Composite MicroRNA Target Predictions and Comparisons of Several Prediction Algorithms

Jin Zhou[1], Shili Lin[1,2], Vince Melfi[1,3], and Joe Verducci[1,2]

[1]Mathematical Biosciences Institute, and
[2]Department of Statistics, Ohio State University.
[3]Department of Statistics and Probability, Michigan State University.

April 18, 2006

Abstract

Motivation: MicroRNAs (or miRNAs) are short noncoding RNAs whose primary role is to repress translations by negatively regulating gene expressions post-transcriptionally through binding to their mRNA targets. Currently, there exist a few computational algorithms for human miRNA target predictions, but their results may vary widely. Therefore, it would be useful to consolidate and filter through these discrepant results so that researchers may have a greater degree of certainty in the validity of these targets before engaging in costly experiments.

Results: We studied three of the most popular target prediction algorithms, miRanda, TargetScan, and PicTar, systematically through the use of three measures of similarity/distance and a statistical test on the gene ontology categories that are enriched in the target lists. Furthermore, two composite statistics were devised to combine and rank the composite target list. We applied the methods developed to all human miRNAs that had been identified. Our results indicate that TargetScan and PicTar tend to have a greater degree of similarity, based on all measures considered. We also demonstrate that our composite statistics and the program implementing them can be useful tools for filtering through large target sets to short list genes for downstream experiments.

Contact: shili@stat.ohio-state.edu.

Supplementary Information: The program implementing the composite statistics (R codes) and supplementary materials are at
`http://www.stat.ohio-state.edu/~statgen/SOFTWARE/miRComp`.

1 Introduction

MicroRNAs (miRNAs, or miRs) are short, 21-22 nucleotides (nt), noncoding RNAs whose primary role, as currently understood, is to repress translations by negatively regulating gene expressions post-transcriptionally through binding to mRNA targets (Bartel, 2004). These tiny RNAs are highly conserved over species, both in plants and humans, with some conserved from *C. elegans*, through *Drosophila*, to *Homo Sapiens* (Bartel, 2004). The discoveries of the two founding members of the miRNA family were made in *C. elegans* through the prototype stRNAs (small temporal RNAs) that regulate developmental transitions. The first microRNA, *lin-4*, was discovered more than 10 years ago in 1993 by a group of scientists interested in studying the timing of *C. elegans*'s larval development (Lee *et al.*, 1993). The precursor of the *lin-4* miRNA is folded into a stem loop (also known as a lady's hairpin structure), with sequence complementarity between the 22nt mature part and the 3' untranslated region (UTR) of *lin-14 mRNA*. Not until seven years later were the hairpin structure, the 22nt mature segment, and its complementarity to 3' UTR mRNAs exemplified once again when the second founding member, *let-7*, was discovered (Reinhart *et al.*, 2000; Slack *et al.*, 2000). The *let-7* miRNA performs its function by binding to the 3' UTR of *lin-41* and *lin-57* and inhibiting their translation. Soon after, *let-7* was detected in *Homo Sapiens*, *Drosophila*, and many other species of animals (Pasquinelli *et al.*, 2000). Since then, the number of miRNAs has grown exponentially, with hundreds identified in worms, flies, plants, vertebrates, and mammals.

It is well documented that a super majority of the miRNA genes are derived from independent transcription units (Lagos-Quintana *et al.*, 2001); they are novel genes not previously annotated. Only about 25% of them are in the introns of pre-mRNAs (Lau *et al.*, 2001).

In addition to the regulation of developmentally timed events, a hallmark of the two founding members of this abundant family of regulatory genes, miRNAs play many other key roles in a variety of biological processes, including the control of cell proliferation, cell death, and fat storage. More recently, a great deal of attention has been focused on the involvement of miRNAs in tumorigenesis (Calin *et al.*, 2004; He *et al.*, 2005; Lu *et al.*, 2005; O'Donnell *et al.*, 2005). The functional role of a miRNA is typically inferred from the functions of its mRNA targets, which have a portion of their 3' UTR sequence complementary to the miRNA. The complementarities need not be exact in animals (Ambros, 2004), and various computer programs employ different algorithms for assessing the strength of the matching. Among the most popular computer programs are miRanda (Enright *et al.*, 2003; John *et al.*, 2004), TargetScan (Lewis *et al.*, 2003, 2005), and PicTar (Krek *et al.*, 2005). Vatolin *et al.* (2006), and Wang and Wang (2006) are the most recent additions to the list of computational methods for target identification. A common theme among the majority of such algorithms is the ranking of the targets based on some measure of strength/significance of the putative targets.

According to the Sanger registry (`http://microrna.sanger.ac.uk`), researchers have already submitted evidence for more than 30,000 putative mRNA targets of miRNAs in *Mus Musculus* and nearly as many in *Homo Sapiens*. Clearly, it would be useful for researchers to have some degree of certainty in the validity of these targets before engaging in costly experiments. Currently, there is only modest agreement among the various computational methods. This notion is exemplified by a 2005 competition sponsored by the New York

Academy of Sciences, in which the number of targets predicted for two miRNAs ranged from 1 to more than 500 from different algorithms (`http://www.nyas.org/ebriefreps/main.asp?intSubsectionID=3392`). Thus it is of importance to compare these algorithms in a systematic fashion. Furthermore, because of low powers yet high false positive rates of these prediction programs due to imperfect understanding of the functions of miRNAs and how they achieve these functions, various filtering steps are necessary to screen through the various lists to arrive at a "shortlist" of putative targets for experimental validation.

In this paper, by analyzing the lists of putative targets predicted by the three aforementioned established computer programs, we measure the extent of agreement according to three similarity measures. We also study the similarities in the gene ontology categories that are enriched in the target lists. Furthermore, we have developed and implemented an algorithm to sort through the multiple lists to identify genes that are common targets of multiple prediction programs and are of relatively high likelihoods of being true targets. Two composite statistics are devised to rank the combined predicted targets. This effort can be viewed as a filtering step to short list genes for future experiments. The number of genes to be validated can be chosen by the investigator with the aid of the statistically ranked list. Our program (R code) can be downloaded freely from our web site.

2 Algorithms for miRNA target predictions

The three target prediction algorithms we study, miRanda (Enright *et al.*, 2003; John *et al.*, 2004), PicTar (Krek *et al.*, 2005), and TargetScan (Lewis *et al.*, 2003, 2005), have somewhat similar criteria at the large scale, but differ greatly in their details. For example, all require (complementary) matching of a portion of the target to the miRNA, but the size and location of the required match differ, as do the requirements for the accuracy of the match. Similarly, all require conservation across several organisms, but the organisms and the specific details of the required conservation differ. The algorithms also differ in whether anything beyond seed matching and conservation are required, and in what the additional requirements are. Brief descriptions of the three algorithms are given below.

