRNA sequences. This suggests that inside the families new miRNA genes appear by local duplications. MiRNA cluster formation by gene duplication has been observed in animals and probably dominates the evolution of plant.<sup>33</sup> Second, there are indications that some miRNA genes in mammals resulted from transposons while convincing evidences are absent for plants.<sup>34</sup> Finally, miRNA genes appear de novo from the many thousands RNA transcripts from non-coding DNA regions (more often from non-coding protein introns).<sup>35</sup> Some of these RNA transcripts are expressed at low levels processed imprecisely and lack targets. Apparently they are subject to weak or no purifying selection and as a result easily form different secondary structures. It seems that these RNA transcripts form dynamical pool with high secondary structure plasticity for selection to act upon. It is proposed that in plants new miRNA genes emerge via mechanisms of inverted duplication events, while in animals new miRNA genes evolve gradually from intermediate non-functional hairpins structures.<sup>36</sup> Once a functional hairpin structure is generated, it will start to interact with the miRNA processing machinery and different targets. At the beginning this interactions will be very weak but gradually they will increase together with the expression levels of the new miRNA gene. This will give enough time for sequence-structure synchronization of miRNA transcript to adapt its sequence and kinetics of folding to provide specificity for every step of processing between transcription and integration of mature miRNA in particular biological pathway.

## Finding, folding and sequence comparison

Technological advances such as bioinformatics and next-generation sequencing allowed the identification of a great number of miRNAs in different organisms in recent years. To date, over 2,000 human miRNAs have been identified and deposited in miRBase, while one to two hundred miRNAs are expressed in lower metazoans and plants.<sup>37</sup>

Given the increased number of sequenced whole genomes and the fact that much of these genomes are transcribed research on miRNAs especially finding and prediction of their target genes and the mechanism by which they repress their mRNA constitutes one of the frontiers in the study of post-transcriptional gene regulation.

miRNA transcripts pass through a few processing steps before mature miRNA is excised and loaded on the RISC complex. These processing steps require miRNA gen transcripts to have structures and sequences with certain biochemical characteristics. Therefore, finding of miRNA genes include discovery of pre-miRNA sequences which in equilibrium thermodynamic conditions can fold in known or new hairpin structure with certain characteristics that are important for proper biochemical processing.

RNA molecules at equilibrium have an ensemble of structures that they visit with different probabilities. How to find those with

proper biochemical characteristics? Experimentally it has been shown that RNA transcripts of miRNA genes relative to other ncRNA transcripts and mRNA have unusual thermodynamically stable long hairpin structures.<sup>38–40</sup> It seems that during the process of adaptation of structures and kinetics of folding of miRNA transcripts so that they could provide specificity for every step of processing between transcription and integration into particular biological pathway, a special set of sequences have been selected. As a result miRNA transcripts with selected sequences: have stable hairpin structures as their free energy minimum (fem) conformations; are separated by free energy gap from the other low free energy conformations; have conformations very different and separated with high free energy barriers from the fem conformation. This is not necessarily true for other ncRNA. For example, depending on their processing and biological functions, at thermodynamic equilibrium instead of one fem conformation (and structurally closed to the fem fluctuations) other ncRNA could be represented with different sets of inter converting secondary structures (switches). Sequences that can fulfill the thermodynamics requirements of these switches will be very different from sequences that fulfill requirements for fem conformations.

Therefore, we can formulate the problem of finding miRNA genes as the problem of finding the base-pairing that gives the lowest free energy change in going from the unfolded to folded state of their RNA transcript sequences. Also, we need to combine sequence, folding and structure considerations in a single integrated approach. Some additional filtering based on hairpin biochemical characteristics have to be done too like – sequence and structure requirements for Drosha and Dicer processing for example.<sup>41–43</sup>

Currently, RNA sequence dependence of free energy changes during the formation of secondary structures is not completely understood. Together with estimated 1.8n possible secondary structures for sequence with n nucleotides, this makes the finding of miRNA genes very difficult task.<sup>44</sup> Nevertheless, dynamic programming algorithm, a type of recursive algorithm commonly used for optimization problems in biology, can search for polynomial time the entire set of possible RNA secondary structures and find the lowest free energy structure or particular sub-set from the ensemble of secondary structures available at equilibrium.<sup>45,46</sup>

