miRTar Hunter: A Prediction System for Identifying Human microRNA Target Sites

Kiejung Park¹, and Ki-Bong Kim*

MicroRNAs (miRNAs) are important regulators of gene expression and play crucial roles in many biological processes including apoptosis, differentiation, development, and tumorigenesis. Recent estimates suggest that more than 50% of human protein coding genes may be regulated by miRNAs and that each miRNA may bind to 300–400 target genes. Approximately 1,000 human miRNAs have been identified so far with each having up to hundreds of unique target mRNAs. However, the targets for a majority of these miRNAs have not been identified due to the lack of large-scale experimental detection techniques. Experimental detection of miRNA target sites is a costly and time-consuming process, even though identification of miRNA targets is critical to unraveling their functions in various biological processes. To identify miRNA targets, we developed *miRTar Hunter*, a novel computational approach for predicting target sites regardless of the presence or absence of a seed match or evolutionary sequence conservation. Our approach is based on a dynamic programming algorithm that incorporates more sequence-specific features and reflects the properties of various types of target sites that determine diverse aspects of complementarities between miRNAs and their targets. We evaluated the performance of our algorithm on 532 known human miRNA:target pairs and 59 experimentally-verified negative miRNA:target pairs, and also compared our method with three popular programs for 481 miRNA:target pairs. *miRTar Hunter* outperformed three popular existing algorithms in terms of recall and precision, indicating that our unique scheme to quantify the determinants of complementary sites is effective at detecting miRNA targets. *miRTar Hunter* is now available at <http://203.230.194.162/~kbbkim>.

INTRODUCTION

Eukaryotic small noncoding RNAs of approximately 21–25 nucleotides function as one of the central regulators of the expression of an extensive repertoire of genes involved in a remarkably wide range of biological processes-ranging from de-

velopment, differentiation, and cell growth and proliferation to apoptosis and tumorigenesis (Chan et al., 2005; Esquela-Karscher and Slack, 2006). These noncoding RNAs belong to at least two general classes: microRNAs (miRNAs) and short interfering RNAs (siRNAs). miRNAs regulate the expression of protein coding genes at the post-transcriptional level by base-pairing with transcripts of their target genes, subsequently leading to translational repression (Bartel, 2004; Esquela-Karscher and Slack, 2006), mRNA cleavage (Llave et al., 2002; Tang et al., 2003; Yekta et al., 2004), or miRNA-induced degradation (Bagga et al., 2005; Lim et al., 2005; Wu et al., 2006). However, these miRNAs have also been reported to activate target genes (Li et al., 2006). They are transcribed as long primary transcripts (pri-miRNAs) in the nucleus and processed into characteristic stem-loop precursor miRNAs (pre-miRNAs) by the enzyme Drosha. The pre-miRNAs are then transported into the cytoplasm, where they are transformed into small, single-stranded miRNAs with the aid of Dicer (Cullen, 2004). One strand of the mature miRNA is incorporated into the RNA-induced silencing complex (RISC) and binds preferentially to the 3'-untranslated region (3'-UTR) of the target mRNA through complementary base-pairing. The 5'-end sequence of the miRNA, which is 6–11 nucleotides in length, is called the "seed". This sequence interacts with the target in an energetically favorable manner and is the determinant of target repression. Mutations in the seed region of a miRNA sequence can lead to inactive interactions (Doench and Sharp, 2004). The miRNA-mRNA interaction controls the expression level of a target gene by a number of mechanisms, including inhibition of translational initiation, inhibition of elongation, and induction of deadenylation, which decreases mRNA stability and increases the rate of mRNA degradation. miRNAs are involved in diverse regulatory pathways, as well as in disease development and progression, and have been used as diagnostic and prognostic markers and to evaluate treatment response (Dweep et al., 2011).