TargetScan (TS): To determine whether an mRNA is a (putative) target of a miRNA, the TargetScan algorithm starts with the miRNA sequence and the 3' UTR from the mRNA from the human, mouse, rat, dog, and chicken. A multiple alignment of the orthologous 3' UTRs from the five organisms is constructed. The algorithm uses the following three ingredients:

1. ***A conserved seed match***, i.e., perfect Watson-Crick complementarity (A:U or G:C matches) between nucleotides 2-7 of the miRNA and a 6 nucleotide section of the 3 UTR of the mRNA. The match is required to occur at corresponding positions for all organisms in the multiple alignment.

2. ***A conserved anchoring A***, which is an A nucleotide on the 3' UTR just downstream of the seed match. This anchoring A is required to be in the UTRs of all the organisms.

3. ***A conserved m8-t8 match***, which is an A:U or G:C match between the eighth nucleotide of the miRNA and the corresponding position on the 3' UTR. Again, this

match must occur on the UTRs of all the organisms.

An mRNA is declared to be a target of a miRNA if there is a conserved seed match, and either a conserved anchoring A or a conserved m8-t8 match (or both).

miRanda (MR): The miRanda algorithm begins by scanning for (partial) sequence complementarity between a miRNA and the 3' UTR of an mRNA. Some implementations also look for matches between the miRNA and the full cDNA sequences. The scanning algorithm is similar to the Smith-Waterman algorithm (Smith and Waterman, 1981); the main difference is that the Smith-Waterman algorithm looks for matching nucleotides such as A:A, while in the present application interest centers on Watson-Crick complementarity, so A:U and G:C matches are highly valued. The G:U "wobble pair" is also allowed.

The precise details of the algorithm are given in Enright *et al.* (2003) and John *et al.* (2004). The basic idea is now sketched. Matched pairs get positive scores (e.g., +5 for G:C and A:U pairs, and +2 for G:U pairs), mismatched pairs get a negative score (e.g., −3), and there is a gap-opening and gap-elongation penalty, (e.g., −8 and −2 respectively). In addition, the individual scores are multiplied by a scaling factor (e.g. 2.0) for nucleotides 1-11 of the miRNA. (This gives extra importance to matching in this area, just as TargetScan requires a seed match in a similar area.) Based on these rules, a score S is computed for an alignment. In addition, the free energy ΔG of the resulting duplex is computed using the RNAlib package (Wuchty *et al.*, 1999). Cutoffs are chosen, such as $S < 90$ and $\Delta G < -17$, to determine which matches to pursue further (The scoring parameters and cutoff values have been changed in the latest version of miRanda).

The target sites on the 3' UTRs that survive the cutoffs are then examined for conservation between organisms. The conservation is not required to be perfect. The positions of the target sites can be shifted slightly (e.g., ±10 residues), and sequence identity is not required to be perfect (e.g., 90% identity may be required). Surviving targets are ranked first by alignment score S and second by free energy ΔG.

PicTar (PT): The PicTar algorithm identifies potential targets via a seed match (similar to the TargetScan algorithm), a free energy calculation, and species conservation. These potential targets are then scored via an algorithm involving a hidden Markov model. Some main features are described next; for exact details see Krek *et al.* (2005).

The algorithm starts with multiple alignments across several organisms of a 3' UTR region and a miRNA. A seven nucleotide segment which has perfect Watson-Crick complementarity between the miRNA and the 3' UTR, starting at either nucleotide one or two of the miRNA, is called a *perfect nucleus*. If a miRNA and a (specific organisms) 3' UTR have a perfect nucleus, then it is checked whether perfect nuclei for the miRNA fall into overlapping positions for all organisms under consideration. If this is true, then the optimal free energy of the duplex is computed and compared to a cutoff. Perfect nuclei that fall into overlapping positions for all organisms and pass the free energy cutoff are called *anchors*.

Anchors are fed into the hidden-Markov-model-based scoring algorithm, which yields a score for each 3' UTR in the multiple alignment, e.g., a score for human, a score for mouse, etc. Scores of the various organisms under consideration are averaged to yield a PicTar score.

3 Methods

3.1 Matching targets

The three algorithms extract 3' UTR information from two databases: PicTar and TargetScan use the UCSC genome browser database (`http://genome.ucsc.edu`), and miRanda uses the Ensembl database (`http://www.ensembl.org`). In addition, different identifiers are used to denote the targets. Specifically, TargetScan, PicTar, and miRanda use HUGO, RefSeq, and Ensembl IDs, respectively. In order to compare and rank the predicted targets from these algorithms, we need to match the different identifiers of the same genes. We chose the HUGO IDs to be the matching identifiers for the following reasons. First, the HUGO ID information of PicTar can be easily extracted from its annotation of RefSeq IDs in the output. Second, miRanda output gives some HUGO ID and RefSeq ID information of Ensembl IDs. For those Ensembl IDs that do not have corresponding HUGO IDs, we downloaded an annotation database from `http://www.ensembl.org/Multi/martview` and built a comprehensive list of Ensembl IDs and their corresponding HUGO and RefSeq IDs. Finally, if several RefSeq IDs correspond to one HUGO ID, we treat them as one target, since they are transcript variants of the same gene.

3.2 Combination of target lists and rank aggregation

To combine the results from these three algorithms, we want to rank the targets by a composite statistic. In what follows we present a method of computing an average rank and a composite p-value that can be used to rank the combined targets. Let S_A^M be the set of potential targets of the miR "M" predicted by algorithm "A", and we use $n_A = ||S_A^M||$ to denote the cardinality of the set S_A^M. Here A can be one of $\{MR, TS, PT\}$, representing the three algorithms miRanda, TargetScan, and PicTar. For simplicity in notation, we omit the superscript representing the particular miR hereafter. Then S, defined by $S = S_{MR} \cup S_{TS} \cup S_{PT}$ is the union of targets predicted by at least one of the algorithms. For each $t \in S$, $R_A(t)$ and $p_A(t)$ denote the rank and "p-value", respectively, of the target t by algorithm A.

Computing the rank is relatively easy. If the target $t \in S$ is not predicted by algorithm A (i.e., $t \notin S_A$), then we set $R_A(t) = n_A + 1$. On the other hand, if the target t is predicted by algorithm A, then the rank $R_A(t)$ is obtained directly from the algorithm, since all three algorithms rank their predicted targets. The p-value is a bit more cumbersome to obtain. If the target t is not predicted by algorithm A, the p-value $p_A(t)$ is simply set to 1.0. For targets t that are predicted by an algorithm A, the p-value is computed as follows:

- For $A = MR$, there is no information given by the algorithm beyond the ranking, so a uniform distribution is assumed, and the p-value is given by $p_A(t) = R_A(t)/(n_A + 1)$. The added 1 in the denominator signifies all the other targets not predicted by the MR algorithm.

- For $A = TS$, the "estimated false discovery rate" given by TargetScan is treated as the p-value, although the published work (Lewis *et al.*, 2003, 2005) is not clear about how this rate is computed.