Computational approaches can be divided in a few groups. In the first group one screen for sequences which can fold in a known experimentally verified pre-miRNA hairpin structure.<sup>17,47</sup> The point here is that functional restrictions enforce more evolutionary pressure to preserve the structure than the sequence. So, the challenge is to minimize false positives. For example, the tested sequence can fold in a desire hairpin structure but is not the true miRNA sequence. Studies have shown that for pre-miRNA with stable thermodynamics hairpin structures the tested sequences nearly always have less favorable folding free energy compared to the true miRNA sequence even.<sup>17,39</sup>

In the second group one find a structure from a set of aligned sequences which is usually obtained based on similarity optimization.<sup>17,48,49</sup> The point here is to identify sites showing correlated mutations. Such mutations can be interpreted as indication for functional or structural constraints. In other words one searches for mutations which preserve particular secondary structure.<sup>50</sup>

The assumption behind the above complementary approaches is that sequences with conserved catalytic activity, common evolutionary ancestry and having not enough time to diverge strongly between each other, fold into very similar structures.<sup>17,51</sup> There exist a single optimal alignment for such set of sequences, which provides an accurate measure of similarity, structure, function and evolutionary history

between them. However, with increasing evolutionary distances between nucleotide sequences, the single optimal alignment method is replaced by an ensemble of alignments of almost equal quality and ensemble of different folded conformations. Recurring difficulties associated with diverged sequence data include alternative alignment possibilities of insertions and deletions, region of length variations in which homology assessment is questionable or impossible, occurrence of localized excessive mutations to the point of saturation and lost of phylogenetic signals. Therefore, for diverged sequences optimizing similarity will not necessarily improve structure, function and evolutionary history assessments.<sup>52,53</sup> This leads to false-positive miRNAs in the search procedures. Finally, in the last group one screen for sequences with new miRNAs. Here, the sequences or the structures of their close relatives are not known. In this case the false positives come from random sequences with no biological functions. It is widely accepted now that sequences which are biologically active are selected in such way that their biologically active conformations have more favorable free energies than random sequences.<sup>17,54</sup>

In the case of single sequence the free energy  $f$  of its fem conformation should be compare with free energy distribution of random sequences obtained from biological one by shuffling in such way that dinucleotide composition is preserved. The mean  $\mu$  and standard deviation  $\sigma$  of fem conformations of a large number of shuffled random sequences is calculated. Statistical significance is expressed in standard deviations from the mean as Z-score. Negative Z-scores indicate that the tested sequence is more stable than the random sequences. However, single sequence methods are of limited statistical significance.

Some additional filtering based on hairpin characteristics have to be done too like – sequence and structure requirements for Drosha and Dicer processing; number and distribution of base pairs; length of the hairpin stem; location of miRNA on the hairpin stem; type of the bulges and their position on the hairpin stem; loop size and others.<sup>17</sup>

The most interesting and general case is connected with the so-called Sankoff's problem for simultaneous optimization of alignment and folding.<sup>55</sup> A simple reason behind the simultaneous optimization of alignment and folding is first synchronization of sequence and structure evolution and second that strong structural consensus among related, but diverged sequences are a good indicator for preserved functional role. Up to now there is no a general solution for this long standing problem. Whereas early methods focused on energy-directed folding of single sequences, now comparative analysis based on structure preserving changes of base pairs has been efficient in improving accuracy, and today this constitutes a key component in genomic screens. The usual approach is based on variations of - first fold than align and oppositely. For example, one can start with the sequences aligned by nucleotide identity and then find the conserved pairs in the given alignment or to predict low free energy structures for each sequence separately and then to sort through the predicted structures to find the structures common to all sequences.<sup>54</sup> A recent approach for the general solution of Sankoff's problem<sup>52,53</sup> has shown that classical alignment algorithm and hybridization without intramolecular base-pairing are mathematically equivalent. A simple analogy based on considerations of the frequency with which different nucleotides {A, C, G, T / U} or gaps occupy or not occupy certain positions in the alignment is shown on (Figure 2). The more substitutions of nucleotides or gaps are at a given position, the more often the formation of mismatches or loops are in these positions in the corresponding hybridization scenarios.

