Prediction of microRNA targets

Pierre Mazière and Anton J. Enright

Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK

Recently, microRNAs (miRNAs) have been shown to be important regulators of genes in many organisms and have already been implicated in a growing number of diseases. MiRNAs are short (21–23 nucleotides) RNAs that bind to the 3′ untranslated regions of target genes. This binding event causes translational repression of the target gene and, evidence now suggests, also stimulates rapid degradation of the target transcript. miRNAs represent a new species of regulator, controlling the levels of potentially large numbers of proteins, many of which might be important drug targets. The expression of miRNAs shows that they are highly differentially expressed, with specific miRNAs active in certain tissues at certain times. In many cancers, miRNA expression is significantly altered, and this has been shown to be a useful diagnostic tool. Several computational approaches have been developed for the prediction of miRNA targets.

Introduction

Currently, there are 474 confirmed microRNs (miRNAs) in humans [1], although there might be many more. miRNAs are expected to have multiple targets; however, few have been confirmed experimentally (only 66 of potentially thousands [2] so far). In the absence of high-throughput experimental techniques to determine the targets of miRNAs, it is vital that computational techniques are developed to unravel their regulatory effects and implications for diseases and diagnostics. Indeed, recent studies have already implicated miRNAs in numerous human diseases, such as colorectal cancer [3], chronic lymphocytic leukaemia [4] and fragile X syndrome [5]. Hence, both the miRNA itself and its regulatory targets are potentially druggable.

The prediction of miRNA targets has been ongoing since the 3′ untranslated regions (3′UTRs) of transcripts were determined to contain binding sites for them [6]. The efficacy of computational approaches to locate and rank potential genomic binding sites is supported by the relatively high degree of miRNA complementarity to experimentally determined binding sites. Despite the later identification of hundreds of miRNAs in a variety of species, through large-scale and sequencing projects [6–9], only a handful of targets had been identified experimentally, for an even smaller number of miRNAs [10–13]. Given the laborious nature of experimental validation of targets, and despite the limited data available, it was imperative that computational approaches be developed that could produce reliable and testable predictions.

Corresponding author: Enright, A.J. (aje@sanger.ac.uk)

Target prediction issues

Researchers initially determined miRNA target transcripts through experiment; they then identified potential sites by manually searching the target transcript for matching locations; and, finally, they confirmed sites through site-directed mutagenesis or other techniques. When the first few target sites had been identified for miRNAs, such as let-7 and lin-4 [12–14], it was obvious that miRNAs had relatively clearly defined patterns of complementarity to the 3′UTRs of their target transcripts. These first few targets detected for *Caenorhabditis elegans* (Figure 1) had enough similarity that it became apparent that clever computational techniques might enable their discovery *in silico*. However, computational analysis is hampered by various issues, which will now be described.

miRNA size

The apparent complementarity between miRNA and target could have been seen as an advantage for computational analysis. However, other features of miRNA–UTR associations make matters more complicated. Conventional sequence alignment algorithms