- For $A = PT$, each predicted target is given a score, with a larger value indicating a greater confidence in the predicted target being a real target. Since the distribution of these scores is quite skewed, we recommend a log-transformation of the scores to obtain the p-value under a normal distribution.

It should be noted that the term "p-value" is being used rather loosely.

One way to combine the ranks from the three lists is through the use of the weighted average of the individual ranks:

$$\bar{R}(t) = \frac{1}{||A||} \sum_{A \in \{MR,TS,PT\}} \frac{1}{n_A + 1} R_A(t), \;\; t \in S,$$

where $||A||$ denote the number of algorithms whose targets lists are being combined. In our application, $||A|| = 3$. These averaged ranks are then reranked to obtain the composite ranks. For obtaining composite p-values, we use the statistic

$$T(t) = -2 \log \prod_{A \in \{MR,TS,PT\}} p_A(t), \;\; t \in S,$$

introduced in Fisher (1925). We then transform $T(t)$ back to obtain a composite p-value according to a chi-square distribution with six degrees of freedom. Based either on the the composite rank or the composite p-value, the investigator can find his or her own short list for validation or further investigation, depending on budgetary considerations or tolerance for "Type I" error. We note again, however, that the calculated p-values should be interpreted as rough guidelines rather than as hard statistical evidence.

Our algorithm and the program implementing the algorithm may be obtained from the website `http://www.stat.ohio-state.edu/~statgen/SOFTWARE/miRComp`. The program is fully documented with several examples. The input files are created from the web servers of the three algorithms. The output includes a spreadsheet of the combined targets and the composite p-value and composite rank for each putative target.

3.3 Three similarity measures of top k lists

Each algorithm yields a target list ranked by the "probability" of a predicted target being a true target of the miR. Since each algorithm uses its own cutoff to determine whether or not to include a gene in the target list, the sizes of the lists vary from algorithm to algorithm and from miR to miR (see Results). To make the matter even more complex, the cutoffs are all based on ad-hoc procedures, and thus the Type I error rate (of including a false target in the list) and the power (of a true target being included) are largely unknown. Therefore, rather than working with a whole list whose statistical properties are unknown, we opt to focus on the top k (k ranging from 1-100) targets of each list and assess the similarity of such top k lists.

The notion of a top k list is widely used in information retrieval and social choice theory (Fagin *et al.*, 2003). Fix a positive integer k. For each algorithm $A \in \{MR, TS, PT\}$ and each $i = 1, \ldots, k$, let X_{Ai} denote the gene id of the i^{th} predicted target of algorithm A

(assuming the targets predicted by the algorithm is ordered from most likely to least likely). For each pair of algorithms (A, A'), define

$$C_k(A, A') = \sum_{i=1}^{k} \sum_{j=1}^{k} I(X_{Ai} = X_{A'j}).$$

Here $I(X_{Ai} = X_{A'j}) = 1$ if X_{Ai} and $X_{A'j}$ are the same gene and is 0 otherwise, so $C_k(A, A')$ counts the number of targets that are predicted by both algorithm A and algorithm A'. Also, define $P_k(A, A') = C_k(A, A')/k$ to be the proportion of commonly predicted targets. The algorithms A and A' are deemed similar if $P_k(A, A')$ is large. In particular, scatter plots of $P_k(A, A')$ versus k should yield useful information on the similarities of the algorithms A and A' in predicting the targets of a particular miR. Relative similarity between two pairs of algorithms can also be assessed from the corresponding proportions.

We use two other well-known distances to gauge the similarities of the top k target lists, Kendall's distance and Spearman's footrule distance (Fagin *et al.*, 2003). Before defining these distances, some additional notation is useful. For $k \geq 1$ let $S_A(k) = \{X_{A1}, \ldots, X_{Ak}\}$ denote the top k targets predicted by algorithm A. For each $t \in S_A(k) \cup S_{A'}(k)$, let $R_A^k(t) = R_A(t)$ if $t \in S_A(k)$, and let $R_A^k(t) = k + 1$ otherwise. Define $R_{A'}^k(t)$ similarly.

Now Kendall's distance may be defined. For any unordered pair $t, u \in S_A(k) \cup S_{A'}(k)$, define

$$K_{tu}^{(p)} = \begin{cases} 0 & \text{if } R_A^k(t) < R_A^k(u),\ R_{A'}^k(t) < R_{A'}^k(u) \text{ or } R_A^k(t) > R_A^k(u),\ R_{A'}^k(t) > R_{A'}^k(u) \\ 1 & \text{if } R_A^k(t) > R_A^k(u),\ R_{A'}^k(t) < R_{A'}^k(u) \text{ or } R_A^k(t) < R_A^k(u),\ R_{A'}^k(t) > R_{A'}^k(u) \\ p & \text{if } R_A^k(t) = R_A^k(u) = k+1 \text{ or } R_{A'}^k(t) = R_{A'}^k(u) = k+1. \end{cases}$$

In words, if the orderings of t and u are the same in both algorithms, $K_{tu}^{(p)}$ is set to 0. If the orderings are different, $K_{tu}^{(p)}$ is set to 1. If both t and u appear in one algorithm's top k list, but neither appears in the other algorithm's top k list, the ranks for the latter algorithm will both be $k + 1$ and $K_{tu}^{(p)}$ is set equal to p. Here $0 \leq p \leq 1$ is a constant. Now define the Kendall's distance to be

$$K_k^{(p)}(A, A') = \sum_{t,u \in S_A(k) \cup S_{A'}(k)} K_{tu}^{(p)}.$$

This definition is equivalent to that given in (Fagin *et al.*, 2003).

Spearman's footrule distance with location parameter $k + 1$, $F^{(k+1)}$, is defined by

$$F^{(k+1)}(A, A') = \sum_{t \in S_A(k) \cup S_{A'}(k)} |R_A^k(t) - R_{A'}^k(t)|.$$

Recall that the proportion of commonly predicted targets, $P_k(A, A')$, is a normalized version of $C_k(A, A')$, since we divide by the maximum value k of $C_k(A, A')$ to obtain $P_k(A, A')$. We will similarly use normalized versions of Kendall's and Spearman's distance measures, by dividing by their maximum values. For Kendall's distance with $p = 0$ (which we used for all of our analysis) the maximum value is k^2 (see Appendix A), whereas for Spearman's footrule with location parameter $k + 1$, the maximum value is $k(k + 1)$. Both of these maxima are achieved when there are no commonly predicted targets. To avoid excessive notation, we

now use the symbols $K_k^{(p)}$ and $F^{(k+1)}$ to denote the normalized versions of Kendall's and Spearman's distances, respectively.

