A PARALLEL APPROACH TO MIRNA TARGET PREDICTION

RAHUL RAJAN THADANI
(*B.Sc.* (*Hons.*), *NUS*)

A THESIS SUBMITTED FOR THE DEGREE OF
MASTER OF SCIENCE

DEPARTMENT OF BIOLOGICAL SCIENCES
NATIONAL UNIVERSITY OF SINGAPORE

2007

I am grateful to Martti, for his steadfast guidance, unstinting encouragement and infectious creativity. Thanks also to Susan, Honcheng, Xie Chao, Sarathi, Faraaz, Fatima and Yongli, without all of whom this would have been a much less fulfilling endeavour.

Summary

Averaging 22 nucleotides in length, microRNAs (miRNAs) are endogenous, post-transcriptional regulators of gene expression. They bind to target messenger RNA transcripts in a sequence specific manner, inducing mRNA degradation, translational repression or endonucleolytic cleavage. Given the fact that only a fraction of the several thousand known miRNAs have well-characterized functions, computational approaches remain an important means of studying miRNA targets. The accurate prediction of a comprehensive set of mRNAs regulated by animal miRNAs remains an open problem. In particular, the prediction of targets that do not possess evolutionarily conserved complementarity to their miRNA regulators is not adequately addressed by current tools.

I describe a novel animal miRNA target prediction algorithm, MicroTar, which is based on miRNA–target complementarity and thermodynamic data. The algorithm uses predicted free energies of unbound mRNA and putative mRNA–miRNA heterodimers, implicitly addressing the accessibility of the mRNA 3′ untranslated region. MicroTar does not rely on evolutionary conservation to discern functional targets and is able to predict both conserved and non-conserved targets. Parallelization makes feasible the use of full-molecule energy computations, rather than the intramolecular-bond-free approximations that are currently used. In addition, a statistical method is applied for determining the significance of target predictions. The algorithm is validated on sets of experimentally-verified targets in three different species; MicroTar achieves better sensitivity than a widely-used target prediction tool in all three cases.

Contents

List of Tables

List of Figures

List of Symbols

G	Gibb's free energy
g_n	Negative normalized free energy
S_i	ith nucleotide in RNA sequence
S_{ij}	Subsequence from S_i to S_j, both inclusive
$W(i,j)$	Minimum free energy (MFE) of all possible structures from S_{ij}
$V(i,j)$	MFE of all possible structures from S_{ij} with S_i and S_j paired
$\mathbf{W}$	Matrix of all $W(i,j)$
$\mathbf{V}$	Matrix of all $V(i,j)$

Chapter 1

Introduction

Averaging 22 nucleotides in length, microRNAs (miRNAs) are endogenous, small RNA regulators of gene expression at the post-transcriptional level. They bind to target messenger RNAs in a sequence-specific manner, inducing mRNA degradation, translational repression or endonucleolytic cleavage. The first miRNA, *lin-4*, was discovered in 1993 in the nematode *Caenorhabditis elegans*, in genetic screens for mutants with disrupted developmental timing [1]. miRNAs, however, languished as something of a worm-specific oddity until the discovery—some seven years later—of *let-7*, a second *C. elegans* miRNA [2], but one that had readily identifiable homologues in the emerging *Drosophila* and human genomes. There has since been an explosion of interest in the field, and the identification of hundreds of miRNAs in organisms as disparate as plants, vertebrates, arthropods, nematodes, and viruses [3] has established miRNAs as pervasive regulators of gene expression (Figure 1.1). miRNAs have been implicated in a diverse array of processes, ranging from organism development to cell differentiation, metabolism, apoptosis, and cancer; they are predicted to regulate a significant fraction of protein-coding genes [4], and have a widespread impact on mammalian mRNA evolution [5].

1.1 Animal miRNA Biogenesis: An Overview

1.1.1 Transcription

MicroRNA genes are found in diverse genomic locations (Figure 1.2). Roughly four-fifths occur in gene deserts—regions devoid of protein-coding genes. A fifth overlap with other transcripts, most commonly with introns of pre-mRNAs, but occasionally also with exons and 3′ untranslated

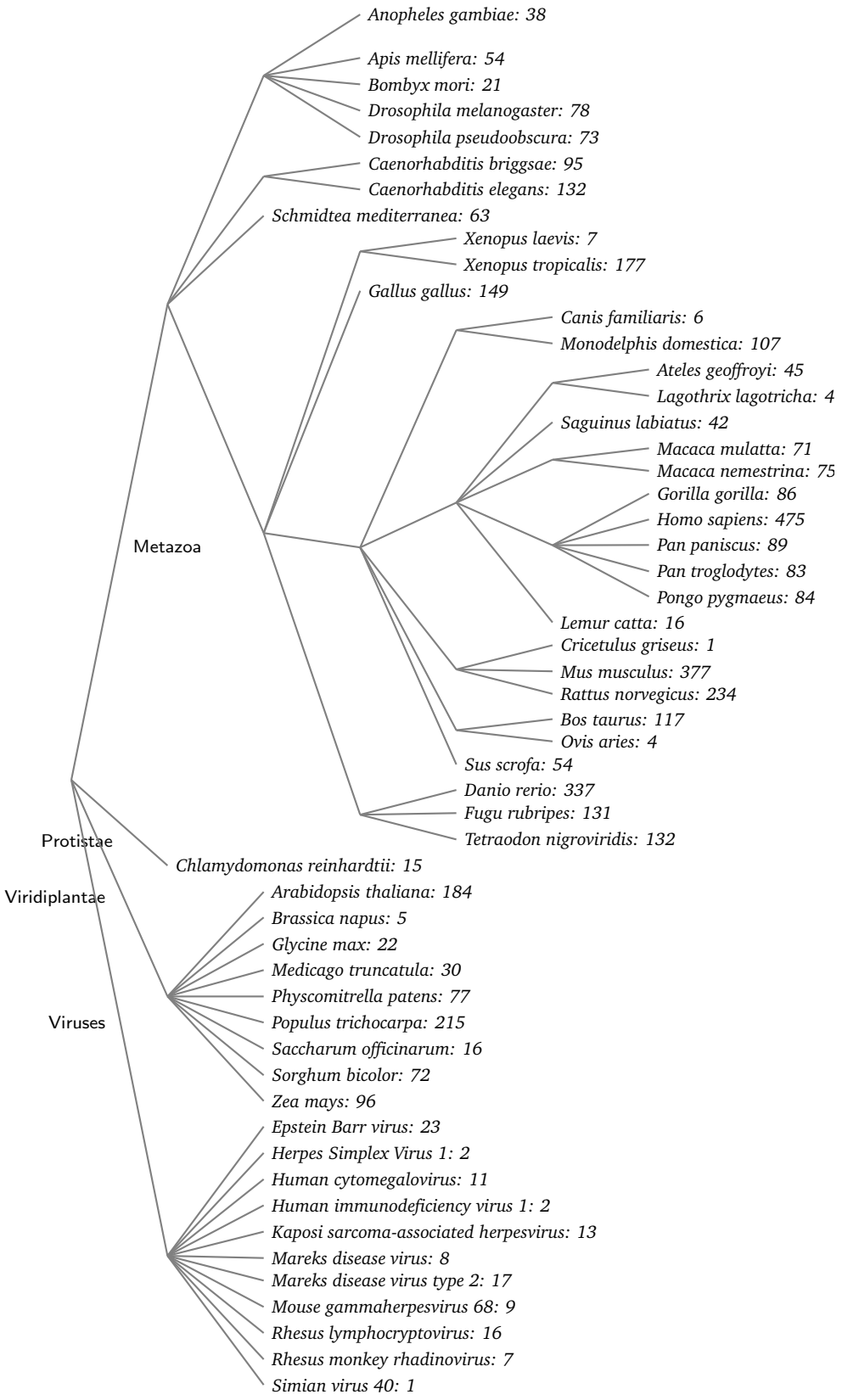

Figure 1.1: Phylogeny and species-level count of known miRNAs; data from miRBase r9.2, May 2007 [3].

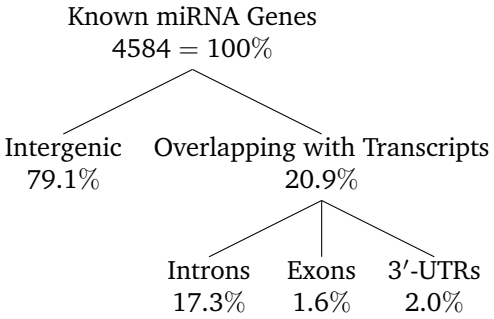

Figure 1.2: Genomic distribution of known miRNA genes; data from miRBase r9.2, May 2007 [3].

regions (UTRs). Intergenic miRNAs frequently occur in clusters with upstream promoters; these are transcribed as a single polycistronic primary transcript (pri-miRNA). miRNAs that overlap with other transcripts are thought to share regulatory elements and a primary transcript with their host genes. Figure 1.3 presents an overview of the entire miRNA biogenesis pathway.

Mounting evidence indicates that Pol II is the principal RNA polymerase for miRNA gene transcription: chromatin immunoprecipitation experiments have demonstrated Pol II to be physically associated with miRNA promoters; pri-miRNAs possess a 5′ 7-methyl guanosine cap and a 3′ polyadenine tail, both hallmarks of Pol II transcription [7]. However, recent results from chromatin immunoprecipitation and cell-free transcription assays implicate Pol III in the transcription of miRNAs interspersed among repetitive Alu elements, and possibly upto a quarter of all human miRNAs [8].

1.1.2 Maturation

The pri-miRNA transcript is cleaved by Drosha, a nuclear RNAse III endonuclease, giving rise to $\sim$70-nucleotide hairpin-shaped miRNA precursors (pre-miRNAs). Drosha works in concert with a cofactor, the DiGeorge syndrome critical region gene 8 (DGCR8) protein in humans (known as Pasha in *D. melanogaster* and *C. elegans*), as part of the Microprocessor complex. In the only model of Microprocessor substrate recognition proposed to date, DGCR8 functions as a molecular anchor that measures distance, for the cut by Drosha, from the base of the hairpin stem at the junction of single- and double-stranded RNA. How Drosha recognizes its substrates, possibly through structural features, is less clearly understood [9].

Cropping by Drosha defines one end of the mature miRNA sequence; further processing requires a Ran-GTP mediated export of the pre-miRNA to the cytoplasm by the nuclear transport factor Exportin-5. In the cytoplasm a second RNAse III, Dicer, further dices the pre-miRNA into

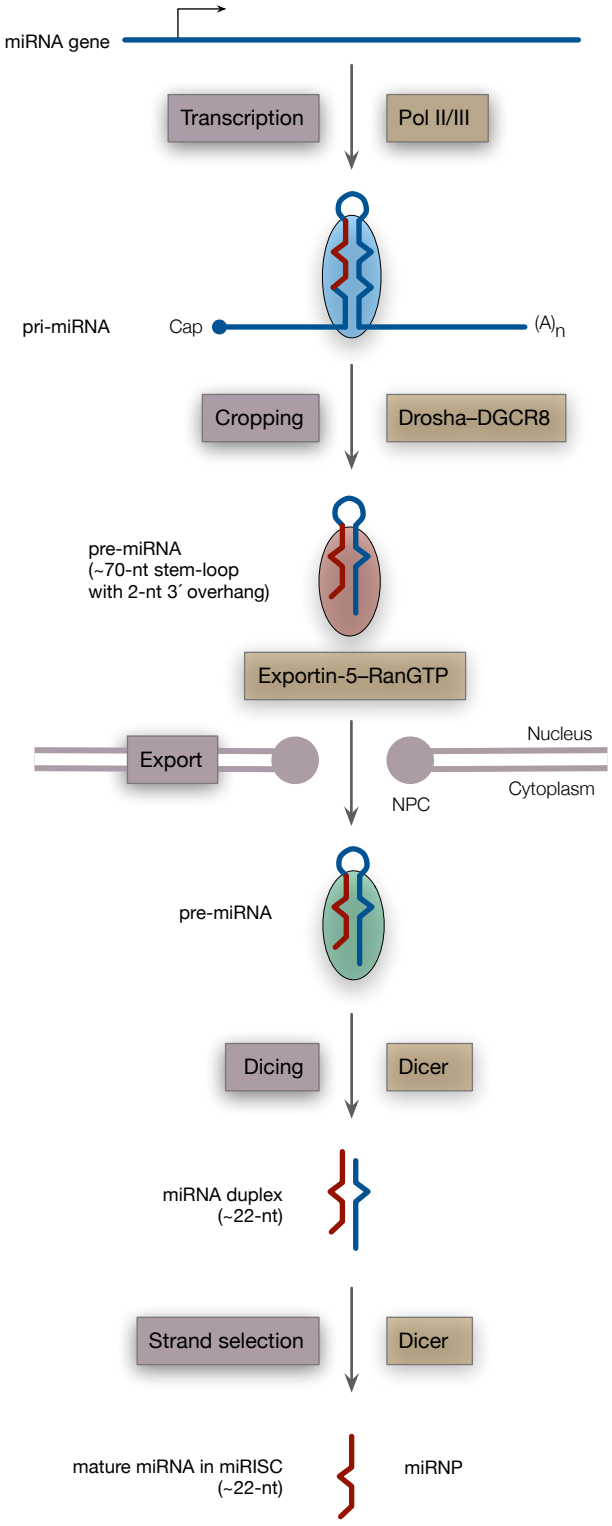

Figure 1.3: An overview of miRNA biogenesis; adapted from [6]. Following transcription by Pol II or III, the primary miRNA is cropped by the Drosha–DGCR8 Microprocessor complex, giving rise to a hairpin-shaped miRNA precursor. The pre-miRNA is exported to the cytoplasm by Exportin-5–RanGTP, where it is further cleaved by Dicer to release a ∼22-nt miRNA duplex. Finally, one of the strands is preferentially incorporated into the miRISC effector complex, which acts on cognate mRNAs in a sequence-specific manner. NPC, nuclear pore complex; miRISC, miRNA-induced silencing complex.

a ∼22-nucleotide RNA duplex. Following strand unravelling by a Helicase, one of the strands—generally the one with the less stable 5′ end—is incorporated into an effector complex called the miRNP or the miRNA-induced silencing complex (miRISC). The other strand, called the miRNA* is thought to be degraded, and is typically found at much lower frequencies in libraries of cloned miRNAs [10]. RISC comprises, at its core, a member of the Argonaute (Ago) protein family, whose members all contain a central PAZ domain (named after the family member proteins Piwi, Argonaute and Zwille), and a carboxy terminal PIWI domain [11].