Although thousands of miRNAs have been identified in animals, plants, and viruses, the targets for the majority of these miRNAs have not been identified due to the lack of large-scale experimental detection techniques. The identification of miRNA targets is critical to unraveling their functions in various biological processes. Many computational approaches have been

Department of Biomedical Technology, Sangmyung University, Cheonan 330-720, Korea, ¹Division of Bio-Medical Informatics, Centers for Genome Science, Korea National Institute of Health, Cheongwon 363-951, Korea
*Correspondence: kbbkim@smu.ac.kr

Received June 21, 2012; revised February 8, 2013; accepted February 8, 2013; published online March 8, 2013

Keywords: computational approach, dynamic programming, miRNA, *miRTar Hunter*, seed region

developed in recent years to predict miRNA targets, for example, miRanda (Betel et al., 2008), TargetScan (Friedman et al., 2009), RNAHybrid (Kruger and Rehmsmeier, 2006), PicTar (Krek et al., 2005), and DIANA-microT (Maragkakis et al., 2009). To predict miRNA targets, these computational approaches typically evaluate three factors: (1) perfect or near-perfect complementarity between the miRNAs and their targets, (2) the thermodynamic stability, *i.e.* the free energy of the miRNA:target duplex, (3) conservation of the miRNA:target duplex across species. Imperfect complementarity between miRNAs and their targets in animals makes target prediction much harder than in plants, whose miRNAs almost always bind their targets with near-perfect complementarity (Doran and Strauss, 2007; Jones-Rhoades and Bartel, 2004). Many existing methods for animals make extensive use of the seed region, but a substantial number of miRNA:target pairs do not have recognizable seed regions. Most existing methods also use patterns of evolutionary conservation to find conserved targets. However, this type of information does not help to identify species-specific targets. In addition, for more than a decade, attempts to study the interactions of miRNAs with their targets have focused on the 3'-UTR region of mRNAs. However, recent studies of miRNA-target interactions have revealed that miRNAs may regulate gene expression by targeting promoters as well as protein coding regions (CDSs). Thus, it is of paramount importance to develop a new approach that can identify putative miRNA target sites not only in the 3'-UTR region, but also in other regions (promoter, 5'-UTR, and CDS) of a gene. In this context, focusing on human miRNAs, we have developed *miRTar Hunter*, a novel computational approach for predicting target sites regardless of the presence or absence of seed matches and across-species patterns of conservation. This approach is based on a dynamic programming algorithm that incorporates sequence-specific features and reflects the properties of various types of target sites that determine diverse aspects of complementarities between miRNAs and their targets. We evaluated the performance of our algorithm on 532 known human miRNA:target pairs and 59 experimentally-verified negative miRNA:target pairs. We also compared our method with three popular programs using a dataset of 481 predicted miRNA:target pairs. Performance evaluation revealed that *miRTar Hunter* outperformed the three popular existing algorithms in terms of recall and precision, indicating that our unique scheme to quantify the determinants of complementary sites is effective at detecting miRNA-target pairs.

MATERIALS AND METHODS

Overall configuration of *miRTar Hunter*

miRTar Hunter is a web-based system that runs on an Apache web server with a Linux operating system. It has a three-tier architecture composed of a client, an application server, and a back-end database. A schematic overview of the *miRTar Hunter* prediction is shown in Fig. 1; the pipeline allows researchers to analyze all possible scenarios to infer the regulatory relationships between miRNAs and targets. The client is a user-friendly web interface for the application server. The application server consists of preprocessing, target scanning, free energy calculation, and post-processing modules. The back-end database is a secondary database that contains a collection of the human miRNAs derived from miRBase database (Kozomara and Griffiths-Jones, 2011). After the parameters/options are specified and miRNAs and mRNAs are input into the system, input data and parameters/options are preprocessed, targets are predicted using a differentially-weighted Smith-Waterman algorithm, the free energy of possible interactions detected by the target scan module are calculated, and the final analysis results are then sorted, ranked, and displayed.

Target scan algorithm

The target prediction component of the *miRTar Hunter* utilizes the Smith-Waterman algorithm (Smith and Waterman, 1981), which is a well-known dynamic programming algorithm that is a core algorithm in many applications. This algorithm performs local sequence alignment to determine regions of similarity between two DNA or protein sequences. We modified it appropriately and created the so-called "differentially-weighted Smith-Waterman" algorithm; this algorithm is optimized to search for maximal local complementarity alignments between a miRNA sequence and all possible positions in the input mRNA sequence. The algorithm consists of three steps: (1) initialize row 0 and column 0, (2) calculate the similarity matrix score, (3) trace back the similarity matrix to search for the optimal local alignment, using dynamic programming. The second step consumes the largest part of the total calculation time. The Smith-Waterman algorithm can be summarized as follows. For two sequences S and T , the length of S is n , $|S|=n$; the length of T is m , $|T|=m$; $V(i,j)$ is the optimal alignment score of two subsequence $S[1]...S[i]$ and $T[1]...T[j]$, and $V(i,j)$ is calculated using Formulas 1 and 2 as shown below:

Fig. 1. Schematic overview of *miRTar Hunter* prediction pipeline. *miRTar Hunter* has a three-tier architecture consisting of a web interface, application server, and back-end database. The back-end database contains a collection of human miRNAs extracted from the miRBase database. The application server consists of a preprocessing module for input data and parameters, a target scan module that searches for complementarity matches between miRNAs and mRNA using a differentially-weighted Smith-Waterman algorithm, a free energy calculation module to calculate the free energy of possible interactions detected by target scan module, and a post-processing module to sort, rank, and display the final analysis results.

	-	S1	S2	S3	S4	S5	...
-	0	0	0	0	0	0	
T1	0	①	②	③	④	⑤	⑥
T2	0	②	③	④	⑤	⑥	⑦
T3	0	③	④	⑤	⑥	⑦	⑧
T4	0	④	⑤	⑥	⑦	⑧	⑨
T5	0	⑤	⑥	⑦	⑧	⑨	⑩
...		⑥	⑦	⑧	⑨	⑩	

Fig. 2. Similarity matrix to calculate sequence and data dependency. While calculating the similarity matrix, the score of any matrix element always depends on the score of three other elements: the up-left neighbor element, the left neighbor, and the up neighbor. The similarity matrix is calculated from the top-left element to the bottom-right element in the direction shown by the arrow.

Initialization:

$$\begin{cases} V(i,0) = 0, 0 \leq i \leq n \\ V(0,j) = 0, 0 \leq j \leq m \end{cases} \quad (1)$$

Recursion relation:

$$V(i,j) = \max \begin{cases} 0 \\ V(i-1,j-1) + (S[i],T[j]) \\ V(i-1,j) + (s[i],-) \\ V(i,j-1) + (-,T[j]) \end{cases} \quad (2)$$

In these formulas, “-” represents a null character or gap; $V(i,0)$ represents the results obtained from comparing each character in S with a gap in T ; $V(0,j)$ is the result obtained from comparing each character in T with a gap in S ; and $(S[i],T[j])$ is the value of the substitution matrix. When calculating the similarity matrix, the score of any matrix element $V(i,j)$ always depends on the score of three other elements: i) the up-left neighbor element $V(i-1, j-1)$, ii) the left neighbor $V(i,j-1)$, iii) the up neighbor $V(i-1,j)$. Therefore, the sequence of the similarity matrix is calculated as shown in Fig. 2. The calculation progresses from the top-left element to the bottom-right element in the direction shown by the arrow in Fig. 2.

The target scan algorithm of *miRTar Hunter* is very similar to the Smith-Waterman algorithm. However, instead of building alignments based on matching nucleotides (A-A or U-U, for example), the scores are based on the complementarity of nucleotides (A=U or G=C). The complementarity scoring matrix used for this analysis also allows G=U ‘wobble’ pairs, which is important for the accurate detection of RNA:RNA duplexes (Wuchty et al., 1999). Table 1 shows the parameters used for complementarity alignments. A score of +7 is assigned to G:C pairs, +5 to A:T pairs, +1 to G:U wobble pairs, and -3 to all other non-complementary pairs. The algorithm uses affine gap penalties (linear in the length of a gap after an initial opening penalty) for gap opening (-8) and gap extension (-2). In addition, following observation of known target sites, complementarity scores (positive and negative values) at the 5’ seed region (positions 1 to 11 counting from the 5’ end of miRNA) and 3’

Table 1. Parameters used for complementarity sequence alignment of miRNAs and mRNAs

Parameters	Symbol	Score	Examples
Complementarity		7	C G or G C
		5	A U or U A
Wobble	:	1	G : U or U : G
		-3	A G G A
Gap	-	-8	A A G G - G
		-2	A C G - - -
Weight	5’ region	4	
	3’ region	2	