To assess whether a distance between two top k lists is significantly smaller (i.e., whether the lists are significantly more similar) than one would expect under random ranking, conditional on the union of the two lists, we generate the empirical null distribution by the following procedure. We randomly permute the rankings in $\{R_A^k(t), t \in S_A(k) \cup S'_A(k)\}$ and in $\{R_{A'}^k(t), t \in S_A(k) \cup S'_A(k)\}$, respectively, to obtain two lists of new rankings. Based on these rankings, we compute the distance measure, for example the normalized Kendall's distance $K_k^{(p)}$. We repeat this procedure N times, and the resulting distances $\{K_{k,n}^{(p)} \colon n = 1, \ldots, N\}$ form an empirical distribution that can be used as an estimate of the null distribution under the hypothesis of random ranking. For all the analysis presented below, we used $N = 1000$. An empirical p-value is calculated in the usual way. For example, with Kendall's distance, the p-value is calculated as $\sum_{n=1}^{N} I(K_{k,n}^{(p)} \leq K_k^{(p)})/N$, the proportion of times that the actual Kendall's distance is at least as large as the simulated Kendall's distance.

3.4 Enriched GO categories

We also compare the predicted targets from different algorithms through the Gene Ontology analysis of the gene products. The idea is that, although two algorithms may predict two different mRNA targets, the functional categories of the two targets may be the same. Thus, comparison of functional categories of the targets provides another avenue for comparing the differences or similarities of the algorithms. Specifically, for each target list S_A, we use the Gene Ontology Tree Machine (GOTM; `http://genereg.ornl.gov/gotm`, Zhang *et al.*, 2004) to annotate the functions of the genes in the list, and to classify them into functional categories.

One particular element of interest is the identification of GO categories that are being enriched in the predicted targets by an algorithm. That is, we want to identify functional categories that harbor more genes in the predicted list than expected if the genes were randomly selected from a reference set. Under the null hypothesis of random selection, the number of genes from the list falling into a particular category follows a hypergeometric distribution, leading to a simple test for the hypothesis. Let E_A denote the set of categories that are significantly enriched in S_A based on the hypergeometric test using all genes in the human genome as the reference set. We use $E = E_{MR} \cup E_{TS} \cup E_{PT}$ to denote the union of all enriched categories from the three target prediction algorithms under consideration. Furthermore, let $V_{A/A'}$ be a binary vector defined on E as follows. For each $e \in E$, if $e \in E_A \cap E_{A'}$, then $V_{A/A'}(e) = 1$; otherwise $V_{A/A'}(e) = 0$. Let $p(A, A')$ be the population proportion of commonly enriched categories from the targets predicted by A and A', and let $p(B, B')$ be the proportion of commonly enriched categories from the targets predicted by algorithms B and B'. We are interested in testing the null hypothesis $H_0 \colon p(A, A') = p(B, B')$ using the Z test for two population proportions. If H_0 is rejected, this provides evidence that the predicted targets from A and A' are significantly more (or less) similar than the predicted targets from B and B'.

Table 1: Top 10 targets of miR-155 based on composite rank

Gene ID	miRanda		TargetScan		PicTar		Composite	
(HUGO)	p-value	rank	p-value	rank	p-value	rank	p-value	rank
BACH1	0.014	1	0.069	2	0.008	2	0.0006	1
CARHSP1	0.189	14	0.28	31	0.092	15	0.099	2
SOCS1	0.338	25	0.22	16	0.112	20	0.144	3
FGF7	0.405	30	0.08	3	0.165	29	0.106	4
MAP3K14	0.216	16	0.26	24	0.370	57	0.257	5
ZIC3	0.541	40	0.13	7	0.114	21	0.140	6
RNF123	0.595	44	0.22	15	0.070	11	0.153	7
CEBPB	0.459	34	0.22	14	0.210	37	0.261	8
ZNF537	1*	74	0.068	1	0.001	1	0.004	9
ASTN2	1*	74	0.25	22	0.019	4	0.097	10

*A p-value of 1 indicates that the target in the combined list is not predicted by the corresponding algorithm

4 Results

4.1 An example of composite ranking

The composite ranking algorithm presented in Section 3.2 is applicable to either entire target lists or top k lists. Here we demonstrate the utility of the algorithm and program using the whole target lists from MR, TS, and PT, for miR-155. In Table 1 we show the top 10 targets based on the composite ranks. The full list can be found in the supplementary material. We observe that most of the top 10 ranked genes are commonly predicted by all three algorithms, but there are two among the top 10 that are only predicted by two of the algorithms. In particular, ZNF537, ranked No. 9 in its composite ranking, was ranked No. 1 in both TS and PT, but not predicted by MR. We may use the composite p-value instead to short list genes for further analysis/validation. In this case, eight of the 10 genes in Table 1 will also appear, but with ZNF537 and ASTN2 (the two genes not predicted by MR) ranked higher, as second and third. Overall, these two composite statistics are highly correlated, with a Spearman correlation of 0.93 for miR-155. We further note that the composite p-value should be treated as just a vehicle for ranking the combined targets; it is not advisable to interpret the p-value in the usual sense and use it as a cutoff for controlling desired (comparison-wise) Type I error rate.

4.2 Number of predicted targets

We considered all 319 human miRNAs that have been identified to date (Release 8.0 of miRBase Sequence Database; `http://microrna.sanger.ac.uk`). Of these, 109 did not yield any predicted targets by any of the three algorithms. In particular, MR and TS did not have any predicted targets for almost 2/3 of the miRs, while PT failed to predict targets

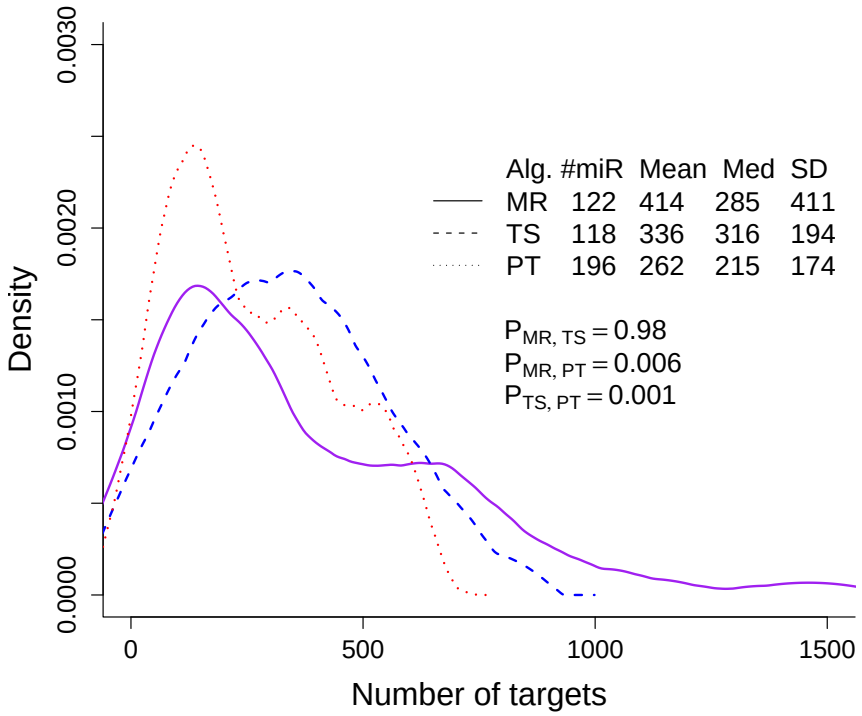

Figure 1: Distribution of the number of targets predicted by each of the three algorithms. In the legend, #miR gives the number of miRs that have any predicted target by the corresponding algorithm. Each $P_{A,A'}$ is the p-value of the Wilcoxon test for the number of predicted targets by the pair of algorithms.