1.1.3 Target Recognition

While mechanisms of target recognition by miRISC are not well understood, loss-of-function mutation studies have demonstrated the core of miRNA sequence specificity to be a heptameric seed sequence at its 5′ end, which is complementary to one or more target sites in its cognate mRNA [4]. Experimentally verified target sites have, thus far, only been found in the 3′-UTRs of mRNAs; *in vitro* tests show target sites in 5′-UTRs and coding regions to be effective downregulators of gene expression [12, 13], but their endogenous occurrence remains undetermined.

1.1.4 Mechanisms of miRNA Action

The extent of complementarity between an miRNA and its target determines the mode of post-transcriptional regulation. A small number of animal miRNAs with sufficient complementarity to their targets induce mRNA endonucleolytic cleavage: slicing between nucleotides 10 & 11 from the 5′ end of the miRNA, as in canonical siRNA-mediated RNA silencing [14]. However, most miRNAs are only partially complementary to their cognate mRNAs, and cause transcript destabilization by other mechanisms such as decapping and deadenylation [15, 16] or translational repression [17, 18].

1.1.5 Expression Patterns

miRNA expression varies considerably in different tissues and at various stages of development, which suggests tissue- or organ-specific functions for miRNAs [19, 20], and potentially critical roles in development to stabilize pathways and increase phenotypic reproducibility [21]. Aberrant miRNA expression is associated with a variety of cancers, and miRNA expression profiles have been used to diagnose and classify cancers [22, 23].

Program	Interface	Reference(s)
miRanda	Web access to predictions, downloadable software http://www.microrna.org/	[27]
PicTar	Web access to predictions http://pictar.bio.nyu.edu/	[28, 29]
TargetScan	Web access to predictions http://www.targetscan.org/	[30]
RNAHybrid	Web submission, Web API, downloadable software http://bibiserv.techfak.uni-bielefeld.de/rnahybrid/	[31]
MicroInspector	Web submission http://mirna.imbb.forth.gr/microinspector/	[32]
DIANA-microT	Web submission http://www.diana.pcbi.upenn.edu/	[33]
Targetboost	Web access to predictions https://demo1.interagon.com/targetboost/	[34]
[Stark *et al.*]	Article supplementary data	[35]
[Robins *et al.*]	Article supplementary data	[36]

Table 1.1: A list of current miRNA target prediction tools, with access details. Note that only RNAHybrid and miRanda provide source code for download.

1.2 miRNA Target Prediction

Functions have only been experimentally assigned to a small fraction of the few thousand known miRNAs [24]. Of the experimental strategies available to investigate miRNA function, stringent genetic tests that link miRNA loss-of-function mutants to misregulated targets, and point mutations in miRNA binding sites to specific phenotypes are impractical on a genomic scale in any animal species [25]. Tissue-culture assays using reporter gene constructs fused to target sequences are an easier alternative, but their reliance on ectopic miRNA expression harbours the danger of measuring what may be a nonphysiological interaction between two molecules with complementary surfaces [26].

Computational approaches are thus likely to remain an important means of studying miRNA targets for the forseeable future, not least as a means of directing wet-lab experiments. These predictions are no doubt hampered by the fact that animal miRNAs—in contrast to plant miRNAs—tend to be only partially complementary to their target mRNAs. This fact, compounded by the small size of these molecules, precludes the use of standard sequence comparison methods.

1.2.1 Current Approaches

Several algorithms have been developed to predict miRNA targets in animal species; these are listed in Table 1.1. A common strategy in several of these programs is to rank target 3′-UTR complementarity by some combination of duplex free energy and/or pairing requirements at the 5′ end (seed region) of the miRNA [25]. For instance, TargetScan [30] combines requirements for conserved perfect Watson-Crick pairing at positions 2–8 of the miRNA with estimates of the free energy of isolated miRNA–target site interactions, ignoring initiation free energy. While *in vitro* tests have shown sites containing G:U base-pairs to be functional but impaired [4], recent *in vivo* experiments have demonstrated them to be efficiently downregulated [26]. Taken together with the presence of a G:U base-pair in the seed region of a functional *let-7* binding site in the *lin-41* 3′-UTR [37], these results make a case for the inclusion of seeds with G:U wobbles in target prediction algorithms.

The PicTar [28, 29] algorithm defines seeds as heptamers with Watson-Crick or G:U pairings at positions 1–7 or 2–8 from the miRNA 5′ end. It combines seed searches with RNA duplex free energy filters, evolutionary conservation requirements, and a probabilistic scoring mechanism to predict targets that are under combinatorial control by co-expressed miRNAs. However, it makes use of RNAHybrid [31], an algorithm that approximates RNA duplex free energies by discarding intramolecular hybridizations in order to achieve linear time complexity.

Robins *et al*. [36] incorporate mRNA secondary structure computed from 3′-UTRs in their target prediction algorithm, but require perfect Watson-Crick complementarity in the seed site. Furthermore, the use of isolated 3′-UTRs is likely to produce structures very different from the structure of 3′-UTRs in folds that use complete mRNA sequences.

While most of the tools listed in Table 1.1 are accessible as web services, only miRanda [27] and RNAHybrid are available as downloadable software that can be modified, extended and run on custom datasets. Most listed algorithms also rely on target conservation across two or more species as a filter. Although this is often necessary to increase the signal-to-noise ratio in genome-wide scans, it results in the unavoidable omission of biologically relevant unconserved targets, as well as those of species-specific miRNAs.

1.2.2 MicroTar: A Novel Approach

This dissertation presents a novel miRNA target prediction algorithm, MicroTar, that does not rely on evolutionary conservation, and is thus not limited to the prediction of conserved targets.

Prediction strategies include the use of partial complementarity of miRNAs to their target messages, the predicted free energies of mRNAs & miRNA–mRNA duplexes, and extreme value statistics. Harnessing the power of parallel computing obviates the need for introducing approximations that discard intramolecular base pairs in estimates of miRNA–mRNA duplex free energy; this has the added advantage of implicitly incorporating the accessibility of 3′-UTRs in the algorithm. The following chapter provides a detailed description of the MicroTar algorithm, and the energy calculations and parallelism it employs.

Chapter 2

Materials and Methods

2.1 The MicroTar Algorithm

The MicroTar algorithm is based on the following assumptions:

- miRNA target specificity is determined by a heptameric seed sequence (beginning at the first or second position from the 5′ end of the miRNA) that is complementary to sites in mRNA 3′-UTRs
- targets are functional if miRNA–mRNA duplex formation is energetically favourable

Beginning with a set of fasta-formatted query (miRNA) sequences and target (mRNA) sequences, the MicroTar algorithm schedules a query–target computation to run on an idle node. Each such computation involves predicting the minimum free energy of the each mRNA molecule, searching for seed sites, and performing a constrained fold where each seed match is, in turn, bound in the miRNA–mRNA heterodimer; the output is a list of putative duplexes more stable than free mRNA. This result is subsequently subjected to a statistical analysis to determine the significance of each miRNA–mRNA match. Figure 2.1 presents a schematic overview of the algorithm.

2.1.1 Overview

Parallelism

MicroTar employs coarse-grained parallelism by scheduling seed search/energy computations for pairs of miRNA–mRNA sequences to be run on nodes as they become available. This balances the

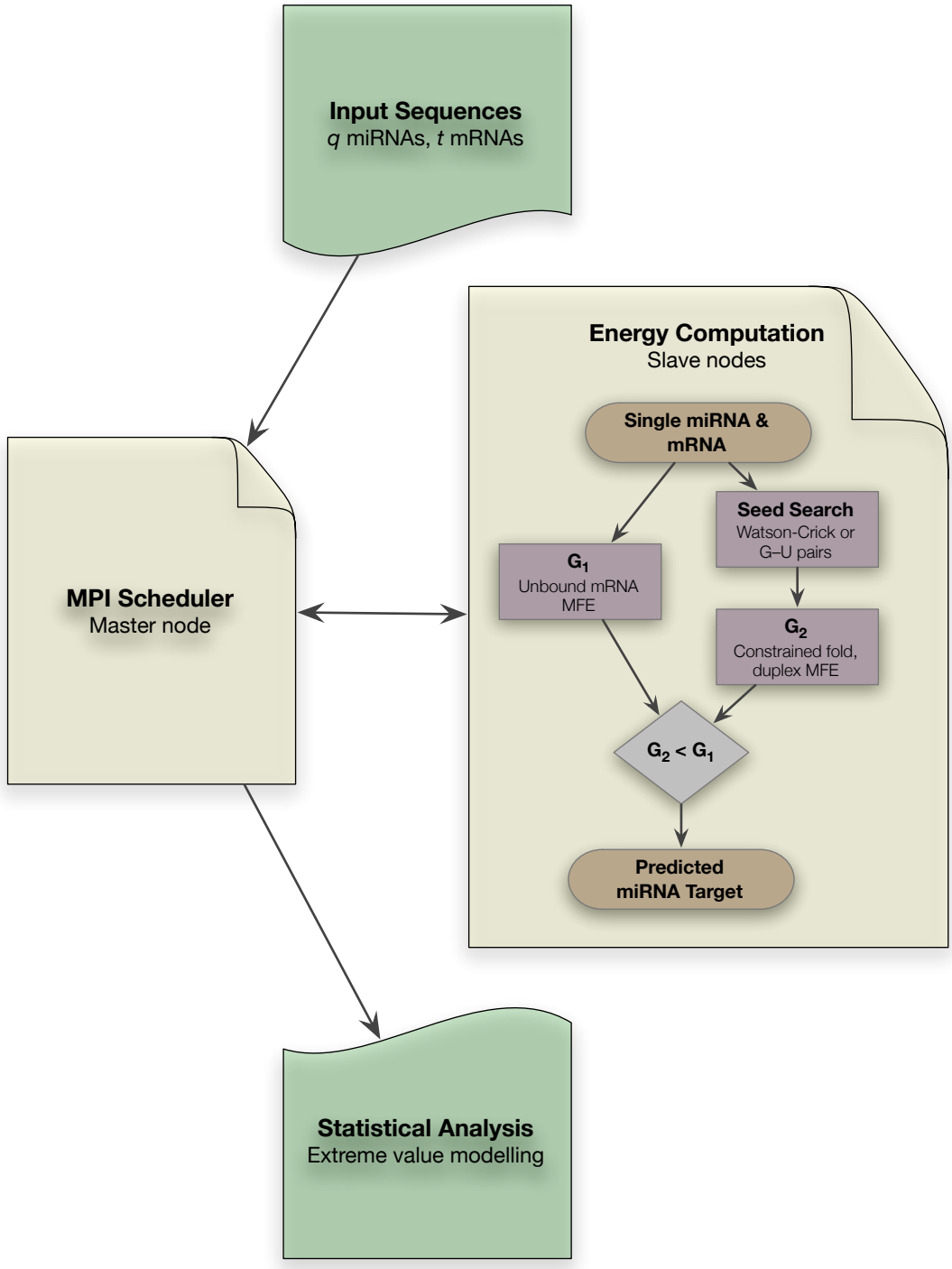

Figure 2.1: Beginning with a set of fasta-formatted query (miRNA) sequences and target (mRNA) sequences, the MicroTar algorithm schedules energy computations on slave nodes as they become available. Each slave node predicts the minimum free energy of the each mRNA molecule, searches for seed sites, and performs a constrained fold where each seed match is, in turn, bound in the miRNA–mRNA heterodimer; the output is a list of putative duplexes more stable than free mRNA. The results are subsequently subjected to a statistical analysis to determine the significance of each miRNA–mRNA match.

speedup from parallelization with the increased communications overhead that would result from finer-grained parallel programming.

Secondary Structure Prediction

The secondary structure and minimum free energy (G_1) of the complete unbound mRNA molecule are predicted using the `fold` routine from the RNAlib library of the ViennaRNA package [38]. This is an implementation of the Zuker & Stiegler dynamic programming algorithm [39], described in Section 2.2 below.

Seed Search

Loss-of-function mutation studies have demonstrated the core of miRNA sequence specificity to be a heptameric seed sequence [4], which the algorithm defines as nucleotides 1–7 or 2–8 at the 5′ end of the miRNA. MicroTar searches each mRNA 3′-UTR (or complete mRNA in the absence of annotations) for sites with Watson-Crick or G–U wobble complementarity to this seed sequence; these hits are called seed matches.

Constrained Fold

The mRNA is now folded with each seed match bound, in turn, to its corresponding miRNA seed sequence. This uses the `cofold` [40] routine from the RNAlib library, also calculating the free energy of the duplex, G_2.

Output

The output is a list of all seed matches, along with predicted energies of the unbound mRNA (G_1), putative mRNA–miRNA heterodimers (G_2), the estimated energy of duplex formation

$$g = G_2 - G_1, \tag{2.1}$$

and optionally, images of the secondary structure of each mRNA before and after miRNA binding (for instance, Figure 2.2).