1359-6446/06/$ - see front matter doi:10.1016/j.drudis.2007.04.002

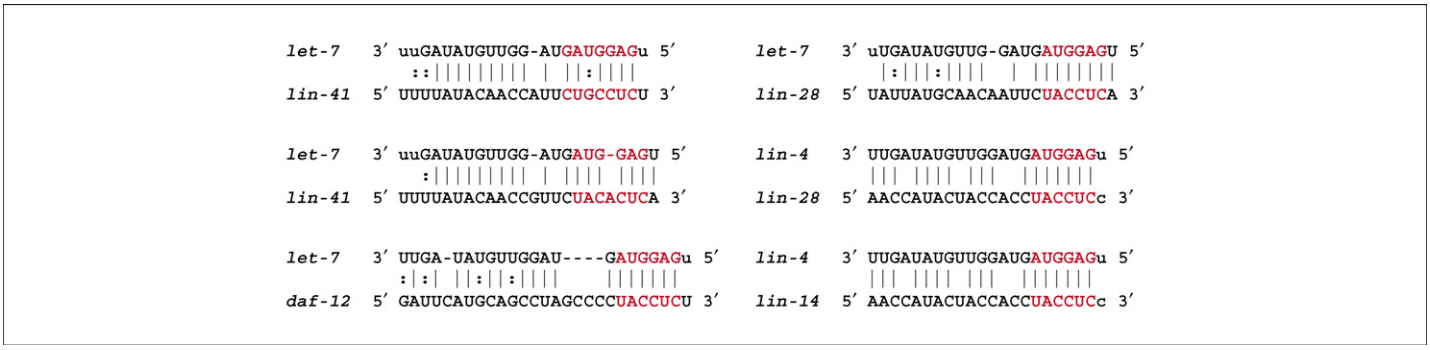

FIGURE 1

Examples of miRNA target duplexes for the first experimentally validated miRNA targets in genes of *C. elegans*. In this case, miRNAs are oriented 3′ to 5′. Red highlighting indicates the 'seed region' of the binding site. Sites are derived using miRanda (version 2.0) on known miRNA target pairs [22].

assume longer sequences than the 20–23 nucleotides of miRNAs. This short length makes ranking and scoring of targets difficult because statistical techniques for sequence matching (such as Karlin–Altschul statistics [15]) require longer sequences. Binding sites actually consist of regions of complementarity, bulges and mismatches [16]. Because standard sequence analysis tools were designed for sequences with longer stretches of matches and fewer gaps, they are much less useful for miRNA target prediction. Recently, position 2–7 of miRNAs, the so-called 'seed' region, has been described as a key specificity determinant of binding, and requires perfect complementarity [17,18]. If one ignores GC content and performs an order of magnitude calculation, then a perfect match for a six-nucleotide seed region of a miRNA should occur approximately once in every 1.3 kilobases in a genome – in other words, on average, almost once in every human 3′UTR. However, it would not seem realistic for a single miRNA to regulate more than a few hundred targets. Effective regulation of transcript translation requires that miRNAs and their targets are located in the same cellular compartments. Hence, most of these theoretical targets correspond to false positives.

Identification of 3′UTRs

To identify miRNA targets in a given species, knowledge of the set of 3′UTRs for this species is a vital step. Despite the accumulation of genome sequences for many species, the location, extent or splice variation of 3′UTRs is still poorly characterized for many mammals. Some species-specific projects, such as the Berkeley *Drosophila* Genome Project (BDGP), produce high-quality transcript information that makes possible the accurate determination of a 3′UTR, from stop codon to polyadenylation site. For other species, such as *Homo sapiens*, some transcripts are well defined, whereas others remain poor in their description [19]. The Ensembl database uses alignment of cDNAs and expressed sequence tags to genomic sequences to extract 3′UTR regions, and so far, evidence is available for human, mouse and zebrafish genomes. However, ~30% of human genes lack definitive 3′UTR boundaries. These regions can be estimated by selecting a downstream flanking sequence of the stop codon, corresponding to the length of an average human 3′UTR (e.g. 1 kilobase). Experimental techniques, such as tiling arrays [20], and ditag or cage tagging, seem to be promising approaches for the generation of high-quality 3′UTR datasets. Attaining reliably annotated and verified 3′UTR datasets will potentially benefit target prediction more than making small improvements to existing prediction methods. In the context of drug discovery, both 3′UTRs and miRNA genes represent drug target candidates through either the generation of synthetic miRNAs or the repression or overexpression of existing miRNAs.

Conservation analysis

Solutions to reduce the number of false positives in target predictions include filtering out those binding sites that do not seem to be conserved across species. The use of predicted binding sites conserved across orthologous 3′UTRs in multiple species are considered more likely to reduce the number of false positives [18,21,22]. However, recently evolved miRNAs, such as miR-430 in zebrafish, might not have conserved targets [6] in the scope of the currently available set of fish genomes. One caveat of conservation analysis concerns the set of species that are compared: looking for conserved targets between humans and chimpanzees will not be helpful, given that at least 99% of the entire transcript will be conserved. Other species might seem more relevant for comparing with human transcripts (e.g. mouse, rat, or dog), but the fact is that genomes are not sequenced according to their evolutionary distances. As a result, the number of false positives can effectively be greatly reduced but this is at the expense of increased false negatives.

Large-scale versus small-scale prediction

Typically, when performing large-scale prediction of targets across a whole genome, the aim is to achieve a higher degree of specificity (few false-positives) compared with sensitivity (few false negatives), to ensure predictions of better quality, albeit fewer of them [23]. Many of the published algorithms and released databases choose such an approach; however, for an individual researcher interested in a single gene or pathway, seeking to investigate a potential role for miRNAs in their system, then sensitivity becomes more important, so that a pool of predicted targets is produced that can be individually tested. This second approach is most probably the one of interest for the development of potential drug therapies based on miRNAs. As will be discussed later, the selection of the target prediction approach should depend on the particular requirements of the researcher.