for just above 1/3 of them. The three curves in Figure 1 each gives the distribution of the number of targets predicted by an algorithm. It is apparent that MR gives large sets of predictions to some of the miRs, reflected in a large mean and standard deviation. Overall, in terms of the number of predicted targets, MR and TS are more similar (also reflected in their medians), while PT predicts the fewest. In fact, a Wilcoxon rank-sum test indicates that the typical number of targets predicted by PT is significantly smaller than the number predicted by either MR or TS. (p-values of 0.006 and 0.001 respectively), while there is no significant difference between the numbers predicted by MR and TS (p-value 0.98).

4.3 Distance/similarity measures

As discussed and shown above, since the statistical properties of the predicted targets from the different algorithms are unknown, and since the actual number of predicted targets varied widely, we study the similarity/differences of these algorithms based on the top k lists, with k ranging from 1 to 100. Out of the total of 319 miRs, 62 of them have predicted targets, and 54 have at least 100 predicted targets, from each of the three algorithms. We used the triplet of measures, $\{P_k(A, A'), K_k^{(0)}(A, A'), F^{(k+1)}(A, A')\}$, to study the similarity of each pair of prediction algorithms A and A'.

Results for the three miRs, *let-7a*, *miR-101*, and *miR-142-3p*, are shown in Figure 2. These three miRs were chosen because their patterns are representative of the entire set, with the results for all 54 miRs given in the supplementary material. In each of the subfigures in Figure 2, we plotted either $P_k(A, A')$ (row 1), $K_k^{(0)}(A, A')$ (row 2), or $F^{(k+1)}(A, A')$ (row 3), versus k, for a miR. The purple solid, blue dashed, and red dotted curves are for $(A, A') = (MR, TS), (MR, PT)$, and (TS, PT), respectively. As can be seen from the figure, all three measures give extremely consistent patterns. Note that $1 - P_k(A, A')$ may also be interpreted as a distance measure like the other two. Furthermore, more than half of the miRs show that TS and PT are more similar than the other two pairs, which show the same degree of similarity. The rest of the minority of miRs show that the three pairs all have comparable similarity. On the average, TS and PT predict 30% common targets among each algorithm's top k lists, while MR and PT, and MR and TS predict only 17.5% and 17% common targets, respectively.

To gain a global view of all the miRs, in Figure 3 we show the density plot for $k = 20, 50$ and 100, on row 1 for the common proportion P_k, row 2 for Kendall's distance $K_k^{(0)}$, and row 3 for Spearman's footrule distance $F^{(k+1)}$. The results reinforce the observations made from Figure 2. The distribution of $P_k(TS, PT)$ in terms of common proportion stochastically dominates those of $P_k(MR, TS)$ and $P_k(MR, PT)$. On the other hand, in terms of distance, the distribution of $K_k^{(0)}(TS, PT)$ or $F^{(k+1)}(TS, PT)$ is stochastically dominated by the distributions of the other two pairs.

To evaluate whether the similarity in the ranked lists observed between a pair of prediction algorithms, especially TS and PT, are significantly more similar than one would expect if the two lists are ranked completely at random, we perform a Monte Carlo simulation as described in Section 3.3, for $k = 20, 50$, and 100. The raw p-values were adjusted to obtain the False Discovery Rates (FDRs) (Benjamini and Hochberg, 1995). Table 2 gives the proportion of miRs whose predicted top k lists are significantly more similar (FDRs < 0.05) in the rankings between two algorithms. The results show that TS and PT have higher proportions of significantly similar rankings than the other two pairs, regardless of the size of the lists, as long as the list is not too small ($k \geq 20$). However, despite less similarity compared to the (TS, PT) pair, the rankings between (MR, TS) and (MR, PT) are still significantly similar for about half of the miRs, with large top k lists ($k \geq 50$).

4.4 Gene Ontology Analysis

Due to the server constraint of GOTM with an upper limit of 500 genes, we could only analyze 29 miRs whose predicted target lists do not exceed the capacity. Table 3 lists the miRs that have at least one p-value smaller than 0.01 across all three tests of equal proportions. The full table with the results for all 29 miRs is available as supplementary materials. As can be seen from the table, $\hat{p}_3 = \hat{p}(TS, PT)$, the proportion of commonly enriched GO categories by the targets of TS and PT, tend to be larger than its two counterparts, indicating that TargetScan and PicTar have more commonly enriched GO categories than the other two pairs. This observation is substantiated by the three tests of equally enriched proportions. Of the 16 miRs in table 3 that have at least one p-value less than 0.01, 15 of them were selected because $p_3 = p(TS, PT)$ is significantly larger than $p_1 = p(MR, TS)$ or $p_2 = p(MR, PT)$, or both (indicated by *).