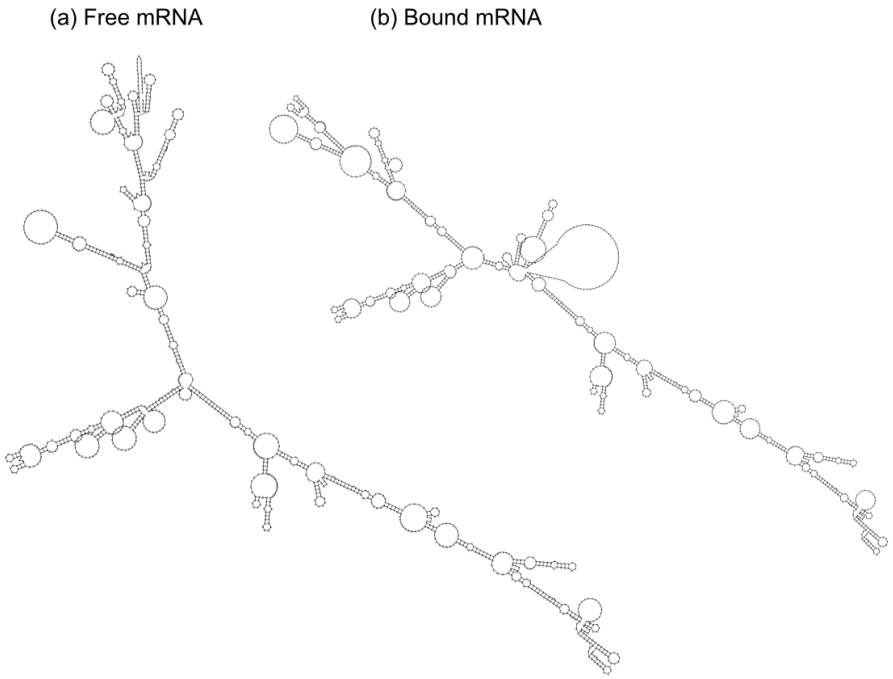

Figure 2.2: Sample output of the *C. elegans* cog-1 [GenBank:NM_001027093] mRNA secondary structure before and after binding with the lsy-6 miRNA. Note the changes in global structure, which cannot be approximated using only 3′-UTRs.

2.1.2 Functional Targets

Seed matches are considered to be functional targets if the relevant miRNA–mRNA heterodimer is more energetically stable than free mRNA, i.e., $g < 0$. It is then possible to estimate the significance of the prediction using extreme value statistics, much in the fashion of Rehmsmeier *et al.* [31] outlined below.

2.1.3 Statistical Analysis of Predicted Targets

Negative normalized free energy

The occurrence of favourable hybridizations of short miRNAs with long mRNAs can frequently be attributed to chance: the longer the mRNA, the more likely the incidence. In order to eliminate the effect of sequence length on our measure of free energy [41, 31], we define the negative normalized free energy

$$g_n = -\frac{g}{\log(mn)} \tag{2.2}$$

where g is defined in equation 2.1, m is the length of the target sequence searched, and n is the length of the miRNA seed.

Extreme Value Statistics

Extreme value distributions (EVDs) are limiting distributions that describe the minimum or maximum of independent random variables [42]. If we consider the miRNA–mRNA duplex energy estimation to be essentially an optimization procedure that produces a minimum, the negative normalized free energy described above is a corresponding maximum, and can be described by an EVD having a distribution function of the form

$$P\left[G \leq t\right] = D(t) = \exp\left(-\exp\left(\frac{a-t}{b}\right)\right). \tag{2.3}$$

A transformation then converts this distribution function into a straight line:

$$\log\left(-\log\left(D\right)\right) = \frac{a-t}{b} = \left(-\frac{1}{b}\right)t + \frac{a}{b}. \tag{2.4}$$

By scanning for targets of random miRNA sequences in the mRNA sequences in the dataset, or those of real miRNAs in shuffled mRNAs, we obtain a set of negative normalized free energies, which we expect will follow an EVD. We then transform the distribution function of the empirical EVD into a straight line, as in Equation 2.4, and estimate the parameters of the EVD by a linear least squares fit to the line $y = mx + c$, obtaining

$$b = -\frac{1}{m} \tag{2.5}$$

and

$$a = cb. \tag{2.6}$$

We can now compute, for each predicted miRNA–mRNA duplex, a p-value, the probability that the same or a more favourable free energy is observed due to chance:

$$P\left[Z \geq g_n\right] = 1 - \exp\left(-\exp\left(\frac{a-g_n}{b}\right)\right) \tag{2.7}$$

where a and b are estimated EVD parameters, and g_n is the negative normalized free energy from Equation 2.2 [31].

2.1.4 Technical details

MicroTar has been written using the C programming language, and makes use of the RNAlib library from the Vienna RNA package [38]. Great care has been taken to make the system suitable for datasets of varying sizes. Sequences are loaded into memory only as required, allowing the handling of virtually any number of sequences. Functions from v2.0 of the Message Passing Interface (MPI) standard [43, 44] are used for parallelization.

2.2 RNA Folding: The Zuker-Stiegler Algorithm

This dynamic programming algorithm for computing minimum free energy (MFE) structures for RNA molecules was proposed by Zuker and Stiegler in a seminal 1981 paper [39] and has become a *de facto* standard. While it has undergone several refinements over the years, including the use of more accurate thermodynamic parameters [45], the core algorithm, as in Zuker and Stiegler's description [39, 46] reproduced below, remains essentially unchanged.

2.2.1 Definitions

Consider an RNA molecule S with its nucleotides numbered $1 \ldots N$ from the 5′ end. S_i denotes the ith nucleotide for $1 \leq i \leq N$, and S_{ij} denotes nucleotides from S_i to S_j, both inclusive. Now imagine the N nucleotides laid out equally spaced on a semicircle (Figure 2.3). In this graph representation, the N nucleotides are *vertices*, the $N-1$ arcs between the bases are *exterior edges* representing phosphodiester bonds, and base pairing is denoted by line segments, or chords (*interior edges*) between vertices. A chord is *admissible* if it connects complementary bases: G–C, A–U or G–U. An admissible structure is then defined to be one whose graph contains only admissible, non-interesecting chords that are never in contact. The no-contact condition ensures that no nucleotide is paired more than once; non-intersection of chords constrains admissible structures to those that are free of pseudoknots[1].

A description of the free energy of the structure, or equivalently, of its graph, completes the picture. A *face* is defined as a planar region of the graph that is bound on all sides. The energy of the structure is then associated with the faces of its graph. A hairpin loop is represented by a face with a single interior edge. A helix consists of interior edges separated by a single exterior edge on either side. A bulge loop has interior edges separated by a single exterior edge on one side,

[1] An RNA structure in which one of the bases inside a hairpin loop pairs with a base outside the hairpin.

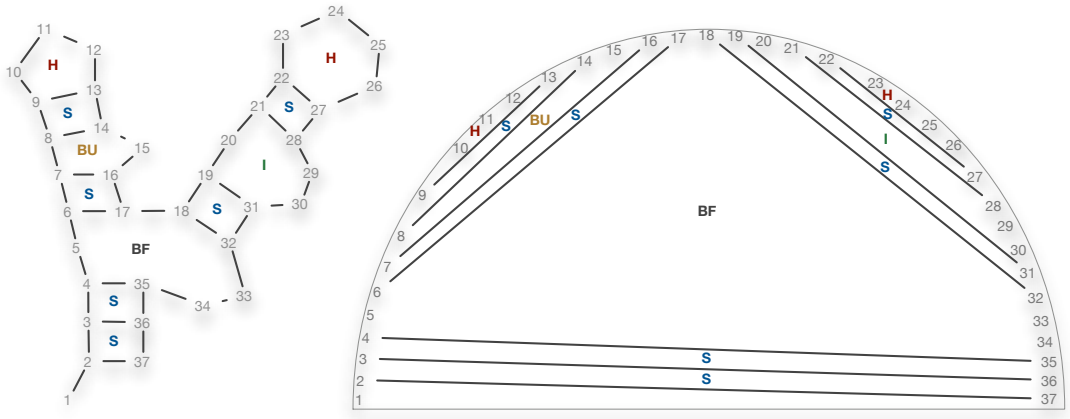

Figure 2.3: Illustrative representations of RNA secondary structure. *Left:* conventional representation; *right:* abstract graph representation with edges and vertices. BF: bifurcation loop, BU: bulge loop, H: hairpin loop, I: interior loop, S: stack.

but more than one exterior edge on the other. Interior loops have more than one exterior edge on either side.

2.2.2 Recursion

The mathematical technique underlying the MFE determination is to compute two potentially different energies for each subsequence S_{ij} of the RNA sequence. For all pairs i, j satisfying $1 \leq i < j \leq N$, let $W(i,j)$ be the MFE of all possible admissible structures formed from the subsequence S_{ij}. Additionally, let $V(i,j)$ be the MFE of all possible admissible stuctures formed from S_{ij} in which S_i and S_j are paired with each other. If S_i and S_j cannot pair, then $V(i,j) \rightarrow \infty$. We note that $W(i,j) \leq V(i,j)$ for all i, j. The numbers $V(i,j)$ and $W(i,j)$ are computed recursively: first for all 5-mers, followed by successively larger subsequences of S.

Boundary conditions for $\mathbf{W}$ and $\mathbf{V}$ are $W(i,j) = V(i,j) \rightarrow \infty$, if $j - i < 4$. Define the energy of exterior loops to be zero, and–for simplicity—also assign zero energies to multiloops. Recursions for $\mathbf{W}$ and $\mathbf{V}$ depend on the energy rules for loops. Let $E_h(i,j)$ be the energy of the hairpin closed by the base pair $i \cdot j$; $E_s(i,j)$ be the energy of the stacked pair $i \cdot j$ and $i+1 \cdot j-1$; and $E_{bi}(i,j,i',j')$ the energy of the bulge or interior loop closed by $i \cdot j$ with $i' \cdot j'$ accessible from $i \cdot j$. Then, for

$1 \leq i < j \leq N$,

$$W(i,j) = \min\{W(i+i,j), W(i,j-1), V(i,j), \min_{i \leq k < j} W(i,k) + W(k+1,j)\} \tag{2.8}$$

$$V(i,j) = \min\{E_h(i,j), E_s(i,j) + V(i+1,j-1), V_{bi}(i,j), V_m(i,j)\} \tag{2.9}$$

$$\text{where } V_{bi}(i,j) = \min_{\substack{i<i'<j'<j \\ i'-i+j-j'<2}} \{E_{bi}(i,j,i',j') + V(i',j')\} \tag{2.10}$$

$$\text{and } V_m(i,j) = \min_{i<k<j-1} \{W(i+1,k) + W(k+1,j-1)\}. \tag{2.11}$$

The recursion continues until reaching $W(1,n)$ which is the minimum folding energy. Finally, computation of the MFE structure—which is equivalent to identifying the interior edges of the associated graph—is achieved by a straightforward traceback through matrices $\mathbf{W}$ and $\mathbf{V}$.

Chapter 3

Results and Discussion

3.1 Parallel Speedup

In general, the speedup achieved by parallelization of an algorithm is defined as

$$S_n = \frac{T_1}{T_n}, \tag{3.1}$$

where n is the number of processors, T_1 is the serial execution time, and T_n is the parallel execution time. While it may seem trivial for speedup to increase with the number of processors used, in practice, the speedup achieved depends on the fraction of program code that can be parallelized and, in distributed memory MPI implementations, the overhead due to communication between processes, relative to the total computation time.

MicroTar employs a coarse-grained parallelization technique: the scheduling of computations for miRNA–mRNA sequence pairs on processors as they become available. This provides for a reasonable amount of computation in parallel, while minimizing communications overhead between processors to achieve a credible amount of speedup (Figure 3.1).

3.2 Validation

On a test using three sets of experimentally verified miRNA targets in *C. elegans*, *D. melanogaster*, and *M. musculus*, from v3.0 of TarBase [24], MicroTar showed better sensitivity than the widely used PicTar [28, 29] algorithm (Table 3.1). miRNA sequences were retrieved from miRBase v9.0 [3]; mRNA sequences from RefSeq entries associated with the corresponding gene entry in the

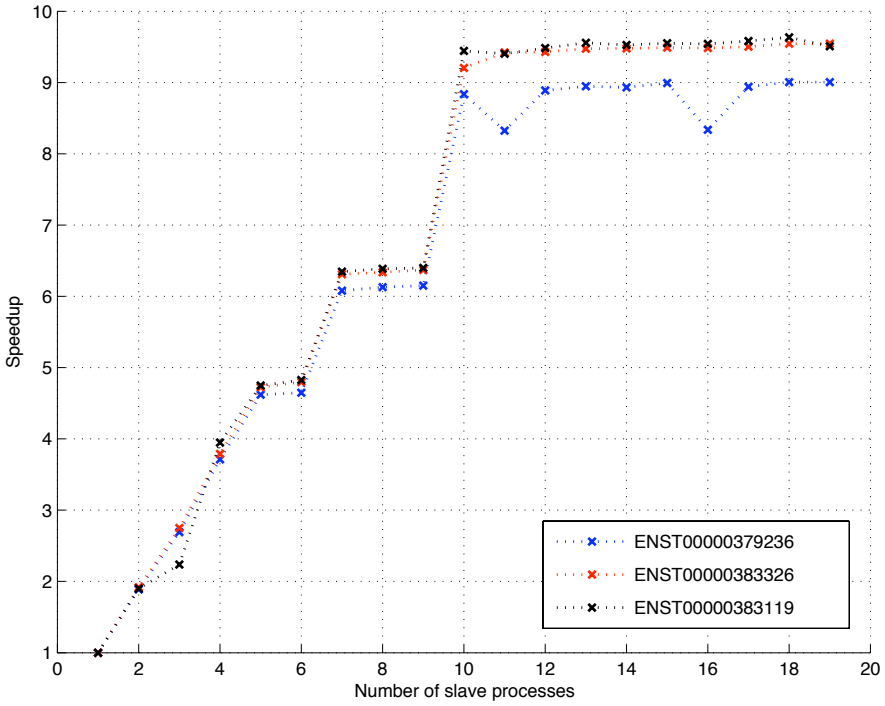

Figure 3.1: Parallel speedup for three MicroTar sample runs on hiv-miR-H1 vs. three human mRNAs: ENST00000383119, ENST00000383326 and ENST00000379236 (in descending order of length), each replicated 20 times for target sequence input; average of 2 repetitions. Note that longer sequences achieve meaningful speedup over a slightly greater range of processors, before communications overhead begins to swamp benefits from parallelization.