Computational target-prediction approaches

Different methods have been developed for computational target prediction (Table 1). These might or might not be made available

TABLE 1

Methods and resources for miRNA target prediction

Method	Type of method	Refs	Method availability	Data availability	Resource
Stark *et al.*	Complementarity	[21]	Online search	Yes	http://www.russell.embl.de/miRNAs/
miRanda	Complementarity	[22]	Download	Yes	http://www.microrna.org/
miRanda miRBase	Complementarity	[1]	Online search	Yes	http://microrna.sanger.ac.uk/
TargetScan	Seed complementarity	[18]	Online search	Yes	http://www.targetscan.org/
TargetScanS	Seed complementarity	[17]	Online search	Yes	http://www.targetscan.org/
DIANA microT	Thermodynamics	[24]	Download	Yes	http://diana.pcbi.upenn.edu/
PicTar	Thermodynamics	[33]		Yes	http://pictar.bio.nyu.edu/
RNAHybrid	Thermodynamics and statistical model	[25]	Download		http://bibiserv.techfak.uni-bielefeld.de/rnahybrid/
miTarget	SVMe	[37]	Online Search		http://cbit.snu.ac.kr/~miTarget/
TarBase	Experimentally validated targets		N/A	Yes	http://diana.pcbi.upenn.edu/tarbase.html

Abbreviation: N/A, not available.

as functional packages but the results are always available, at least as a precomputed set of transcripts, through online resources (see resources section below; Table 1).

Algorithms for miRNA target prediction

The challenge of predicting miRNA targets has resulted in the development of several methods, which fall into different categories [24]. We can distinguish three types of target sites: 5′-dominant canonical, 5′-dominant seed only and 3′-compensatory (Figure 2). These differ in the level of complementarity of miRNA sequences to the site sequences. Therefore, the main approaches look for sequence complementarity and/or for favourable miRNA–target duplex thermodynamics. To increase the signal-to-noise ratio, some methods require strict complementarity between the seed region [17,18] of the miRNA and the predicted target. Conservation of binding sites is also often used as a metric to improve the raw results.

Complementarity searching

The first algorithms did not develop statistical background models to evaluate the significance of each detected hit [25]; rather, they were oriented towards recovery of known targets and the detection of further targets for experimental validation, so that our knowledge of miRNA binding dynamics might be improved. In most