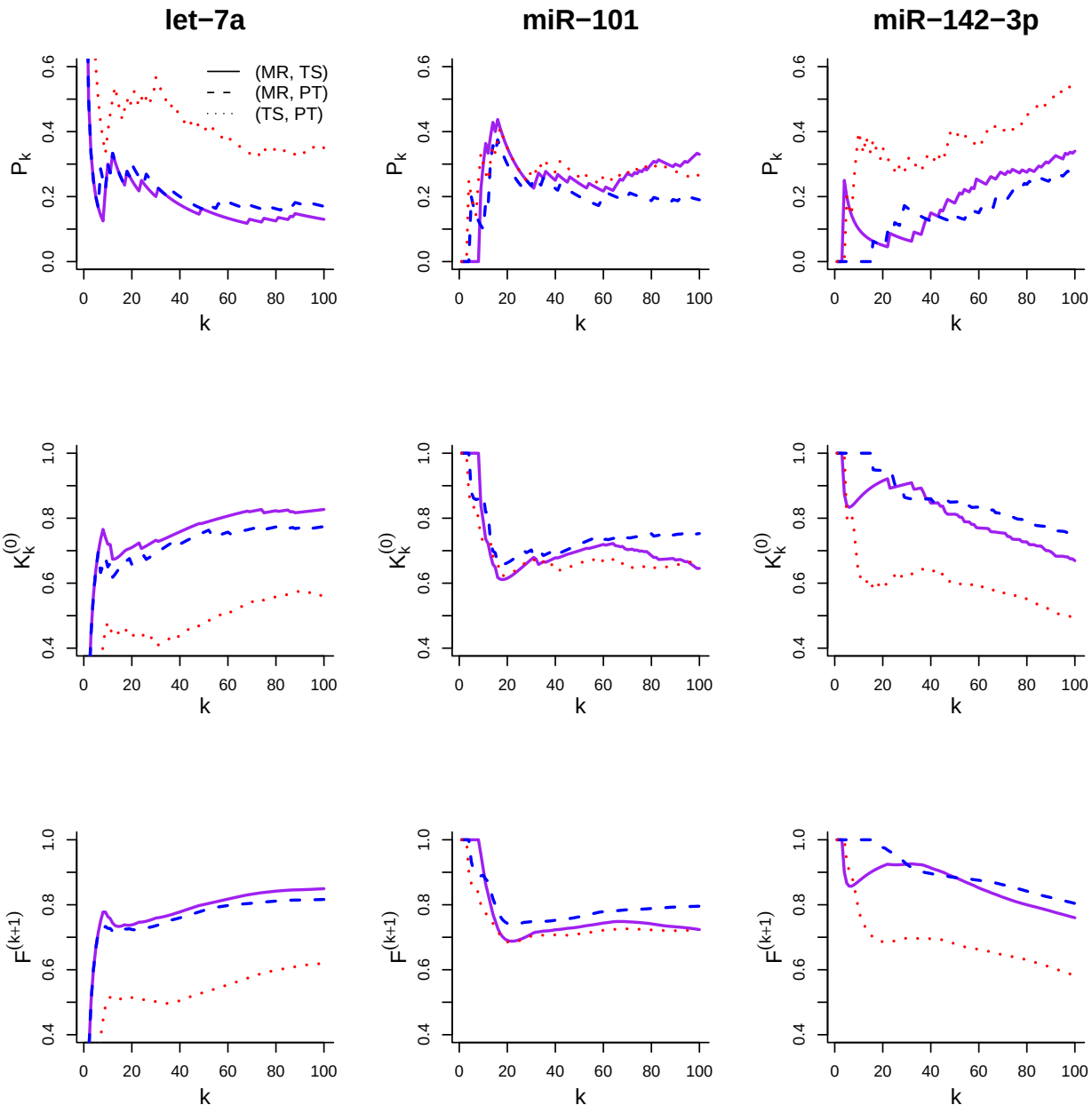

Figure 2: Plot of common proportion (row 1), Kendall's distance (row 2) and Spearman's footrule distance (row 3) vs. k=1 to 100, for three representative miRNAs.

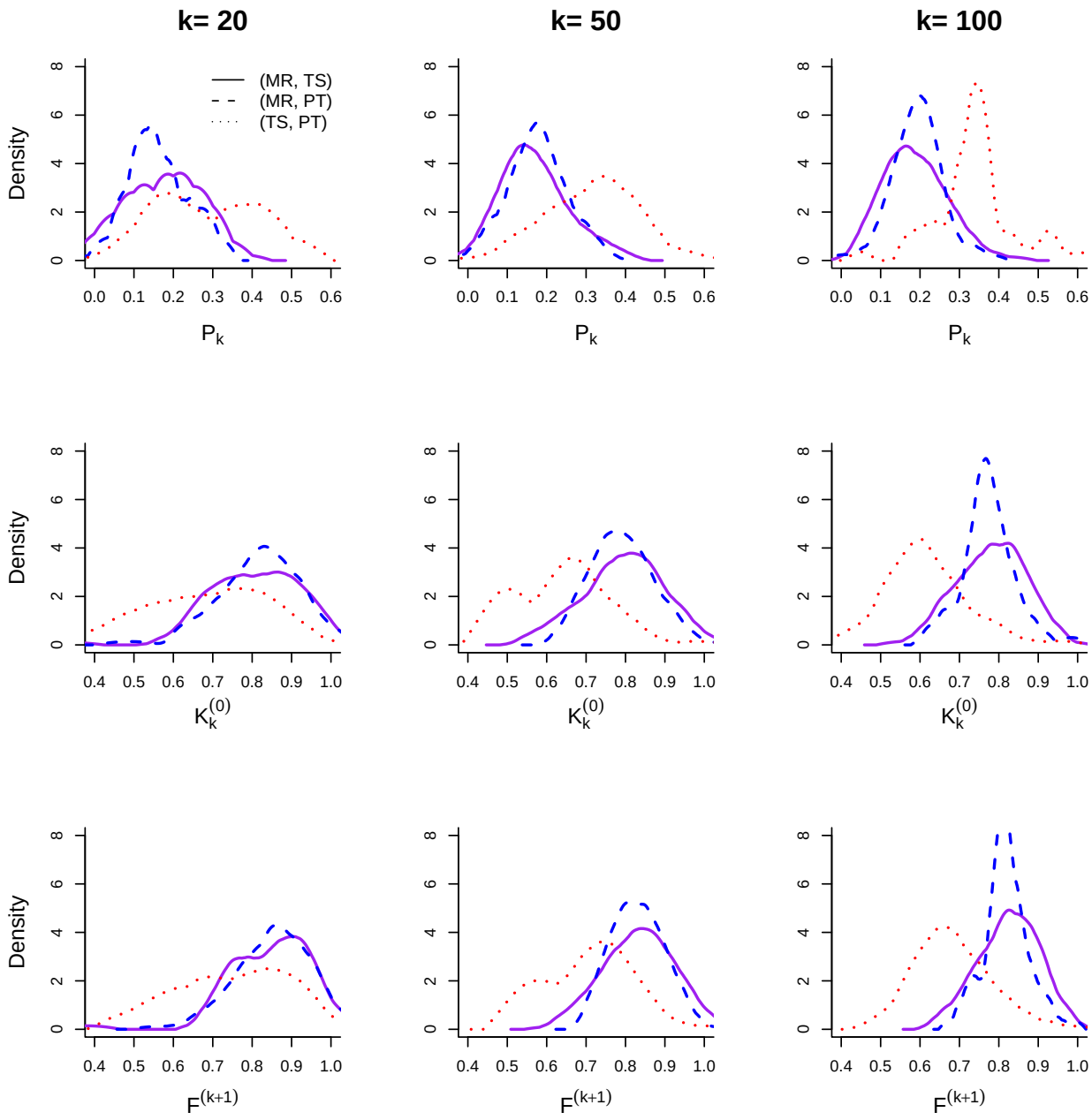

Figure 3: Density plot of common proportion (row 1), Kendall distance (row 2) and Spearman's footrule distance (row 3) for k=20, 50, 100, for the same three representative miRNAs depicted in Figure 2.

Table 2: Proportion of miRs with FDRs smaller than 0.05 for each pair of algorithms when testing the null hypothesis of random ranking.

			Prop. FDR< 0.05		
Distance	k	# of miRs	(MR, TS)	(MR, PT)	(TS, PT)
Kendall	20	62	0.177	0.161	0.306
	50	58	0.517	0.534	0.655
	100	54	0.630	0.704	0.833
Spearman	20	62	0.177	0.177	0.290
	50	58	0.603	0.483	0.603
	100	54	0.556	0.685	0.815

Table 3: Proportion of commonly enriched GO categories from two predicted lists (left half) and p-values of the common proportion tests (right half), where $p_1 = p_{(MR,TS)}, p_2 = p_{(MR,PT)}$, and $p_3 = p_{(TS,PT)}$.