Program	Species	Predicted Targets (TP)	Verified Targets (TP + FN)	Sensitivity TP/(TP + FN)
MicroTar	*D. melanogaster*	39	63	0.62
	C. elegans	8	13	0.62
	M. musculus	24	43	0.56
PicTar	*D. melanogaster*	35	63	0.56
	C. elegans	7	13	0.54
	M. musculus	15	43	0.35

Table 3.1: A comparison of MicroTar and PicTar prediction results on three datasets of experimentally verified miRNA targets; MicroTar achieves better sensitivity in all three cases.

Entrez Gene database. Prediction results are summarized in Figure 3.2, which shows a density plot of free energies of the most stable predicted miRNA–target duplex for each gene-miRNA pair in the three species. It should be emphasized that unverified predicted interactions ought to be viewed as a guide for further experiments and not as false positives. A detailed listing of predicted targets may be found in Appendix B.

3.3 Duplex energy estimation

At the core of the MicroTar algorithm lies a novel approach to the estimation of miRNA–mRNA duplex energy. Interactions are viewed in a global context by predicting folds for the entire mRNA, rather than just its 3′-UTR or seed match. By allowing intramolecular hybridizations, the algorithm implicitly incorporates the accessibility of the 3′-UTR; seed matches in highly inaccessible UTRs are expected to disrupt UTR secondary structure in putative duplexes. Large disruptions in base pairing cannot be compensated for by bond formation during miRNA–mRNA hybridization. This results in a putative duplex with free energy G_2 far greater than that of the unbound mRNA, G_1, and the match is rejected.

Furthermore, parallelization makes feasible the use of full-molecule energy computations, removing the need to introduce approximations that discard intramolecular base pairs in estimates of miRNAmRNA duplex free energy, as is currently done in miRNA target prediction algorithms.

3.4 Significance of predictions

In order to estimate the significance of these predictions, p-values were calculated for the lowest energy duplex for each miRNA–transcript pair, as derived in Equation 2.7. Parameters were

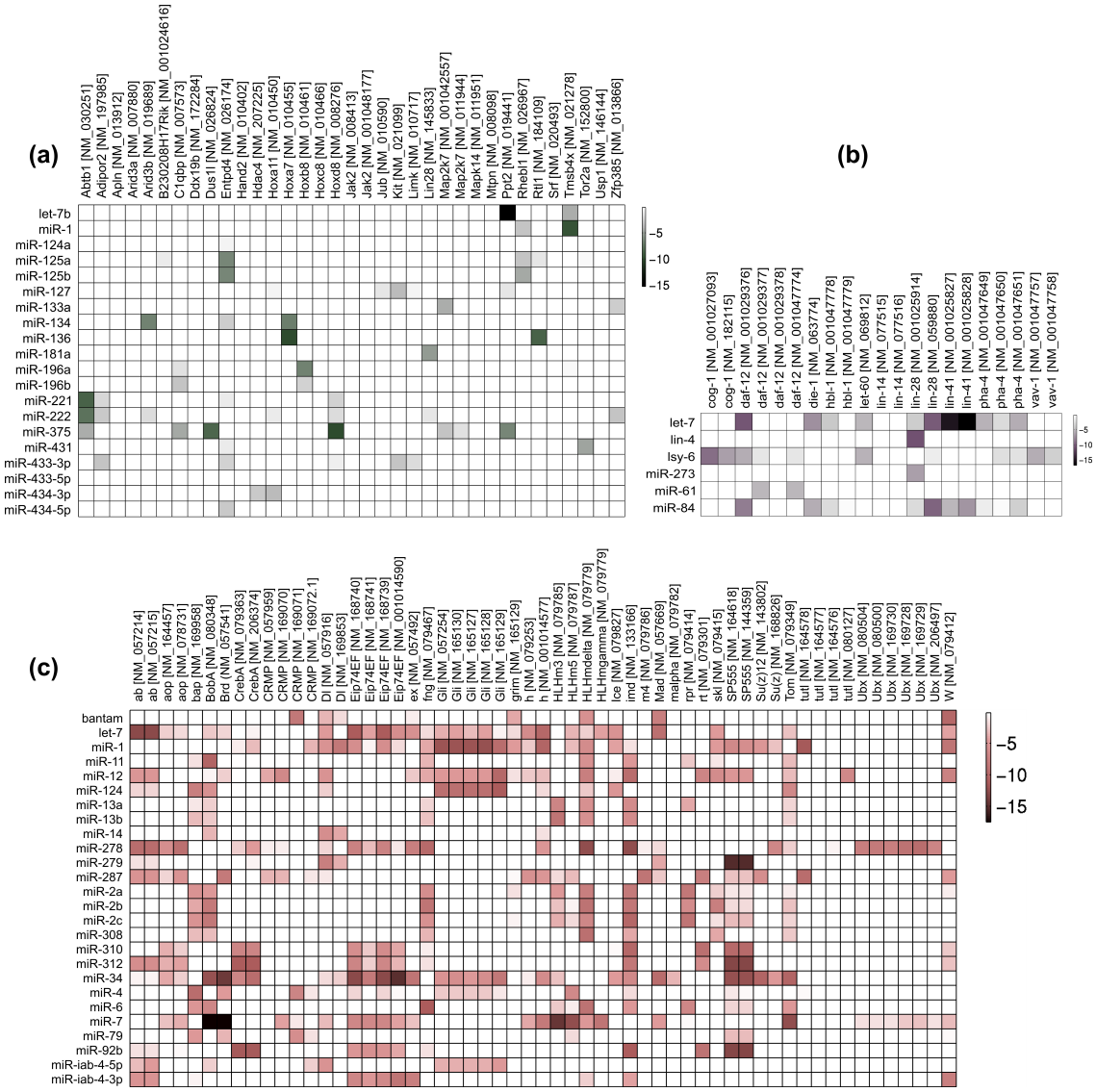

Figure 3.2: A density plot of free energies of the most stable predicted miRNA–target duplex for each gene-miRNA pair in (a) *M. musculus*, (b) *C. elegans*, and (c) *D. melanogaster*, with genes along the x-axis and miRNAs along the y-axis. A more negative free energy indicates a more stable duplex, relative to its unbound mRNA. Darker colours indicate lower free energies, as shown by the scale in the top-right corner of each sub-figure. White squares indicate no predicted interaction.

estimated separately for each species from a distribution computed with miRNAs shuffled using the `shuffleseq` utility from the EMBOSS package [47]), while ensuring a sufficient number of random sequences for approximately 4000 seed matches in each species. Figure 3.3 shows these p-values in a density plot for each miRNA–target pair, as in Figure 3.2.

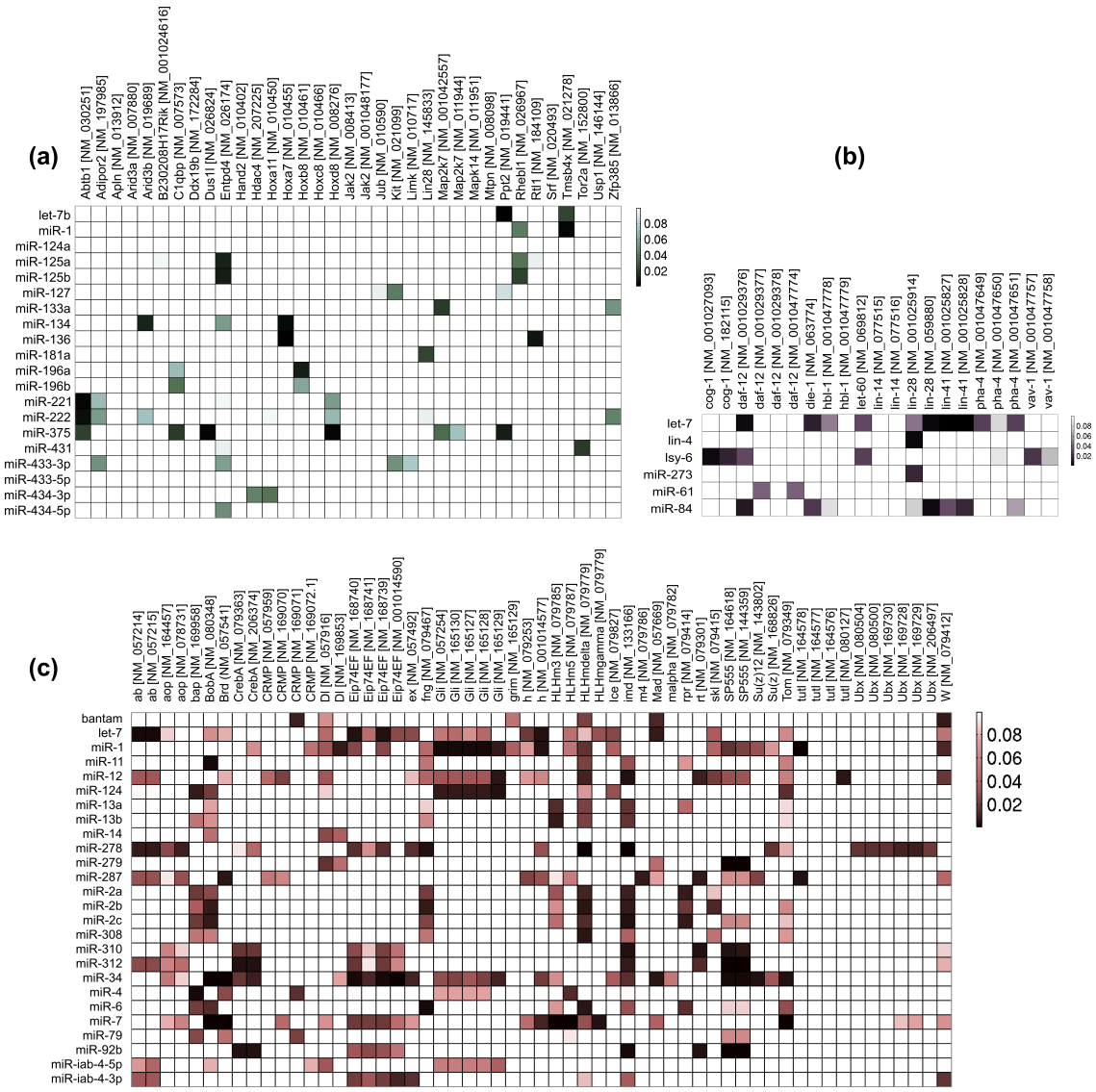

Figure 3.3: A density plot of p-values lower than 0.1, of the most stable predicted miRNA–target duplex for each gene-miRNA pair in (a) *M. musculus*, (b) *C. elegans*, and (c) *D. melanogaster*, with genes along the x-axis and miRNAs along the y-axis. A lower p-value indicates a lower probability of the energy of the duplex (or more favourable energies) occurring due to chance alone. Darker colours indicate lower p-values, as shown by the scale in the top-right corner of each sub-figure. White squares indicate no predicted interaction, or a p-value greater than the cutoff value of 0.1.

Chapter 4

Conclusions

Given the fact that only a fraction of the several thousand known miRNAs have well-characterized functions, computational approaches are likely to remain an important means of studying miRNA targets for the forseeable future. These are especially useful as a means of directing experimental investigations of miRNA function, which remain impractical on a genomic scale in any animal species.

MicroTar is a novel miRNA target prediction algorithm that does not rely on evolutionary conservation to filter predicted targets and is able to address the problem of the prediction of targets that are not conserved across different genomes. Parallel computing makes feasible the use of complex energy prediction algorithms on a large scale, and by using estimates of miRNA–mRNA duplex free energy that allow intramolecular pairings, MicroTar implicitly incorporates the accessibility of 3′-UTRs. In tests on datasets of experimentally verified miRNA targets in *C. elegans*, *D. melanogaster* and *M. musculus*, MicroTar displays greater sensitivity than the widely used PicTar [28, 29] target prediction algorithm.

The limited state of knowledge of miRNA function to date—which necessitates the use of computational methods—is somewhat paradoxically what makes it difficult to assess the performance of miRNA target prediction tools. In the interim, therefore, unverified miRNA targets should be treated not as false positives, but as a guide for further experimentation. With burgeoning interest in the field, miRNA target prediction algorithms are certain to improve in the forseeable future, as physiological details of miRNA–target interaction begin to emerge.

Bibliography

[1] Lee RC, Feinbaum RL, Ambros V (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 75:843–854. doi:10.1016/0092-8674(93)90529-Y.

[2] Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, et al. (2000) The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. Nature 403:901–906. doi:10.1038/35002607.

[3] Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ (2006) miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res 34:D140–D144. doi:10.1093/nar/gkj112.