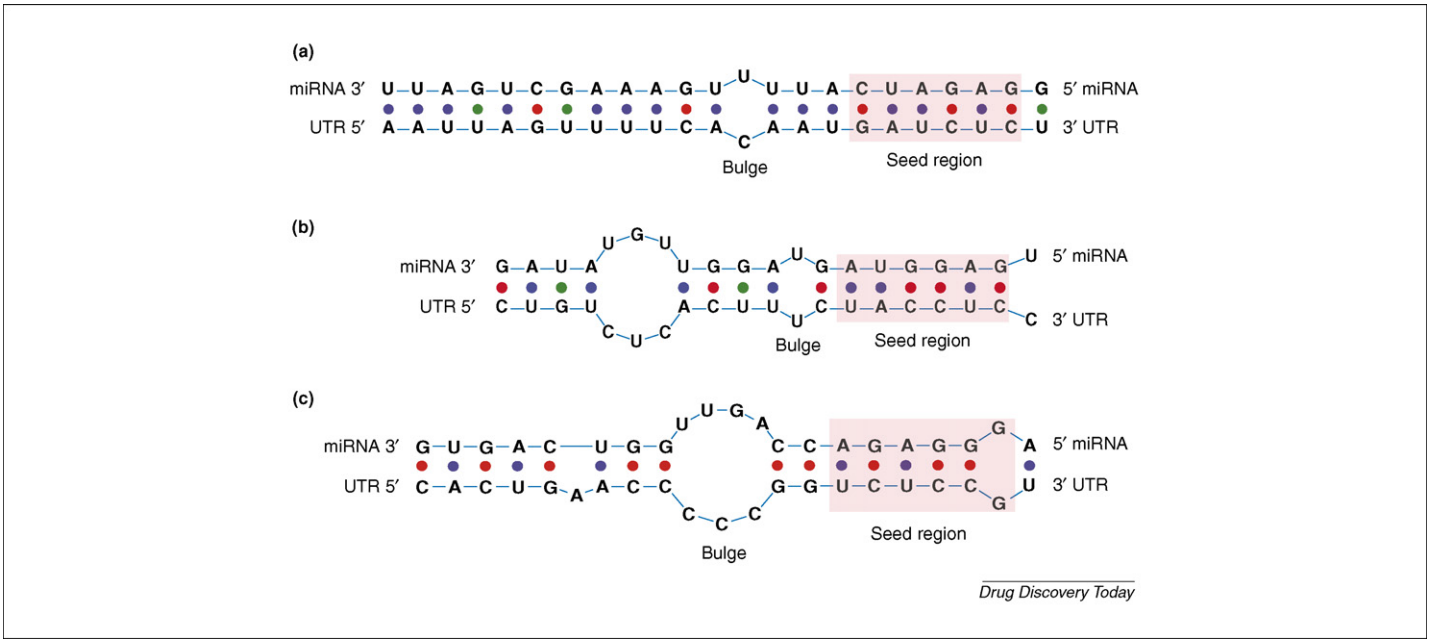

FIGURE 2

Approximate secondary structures of the three main types of target site duplex. **(a)** Canonical sites have good or perfect complementarity at both the 5′ and 3′ ends of the miRNA, with a characteristic bulge in the middle. **(b)** Dominant seed sites have perfect seed 5′ complementarity to the miRNA but poor 3′ complementarity. **(c)** Compensatory sites have a mismatch or wobble in the 5′ seed region but compensate through excellent complementarity at the 3′ end.

cases, they used complementarity initially to identify potential targets, followed by iterative rounds of filtering based on thermodynamics, binding site structure and conservation [18,21–23]. After these filtering steps, a score is typically applied to each detected target; this score can be useful for target ranking. Initial attempts at false-positive rate estimation usually relied on comparing detection methods for real miRNAs and shuffled control miRNAs.

The method of Stark and co-workers

Large-scale prediction of miRNA targets was first successfully published for *Drosophila melanogaster* [21], for which well-annotated and accurate 3′UTRs could be obtained from the BDGP [26]. The sequence search tool HMMer [27] was used to identify the reverse complement miRNA sequences. Similar profiles were built to enable G:U wobble matches. Following the prediction algorithm, the resulting 3′UTRs were filtered for conservation in *Drosophila pseudoobscura* and *Anopheles gambiae*. The detected target sites were scored and used as an input for the MFold algorithm [28], to evaluate the thermodynamic stability of the miRNA–target association. Despite a statistical model based on the normal distribution rather than the extreme value distribution [25] (see below), the method predicted previously validated *D. melanogaster* miRNA binding sites. Many previously unknown binding sites were also predicted, six of which were experimentally validated [21].

miRanda

The miRanda algorithm was the second method to be published [22]. As with the method by Stark *et al.*, miRanda identifies potential binding sites by looking for high-complementarity regions on the 3′UTRs. The scoring matrix used by the algorithm is built so that complementary bases at the 5′ end of the miRNA are rewarded more than those at the 3′ end. Hence, the binding sites exhibiting a perfect or almost-perfect match at the seed region of miRNAs display a better score. The resulting binding sites are then evaluated thermodynamically, using the Vienna RNA folding package [29]. This first version of miRanda successfully predicted many known targets in *D. melanogaster*. The BDGP 3′UTRs dataset was used, and the results were filtered, as described above, to limit predictions to targets conserved in *D. pseudoobscura*. When classified according to gene ontology terms [30], the miRanda-predicted targets were shown to display specific functional patterns for each miRNA. Expression data analysis confirmed this property by suggesting that many individual miRNAs have highly specific roles in particular tissues, processes and pathways [31]. When basic parameter settings are used, the approximated false-positive rate was between 24% and 39%. These values are significantly decreased when multiple sites are considered. Newer miRanda versions [31] implement a strict model for the binding sites that requires almost-perfect complementarity in the seed region with only a single wobble pairing. Other variations of these algorithms are currently under development for a version 3.0 of miRanda [1]. These incorporate a statistical model equivalent to that used by RNAHybrid [25], thereby they efficiently reduce the rate of false-positive predictions. Despite their similar methods and identical input datasets, the scoring and ranking strategies devised by Stark *et al.* [21] and miRanda [22] are different: only 40% of the respective top-ten miRNA targets predicted by both methods overlap. It seems that even small differences in the criteria used for ranking and scoring lead to large differences in the set of predicted targets. The multiplicity of miRNA binding sites on the same 3′UTRs drastically improves their statistical significance in both methods. This is confirmed by experimental evidence showing that multiple sites enhance the silencing effect of miRNAs [32]. However, many miRNAs still seem to operate at a single site on their targets. One given explanation implies that a miRNA exhibiting high complementarity to its single-site target could have the same regulatory effect as a miRNA with a lower level of complementarity but a multiple-site target [32].