				H_0			
miR	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$	$p_1 = p_2$	$p_1 = p_3$	$p_2 = p_3$	
miR-139	0.164	0.164	0.598	1	0	0	*
miR-155	0.328	0.293	0.586	0.688	0.004	0.0009	*
miR-203	0.023	0.031	0.465	0.701	0	0	*
miR-215	0.037	0.073	0.232	0.302	0.0001	0.004	*
miR-216	0.067	0.022	0.489	0.303	0	0	*
miR-218	0.204	0.19	0.352	0.765	0.005	0.0015	*
miR-219	0.024	0.182	0.08	0.0003	0.084	0.042	
miR-302b	0.397	0.405	0.664	0.893	0	0	*
miR-302c	0.366	0.435	0.565	0.255	0.001	0.035	
miR-302d	0.236	0.252	0.593	0.766	0	0	*
miR-32	0.232	0.271	0.496	0.473	0	0.0001	*
miR-33	0.193	0.199	0.357	0.892	0.0006	0.0009	*
miR-369-3p	0.009	0	0.395	0.315	0	0	*
miR-373	0.286	0.286	0.471	1	0.001	0.001	*
miR-375	0	0	0.491	NaN	0	0	*
miR-7	0.262	0.295	0.426	0.568	0.006	0.031	

5 Discussion

In this paper we compare the similarity of three miRNA target prediction algorithms, miRanda, TargetScan, and PicTar, using proportion of common targets, and modifications of Kendall's distance and Spearman's footrule. By comparing the top k lists of 54 miRs, all three measures show that TargetScan and PicTar are more similar in their target lists. This by no means invalidates miRanda, since that system may help identify targets that are not found using either TargetScan or PicTar. As relatively few putative targets of computational programs have been experimentally verified, the ultimate accuracies of the methods remain in question. We suggest only that targets identified by multiple algorithms would appear to be the better candidates for verification.

Greater similarity between TargetScan and PicTar is also observed by investigating the proportions of commonly enriched gene ontology categories in the target genes by a pair of algorithms. Out of the 29 miRNAs analyzed, 25 of them have estimated common proportions between TargetScan and PicTar at least as large as those of the other two pairs, miRanda and TargetScan, and miRanda and PicTar. This observation is further substantiated by the results of the three tests on equal common proportions.

Rank aggregation of top k lists is a useful technique for quantifying the relative likelihood of targets. In this paper we use the Borda method (weighted average of ranks) and a composite p-value to combine the three top k lists. We wish to re-emphasize that the so-called p-values should not be interpretated in the usual manner; we would suggest that they only be used as a means for ranking the relative likelihoods of the putative targets being true targets. A more sophisticated method would be to find the optimal top k list τ that minimizes $\sum_i d(\tau, \tau_i)$, where d is a distance measure and $\tau_i's$ are the existing full lists or top k lists. Although an NP-hard problem for some distances, often a solution may be found by searching near computable optima for other distances. This is undoubtedly an interesting problem, but it is out of the scope of the current paper.

In the example shown for composite prediction, we see that the top 10 rankings based on either the composite rank or the composite p-value have eight targets in common. In fact, the two ranking systems are highly correlated, although the composite rank statistic tends to favor targets that are commonly predicted by all algorithms while its composite p-value counterpart prefers those ranked near the top even thought they may have been predicted by only a subset of the algorithms that includes TargetScan. This is not suprising, as the range of the p-values (estimated false discovery rates) for those predicted by TargetScan tends to be smaller than those of the other two algorithms. One may rescale the range of the p-values through a linear transformation so that all three are comparable, if so desired.

Also, in the example, we used the whole lists from the algorithms. Had we used only top k lists (e.g. top 50 lists), the results would have been different, and the results will also vary depending on k. Whether to use the whole lists or top k lists for an appropriate k is problem/situation dependent, because the statistical properties of the predicted target set are largely unknown. If an investigator is interested in only a small number (say 10) of putative targets for further investigation/validation, then composite ranking using top k (say $k = 20$) would probably be more sensible. The issue of p-value scale difference may be addressed in this case by a procedure similar to the "quantile normalization" approach frequently employed in gene expression analysis (Bolstad *et al.*, 2003).

Another miRNA target prediction algorithm, DIANA-MicroT (Kiriakidou et. al. 2004) has also been used in several studies, but this algorithm requires inputing the gene sequence or the miRNA sequence to make a prediction, and the predicted targets for each miR are usually very few. We tried all the targets of miR-155 predicted by miRanda, TargetScan, and PicTar, and found that only one target, CEBPB, was also predicted by DIANA-MicroT.

Finally, although the paper uses the three most popular algorithms for demonstrating the methods proposed for comparisons and rank aggregation, the methods are applicable to a different set of, or including additional, algorithms, such as the recent additions from Vatolin *et al.* (2006) and Wang and Wang (2006).

Acknowledgements

This work was supported in part by NSF grants DMS-0112050 (to the MBI) and DMS-0306800 (to S.L.). The authors would like to thank Dr. Carlo Croce for introducing them to microRNA, and Drs. George Calin and Tom Schmittgen for helpful discussions.

References

Alberts,B., Johnson,A, Lewis,J., Raff,M., Roberts,K., Walter,P. (2002) *Molecular Biology of the Cell*, fourth edition. Garland Science.

Ambros,V. (2004) The functions of animal microRNAs. *Nature*, **431**, 350-355.

Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281-297.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. B*, **57**, 289-300.

Bolstad,B.M., Irizarry,R.A., Astrand,M., Speed,T.P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioiformatics*, **19**, 185-193.

Calin,G.A., Sevignani,C., Dumitru,C.D., Hyslop,T., Noch,E., Yendamuri,S., Shimizu,M., Rattan,S., Bullrich,F., Negrini,M., Croce,C.M. (2004) Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. *PNAS*, **101**, 2999-3004.

Eis,P.S., Tam,W., Sun,L., Chadburn,A., Li,Z., Gomez,M.F., Lund,E., Dahlberg,J.E. (2005) Accumulation of miR-155 and BIC RNA in human B cell lymphomas. *PNAS*, **102**, 3627-3632.

Enright,A.J., John,B., Gaul,U., Tuschl,T., Sander,C., Marks,D.S. (2003) MicroRNA targets in *Drosophila. Genome Biol.*, **5**, R1.

Fagin,R., Kumar,R., and Sivakumar,D. (2003) Comparing top k lists. *SIAM J. Discrete Math.*, **17**, 134-160.

Fisher,R.A. (1925) *Statistical methods for research workers*. London: Oliver and Boyd.