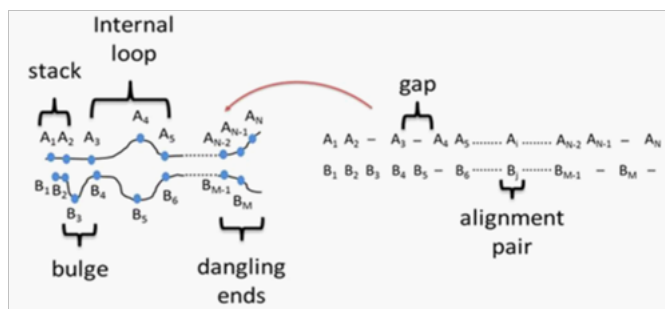


Figure 2 Correspondence between alignment and hybridization.

The equivalence between alignment and hybridization allows integrating both the energy-based and evolutionary-based information of sequence/structure covariations to predict the simultaneous folding/co-folding and alignment of RNA/DNA sequences. This can be done by integrating evolutionary and thermodynamic partition functions in the frame of dynamic programming algorithm (Figure 3).<sup>52</sup>

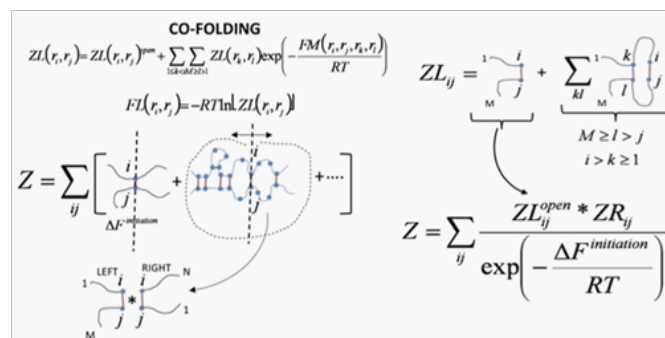


Figure 3 Calculation of statistical sum Z using the method of dynamic programming. ZLij and ZRij are the left and right statistical sums for a given intermediate base-pair (Ai, Bj).

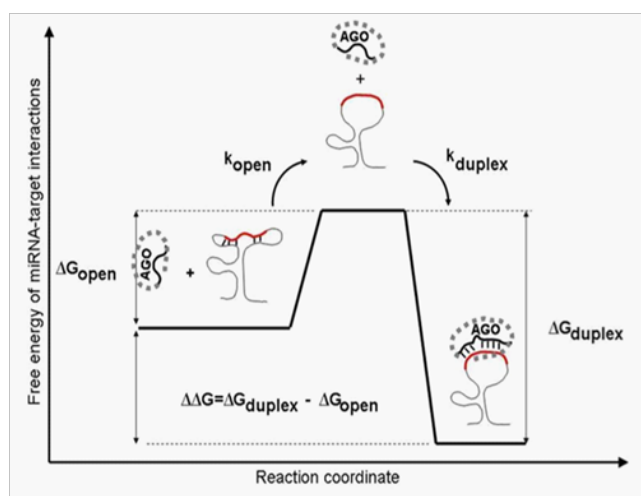
## Target prediction

The regulatory information for target recognition and the degree of repression is encoded in the nucleotide sequence of miRNA genes. A key factor for inferring the regulatory information for target recognition is the analysis of miRNAs hybridization with mRNA of their target genes together with precise target-site localization.

The computational studies of the determinants of miRNA targeting specificity can be divided into two main classes: those that emphasize sequence features, and those that emphasize structural aspects.

Structural aspects include computational approaches for co-folding of two interacting nucleotide sequences including all possible conformational states like – matches, mismatches, bulges, symmetric and asymmetric interior loops, multi-branched loops and kissing loops in the regions where the two sequences fold separately and in the regions where they interact. Interaction regions and their conformational states are important for miRNA-mRNA interactions.<sup>56,57</sup> Previous work revealed that 7-8 nucleotides of the so cold seed region at the 5' end of the miRNA are responsible for target recognition.<sup>58</sup> The seed region is small so is the target side, but it can be part of a bigger folded region of mRNA sequence. A portion of the mRNA where target site resides and is involved in intramolecular base-pairing has to become locally accessible to a targeting miRNA. The overall free energy to go from folded to unfolded state of this portion of mRNA and the hybridization of the target site of the