TargetScan and TargetScanS

Although the previously mentioned methods attempt to find all potentially complementary sites and then filter them according to different criteria, TargetScan [18] uses a different approach. This method requires perfect complementarity to the seed region of a miRNA and then extends these regions to unravel complementarity outside the region. This aims at filtering many false positives from the beginning of the prediction process. For the same purpose, the conservation criteria, based on the presence of the seed region in an island of conservation, are introduced early in the process by using groups of orthologous 3′UTRs as input data. The following step is common to the other methods: the predicted binding sites are tested for their thermodynamic stability, in this case with RNAFold from the Vienna Package [29]. TargetScan was the first method to be applied for human miRNA target prediction, using mouse, rat and fish genomes for conservation analysis [18]. Shuffled sequences, with maintained dinucleotide compositions that mimic real 3′UTRs, were used to determine the significance of binding sites. The estimated false-positive rate varies between 22% and 31%. The method was shown to predict not only known miRNA binding sites but also novel sites. Luciferase reporter constructs validated 11 of the 15 tested sites [18]. TargetScanS simplified the TargetScan method and improved the target prediction fidelity [17]. The miRNA complementarity is now limited to a six-nucleotide seed, followed by an additional 3′ match of an adenosine anchor at position 1. No other criteria are required once the previous conditions are met; contrary to previous algorithms, single-site 3′UTRs are sufficient for a reliable prediction. TargetScan and TargetScanS feature an efficient reduction in the false-positive rate but, because of the required strict complementarity in the seed region, loosely conserved targets and those containing wobble pairings are more likely to be missed, including 3′-compensatory sites.

PicTar

The PicTar algorithm uses a group of orthologous 3′UTRs from multiple species as the input dataset [33]. The algorithm scans the alignments of 3′UTRs for those displaying seed matches to miRNAs. The retained alignments are then filtered according to their thermodynamic stability. Each predicted target is scored by using a Hidden Markov Model (HMM) maximum-likelihood fit approach. PicTar is the first method that uses the criteria of co-expression in space and time of miRNAs and their targets. The experimental validation of seven out of 13 tested predicted targets, as well as the confirmation of eight of nine previously known targets, demonstrates the efficiency of the algorithm.

Thermodynamic-based algorithms

Interestingly, it is still not clear whether sequence or structure is the better predictor of a miRNA binding site. Although the above methods focus on first finding complementarity and then analysis of thermodynamics, other methods use thermodynamics as the initial indicator of miRNA-binding site potential.

DIANA-microT

The DIANA-microT method uses a 38-nucleotide window that is progressively moved across a 3′UTR sequence [24]. Using dynamic programming, the free energy (ΔG kcal/mol) of the potential binding site is calculated at each step and compared with the results obtained from shuffled sequences with the same dinucleotide composition as real 3′UTRs. Contrary to sequence complementarity-based methods, DIANA-microT demands 3′ complementarity to the miRNA and does not bother with site multiplicity. Using this technique, all currently known *C. elegans* miRNA binding sites were predicted successfully, with false-positive rates similar to those found in previously described methods.