He,L., Thomson,J.M., Hemann,M.T., Hernando-Monge,E., Mu,D., Goodson,S., Powers,S., Cordon-Cardo,C., Lowe,S.W., Hannon,G.J., Hammond,S.M. (2005) A microRNA polycistron as a potential human oncogene. *Nature*, **435**, 828-834.

John,B., Enright,A.J., Aravin,A., Tuschl,T., Sander,C., Marks,D.S. (2004) Human microRNA targets. *PLOS Biology*, **2**, 1862-1879.

Kiriakidou,M., Nelson,P.T., Kouranov,A., Fitziev, P., Bouyioukos,C., Mourelatos,Z. and Hatzigeorgiou,A. (2004) A combined computational-experimental approach predicts human microRNA targets. *Genes Dev.*, **18**, 1165-1178.

Krek,A., Grun,D., Poy,M.N., Wolf,R., Rosenberg,L., Epstein,E.J., MacMenamin,P., da Piedade,I., Gunsalus,K.C., Stoffel,M., Rajewsky,N. (2005) Combinatorial microRNA target predictions. *Nature Genetics*, **37**, 495-500.

Lagos-Quintana,M., Rauhut,R., Lendeckel,W., Tuschl,T. (2001) Identification of novel genes coding for small expressed RNAs. *Science*, **294**, 853-858.

Lau,N.C., Lim,L.P., Weinstein,E.G., Bartel,D.P. (2001) An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science*, **294**, 858-862.

Lee,R.C., Feinbaum,R.L., Ambros,V. (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, **75**, 843-854.

Lewis,B.P., Shih,I.H., Jones-Rhoades,M.W., Bartel,D.P., Burge,C.B. (2003) Prediction of mammalian microRNA targets. *Cell*, **115**, 787-798.

Lewis,B.P., Burge,C.B., Bartel,D.P. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**, 15-20.

Lu,J., Getz,G., Miska,E.A., Alvarez-Saavedra,E., Lamb,J., Peck,D., Sweet-Cordero,A., Ebert,B.L., Mak,R.H., Ferrando,A.A., Downing,J.R., Jacks,T., Horvitz,H.R., Golub,T.R. (2005) MicroRNA expression profiles classify human cancers. *Nature*, **435**, 834-838.

O'Donnell,K.A., Wentzel,E.A., Zeller,K.I., Dang,C.V., Mendell,J.T. (2005) c-Myc-regulated microRNAs modulate E2F1 expression. *Nature*, **435**, 839-843.

Pasquinelli,A.E., Reinhart,B.J., Slack,F., Martindale,M.Q., Kuroda,M.I., Maller,B., Hayward,D.C., Ball,E.E., Degnan,B., Muller,P., Spring,J., Srinivasan,A., Fishman,M., Finnerty,J., Corbo,J., Levine,M., Leahy,P., Davidson,E., Ruvkun,G. (2000) Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*, **408**, 86-89.

Reinhart,B.J., Slack,F.J., Basson,M., Pasquinelli,A.E., Bettinger,J.C., Rougvie,A.E., Horvitz,H.R., Ruvkun,G. (2000) The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature*, **403**, 901-906.

Slack,F.J., Basson,M., Liu,Z., Ambros,V., Horvitz,H.R., Ruvkun,G. (2000) The lin-41 RBCC gene acts in the C. elegans heterochronic pathway between the let-7 regulatory RNA and the lin-29 transcription factor. *Mol. Cell*, **5**, 659-69.

Smith,T.F. and Waterman, M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195-197.

Vatolin,S., Navaratne,K., and Weil,R.J. (2006) A novel method to detect functional microRNA targets. *J. Mol. Biol.*, to appear.

Wang,X.J., Reyes,J.L., Chua,N.H. and Gaasterland,T. (2004) Prediction and identification of *Arabidopsis thaliana* microRNAs and their mRNA targets. *Genome Biol.*, **5**, R65.

Wang,X. and Wang,X. (2006) Systematic identification of microRNA functions by combining target prediction and expression profiling. *Nucleic Acids Research*, **34**, 1646-1652.

Wuchty,S., Fontana,W., Hofacker,I.L., Schuster,P. (1999) Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, **49**, 145-165.

Zhang,B., Schmoyer,D., Kirov,S. and Snoddy,J. (2004) GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. *BMC Bioinformatics*, **5**, 16.

Appendix A

Proposition. For any $p \in [0,1]$, the maximum of $K_k^{(p)}$ occurs when the two top k lists from algorithms A and A' are disjoint. In particular, for $p = 0$, the maximum is k^2.

Proof. Let $S_A(k)$ and $S_{A'}(k)$ be the top k targets predicted by algorithms A and A', respectively, as defined in the main text. Suppose $S_A(k)$ and $S_{A'}(k)$ have $k-m$ common targets, $m \in \{0, 1, \cdots, k\}$, denoted by $S_C = S_A(k) \cap S_{A'}(k)$. We also use $S_{A\setminus C}$ ($S_{A'\setminus C}$) to denote the set that comprises of targets predicted by A (A') but not commonly predicted by both. For each $t \in S_A(k) \cup S_{A'}(k)$, define $R_A^k(t) = R_A(t)I(t \in S_A(k)) + (k+1)I(t \in S_{A'\setminus C})$ and $R_{A'}^k(t) = R_{A'}(t)I(t \in S_{A'}(k)) + (k+1)I(t \in S_{A\setminus C})$, where $R_A(t)$ and $R_{A'}(t)$ are the rankings of the predicted targets as defined in the main text, and I is the usual 0-1 indicator function. Then,

$$\begin{aligned} K_k^{(p)} &= m(m-1)p + m^2 + \sum_{t\in S_{A\setminus C}, u\in S_C} I(R_A^k(t) < R_A^k(u)) + \sum_{t\in S_{A'\setminus C}, u\in S_C} I(R_{A'}^k(t) < R_{A'}^k(u)) + \\ &\quad \sum_{t<u, t,u\in S_C} [I(R_A^k(t) < R_A^k(u), R_{A'}^k(t) > R_{A'}^k(u)) + I(R_A^k(t) > R_A^k(u), R_{A'}^k(t) < R_{A'}^k(u))] \\ &\leq f(m) = m(m-1)p + m^2 + 2m(k-m) + \frac{1}{2}(k-m)(k-m-1), \end{aligned}$$

with equality achieved when all the rankings in $\{R_A^k(t), t \in S_{A\setminus C}\}$ $\left(\{R_{A'}^k(t), t \in S_{A'\setminus C}\}\right)$ are smaller than those in $\{R_A^C(t), t \in S_C\}$ $\left(\{R_{A'}^C(t), t \in S_C\}\right)$ and all the rankings in S_C for

A and A' are in opposite directions. Since $f(m)$ is an increasing function in the range of $m \in \{0, 1, \cdots, k\}$, the maximum of $f(m)$ occurs when $m = k$, corresponding to the situation that the two top k lists are disjoint. Plugging $p = 0$ into $f(k)$, we obtain the maximum of k^2 for $K_k^{(0)}$.