unfolded portion with the 5' seed region of miRNA is the driving force of this reaction. Therefore, the overall successes of miRNA binding to mRNA depend on mRNA length, its available secondary structures at thermodynamic equilibrium, binding site accessibility and thermodynamic stability of miRNA-mRNA duplex (Figure 4).<sup>56,59-61</sup> One has also to include competition between different miRNAs or between miRNAs and proteins for the same target side. In this case concentrations of mRNA, miRNAs and proteins play crucial role. The concentration of individual miRNAs in cells may vary by four orders of magnitude.<sup>62</sup> This is even more important when there is more than one target side. Depending on the distance between target sides and miRNAs concentrations cooperative or competing mode of action have been reported to be predictive for the functionality of miRNA target sites. Effective modeling of saturation of binding and competition on binding sites is still in its infancy.<sup>63</sup> Interestingly, recently it was found that miRNA genes that produce identical or nearly identical mature miRNAs have distinct biological activities that are controlled by their pri/pre-miRNA loop sequences. These raise the possibility of pri-/pre-miRNAs as target-recognizing RNA species.<sup>64</sup> Sequence aspects include the sequence composition of the 3' UTRs or of the immediate environment of the putative target sites to make interaction of target side with other portions of mRNA less effective and make target sides more accessible,<sup>59,65</sup> the base-pairing pattern in the 3' region of the miRNA;<sup>66</sup> conservation of the miRNA binding site – it has been shown that conserved perfect seed matches to miRNAs have reduced SNP density which indicate that target sites are under purifying selection.<sup>67</sup> It was also shown that nucleotide words complementary to miRNA seed regions are better conserved in 3' UTRs than “random words” of the same length.<sup>17</sup> However, there are experiments which question whether a seed match is either necessary or sufficient for miRNA repression: showing that perfect base pair matching does not guarantee interaction between miRNA and target gene, and wobble G:U base pairs are often tolerated in target sites.<sup>68</sup> Taking advantage of miRNAs conservation we can use known mature miRNA sequence in one species to detect the probable location of its orthologue in another species, or further paralogues in the original species, by combining sequence analysis, sequence alignment and RNA secondary structure prediction.<sup>69</sup> However, given the numerous rules and exceptions one needs an integrated statistical model to determine the most optimal prediction scenario.



**Figure 4** Illustration of free energy and kinetics changes for miRNA-target interactions.

While for finding miRNAs there is lot of attempts to approach the solution of Sankoff's problem for simultaneous alignment and

folding implemented in numerous software tools there are almost no attempts to consider simultaneous alignment and co-folding in the frame of comparative genomics. Co-folding is an integral part of our comparative approach.<sup>52,53</sup> More importantly, concentrations of miRNAs, mRNA and proteins can be easily included too. This is very important for modeling the saturation of binding and competition on binding sites, and the extension of this approach is the focus of current research.

## Concluding remarks

With almost 2000<sup>47</sup> reported miRNAs in human, and about 10 000 candidate genes, the network of interactions is extremely complex. Moreover, details about local concentrations and competitions among miRNAs in living cell are missing. Despite the advances in the recent years the mechanistic details of the function of miRNAs in repressing protein synthesis are still poorly understood<sup>70-72</sup> and this reflects on quality of computational predictions. Thus, the degree of overlap between their lists of predicted targets is very poor and sometimes null.<sup>73</sup>

To improve computational miRNA target recognition approaches we need a deeper quantitative understanding of the miRNA-induced response and network regulation. The major problem with computational approaches is that it is difficult to predict, without sequencing data, when and in what type of cells a potential genomic hairpin will be expressed as RNA, and be available for processing.

This problem is address by the new technology RNA-seq.<sup>74</sup> This technology enables not only the assessment of the expression level for a given miRNA loci, but also determines the exact mature sequence for that miRNA loci with a single base-pair resolution. This makes it possible to distinguish between miRNAs that share the same set of targets, but which can have small differences in their mature sequence.<sup>75</sup>

Based on RNA-seq It has been found also that:<sup>76</sup>

- mRNAs are in general expressed at higher levels prior to miRNA expression
- mRNAs not expressed in the same tissue as the miRNA tend to have more non conserved target sites
- mRNAs with conserved targets are expressed in the tissue in which the miRNA is expressed, but at a lower level than compared to adjacent tissue
- mRNAs expressed in the same tissue as an miRNA tend to avoid having target sites for that miRNA
- Housekeeping genes have shorter UTRs and fewer target sites than other mRNAs, probably to avoid targeting by miRNAs

Some kinetic factors should be taken in consideration too. Such factors are transcription and turnover rate for RNA and protein molecules.<sup>77</sup> Thus, the half-life of different proteins can vary from minutes to days whereas the degradation rate of mRNA would fall within 2-7hrs. Also, the rate of mRNA transcription is lower compared to protein translation in mammalian cells - 2 mRNAs transcribed per hour versus dozens of proteins per mRNA per hour.