The various parameters were designed based on known mechanisms of miRNA target recognition and knowledge of diverse types of target sites.

region (positions 12 to 15) are multiplied by scaling factors of 4 and 2, respectively, so as to reflect the types of miRNA target sites (observed 5’-3’ asymmetry), including 7mer-A1 sites, 7mer-m8 sites, 8mer sites, 6mer sites, offset 6mer sites, 3’-supplementary sites, and 3’-compensatory sites. Seven of these types of target sites were clearly defined by Bartel’s group (Bartel, 2009). The scaling factor is an adjustable parameter that can be optimized as new experimental information becomes available regarding the types of target sites or seed regions. Furthermore, we applied four empirical rules to ensure that the proposed miRNA:mRNA duplexes followed experimentally determined patterns: no mismatches at positions 2 to 4 (counting from the 5’ end); fewer than five mismatches between positions 3-12; at least one mismatch between positions 9 and $L-5$ (where L is the total alignment length); and fewer than two mismatches in the last five positions of the alignment. Using these parameters, the target scan algorithm optimizes the complementarity alignment score between a miRNA sequence and a mRNA sequence, sums over all aligned positions, and creates a ranking of all non-overlapping complementarity alignments in decreasing order of complementarity alignment score down to some threshold value (default value 100). The detection of suboptimal alignments follows heuristics previously used in sequence alignment (Schneider and Sander, 1996; Waterman and Eggert, 1987). The key extension of the *miRTar Hunter* algorithm compared to the Smith-Waterman sequence alignment algorithm is the addition of weighted scores for certain positions and regions in the complementarity alignments according to the four empirical rules and seven types of target sites described above.

To analyze the time complexity of our algorithm, we analyzed each individual part of the algorithm. To initialize the matrix, the scores of row 0 and column 0 need to be input. This has a time complexity of $O(m+n)$. To fill the matrix, each cell of the matrix is traversed and a constant number of operations is performed in each cell, and thus the time complexity of this part is $O(mn)$.

The screenshot shows the miRTar Hunter web interface. At the top, there are logos for Sangmyung University and the Ministry of Health, Welfare and Family Affairs. The main content area is divided into five sections, each with a green header:

- Enter microRNA(s):** Includes a text input for 'Multi-FASTA Format', a 'sample' button, and a 'File Upload' button with a '찾아보기...' (Find) button.
- Sequence Submission:** Includes a text input for 'FASTA Format Sequence', a 'sample' button, and a 'File Upload' button with a '찾아보기...' (Find) button.
- Core Algorithm Options:** Includes input fields for 'Complementarity Score Threshold' (set to 100), 'Free Energy Threshold' (set to 1 kcal/mol), 'Weight for Seed' (set to 4), and 'Weight for 3'-Region' (set to 2).
- Alignment Options:** Includes input fields for 'Gap Opening Penalty' (set to -9) and 'Gap Extension Penalty' (set to -4).
- Types of microRNA Target Sites:** Includes a 'View type detail' link and a list of checkboxes for target site types: '7mer-A1 site', '7mer-m8 site', '8mer site', '6mer site', 'offset 6mer site', '3'-supplementary site', and '3'-compensatory site'.

Fig. 3. User-friendly web interface of *miRTar Hunter*. This web interface has five main sections: “Enter microRNAs”, “Sequence Submission”, “Core Algorithm Options”, “Alignment Options”, and “Types of microRNA Target Sites”. The web interface is designed to allow customization of various options/parameters for tailored analyses.

In the traceback, the algorithm requires the maximum cell be found, and this must be done by traversing the entire matrix, making the time complexity for the traceback $O(mn)$. It is also possible to keep track of the largest cell during the matrix filling segment of the algorithm, although this will not change the overall complexity. Thus the total time complexity of our algorithm is $O(m+n) + O(mn) + O(mn) = O(mn)$. Because this algorithm fills a single matrix of size mn and stores at most n positions for the traceback, the total space complexity of this algorithm is $O(mn) + O(n) = O(mn)$.