[4] Brennecke J, Stark A, Russell RB, Cohen SM (2005) Principles of microRNA–target recognition. PLoS Biol 3:e85. doi:10.1371/journal.pbio.0030085.

[5] Farh KKH, Grimson A, Jan C, Lewis BP, Johnston WK, et al. (2005) The widespread impact of mammalian microRNAs on mRNA repression and evolution. Science 310:1817–1821. doi:10.1126/science.1121158.

[6] Kim V (2005) microRNA biogenesis: coordinated cropping and dicing. Nat Rev Mol Cell Biol 6:376–385. doi:10.1038/nrm1644.

[7] Lee Y, Kim M, Han J, Yeom KH, Lee S, et al. (2004) MicroRNA genes are transcribed by RNA polymerase II. EMBO J 23:4051–4060. doi:10.1038/sj.emboj.7600385.

[8] Borchert GM, Lanier W, Davidson BL (2006) RNA polymerase III transcribes human microRNAs. Nat Struct Mol Biol 13:1097–1101. doi:10.1038/nsmb1167.

[9] Han J, Lee Y, Yeom KH, Nam JW, Heo I, et al. (2006) Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. Cell 125:887–901. doi:10.1016/j.cell.2006.03.043.

[10] Bartel DP (2004) MicroRNAs: Genomics, biogenesis, mechanism, and function. Cell 116:281–297. doi:10.1016/S0092-8674(04)00045-5.

[11] Du T, Zamore PD (2005) microPrimer: the biogenesis and function of microRNA. Development 132:4645–4652. doi:10.1242/10.1242/dev.02070.

[12] Lytle JR, Yario TA, Steitz JA (2007) Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5′UTR as in the 3′UTR. Proc Natl Acad Sci U S A 104:9667–9672. doi:10.1073/pnas.0703820104.

[13] Kloosterman WP, Wienholds E, Ketting RF, Plasterk RH (2004) Substrate requirements for let-7 function in the developing zebrafish embryo. Nucleic Acids Res 32:6284–6291. doi:10.1093/nar/gkh968.

[14] Yekta S, Shih IH, Bartel DP (2004) MicroRNA-directed cleavage of HOXB8 mRNA. Science 304:594–596. doi:10.1126/science.1097434.

[15] Wu L, Fan J, Belasco JG (2006) MicroRNAs direct rapid deadenylation of mRNA. Proc Natl Acad Sci U S A 103:4034–4039. doi:10.1073/pnas.0510928103.

[16] Giraldez AJ, Mishima Y, Rihe J, Grocock RJ, Dongen SV, et al. (2006) Zebrafish miR-430 promotes deadenylation and clearance of maternal mRNAs. Science 312:75–79. doi:10.1126/science.1122689.

[17] Olsen PH, Ambros V (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. doi:10.1006/dbio.1999.9523.

[18] Ambros V, Chen X (2007) The regulation of genes and genomes by small RNAs. Development 134:1635–1641. doi:10.1242/dev.002006.

[19] Sempere LF, Freemantle S, Pitha-Rowe I, Moss E, Dmitrovsky E, et al. (2004) Expression profiling of mammalian microRNAs uncovers a subset of brain-expressed microRNAs with possible roles in murine and human neuronal differentiation. Genome Biol 5:R13.

[20] Lagos-Quintana M, Rauhut R, Yalcin A, Meyer J, Lendecke W, et al. (2002) Identification of tissue-specific microRNAs from mouse. Current Biol 9:735–739. doi:10.1016/S0960-9822(02)00809-6.

[21] Hornstein E, Shomron N (2006) Canalization of development by microrna. Nat Genet 38:S20–S24. doi:10.1038/ng1803.

[22] Dalmay T, Edwards D (2006) MicroRNAs and the hallmarks of cancer. Oncogene 25:6170–6175. doi:10.1038/sj.onc.1209911.

[23] Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, et al. (2005) MicroRNA expression profiles classify human cancers. Nature 435:834–838. doi:10.1038/nature03702.

[24] Sethupathy P, Corda B, Hatzigeorgiou AG (2006) Tarbase: A comprehensive database of experimentally supported animal microrna targets. RNA 12:192–197. doi:10.1261/rna.2239606.

[25] Lai EC (2004) Predicting and validating microRNA targets. Genome Biol 5:115. doi:10.1186/gb-2004-5-9-115.

[26] Didiano D, Hobert O (2006) Perfect seed pairing is not a generally reliable predictor for miRNA-target interactions. Nat Struct Mol Biol 13:849–851. doi:10.1038/nsmb1138.

[27] John B, Enright AJ, Aravin A, Tuschl T, Sander C, et al. (2004) Human microRNA targets. PLoS Biol 2:e363. doi:10.1371/journal.pbio.0020363.

[28] Krek A, Grün D, Poy MN, Wolf R, Rosenberg L, et al. (2005) Combinatorial microRNA target predictions. Nat Genet 37:495–500. doi:10.1038/ng1536.

[29] Lall S, Grün D, Krek A, Chen K, Wang YL, et al. (2006) A genome-wide map of conserved microRNA targets in C. elegans. Curr Biol 16:460–471. doi:10.1016/j.cub.2006.01.050.

[30] Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP (2003) Prediction of mammalian microRNA targets. Cell 115:787–798. doi:10.1016/S0092-8674(03)01018-3.

[31] Rehmsmeier M, Steffen P, Höchsmann M, Giegerich R (2004) Fast and effective prediction of microRNA/target duplexes. RNA 10:1507–1517. doi:10:1507-1517.

[32] Rusinov V, Baev V, Minkov IN, Tabler M (2005) MicroInspector: a web tool for detection of miRNA binding sites in an RNA sequence. Nucleic Acids Res 33:W696–W700. doi:10.1093/nar/gki364.

[33] Kiriakidou M, Nelson PT, Kouranov A, Fitziev P, Bouyioukos C, et al. (2004) A combined computational-experimental approach predicts human microRNA targets. Genes Dev 18:1165–1178. doi:10.1101/gad.1184704.

[34] Sætrom O, Ola Snøve J, Sætrom P (2005) Weighted sequence motifs as an improved seeding step in microRNA target prediction algorithms. RNA 11:995–1003. doi:10.1261/rna.7290705.

[35] Stark A, Brennecke J, Russell RB, Cohen SM (2003) Identification of Drosophila microRNA targets. PLoS Biol 1:e60. doi:10.1371/journal.pbio.0000060.

[36] Robins H, Li Y, Padgett RW (2005) Incorporating structure to predict microRNA targets. Proc Natl Acad Sci U S A 102:4006–4009. doi:10.1073/pnas.0500775102.

[37] Vella MC, Choi EY, Lin SY, Reinert K, Slack FJ (2004) The C. elegans microRNA let-7 binds to imperfect let-7 complementary sites from the lin-41 3′UTR. Genes Dev 18:132–137. doi:10.1101/gad.1165404.

[38] Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, et al. (1994) Fast folding and comparison of RNA secondary structures. Monatsh Chem 125:167–188. doi:10.1007/BF00818163.

[39] Zuker M, Stiegler P (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res 9:133–148.

[40] Bernhart SH, Tafer H, Mückstein U, Flamm C, Stadler PF, et al. (2006) Partition function and base pairing probabilities of RNA heterodimers. Algorithms Mol Biol 1:3. doi:10.1186/1748-7188-1-3.

[41] Karlin S, Altschul SF (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.

[42] Gumbel EJ (1958) Statistics of Extremes. New York, USA: Columbia University Press.

[43] Message Passing Interface Forum (2003) MPI: A message-passing interface standard. Technical report. URL `http://www.mpi-forum.org/docs/mpi-11.ps`.

[44] The MPI Forum (1993) MPI: a message passing interface. In: Proceedings of the 1993 ACM/IEEE conference on Supercomputing. SIGARCH: ACM Special Interest Group on Computer Architecture, New York, USA: ACM Press, pp. 878–883. doi:10.1145/169627.169855.

[45] Mathews DH, Sabina J, Zuker M, Turner DH (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J Mol Biol 288:911–940. doi:10.1006/jmbi.1999.2700.

[46] Zuker M (1989) The use of dynamic programming algorithms in RNA secondary structure prediction, Boca Raton, FL: CRC Press, chapter 7. pp. 159–184.

[47] Rice P, Longden I, Bleasby A (2000) EMBOSS: The european molecular biology open software suite. Trends Genet 16:276–277. doi:10.1016/S0168-9525(00)02024-2.

Appendix A

Reports

1. **Thadani R**, Tammi MT (2006) MicroTar: predicting microRNA targets from RNA duplexes. *BMC Bioinformatics* **7**(Suppl 5):S20. doi:10.1186/1471-2105-7-S5-S20.

2. Chao X, **Thadani R** and Tammi MT (2007) Single nucleotide polymorphisms mediate the differential microRNA regulation of domestic chicken breeds. *In preparation*.

BMC Bioinformatics

BioMed Central

Proceedings

Open Access

MicroTar: predicting microRNA targets from RNA duplexes

Rahul Thadani[1] and Martti T Tammi*[1,2,3]

Address: [1]Department of Biological Sciences, National University of Singapore, 14 Science Drive 4, Singapore 117543, [2]Department of Biochemistry, National University of Singapore, 8 Medical Drive, Singapore 117597 and [3]Karolinska Institutet, Department of Microbiology, Tumor and Cell Biology, Stockholm, Sweden

Email: Rahul Thadani - rahul.thadani@nus.edu.sg; Martti T Tammi* - martti@nus.edu.sg

* Corresponding author

from International Conference in Bioinformatics – InCoB2006
New Dehli, India. 18–20 December 2006

Published: 18 December 2006

BMC Bioinformatics 2006, **7**(Suppl 5):S20 doi:10.1186/1471-2105-7-S5-S20

Abstract

Background: The accurate prediction of a comprehensive set of messenger RNAs (targets) regulated by animal microRNAs (miRNAs) remains an open problem. In particular, the prediction of targets that do not possess evolutionarily conserved complementarity to their miRNA regulators is not adequately addressed by current tools.

Results: We have developed MicroTar, an animal miRNA target prediction tool based on miRNA-target complementarity and thermodynamic data. The algorithm uses predicted free energies of unbound mRNA and putative mRNA-miRNA heterodimers, implicitly addressing the accessibility of the mRNA 3' untranslated region. MicroTar does not rely on evolutionary conservation to discern functional targets, and is able to predict both conserved and non-conserved targets. MicroTar source code and predictions are accessible at http://tiger.dbs.nus.edu.sg/microtar/, where both serial and parallel versions of the program can be downloaded under an open-source licence.

Conclusion: MicroTar achieves better sensitivity than previously reported predictions when tested on three distinct datasets of experimentally-verified miRNA-target interactions in *C. elegans*, *Drosophila*, and mouse.

Background

MicroRNAs (miRNAs) are a class of endogenous, small regulatory RNA averaging 22 nucleotides in length that mediate the post-transcriptional regulation of messenger RNAs. They bind to target messages in a sequence-specific manner, and induce translational repression or endonucleolytic cleavage. The first two miRNAs, *lin-4* and *let-7* were discovered some seven years apart in the worm *C. elegans*, in genetic screens for mutants with disrupted developmental timing [1,2]. There has since been an explosion of interest in the field, and the identification of hundreds of miRNAs in metazoans as disparate as vertebrates, arthropods, nematodes, and viruses [3] has established miRNAs as pervasive regulators of gene expression. For recent reviews, see [4-6].

Functions have only been experimentally assigned to a small fraction of the few thousand known miRNAs [7]. Of the experimental strategies available to investigate miRNA function, stringent genetic tests that link miRNA loss-of-function mutants to misregulated targets, and point mutations in miRNA binding sites to specific phenotypes are

impractical on a genomic scale in any animal species [8]. Tissue-culture assays using reporter gene constructs fused to target sequences are an easier alternative, but their reliance on ectopic miRNA expression harbours the danger of measuring what may be a nonphysiological interaction between two molecules with complementary surfaces [9]. Computational approaches are thus likely to remain an important means of studying miRNA targets for the forseeable future, not least as a means of directing wet-lab experiments. These predictions are no doubt hampered by the fact that animal miRNAs – in contrast to plant miRNAs – tend to be only partially complementary to their target mRNAs. This fact, compounded by the small size of these molecules, precludes the use of standard sequence comparison methods.

Several algorithms have been developed to predict miRNA targets in animal species; these are listed in Table 1. A common strategy in several of these programs is to rank target 3' untranslated region (UTR) complementarity by some combination of duplex free energy and/or pairing requirements at the 5' end (seed region) of the miRNA [8]. For instance, TargetScan [10] combines requirements for conserved perfect Watson-Crick pairing at positions 2–8 of the miRNA with estimates of the free energy of isolated miRNA-target site interactions, ignoring initiation free energy. While *in vitro* tests have shown sites containing G:U base-pairs to be functional but impaired [11], recent *in vivo* experiments have demonstrated them to be efficiently downregulated [9]. Taken together with the presence of a G:U base-pair in the seed region of a functional *let-7* binding site in the *lin-41 3*'-UTR [12], these results make a case for the inclusion of seeds with G:U wobbles in target prediction algorithms.

The PicTar [13,14] algorithm defines seeds as heptamers with Watson-Crick or G:U pairings at positions 1–7 or 2–8 from the miRNA 5' end. It combines seed searches with RNA duplex free energy filters, evolutionary conservation requirements, and a probabilistic scoring mechanism to predict targets that are under combinatorial control by co-expressed miRNAs. However, it makes use of RNAHybrid [15], an algorithm that approximates RNA duplex free energies by discarding intramolecular hybridizations in order to achieve linear time complexity.