RNAHybrid

The lack of strong statistical models is one of the main criticisms that can be levelled at the methods previously described here. RNAHybrid was the first method to address this issue by developing a model as robust as those used for large-scale sequence comparison [25]. Contrary to tools such as MFold and Vienna, which are designed for single-sequence folding and therefore require an artificial linker between the miRNA and its potential binding site [33], RNAHybrid identifies regions in the 3′UTRs that have the potential to form a thermodynamically favourable duplex with a specific miRNA. The maximum free energy of a miRNA is calculated for every 3′UTRs of a set of shuffled 3′UTR sequences with maintained dinucleotide frequencies. Normalisation for both 3′UTR and miRNA length using $S_{norm} = \log(S/mn)$ is applied to these energies. Random energies derived in this manner should exhibit an extreme value distribution (EVD). Subsequently, the parameters of the EVD that best describe the data for a given miRNA are empirically calculated using the derived distribution from shuffled sequences. Each hit to any 3′UTR for this miRNA is then assigned a *P* value calculated directly from these parameters. Hence, at the scanning stage, miRNAs are scanned against a database of real 3′UTRs, and each hit is compared with the expected distribution and assigned a *P* value.

Moreover, the statistical model implemented in RNAHybrid takes into account multiple sites and conserved sites, by respectively combining individual *P* values using Poisson statistics and calculating conservative *P* values for conserved sites. A statistical fitting approach corrects for highly conserved 3′UTRs by evaluating the overall conservation in the group of sequences compared with the conservation at the site. The resulting statistics cover individual site quality, quantity of sites, whether they are conserved and how significant this conservation is, given the input sequences. The method was successfully tested to predict known targets in *D. melanogaster*, with a low false-positive rate. The association of *P* values with predicted targets is an appreciable asset for directly comparing predicted binding sites.

Motif-mining approaches

The motif-mining approaches work from an opposite angle compared with previous sequence-based methods, by looking in genomic 3′UTR sequences for overrepresented motifs of six, seven or eight nucleotides. These motifs are then compared with miRNA seed regions to predict potential miRNA targets.

Whole genome

Xie *et al.* [34] published a large-scale analysis of common motifs in 3′UTR sequences. Although their primary goal was not the detection of miRNA binding sites, a large number of eight-nucleotide conserved motifs matching the seed region of known mammalian miRNAs was identified in humans and other mammals. Other motifs indicated the presence of new miRNAs identified by the authors.

Gene expression-based studies

In a similar approach, Giraldez *et al.* [35] used microarray expression studies to identify 3′UTR binding sites for miRNAs. The main idea was to suppress miRNA regulatory effects in specific cells or tissues and to measure the resulting changes in transcript expression. This was done by measuring the expression profiles of a zebrafish embryo, whose miRNAs were not able to mature because of a mutation of the Dicer protein (MZDicer) [36]. These expression data were then compared with the expression profiles of another MZDicer fish, which had been injected with miRNA miR-430, and with wild-type fish. This injection was expected to result in a degradation of the targeted transcripts measurable on an array. A measurement greater than a twofold underexpression in MZDicer$^+$ miR430 compared with wild-type for a given transcript can be considered as an experimental prediction of a miRNA target. The question remains as to whether the underexpression of a transcript is the direct effect of the miRNA targeting this transcript, the consequence of the degradation of another transcript, or just noise. A strong signal for six-, seven- and eight-nucleotide patterns directly complementary to miR-430 confirmed that most identified transcripts seem to be direct targets of this miRNA. Green fluorescent protein reporter constructs enabled the confirmation of a large number of these targets. This study led to the conclusion that a single six-nucleotide match is not required to imply miRNA regulation, but that multiple six-nucleotide sites or a single eight-nucleotide match could explain at least 90% of the validated targets [35].

Support vector machine method

An alternative way of predicting miRNA targets would be to gather all of the known miRNA–target associations, to determine a certain number of features describing these associations and to build a statistical model that would fit these features. This was the method used to develop the support vector machine (SVM) classifier miTarget [37]. The SVM was trained to recognize 41 features, based on results produced by the Vienna RNA package [29]. These features can be categorized in three elements: structural features, thermodynamic features and position-based features. Although the first two categories correspond to properties already described, the position-based features attempt to describe more accurately the mechanism of the seed pairing. The introduction of these position-based features could potentially increase the specificity of miTarget compared with other algorithms. However, for SVMs to

work well, they normally require a large negative training set, which is not currently available for miRNA targets. Considering that this limitation might be rapidly overcome, SVM-based approaches remain a promising solution.