Free energy calculation

To estimate the thermodynamic properties of the predicted miRNA:mRNA duplex from each complementarity alignment, *miRTar Hunter* uses the minimum free energy folding routine (MFE Fold), one of folding routines from the Vienna 1.3 RNA secondary structure programming library (RNAlib) (Hofacker, 2003). The calculation of minimum free energy structures is based on dynamic programming algorithm originally developed by Zuker and Stiegler (1981). The expanded thermodynamic parameters are more computationally intensive than the initial target scan, but allow potential hybridization sites to be scored according to their respective folding energies. The miRNA sequence and mRNA sequence from a complementarity alignment are joined into a single sequence with an eight base sequence linker containing artificial ‘X’ bases that cannot base pair. This strand-linker-strand configuration assumes the phase space entropy of strand-strand association is constant for all miRNA-target complementarities (Hofacker, 2003; Schneider and Sander, 1996). The minimum energy of this structure with the last complementary base pair (from the initial sequence alignment) constrained is then calculated using MFE Fold and checked against a threshold value. After target scanning and free energy calculation, predicted target sites for each miRNA are classified into types of target sites and sorted first according to complementarity alignment score and then according to free energy.

Data collection

As stated above, we created a secondary database for human miRNAs obtained from the miRBase database. The secondary database contains 1,448 human miRNA entries hyperlinked to miRBase through the miRNA name and identifier number. Human miRNAs can be retrieved from this database by their names, accession numbers, and mature sequences. The retrieved miRNAs can be used directly as input data for target prediction analysis.

To assess the performance of *miRTar Hunter*, we collected 481 predicted miRNA:mRNA target pairs from the miRNAMap database (Hsu et al., 2008), 532 experimentally-verified positive miRNA:target pairs from the TarBase database (Papadopoulos et al., 2009), and 59 experimentally-verified negative miRNA:target pairs from the web site (http://www.isical.ac.in/~bioinfo_miu/download20.htm) of the Machine Intelligence Unit of the Indian Statistical Institute. The miRNA:target pair of the web site consists of a miRNA name and the GenBank accession number of a target gene. Mature sequences of the miRNAs in the pairs were retrieved from miRBase by their names and transcript sequences of the target genes in the pairs were retrieved from GenBank by their accession numbers. The miRNA:target pair of the TarBase database consists of a mature miRNA sequence and the name of a target gene. In this case, the transcript sequence of a target gene was retrieved from GENE database in National Center for Biotechnology and Information (NCBI) by its gene name. The experimentally-verified positive and negative <mature miRNA sequence, target transcript sequence> pair data were employed to evaluate the performance of our method in terms of recall and precision. The target mRNA sequences from miRNAMap has been identified by three representative computational tools, namely miRanda, TargetScan, and RNAHybrid. We used predicted target data to indirectly compare our method with the three popular programs.

Fig. 4. Performance of the target scan algorithm implemented in *miRTar Hunter* based on 532 experimentally-verified positive miRNA: target pairs and 59 experimentally-verified negative miRNA:target pairs. Our method had sensitivities of 100%, 98.12%, and 77.4% for threshold values of 100, 120, and 140, respectively, which is much higher than that of other existing programs. In previous work, most existing programs had almost the same recall of about 60–65% (Sethupathy et al., 2006). Similar to other programs, our method had a low precision of 42.3% for the threshold value of 100. However, it had relatively high precisions of 78.8% and 96.2% for the threshold values of 120 and 140, respectively.

RESULTS

User-friendly Web Interface and diverse options/parameters for customized analysis

miRTar Hunter is implemented in C, PHP (Personal Hypertext Preprocessor), and SQL (Structured Query Language) languages on a MySQL database management system and a Linux operating system. Figure 3 shows the first page of the *miRTar Hunter* web interface, which is composed of five main sections: “Enter microRNAs”, “Sequence Submission”, “Core Algorithm Options”, “Alignment Options”, and “Types of microRNA Target Sites”. There are three different options for user input of microRNA(s): “Enter microRNAs” is for entering either the microRNAs in multi-FASTA format directly into a textbox or for uploading a file of microRNA(s) or retrieving microRNA(s) from the human miRNA database. The “Sequence Submission” box allows users to input mRNA sequences in FASTA format directly into a textbox or to upload a file containing the input mRNA. The “Core Algorithm Options” has four parameters: complementarity score threshold, free energy threshold, the weight for the 5' region (or seed region), and the weight for the 3' region. The gap opening penalty and gap extension penalty can be altered in the “Alignment Options” section. All options and parameters provided can be specified by the user to be tailored to his/her intent. The “Types of microRNA Target Sites” section allows the user to select multiple items among 7mer-A1 sites, 7mer-m8 sites, 8mer sites, 6mer sites, offset 6mer sites, 3'-supplementary sites, and 3'-compensatory sites. The web interface was designed to allow users to perform customized analyses.