Robins *et al*. [16] incorporate mRNA secondary structure computed from 3'-UTRs in their target prediction algorithm, but require perfect Watson-Crick complementarity in the seed site. Furthermore, the use of isolated 3'-UTRs is likely to produce structures very different from the structure of 3'-UTRs in folds that use complete mRNA sequences.

While most of the tools listed in Table 1 are accessible as web services, only miRanda [17] and RNAHybrid are available as downloadable software that can be modified, extended and run on custom datasets. Most listed algorithms also rely on target conservation across two or more species as a filter. While this is necessary to distinguish functional targets from a vast array of candidates, it results in the unavoidable omission of real targets that are not thus conserved.

Here we present MicroTar, an miRNA target prediction program that does not rely on evolutionary conservation. Through the use of the partial complementarity of miRNAs to their target messages, and the predicted free energy of complete mRNA molecules, we are able to address the problem of the prediction of targets that are not conserved across different genomes. Moreover, harnessing the power of parallel computing obviates the need for introducing approximations that discard intramolecular base pairs in estimates of miRNA-mRNA duplex free energy; we thus implicitly incorporate the accessibility of 3'-UTRs in the algorithm. MicroTar source code – available under an

Table 1: miRNA target prediction tools. A list of current miRNA target prediction tools, with access details. Note that only RNAHybrid and miRanda provide source code for download.

Program	Interface	Reference(s)
miRanda	Web access to predictions, downloadable software http://www.microrna.org/	[17]
PicTar	Web access to predictions http://pictar.bio.nyu.edu/	[13,14]
TargetScan	Web access to predictions http://www.targetscan.org/	[10]
RNAHybrid	Web submission, Web API, downloadable software http://bibiserv.techfak.uni-bielefeld.de/rnahybrid/	[15]
MicroInspector	Web submission http://mirna.imbb.forth.gr/microinspector/	[25]
DIANA-microT	Web submission http://www.diana.pcbi.upenn.edu/	[26]
Targetboost	Web access to predictions https://demo1.interagon.com/targetboost/	[27]
[Stark *et al.*]	Article supplementary data	[28]
[Robins *et al.*]	Article supplementary data	[16]

open-source licence – and predictions can be accessed at the MicroTar website [18].

Implementation

Overview

The MicroTar algorithm is based on the following assumptions:

• miRNA target specificity is determined by a heptameric seed sequence (beginning at the first or second position from the 5' end of the miRNA) that is complementary to sites in mRNA 3'-UTRs

• targets are functional if miRNA-mRNA duplex formation is energetically favourable

Beginning with a set of fasta-formatted query (miRNA) sequences and target (mRNA) sequences, the MicroTar algorithm predicts the minimum free energy of the each mRNA molecule, searches for seed sites, and performs a constrained fold where each seed match is, in turn, bound in the miRNA-mRNA heterodimer; the output is a list of putative duplexes more stable than free mRNA, along with images of bound and unbound mRNA secondary structure. This result is subsequently subjected to a statistical analysis to determine the significance of each miRNA-mRNA match. Figure 1 presents a schematic overview of this algorithm.

Secondary structure prediction

The secondary structure and minimum free energy of the complete unbound mRNA molecule are predicted using the fold routine from the RNAlib library of the ViennaRNA package [19]. This is an implementation of the Zuker & Stiegler dynamic programming algorithm [20]. We denote the predicted free energy of unbound mRNA as G_1.

Seed search

Loss-of-function mutation studies have demonstrated the core of miRNA sequence specificity to be a heptameric seed sequence [11], which we define as nucleotides 1–7 or 2–8 at the 5' end of the miRNA. MicroTar searches each mRNA 3'-UTR (or complete mRNA in the absence of annotations) for sites with Watson-Crick or G–U wobble complementarity to this seed sequence; we refer to these hits as seed matches.

Constrained fold

For each seed match above, the mRNA is again folded under the constraint that the miRNA seed is bound to its corresponding match. This uses the cofold [21] routine from the RNAlib library. We denote the free energy of the duplex as G_2.

Output

The output is a list of all seed matches, along with predicted energies of the unbound mRNA (G_1), putative mRNA-miRNA heterodimers (G_2), the estimated energy of duplex formation ($g = G_2 - G_1$), and optionally, images of the secondary structure of each mRNA before and after miRNA binding (see e.g., Figure 2).

Functional targets

Seed matches are considered functional targets if the relevant miRNA-mRNA heterodimer is more energetically stable than free mRNA, i.e., $g < 0$. We then estimate the significance of the prediction using extreme value statistics, much in the fashion of Rehmsmeier *et al.* [15]. This procedure is outlined below.

Statistical analysis of predicted targets

Negative normalized free energy

The occurrence of favourable hybridizations of short miRNAs with long mRNAs can frequently be attributed to chance: the longer the mRNA, the more likely the incidence. In order to eliminate the effect of sequence length on our measure of free energy [15,22], we define the negative normalized free energy

$$g_n = -\frac{g}{\log(mn)} \qquad (1)$$

where m is the length of the target sequence searched, and n is the length of the miRNA.

Extreme value statistics

Extreme value distributions (EVDs) are limiting distributions that describe the minimum or maximum of independent random variables [23]. If we consider the miRNA-mRNA duplex energy estimation to be essentially an optimization procedure that produces a minimum, the negative normalized free energy described above is a corresponding maximum, and can be described by an EVD having a distribution function of the form

$$P[G \le t] = D(t) = \exp\left(-\exp\left(\frac{a-t}{b}\right)\right). \qquad (2)$$

A transformation then converts this distribution function into a straight line:

$$\log(-\log(D)) = \frac{a-t}{b} = \left(-\frac{1}{b}\right)t + \frac{a}{b}. \qquad (3)$$

By scanning for targets of random miRNA sequences in the mRNA sequences in the dataset, we obtain a set of negative normalized free energies, which we expect will follow an EVD. We then transform the distribution function of the empirical EVD into a straight line, as in Equation 3,

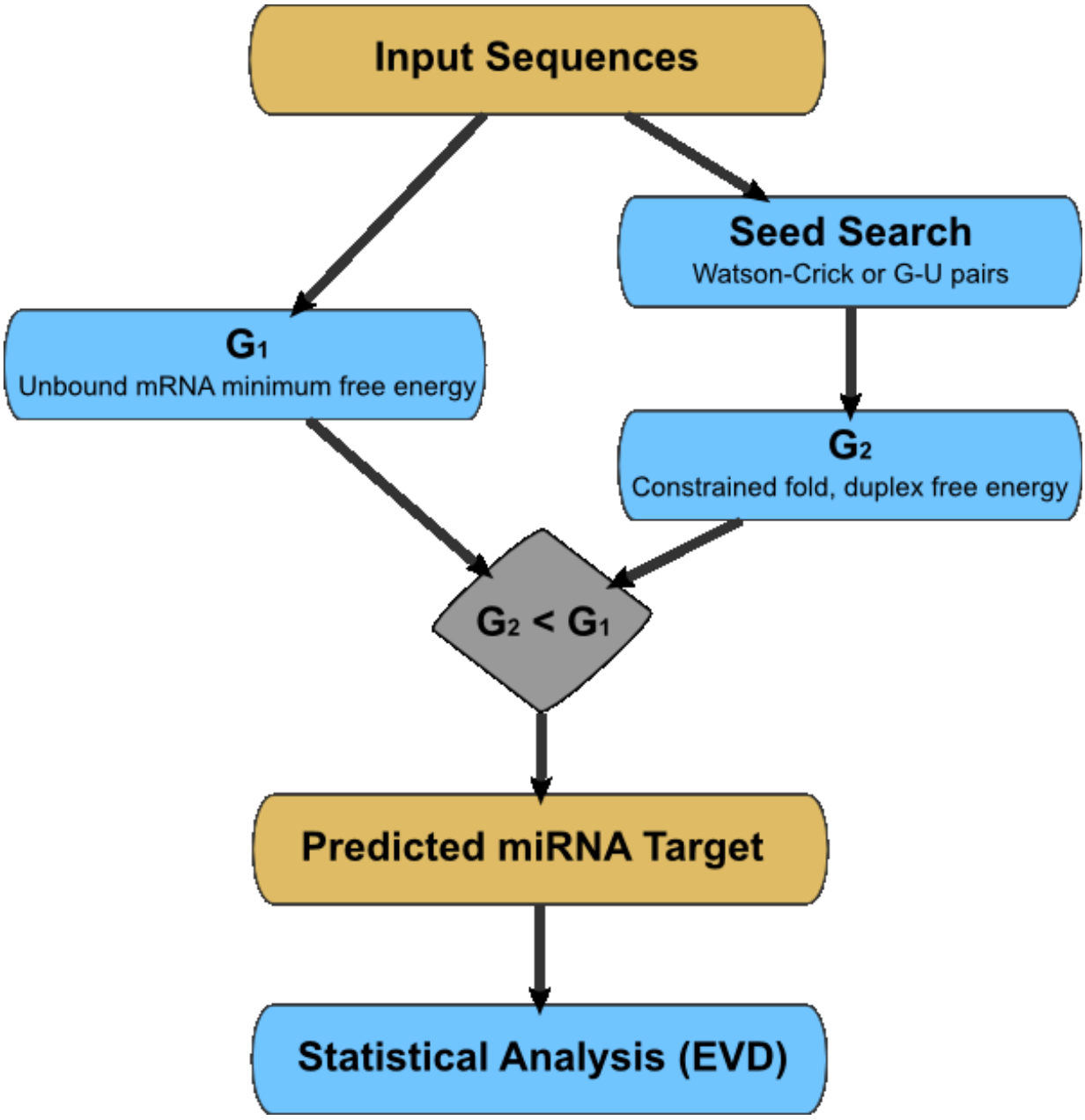

Figure 1
MicroTar algorithm. Beginning with a set of fasta-formatted query (miRNA) sequences and target (mRNA) sequences, the MicroTar algorithm predicts the minimum free energy of the each mRNA molecule, searches for seed sites, and performs a constrained fold where each seed match is, in turn, bound in the miRNA-mRNA heterodimer; the output is a list of putative duplexes more stable than free mRNA, along with images of bound and unbound mRNA secondary structure. This result is subsequently subjected to a statistical analysis to determine the significance of each miRNA-mRNA match.

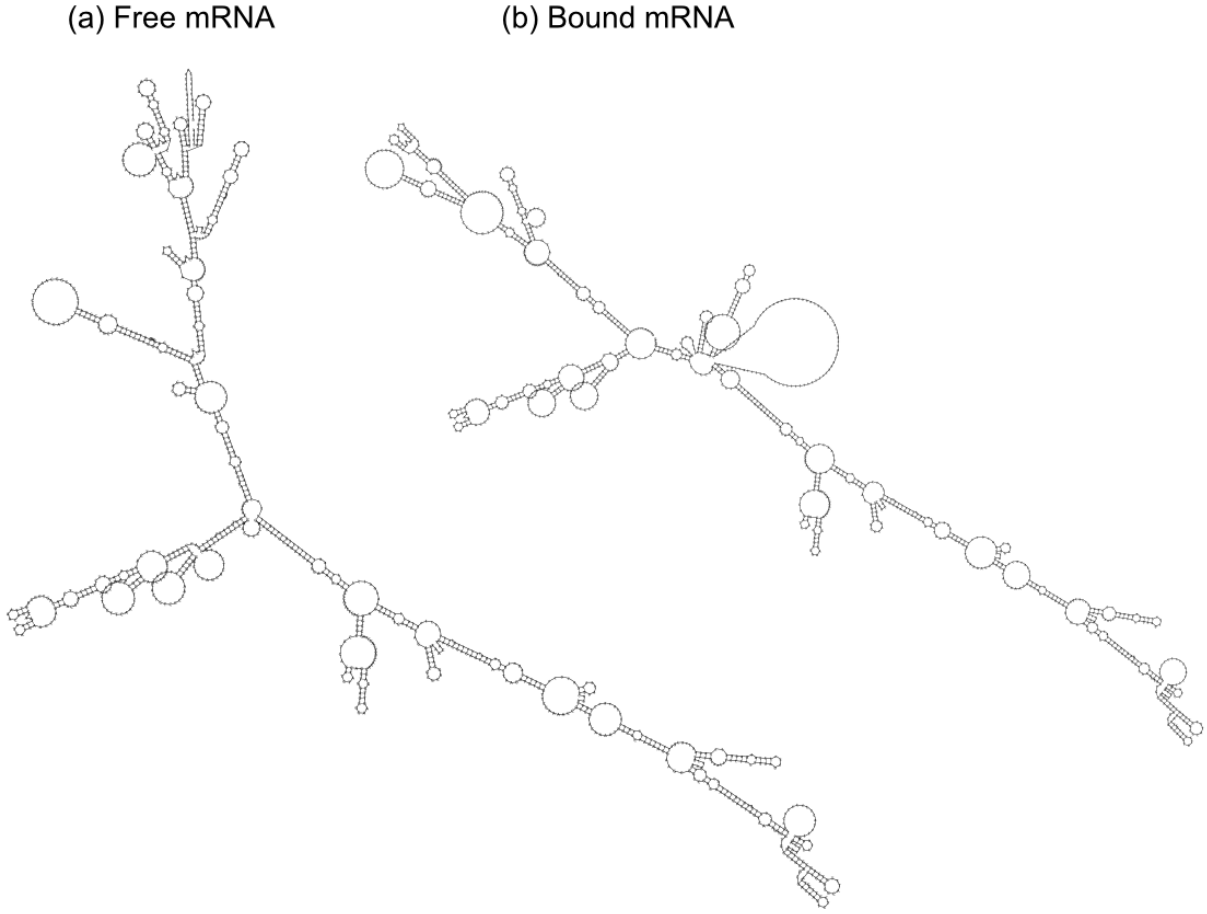

Figure 2
mRNA secondary structure. Sample output of the *C. elegans*. cog-1 [GenBank:NM_001027093] mRNA secondary structure before and after binding with the lsy-6 miRNA. Note the changes in global structure, which cannot be approximated using only 3'-UTRs.

and estimate the parameters of the EVD by a linear least squares fit to the line $y = mx + c$, obtaining

$$b = -\frac{1}{m} \qquad (4)$$

and

$$a = cb. \qquad (5)$$

We can now compute, for each predicted miRNA-mRNA duplex, a *p*-value, the probability that the same or a more favourable free energy is observed due to chance:

$$P[Z \geq g_n] = 1 - \exp\left(-\exp\left(\frac{a - g_n}{b}\right)\right) \qquad (6)$$

where a and b are estimated EVD parameters, and g_n is the negative normalized free energy from Equation 1 [15].