Performance of target-prediction methods against validated targets

It remains difficult to assess accurately the performance of many of the methods listed above. Traditionally, this is because few validated miRNA targets are known. Hence, although the methods published tend to be able to predict the few known targets, these constitute a small proportion of target predictions overall. As a result, initial estimates of false-positive rates tend to use sequence-shuffling approaches to approximate error rates. Recently, larger numbers of validated target predictions have become available [2], enabling such analyses to be performed. However, the problem remains that many methods are not available to be downloaded for inclusion in independent testing on a common dataset. The datasets used by these methods, to prove their efficiency, are highly diverse; therefore, a direct comparison of these assessments is not possible.

Recent work [2] compared the performance of five methods, namely miRanda, TargetScan, TargetScanS, PicTar and DIANA-microT, using a set of 84 miRNA–target gene interactions, involving 32 miRNAs. The sensitivity and specificity of the five methods were measured. The conclusion of this work describes the sensitivity and specificity tradeoff in target prediction algorithms – that is, an increased sensitivity can be obtained at the expense of specificity. TargetScanS, PicTar and miRanda seem to be the best methods, with sensitivity values ranging between 65% and 68%. The latter seems to be slightly less efficient in term of specificity, although this issue is tackled for the next versions to be released.

Resources for miRNA target prediction

For a project requiring information about miRNA targets, the possibilities are limited to either running one of the above algorithms, if they are available, or looking for sets of precomputed targets. The first option gives the user more flexibility in setting crucial parameters and thresholds, whereas the second is a more straightforward solution when databases offer a ready-to-use dataset for the gene of interest.

Using stand-alone packages

Although some of the above algorithms are not stand-alone packages, or might be difficult to use, some of them are well documented and available to be downloaded (Table 1). The strategies of new drug development greatly benefit from these freely accessible packages because they enable the research teams to exert fine control on parameters and guarantee the privacy of the resulting predicted targets. The miRanda package [22] [Wellcome Trust Sanger Institute (http://www.sanger.ac.uk/and Memorial Sloan–Kettering Cancer Center (http://www.mskcc.org/)] and RNA Hybrid [25] (University of Bielefeld; http://www.uni-bielefeld.de/International/) are two examples of privacy guaranteed packages. The first of these aims at giving the user an overview of the miRNA that might target a specific gene. The sequences of 3′UTRs and miRNA are supplied in FASTA format, and the algorithm scans each miRNA against each 3′UTR; it then reports potential binding sites. Default settings result in a relatively loose set of predictions that might contain a large number of false positives. This can be avoided by choosing a more stringent set of parameters. More complex statistical analysis, with a probabilistic score for each prediction, is expected for the future versions of miRanda.

Using web resources

Several online resources have been published (Table 1). Some of these reflect the results of studies published in different journals and are not supposed to be updated; others are regularly modified to include new miRNA sequences or improved sequenced 3′UTRs. The latter resources are organized in such a way that enables the user either to look for a specific miRNA (miRNA-centric) or to identify a binding site on the gene of interest (gene-centric). This information must be considered cautiously, given the rate of false positives indicated for the different algorithms described earlier. We expect that, as larger numbers of miRNA targets are validated experimentally, these resources will enable users to query both computational predictions and experimentally validated miRNA targets.

Conclusion

The rapid development of computational methods for miRNA target prediction is promising for future research. As the understanding of miRNA binding biology increases, it can be expected that existing algorithms will become progressively more accurate. New types of algorithms exploiting this novel information will undoubtedly be released. In addition, online resources are expected to be extended, providing researchers with useful tools and data for assessing the impact of miRNAs on the gene or biological process of interest. As such, the miRNA target prediction field is already a promising asset for new drug discovery. In parallel, improvements to the underlying sequence data will enable better delineation of 3′UTRs and splicing events that will further improve computational approaches for miRNA target prediction.

References

1 Griffiths-Jones, S. *et al.* (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.* 34 (Database issue), D140–D144

2 Sethupathy, P. *et al.* (2006) TarBase: a comprehensive database of experimentally supported animal microRNA targets. *RNA* 12, 192–197

3 Michael, M.Z. *et al.* (2003) Reduced accumulation of specific microRNAs in colorectal neoplasia. *Mol. Cancer Res.* 1, 882–891

4 Calin, G.A. *et al.* (2002) Frequent deletions and down-regulation of micro-RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukaemia. *Proc. Natl. Acad. Sci. U.S.A.* 99, 15524–15529

5 Caudy, A.A. *et al.* (2002) Fragile X-related protein and VIG associate with the RNA interference machinery. *Genes Dev.* 16, 2491–2496