Kinetic factors are important for cells which need to carry out a given function within a limited space of time. For such cells it is easier to find a direct correlation between the mRNA and protein levels than in less active cells. Therefore, when the aim is to find what genes are regulated by a given miRNA and how this explain a certain cell type

function target predictions are necessary to guide target discovery. Moreover, connecting miRNA and their mRNA target genes to different biological response and functional pathways provides a unique opportunity to decipher the evolutionary forces involved in shaping miRNA network regulatory potentials.

## Acknowledgement

None

## Conflict of interest

Authors declare that there is no conflict of interest.

## References

- Baek D, Villen J, Shin C, et al. The impact of microRNAs on protein output. *Nature*. 2008;455(7209):64–71.
- Lim LP, Lau NC, Garrett-Engele P, et al. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature*. 2005;433(7027):769–773.
- Valencia-Sanchez MA, Liu J, Hannon GJ, et al. Control of translation and mRNA degradation by miRNAs and siRNAs. *Genes Dev*. 2006;20(5):515–524.
- Ameres SL, Zamore PD. Diversifying microRNA sequence and function. *Nat Rev Mol Cell Biol*. 2013;14(8):475–488.
- Inui M, Martello G, Piccolo S. MicroRNA control of signal transduction. *Nat Rev Mol Cell Biol*. 2010;11(4):252–263.
- Baumjohann D, Ansel KM. MicroRNA-mediated regulation of T helper cell differentiation and plasticity. *Nat Rev Immunol*. 2013;13(9):666–678.
- Yoo AS, Staahl BT, Chen L, et al. MicroRNA-mediated switching of chromatin-remodeling complexes in neural development. *Nature*. 2009;460(7255):642–646.
- Bader AG, Brown D, Stoudemire J, et al. Developing therapeutic microRNAs for cancer. *Gene Ther*. 2011;18(12):1121–1126.
- Ling H, Fabbri M, Calin GA. MicroRNAs and other non-coding RNAs as targets for anticancer drug development. *Nat Rev Drug Discov*. 2013;12(11):847–865.
- Lujambio A, Lowe SW. The microcosmos of cancer. *Nature*. 2012;482(7385):347–355.
- Rajewsky N. MicroRNA target predictions in animals. *Nat Genet*. 2006;38(Suppl):S8–S13.
- Min H, Yoon S. Got target? Computational methods for microRNA target prediction and their extension. *Exp Mol Med*. 2010;42(4):233–244.
- Orom UA, Lund AH. Experimental identification of microRNA targets. *Gene*. 2010;451(1–2):1–5.
- Allmer J, Yousef M. Computational methods for ab initio detection of microRNAs. *Front Genet*. 2012;3:209.
- Gomes CP, Cho JH, Hood L, et al. A review of computational tools in microRNA discovery. *Front Genet*. 2013;4:81.
- Thomson DW, Bracken CP, Goodall GJ. Experimental strategies for microRNA target identification. *Nucleic Acids Res*. 2011;39(16):6845–6853.
- Lindow M, Gorodkin J. Principles and limitations of computational microRNA gene and target finding. *DNA Cell Biol*. 2007;26(5):339–351.
- Li L, Xu J, Yang D, et al. Computational approaches for microRNA studies: a review. *Mamm Genome*. 2010;21(1–2):1–12.
- Wang C, Wei L, Guo M, et al. Computational approaches in detecting non-coding RNA. *Curr Genomics*. 2013;14:371–377.