Performance evaluation of *miRTar Hunter*

A variety of quantitative measures have been proposed to characterize the accuracy of prediction methods in the bioinformatics field. A common way of evaluating the performance of a prediction method is using recall and precision. These measures are named for their origin in information retrieval. Recall and precision are defined as

Fig. 5. The distribution of complementarity alignment scores of miRNA-target pairs predicted by miRanda, TargetScan, and RNA-Hybrid. The predicted data were extracted from the miRNAmap database. Our method yielded complementarity alignment scores above the threshold value of 100 for the predicted data. This result indicates that *miRTar Hunter* can detect any target sites predicted by the three popular programs with the threshold value set to 100.

$$recall = \frac{TP}{TP + FN} \quad \text{and} \quad precision = \frac{TN}{TN + FP} \quad (1)$$

Fundamentally, accuracy is related to the degree of concordance between predicted and actual miRNA targets. The outcome of a prediction experiment results in a number of correctly predicted data from both classes, i.e., “true positives (TP)” and “true negatives” (TN). In the same sense, mispredicted data are called “false positives” (FP) and “false negatives” (FN). In the context of miRNA target prediction programs, recall is the proportion of actual miRNA targets that have been correctly predicted as miRNA targets and precision or confidence (as it is called in data mining) denotes the proportion of predicted targets that are actually targets.

As stated above, to assess the predictive accuracy of our target scan algorithm in terms of recall and precision, we used 532 experimentally-verified positive miRNA:target pairs from the TarBase and 59 experimentally-verified negative miRNA: target pairs from the web site (http://www.isical.ac.in/~bioinfo_miu/download20.htm) of the Machine Intelligence Unit of the Indian Statistical Institute. The results of the performance test are shown in Fig. 4. Our target scan algorithm displayed much higher recall than the programs TargetScan, miRanda and RNAHybrid. Sethupathy and colleagues evaluated mammalian target prediction programs (Sethupathy et al., 2006) based on experimentally-supported miRNA-target interactions provided by TarBase, and estimated the performance of each program by determining the program's recall. They reported that most programs had a relatively low recall of about 60–65%. In contrast, our method showed a relatively high recall of 100%, 98.12%, and 77.4% for threshold values of 100, 120, and 140, respectively. However, false positive data are major problems in miRNA target predictions. Similar to the existing programs, our method had a low precision of 42.3% for a threshold value of 100. However, it showed relatively high precisions of 78.8% and 96.2% for threshold values of 120 and 140, respectively. It is difficult to avoid a high rate of false positives in miRNA target prediction in animals because of the partial complementarity of miRNAs and their targets in animals. We also indirectly compared the performance of *miRTar Hunter* with that of TargetScan, miRanda, and RNAHybrid. For this indirect comparison, we selected 481 predicted miRNA:target pairs from the miRNAmap database (Hsu et al., 2008) that had been predicted by TargetScan, miRanda, and RNAHybrid. These pairs were input into *miRTar Hunter*, and the complementarity alignment score for

each pair was calculated. Figure 5 shows the distribution of complementarity alignment scores, which were all above the threshold value of 100. This result indicates that *miRTar Hunter* can detect any target sites predicted by the three popular programs with a threshold value of 100. A direct comparison of our algorithm with existing algorithms is challenging, because the different published methods are based on different principles, making it difficult to an accurate comparison. In addition, many methods are not available for download for independent testing on a common dataset, whereas the datasets used by these methods are highly diverse, which is why we only indirectly compared these methods.