6 Lee, R.C. *et al.* (1993) The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell* 75, 843–854

7 Lagos-Quintana, M. *et al.* (2002) Identification of tissue-specific microRNAs from mouse. *Curr. Biol.* 12, 735–739

8 Lagos-Quintana, M. *et al.* (2001) Identification of novel genes coding for small expressed RNAs. *Science* 294, 853–858

9 Lee, R.C. *et al.* (2001) An extensive class of small RNAs in *Caenorhabditis elegans*. *Science* 294, 862–864

10 Lee, S.S. *et al.* (2003) Daf-16 target genes that control *C. elegans* life-span and metabolism. *Science* 300, 644–647
11 Brennecke, J. *et al.* (2003) *bantam* encodes a developmentally regulated microrna that controls cell proliferation and regulates the proapoptotic gene *hid* in *Drosophila*. *Cell* 113, 25–36
12 Slack, F.J. *et al.* (2000) The *lin-41* RBCC gene acts in the *C. elegans* heterochronic pathway between the *let-7* regulatory RNA and the LIN-29 transcription factor. *Mol. Cell* 5, 659–669
13 Reinhart, B.J. *et al.* (2000) The 21-nucleotide *let-7* RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature* 403, 901–906
14 Olsen, P.H. *et al.* (1999) The *lin-4* regulatory RNA controls developmental timing in *Caenorhabditis elegans* by blocking lin-14 protein synthesis after the initiation of translation. *Dev. Biol.* 216, 671–680
15 Karlin, S. *et al.* (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. U. S. A.* 87, 2264–2268
16 Brennecke, J. *et al.* (2005) Principles of microRNA–target recognition. *PLoS Biol.* 3, e85
17 Lewis, B.P. *et al.* (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* 120, 15–20
18 Lewis, B.P. *et al.* (2003) Prediction of mammalian microRNA targets. *Cell* 11, 787–798
19 Hubbard, T. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res.* 30, 38–41
20 David, L. *et al.* (2006) A high-resolution map of transcription in the yeast genome. *Proc. Natl. Acad. Sci. U. S. A.* 103, 5320–5325
21 Stark, A. *et al.* (2003) Identification of *Drosophila* microRNA targets. *PLoS Biol.* 1, E60
22 Enright, A.J. *et al.* (2003) MicroRNA targets in *Drosophila*. *Genome Biol.* 5, R1
23 Lai, E.C. (2004) Predicting and validating microRNA targets. *Genome Biol.* 5, 115
24 Kiriakidou, M. *et al.* (2004) A combined computational–experimental approach predicts human microRNA targets. *Genes Dev.* 18, 1165–1178
25 Rehmsmeier, M. *et al.* (2004) Fast and effective prediction of microRNA/target duplexes. *RNA* 10, 1507–1517
26 Gelbart, W.M. *et al.* (1997) FlyBase: a *Drosophila* database. The FlyBase consortium. *Nucleic Acids Res.* 25, 63–66
27 Eddy, S.R. (1995) Multiple alignment using hidden Markov models. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 3, 114–120
28 Zuker, M. (1989) On finding all suboptimal foldings of an RNA molecule. *Science* 244, 48–52
29 Wuchty, S. *et al.* (1999) Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers* 49, 145–165
30 Ashburner, M. *et al.* (2002) On ontologies for biologists: the gene ontology – untangling the web. *Novartis Found. Symp.* 247, 66–80 discussion 80–83; 84–90; 244–252
31 John, B. *et al.* (2004) Human microRNA targets. *PLoS Biol.* 2, e363
32 Doench, J.G. *et al.* (2004) Specificity of microRNA target selection in translational repression. *Genes Dev.* 18, 504–511
33 Krek, A. *et al.* (2005) Combinatorial microRNA target predictions. *Nat. Genet.* 3, 495–500
34 Xie, X. *et al.* (2005) Systematic discovery of regulatory motifs in human promoters and 3′UTRs by comparison of several mammals. *Nature* 434, 338–345
35 Giraldez, A.J. *et al.* (2006) Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. *Science* 312, 75–79
36 Giraldez, A.J. *et al.* (2005) MicroRNAs regulate brain morphogenesis in zebrafish. *Science* 308, 833–838
37 Kim, S.-K. *et al.* (2006) miTarget: microRNA target gene prediction using a support vector machine. *BMC Bioinformatics* 7, 411