- Leclercq M, Diallo AB, Blanchette M. Computational prediction of the localization of microRNAs within their pre-miRNA. *Nucleic Acids Res*. 2013;41(15):7200–7211.
- Barbato C, Arisi I, Frizzo ME, et al. Computational challenges in miRNA target predictions: to be or not to be a true target? *J Biomed Biotechnol*. 2009;2009:803069.
- Winter J, Jung S, Keller S, et al. Many roads to maturity: microRNA biogenesis pathways and their regulation. *Nat Cell Biol*. 2009;11(3):228–234.
- Lee Y, Jeon K, Lee JT, et al. MicroRNA maturation: stepwise processing and subcellular localization. *EMBO J*. 2011;21(17):4663–4670.
- Berezikov E. Evolution of microRNA diversity and regulation in animals. *Nat Rev Genet*. 2011;12(12):846–860.
- Guo L, Zhao Y, Zhang H, et al. Integrated evolutionary analysis of human miRNA gene clusters and families implicates evolutionary relationships. *Gene*. 2014;534(1):24–32.
- Neilson JR, Zheng GX, Burge CB, et al. Dynamic regulation of miRNA expression in ordered stages of cellular development. *Genes Dev*. 2007;21(5):578–589.
- Zhu JK. Reconstituting plant miRNA biogenesis. *Proc Natl Acad Sci U S A*. 2008;105(29):9851–9852.
- Thakur V, Wanchana S, Xu M, et al. Characterization of statistical features for plant microRNA prediction. *BMC Genomics*. 2011;12:108.
- Zhou X, Ruan J, Wang G, et al. Characterization and Identification of microRNA core promoters in four model species. *Plos Comput Biol*. 2007;3(3):e37.
- Merchan F, Boualem A, Crespi M, et al. Plant polycistronic precursors containing non-homologous microRNAs target transcripts encoding functionally related proteins. *Genome Biol*. 2009;10(12):R136.
- Guerra-Assuncao JA, Enright AJ. Large-scale analysis of microRNA evolution. *BMC Genomics*. 2012;13:218.
- Li A, Mao L. Evolution of plant microRNA gene families. *Cell Res*. 2007;17(3):212–218.
- Maher C, Stein L, Ware D. Evolution of Arabidopsis microRNA families through duplication events. *Genome Res*. 2005;16(4):510–519.
- Li Y, Li C, Xia J, et al. Domestication of transposable elements into microRNA genes in plants. *PLoS One*. 2011;6(5):e19212.
- Marco A, Ninova M, Ronshaugen M, et al. Clusters of microRNAs emerge by new hairpins in existing transcripts. *Nucleic Acids Res*. 2013;41(16):7745–7752.
- Axtell MJ, Westholm JO, Lai EC. Vive la difference: biogenesis and evolution of microRNAs in plants and animals. *Genome Biol*. 2011;12(4):221.
- Griffiths-Jones S, Saini HK, van Dongen S, et al. miRBase: tools for microRNA genomics. *Nucleic Acids Res*. 2008;36(Database issue):D154–D158.
- Bonnet E, Wuyts J, Rouze P, et al. Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics*. 2004;20(17):2911–2917.
- Borenstein E, Ruppin E. Direct evolution of genetic robustness in microRNAs. *Proc Natl Acad Sci U S A*. 2006;103(17):6593–6598.
- Szollósi GJ, Derenyi I. Congruent evolution of genetic and environmental robustness in micro-RNA. *Mol Biol Evol*. 2009;26(4):867–874.
- Park JE, Heo I, Tian Y, et al. Dicer recognizes the 5' end of RNA for efficient and accurate processing. *Nature*. 2013;475(7355):201–205.