DISCUSSION

Experimental identification of miRNA target sites is a costly and time-consuming process. While recent estimates suggest that more than 50% of human protein coding genes may be regulated by miRNAs and that each miRNA may bind to hundreds of target genes, the latest release of the TarBase database contains information on only 995 human *in vivo* miRNA-target gene interactions involving 103 distinct miRNAs and 825 distinct genes, a far cry from the actual extent of miRNA targeting (Bartel, 2009; Papadopoulos et al., 2009). Computational prediction of miRNA-target gene interactions is a valuable tool for guiding wet-lab experiments, and it remains the only option for systematic genome-wide analysis. It is also a challenging task because of the difficulty of distinguishing true miRNA-mRNA interactions from the noisy background of millions of possible miRNA-mRNA combinations and, more generally, because the basic mechanisms of miRNA target recognition remain largely unknown. In recent years, several target prediction algorithms based on different principles have been developed (Bartel, 2009). However, the two recurring parameters used by the available methods are the existence of a seed match and evolutionary conservation of the target site across multiple species. Utilization of these powerful constraints in prediction algorithms leads to more reliable detection of functional duplexes, but at the same time, limits our ability to identify biologically relevant miRNA target sites that do not fulfill these requirements. By definition, organism-specific or simply poorly conserved sites cannot be predicted at all if a conservation filter is applied. It has also been suggested that the seed match requirement may be too stringent, and that several non-canonical types of target sites exist that cannot be detected by seed match based methods. Furthermore, many potential miRNA-target interactions that do involve conserved seed regions may be non-functional in a physiological context.

In light of the above, we sought to develop a computational technique free from both the seed requirement and the conservation filter and thus capable of predicting species-specific and non-canonical as well as canonical target sites. While *miRTar Hunter* does not require a perfect seed match, it weights the seed region and the non-seed region differentially according to various types of target sites. However, our algorithm does have flaws that can be improved upon. If multiple miRNAs target the same site on a transcript, only the miRNA with the highest complementarity alignment score and lowest energy score is reported for that site. This is a possible source of false negatives because different miRNAs are expressed at different times in cell development. It is possible that multiple miRNAs can bind to overlapping sites, but because they are never expressed at the same time in the cell, they do not usually compete for binding. Furthermore, even though we modified the Smith-Water-

man alignment algorithm specifically for miRNA/mRNA sequence comparisons, it can be argued that the Smith-Waterman alignment was originally designed to compare evolutionarily-related sequences, and that miRNAs and their mRNA targets do not fall into this category. Finally, the parameter values and penalties assigned throughout our algorithm are arbitrary and should be modified as greater knowledge regarding the biology of miRNA targeting accumulates. Ultimately, we intend to further refine the rules for miRNA-target interactions as more and more miRNA interactions are characterized and validated.

ACKNOWLEDGMENTS

This work was supported by a grant from the Korea Centers for Disease Control and Prevention.