Technical details

MicroTar has been written using the C programming language, and makes use of the RNAlib library from the Vienna RNA package [19]. Great care has been taken to make the system suitable for datasets of varying sizes. Sequences are loaded into memory only as required, allowing the handling of virtually any number of sequences. The parallel version uses functions from v2.0 of the Message Passing Interface (MPI).

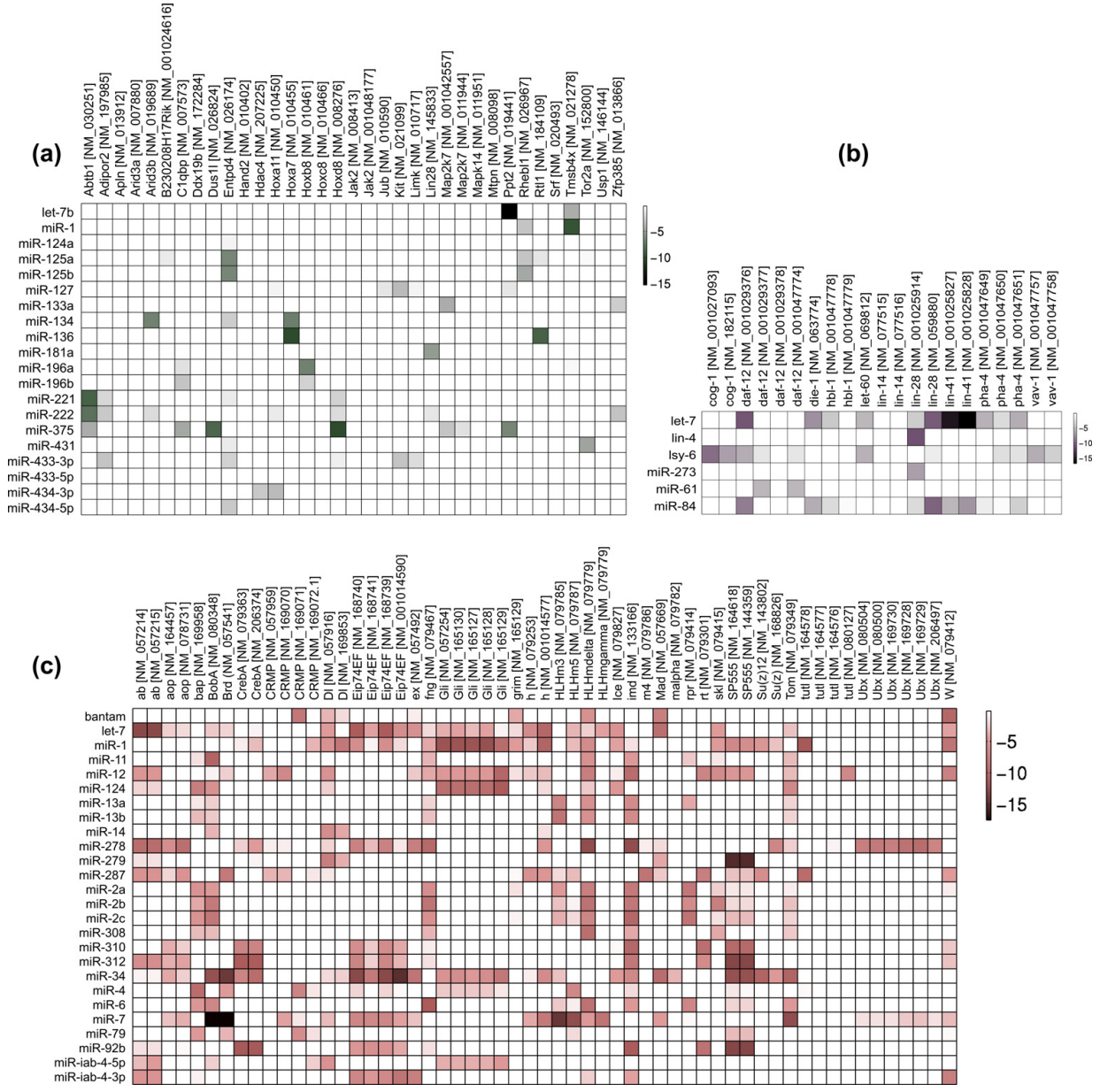

Figure 3

Energies of predicted miRNA targets. A density plot of free energies of the most stable predicted miRNA-target duplex for each gene-miRNA pair in (a) mouse, (b) *C. elegans*, and (c) *Drosophila*, with genes along the x-axis and miRNAs along the *y*-axis. A more negative free energy indicates a more stable duplex, relative to its unbound mRNA. Darker colours indicate lower free energies, as shown by the scale in the top-right corner of each sub-figure. White squares indicate no predicted interaction.

MicroTar should compile and run under Linux and most flavours of UNIX. It has been tested under Fedora Core 4 & 5 and CentOS 4.4 Linux distributions, on both 32 and 64 bit platforms.

Results and Discussion

Validation

We performed a test of MicroTar on three sets of experimentally verified miRNA targets in *C. elegans*, *Drosophila*,

Table 2: MicroTar target predictions compared to PicTar. A comparison of MicroTar and PicTar prediction results on three datasets of experimentally verified miRNA targets; MicroTar achieves better sensitivity in all three cases.

Program	Species	**Targets Predicted** (TP)	**Targets in Dataset** (TP + FN)	**Sensitivity** TP/(TP + FN)
MicroTar	*D. melanogaster*	39	63	0.62
	C. elegans	8	13	0.62
	M. musculus	24	43	0.56
PicTar	*D. melanogaster*	35	63	0.56
	C. elegans	7	13	0.54
	M. musculus	15	43	0.35

and mouse, from v3.0 of TarBase [7]. miRNA sequences were retrieved from miRBase v9.0 [3]; mRNA sequences from RefSeq entries associated with the corresponding gene entry in the Entrez Gene database. In the absence of 3'-UTR annotations, the entire mRNA sequence was scanned for seed matches by MicroTar. These results are summarized in Figure 3, which shows a density plot of free energies of the most stable predicted miRNA-target duplex for each gene-miRNA pair in the three species.

Furthermore, we compared our predictions to the widely-used PicTar algorithm, which was recently updated and applied to miRNAs in *C. elegans*. This comparison is shown in Table 2, where we note that MicroTar achieves better sensitivity in all three cases. We emphasize that unverified predicted interactions should be viewed as a guide for further experiments and not as false positives. Detailed lists of targets predicted are available as supplementary data (see Additional File 1 – MicroTar target predictions compared to PicTar), and on the MicroTar website [18].

Duplex energy estimation

At the core of the MicroTar algorithm lies a novel approach to the estimation of miRNA-mRNA duplex energy. Interactions are viewed in a global context by predicting folds for the entire mRNA, rather than just its 3'-UTR or seed match. By allowing intramolecular hybridizations, we implicitly incorporate the accessibility of the 3'-UTR; seed matches in highly inaccessible UTRs are expected to disrupt UTR secondary structure in putative duplexes. Large disruptions in base pairing cannot be compensated for by bond formation during miRNA-mRNA hybridization. This results in a putative duplex with free energy G_2 far greater than that of the unbound mRNA, G_1, and the match is rejected.

Significance of predictions

In order to estimate the significance of our predictions, we calculated the p-value for the lowest energy duplex for each miRNA-transcript pair, as derived in Equation 6. The parameters were estimated separately for each species from a distribution computed with random miRNAs. We shuffled miRNAs using the shuffleseq utility from the EMBOSS package [24], ensuring that there were a sufficient number of random sequences for approximately 4000 seed matches in each species. Figure 4 shows these *p*-values in a density plot for each miRNA-target pair, as in Figure 3.

Conclusion

MicroTar does not rely on evolutionary conservation to filter predicted targets and is able to address the problem of the prediction of targets that are not conserved across different genomes. Parallel computing makes feasible the use of complex energy prediction algorithms on a large scale, and by using estimates of miRNA-mRNA duplex free energy that allow intramolecular pairings, MicroTar implicitly incorporates the accessibility of 3'-UTRs. In tests on three datasets of experimentally verified miRNA targets in *C. elegans*, *Drosophila* and mouse, MicroTar displays greater sensitivity than previously developed target prediction programs.

Availability and Requirements

Project name: MicroTar

Project home page: http://tiger.dbs.nus.edu.sg/microtar/

Operating systems: Linux, UNIX

Programming language: C

Other requirements: GNU autoconf/automake

Licence: New BSD licence

Any restrictions to use by non-academics: None (check ViennaRNA licence, however)

Authors' contributions

MTT and RT planned the project. RT acquired the data and implemented the algorithm. Both authors prepared and approved the final manuscript.

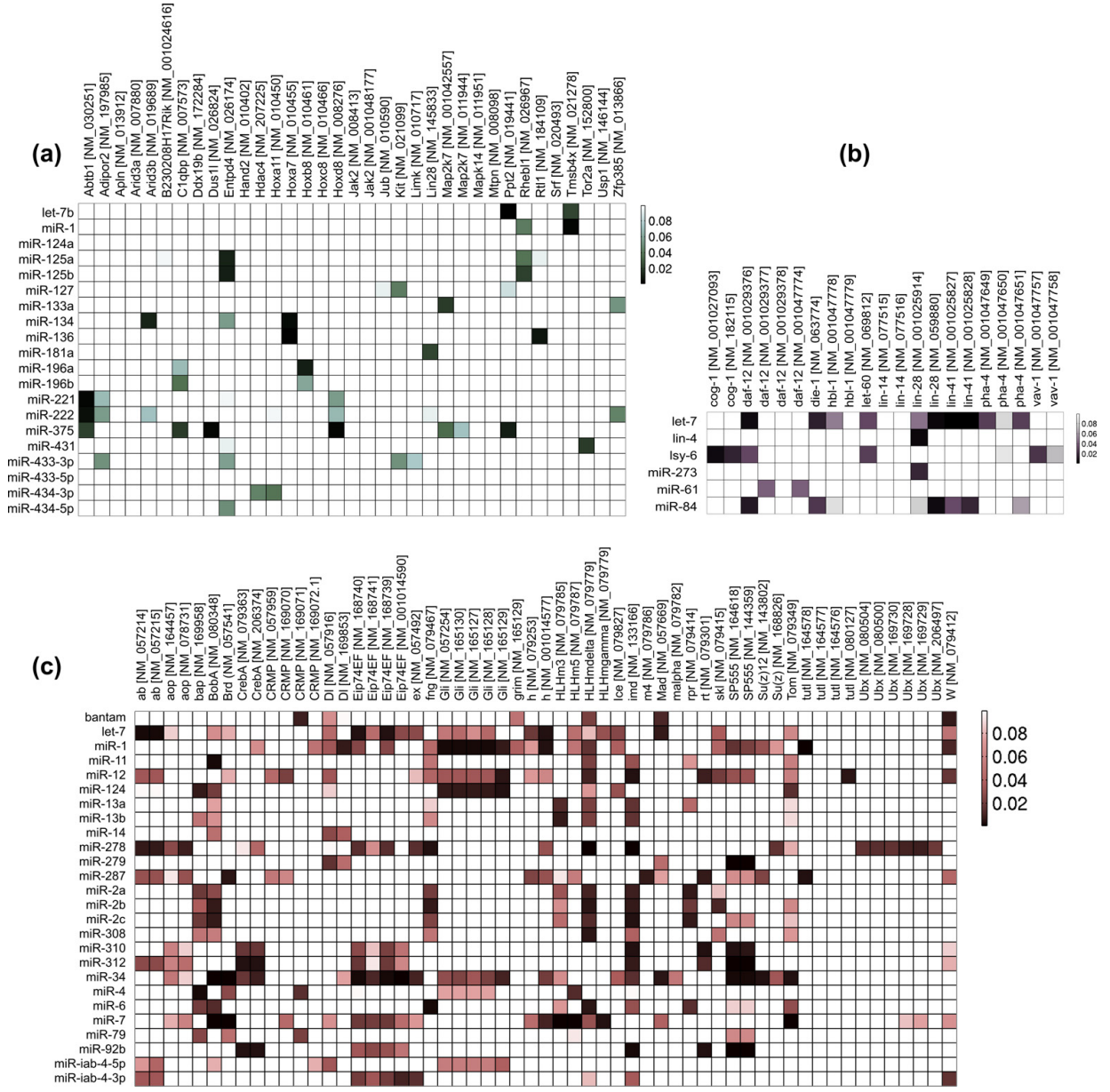

Figure 4

p-values of predicted miRNA targets. A density plot of p-values lower than 0.1, of the most stable predicted miRNA-target duplex for each gene-miRNA pair in (a) mouse, (b) *C. elegans*, and (c) *Drosophila*, with genes along the *x*-axis and miRNAs along the *y*-axis. A lower *p*-value indicates a lower probability of the energy of the duplex (or more favourable energies) occurring due to chance alone. Darker colours indicate lower *p*-values, as shown by the scale in the top-right corner of each sub-figure. White squares indicate no predicted interaction, or a *p*-value greater than the cuto3 value of 0.1.