42. Auyeung VC, Ulitsky I, McGeary SE, et al. Beyond secondary structure: primary–sequence determinants license pri–miRNA hairpins for processing. *Cell*. 2013;152(4):844–858.
43. Burke JM, Kelenis DP, Kincaid RP, et al. A central role for the primary microRNA stem in guiding the position and efficiency of Drosha processing of a viral pri–miRNA. *RNA*. 2014;20(7):1068–1077.
44. Zuker M, Sankoff D. RNA secondary structure and their prediction. *Bulletin of Mathematical Biology*. 1984;46(4):591–621.
45. Dreyfus SE, Law AM. *The art and theory of dynamic programming*. 1st ed. Academic Press ISBN 978–0–12–221860–6; 1977.
46. Eddy SR. What is dynamic programming? *Nat Biotechnol*. 2004;22(7):909–910.
47. Garcia–Martin JA, Clote P, Dotu I. RNAiFOLD: A constraint programming algorithm for RNA inverse folding and molecular design. *J Bioinform Comput Biol*. 2013;11(2):1350001.
48. Hofacker IL, Fekete M, Stadler PF. Secondary structure prediction for aligned RNA sequences. *J Mol Biol*. 2002;319(5):1059–1066.
49. Spirollari J, Wang JT, Zhang K, et al. Predicting consensus structures for RNA alignments via pseudo–energy minimization. *Bioinform Biol Insights*. 2009;3:51–69.
50. Washietl S, Hofacker IL. Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. *J Mol Biol*. 2004;342(1):19–30.
51. Washietl S, Will S, Hendrix DA, et al. Computational analysis of noncoding RNAs. *Wiley Interdiscip Rev RNA*. 2012;3(6):759–778.
52. Dimitrov R, Gouliamova D. New method for sequence alignment based on probabilities of nucleotide correspondences. *Biotechnology & Biotechnological Equipment*. 2012;26(Suppl 1):218–223.
53. Dimitrov RA, Gouliamova DE. Biological sequence comparison, molecular evolution and phylogenetics. *Biotechnology & Biotechnological Equipment*. 2012;26(1):209–217.
54. Gorodkin J, Hofacker IL. From Structure prediction to genomic screens for novel non–coding RNAs. *PLoS Comput Biol*. 2011;7(8):e1002100.
55. Sankoff D. Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM Journal of Applied Mathematics*. 1985;45:810–825.
56. Dimitrov RA. A New Formalism for Calculation of the Partition Function of Single Stranded Nucleic Acids Bulg. *J Phys*. 2005;32:220–235.
57. Muckstein U, Tafer H, Hackermuller J, et al. Thermodynamics of RNA–RNA binding. *Bioinformatics*. 2006;22(10):1177–1182.
58. Wang X. Composition of seed sequence is a major determinant of microRNA targeting patterns. *Bioinformatics*. 2014;30(10):1377–1383.
59. Kertesz M, Iovino N, Unnerstall U, et al. The role of site accessibility in microRNA target recognition. *Nat Genet*. 2007;39(10):1278–1284.
60. Lange SJ, Maticzka D, Mohl M, , et al. Global or local? Predicting secondary structure and accessibility in mRNAs. *Nucleic Acids Res*. 2012;40(12):5215–5226.
61. Marin RM, Vanicek J. Efficient use of accessibility in microRNA target prediction. *Nucleic Acids Res*. 2011;39(1):19–29.
62. Willenbrock H, Salomon J, Sokilde R, et al. Quantitative miRNA expression analysis: comparing microarrays with next–generation sequencing. *RNA*. 2009;15(11):2028–2034.
63. Linsley PS, Schelter J, Burchard J, et al. Transcripts targeted by the microRNA–16 family cooperatively regulate cell cycle progression. *Mol Cell Biol*. 2007;27(6):2240–2252.
64. Chen CZ. An unsolved mystery: the target–recognizing RNA species of microRNA genes. *Biochimie*. 2013;95(9):1663–1676.
65. Robins H, Press WH. Human microRNAs target a functionally distinct population of genes with AT–rich 3' UTRs. *Proc Natl Acad Sci U S A*. 2005;102(43):15557–15562.
66. Maragkakis M, Alexiou P, Papadopoulos GL, et al. Accurate microRNA target prediction correlates with protein repression levels. *BMC Bioinformatics*. 2009;10:295.
67. Chen K, Rajewsky N. The evolution of gene regulation by transcription factors and microRNAs. *Nat Rev Genet*. 2007;8(2):93–103.
68. Didiano D, Hobert O. Perfect seed pairing is not a generally reliable predictor for miRNA–target interactions. *Nat Struct Mol Biol*. 2006;13(9):849–851.
69. Guerra–Assuncao JA, Enright AJ. MapMi: automated mapping of microRNA loci. *BMC Bioinformatics*. 2010;11:133.
70. Gu S, Kay MA. How do miRNA mediate translational repression? *Silence*. 2010;1(1):11.
71. Hafner M, Ascano M Jr, Tuschl T. New insights in the mechanism of microRNA–mediated target repression. *Nat Struct Mol Biol*. 2011;18(11):1181–1182.
72. Pasquinelli AE. MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship. *Nat Rev Genet*. 2012;13(4):271–282.
73. Sethupathy P, Megraw M, Hatzigeorgiou AG. A guide through present computational approaches for the identification of mammalian microRNA targets. *Nat Method*. 2006;3(11):881–886.
74. Wang Z, Gerstein M, Snyder M. RNA–Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*. 2009;10(1):57–63.
75. Neilsen CT, Goodall GJ, Bracken CP. IsomiRs – the overlooked repertoire in the dynamic microRNAome. *Trends Genet*. 2012;28(11):544–549.
76. Kloosterman WP, Plasterk RH. The diverse functions of microRNAs in animal development and disease. *Dev Cell*. 2006;11(4):441–450.
77. Vogel C, Marcotte EM. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat Rev Genet*. 2012;13(4):227–232.