REFERENCES

- Bagga, S., Bracht, J., Hunter, S., Massirer, K., Holtz, J., Eachus, R., and Pasquinelli, A.E. (2005). Regulation by let-7 and lin-4 mirnas results in target mRNA degradation. *Cell* 122, 553-563.
- Bartel, D.P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116, 281-297.
- Bartel, D.P. (2009). MicroRNAs: target recognition and regulatory functions. *Cell* 136, 215-233.
- Betel, D., Wilson, M., Gabow, A., Marks, D.S., and Sander, C. (2008). The microRNA.org resource: targets and expression. *Nucleic Acids Res.* 36, D149-153.
- Chan, J.A., Krichevsky, A.M., and Kenneth, S.K. (2005). MicroRNA-21 is an antiapoptotic factor in human glioblastoma cells. *Cancer Res.* 65, 6029-6033.
- Cullen, B.R. (2004). Transcription and processing of human microRNA precursors. *Mol. Cell* 16, 861-865.
- Doench, J.G., and Sharp, P.A. (2004). Specificity of microRNA target selection in translational repression. *Genes Dev.* 18, 504-511.
- Doran, J., and Strauss, W.M. (2007). Bioinformatic trends for the determination of miRNA-target interactions in mammals. *DNA Cell Biol.* 26, 353-350.
- Dweep, H., Sticht, C., Pandey, P., and Gretz, N. (2011). miRWalk - Database: Prediction of possible miRNA binding sites by "walking" the genes of three genomes. *J. Biomed. Inform.* 44, 839-847.
- Esquela-Kerscher, A., and Slack, F.J. (2006). Oncomirs - microRNAs with a role in cancer. *Nat. Rev. Cancer* 6, 259-269.
- Friedman, R.C., Farh, K.K., Burge, C.B., and Bartel, D.P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.* 19, 1-11.
- Hofacker, I. (2003). Vienna RNA secondary structure server. *Nucleic Acids Res.* 31, 3429-3431.
- Hsu, S.D., Chu, C.H., Tsou, A.P., Chen, S.J., Chen, H.C., Hsu, P., Wong, Y.H., Chen, Y.H., Chen, G.H., and Huang, H.D. (2008). miRMAP 2.0: genomic maps of microRNAs in metazoan genomes. *Nucleic Acids Res.* 36, D165-169.
- Jones-Rhoades, M.W., and Bartel, D.P. (2004). Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. *Mol. Cell* 14, 787-799.
- Kozomara, A., and Griffiths-Jones, S. (2011). miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.* 39, D152-157.
- Krek, A., Grun, D., Poy, M.N., Wolf, R., Rosenberg, L., Epstein, E., MacMenamin, P., Piedade, I.D., Gunsalus, K.C., Stoffel, M., et al. (2005). Combinatorial microRNA target predictions. *Nat. Genet.* 37, 495-500.
- Kruger, J., and Rehmsmeier, M. (2006). RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic Acids Res.* 34, W451-454.
- Li, L.C., Okino, S.T., Zhao, H., Pookot, D., Place, R.F., Urakami, S., Enokida, H., and Dahiya, R. (2006). Small dsRNAs induce transcriptional activation in human cells. *Proc. Natl. Acad. Sci. USA* 103, 17337-17342.
- Lim, L.P., Lau, N.C., Garrett-Engele, P., Grimson, A., Schelter, J.M., Castle, J., Bartel, D.P., Linsley, P.S., and Johnson, J.M. (2005). Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* 433, 769-773.

- Llave, C., Xie, Z., Kasschau, K.D., and Carrington, J.C. (2002). Cleavage of scarecrow-like mRNA targets directed by a class of Arabidopsis miRNA. *Science* 297, 2053-2056.
- Maragkakis, M., Reczko, M., Simossis, V.A., Alexiou, P., Papadopoulos, G.L., Dalamagas, T., Giannopoulos, G., Goumas, G., Koukis, E., Kourtis, K., et al. (2009). DIANA-microT web server: elucidating microRNA functions through target prediction. *Nucleic Acids Res.* 37, W273-276.
- Papadopoulos, G.L., Reczko, M., Simossis, V.A., Sethupathy, P., and Hatzigeorgiou, A.G. (2009). The database of experimentally supported targets: a functional update of TarBase. *Nucleic Acids Res.* 37, D155-158.
- Schneider, R., and Sander, C. (1996). The HSP database of protein structure-sequence alignments. *Nucleic Acids Res.* 24, 201-205.
- Sethupathy, P., Megraw, M., and Hatzigeorgiou, A.G. (2006). A guide through present computational approaches for the identification of mammalian microRNA targets. *Nat. Methods* 3, 881-886.
- Smith, T.F., and Waterman, M.S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* 147, 195-197.
- Tang, G., Reinhart, B.J., Bartel, D.P., and Zamore, P.D. (2003). A biochemical framework for RNA silencing in plants. *Genes Dev.* 17, 49-63.
- Waterman, M.S., and Eggert, M. (1987). A new algorithm for best subsequence alignments with application to tRNA-rRNA comparisons. *J. Mol. Biol.* 197, 723-728.
- Wu, L., Fan, J., and Belasco, J.G. (2006). MicroRNAs direct rapid deadenylation of mRNA. *Proc. Natl. Acad. Sci. USA* 103, 4034-4039.
- Wuchty, S., Fontana, W., Hofacker, I.L., and Schuster, P. (1999). Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers* 49, 145-165.
- Yekta, S., Shih, I.H., and Bartel, D.P. (2004). MicroRNA-directed cleavage of HOXB8 mRNA. *Science* 304, 594-596.
- Zuker, M., and Stiegler, P. (1981). Optimal computer folding of large RNA sequences using thermodynamic and auxiliary information. *Nucleic Acids Res.* 9, 133-148.