Additional material

Additional File 1

MicroTar target predictions compared to PicTar. *A list of all experimentally verified targets in the three datasets used (*C. elegans, Drosophila *and mouse), with a comparison of those predicted by MicroTar and those found on the PicTar website.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-7-S5-S20-S1.xls]

Acknowledgements

This work was supported in part by grant R-154-000-265-112 from the National University of Singapore.

RT acknowledges support from the National University of Singapore Research Scholarship.

This article has been published as part of *BMC Bioinformatics* Volume 7, Supplement 5, 2006: APBioNet – Fifth International Conference on Bioinformatics (InCoB2006). The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2105/7?issue=S5

References

1. Lee RC, Feinbaum RL, Ambros V: **The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14.** *Cell* 1993, **75:**843-854.
2. Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, Rougvie AE, Horvitz HR, Ruvkun G: **The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans.** *Nature* 2000, **403:**901-906.
3. Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ: **miRBase: microRNA sequences, targets and gene nomenclature.** *Nucleic Acids Res* 2006, **34:**D140-D144.
4. Bartel DP: **MicroRNAs: Genomics, Biogenesis, Mechanism, and Function.** *Cell* 2004, **116:**281-297.
5. Du T, Zamore PD: **microPrimer: the biogenesis and function of microRNA.** *Development* 2005, **132:**4645-4652.
6. Kim VN, Nam JW: **Genomics of microRNA.** *Trends Genet* 2006, **22:**165-173.
7. Sethupathy P, Corda B, Hatzigeorgiou AG: **TarBase: A comprehensive database of experimentally supported animal microRNA targets.** *RNA* 2006, **12:**192-197.
8. Lai EC: **Predicting and validating microRNA targets.** *Genome Biol* 2004, **5:**115.
9. Didiano D, Hobert O: **Perfect seed pairing is not a generally reliable predictor for miRNA-target interactions.** *Nat Struct Mol Biol* 2006, **13:**849-851.
10. Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP: **Prediction of Mammalian MicroRNA Targets.** *Cell* 2003, **115:**787-798.
11. Brennecke J, Stark A, Russell RB, Cohen SM: **Principles of MicroRNA-Target Recognition.** *PLoS Biol* 2005, **3:**e85.
12. Vella MC, Choi EY, Lin SY, Reinert K, Slack FJ: **The C. elegans microRNA let-7 binds to imperfect let-7 complementary sites from the lin-41 3'UTR.** *Genes Dev* 2004, **18:**132-137.
13. Krek A, Grün D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, MacMenamin P, da Piedade I, Gunsalus KC, Stoffel M, Rajewsky N: **Combinatorial microRNA target predictions.** *Nat Genet* 2005, **37:**495-500.
14. Lall S, Grün D, Krek A, Chen K, Wang YL, Dewey CN, Sood P, Colombo T, Bray N, MacMenamin P, Kao HL, Gunsalus KC, Pachter L, Piano F, Rajewsky N: **A genome-wide map of conserved microRNA targets in C. elegans.** *Curr Biol* 2006, **16:**460-471.
15. Rehmsmeier M, Steffen P, Höchsmann M, Giegerich R: **Fast and effective prediction of microRNA/target duplexes.** *RNA* 2004, **10:**1507-1517.
16. Robins H, Li Y, Padgett RW: **Incorporating structure to predict microRNA targets.** *Proc Natl Acad Sci U S A* 2005, **102:**4006-4009.
17. John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS: **Human MicroRNA Targets.** *PLoS Biol* 2004, **2:**e363.
18. **MicroTar: microRNA target prediction** [http://tiger.dbs.nus.edu.sg/microtar/]
19. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P: **Fast folding and comparison of RNA secondary structures.** *Monatsh Chem* 1994, **125:**167-188.
20. Zuker M, Stiegler P: **Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information.** *Nucleic Acids Res* 1981, **9:**133-148.
21. Bernhart SH, Tafer H, Mückstein U, Flamm C, Stadler PF, Hofacker IL: **Partition function and base pairing probabilities of RNA heterodimers.** *Algorithms Mol Biol* 2006, **1:**3.
22. Karlin S, Altschul SF: **Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes.** *Proc Natl Acad Sci U S A* 1990, **87:**2264-2268.
23. Gumbel EJ: *Statistics of Extremes New York: Columbia University Press*; 1958.
24. Rice P, Longden I, Bleasby A: **EMBOSS: The European Molecular Biology Open Software Suite.** *Trends Genet* 2000, **16:**276-277.
25. Rusinov V, Baev V, Minkov IN, Tabler M: **MicroInspector: a web tool for detection of miRNA binding sites in an RNA sequence.** *Nucleic Acids Res* 2005, **33:**W696-W700.
26. Kiriakidou M, Nelson PT, Kouranov A, Fitziev P, Bouyioukos C, Mourelatos Z, Hatzigeorgiou A: **A combined computational-experimental approach predicts human microRNA targets.** *Genes Dev* 2004, **18:**1165-1178.
27. Sætrom O, Ola Snøve J, Sætrom P: **Weighted sequence motifs as an improved seeding step in microRNA target prediction algorithms.** *RNA* 2005, **11:**995-1003.
28. Stark A, Brennecke J, Russell RB, Cohen SM: **Identification of Drosophila microRNA targets.** *PLoS Biol* 2003, **1:**e60.

Appendix B

MicroTar: miRNA Target Predictions

B.1 Caenorhabditis elegans

TarBase Id	miRNA	Gene	NCBI GeneID	Predicted by	
				PicTar	MicroTar
9	lsy-6	cog-1	175149	**Y**	**Y**
5	let-7	daf-12	181263	**Y**	**Y**
8	let-7	die-1	174569	**N**	**Y**
10	miR-273	die-1	174569	**N**	**N**
4	let-7 family	hbl-1	180848	**Y**	**Y**
13	lin-4	hbl-1	180848	**N**	**N**
11	let-7	let-60	178104	**N**	**Y**
12	miR-84	let-60	178104	**N**	**N**
1	lin-4	lin-14	181337	**Y**	**N**
2	lin-4	lin-28	172626	**Y**	**Y**
3	let-7	lin-41	172760	**Y**	**Y**
6	let-7	pha-4	180357	**N**	**Y**
14	miR-61	vav-1	181153	**Y**	**N**
		Targets Predicted: TP		**7**	**8**
		Total Targets: TP+FN		**13**	**13**
		Sensitivity: TP/(TP+FN)		**0.538461538**	**0.615384615**

B.2 Drosophila melanogaster

TarBase Id	miRNA	Gene	NCBI GeneID	Predicted by	
				PicTar	MicroTar
5	bantam	Mad	33529	**N**	**Y**

21	bantam	W	40009	**Y**	**Y**
64	iab-4	Ubx	42034	**N**	**N**
1	let-7	ab	34560	**Y**	**Y**
63	miR-1	Dl	42313	**N**	**Y**
10	miR-1	tutl	46015	**N**	**Y**
45	miR-11	BobA	50281	**N**	**Y**
58	miR-11	grim	40014	**N**	**N**
43	miR-11	HLHmdelta	43150	**Y**	**Y**
44	miR-11	m4	43157	**N**	**N**
42	miR-11	malpha	43153	**N**	**N**
59	miR-11	skl	40016	**N**	**Y**
12	miR-12	rt	39297	**N**	**Y**
13	miR-124	Gli	34927	**Y**	**Y**
56	miR-13	grim	40014	**N**	**N**
55	miR-13	rpr	40015	**Y**	**Y**
57	miR-13	skl	40016	**Y**	**N**
62	miR-14	Ice	43514	**N**	**N**
18	miR-2	grim	40014	**N**	**Y**
20	miR-2	rpr	40015	**Y**	**Y**
19	miR-2	skl	40016	**Y**	**Y**
66	miR-278	ex	33218	**N**	**Y**
8	miR-279	SP555	53471	**Y**	**Y**
6	miR-287	CRMP	40675	**N**	**Y**
41	miR-2a-1	HLHmdelta	43150	**Y**	**Y**
40	miR-2a-1	malpha	43153	**N**	**N**
60	miR-308	grim	40014	**Y**	**N**
61	miR-308	skl	40016	**Y**	**Y**
9	miR-310	imd	44339	**Y**	**Y**
3	miR-312	CrebA	39682	**Y**	**Y**
4	miR-34	Eip74EF	39962	**Y**	**Y**
11	miR-34	Su(z)12	48071	**Y**	**Y**
49	miR-4	bap	42537	**Y**	**Y**
47	miR-4	BobA	50281	**N**	**N**
34	miR-4	Brd	39620	**Y**	**Y**
35	miR-4	HLHm5	43158	**Y**	**Y**
30	miR-4	HLHmdelta	43150	**Y**	**N**
31	miR-4	HLHmgamma	43151	**N**	**N**
33	miR-4	m4	43157	**N**	**N**
32	miR-4	malpha	43153	**Y**	**N**
29	miR-4	Tom	39619	**Y**	**N**
53	miR-6	grim	40014	**N**	**N**
52	miR-6	rpr	40015	**Y**	**Y**
54	miR-6	skl	40016	**Y**	**N**

51	miR-6	W	40009	**N**	**N**
65	miR-7	aop	33392	**Y**	**Y**
27	miR-7	BobA	50281	**Y**	**Y**
24	miR-7	Brd	39620	**Y**	**Y**
14	miR-7	fng	40314	**N**	**N**
15	miR-7	h	38995	**Y**	**Y**
16	miR-7	HLHm3	43156	**Y**	**Y**
7	miR-7	HLHm5	43158	**Y**	**Y**
46	miR-7	HLHmdelta	43150	**N**	**Y**
23	miR-7	HLHmgamma	43151	**Y**	**Y**
25	miR-7	m4	43157	**N**	**N**
26	miR-7	Tom	39619	**Y**	**Y**
50	miR-79	bap	42537	**Y**	**Y**
48	miR-79	BobA	50281	**N**	**N**
39	miR-79	HLHm5	43158	**Y**	**Y**
36	miR-79	HLHmgamma	43151	**N**	**N**
38	miR-79	m4	43157	**N**	**N**
37	miR-79	malpha	43153	**N**	**N**
2	miR-92b	CrebA	39682	**Y**	**Y**
			Targets Predicted: TP	**35**	**39**
			Total Targets: TP+FN	**63**	**63**
			Sensitivity: TP/(TP+FN)	**0.56**	**0.62**

B.3 Mus musculus

TarBase Id	miRNA	Gene	NCBI GeneID	Predicted by	
				PicTar	MicroTar
3	let-7b	Mtpn	14489	**Y**	**Y**
26	miR-1	Hand2	15111	**N**	**N**
60	miR-1	Hdac4	208727	**N**	**N**
27	miR-1	Tmsb4x	19241	**N**	**Y**
45	miR-124	Mapk14	26416	**Y**	**N**
1	miR-124	Mtpn	14489	**Y**	**Y**
672	miR-125a	Lin28	83557	**Y**	**Y**
688	miR-125b	Abtb1	80283	**Y**	**Y**
687	miR-125b	Apln	30878	**N**	**N**
682	miR-125b	Arid3a	13496	**N**	**Y**
683	miR-125b	Arid3b	56380	**Y**	**Y**
684	miR-125b	B230208H17Rik	227624	**Y**	**N**
680	miR-125b	Ddx19b	234733	**N**	**Y**
681	miR-125b	Dus1l	68730	**Y**	**N**

685	miR-125b	Entpd4	67464	**Y**	**Y**
692	miR-125b	Jub	16475	**N**	**N**
673	miR-125b	Lin28	83557	**Y**	**N**
693	miR-125b	Map2k7	26400	**N**	**N**
690	miR-125b	Ppt2	54397	**N**	**Y**
689	miR-125b	Rhebl1	69159	**N**	**Y**
691	miR-125b	Tor2a	30933	**N**	**N**
686	miR-125b	Zfp385	29813	**Y**	**N**
676	miR-127	Rtl1	353326	**N**	**Y**
61	miR-133	Srf	20807	**N**	**Y**
63	miR-134	Limk1	16885	**N**	**N**
671	miR-136	Rtl1	353326	**N**	**Y**
64	miR-181a	Hoxa11	15396	**Y**	**Y**
21	miR-196	Hoxa7	15404	**N**	**N**
22	miR-196	Hoxb8	15416	**N**	**Y**
23	miR-196	Hoxc8	15426	**N**	**N**
24	miR-196	Hoxd8	15437	**N**	**Y**
58	miR-221	Kit	16590	**Y**	**Y**
59	miR-222	Kit	16590	**Y**	**Y**
44	miR-375	Adipor2	68465	**N**	**Y**
42	miR-375	C1qbp	12261	**N**	**Y**
41	miR-375	Jak2	16452	**N**	**N**
2	miR-375	Mtpn	14489	**N**	**N**
43	miR-375	Usp1	230484	**Y**	**N**
679	miR-431	Rtl1	353326	**N**	**Y**
677	miR-433-3p	Rtl1	353326	**N**	**N**
678	miR-433-5p	Rtl1	353326	**N**	**N**
674	miR-434-3p	Rtl1	353326	**N**	**Y**
675	miR-434-5p	Rtl1	353326	**N**	**Y**
			Targets Predicted: TP	**15**	**24**
			Total Targets: TP+FN	**43**	**43**
			Sensitivity: TP/(TP+FN)	**0.348837209**	**0.558139535**