

# MiRNA Prediction with a Novel Ranking Algorithm Based on Random Walks

Yunpen Xu\*, Xuefeng Zhou\* and Weixiong Zhang<sup>1†</sup>

Department of Computer Science and Engineering

<sup>1</sup>Department of Genetics

Washington University in Saint Louis

Saint Louis, MO 63130-4899, USA

## Abstract

MicroRNA (miRNAs) play very important roles in diverse cellular and physiological processes through the post-transcriptional gene regulatory pathway. Several proposed computational approaches have been shown to complement experimental methods in discovery of miRNAs whose expression is restricted to nonabundant cell types or specific environmental conditions. These computational methods rely on enough characterized miRNAs as training samples, and resort to genome annotations to minimize the miRNA candidate pools. However, many sequenced genomes have only very few identified miRNAs, and most of these genomes are not well annotated. The existing tools are not feasible in these situations.

In this work, we propose a novel ranking algorithm based on random walks, for identification of miRNAs. This algorithm aims at predicting miRNAs from genomes with few known miRNA reported and/or less annotation information. Our algorithm is first tested on *H. sapiens* data. With very few known miRNAs as labeled samples, we can obtain a prediction accuracy of more than 95%. We employed our approach to conduct a study on *A. gambiae*. 200 novel miRNA precursors are predicted. 78 of the putative miRNA precursors encode mature miRNAs that are conserved in at least one animal species.

---

\*These authors contributed equally to this research, and are listed by the order of their last names.

†Corresponding author: zhang@cse.wustl.edu, phone: (314)935-8788, fax: (314)935-7302.

# 1 Introduction

MicroRNAs (miRNAs) are defined as endogenous single-stranded noncoding RNAs of 22 nt in length derived from long precursors that fold into hairpin structures [2, 11]. MicroRNAs are crucial in the development of animals and plants, and have been shown to play an important role in genetic diseases, including cancer. In animal, most miRNAs bind to 3' untranslated regions of their target mRNAs and repress the translation of the targets [2]. However, in plants, mature miRNAs directly base-pair with complementary sites in the coding regions of target mRNAs, resulting in the target mRNAs to be subsequently cleaved or degraded [2, 11].

Identification of novel miRNA genes is one of the most imminent problems towards the understanding of post-transcriptional gene regulation. Two major strategies for identifying novel miRNAs are experimentally cloning and in silico prediction [2, 3, 11]. In the cloning-based approaches, distinct 22nt RNA transcripts are first isolated and intensive cloned, and the sequenced. However, These methods are highly biased toward abundantly and/or ubiquitously expressed miRNAs. Only abundant miRNA genes can be easily detected [2, 3, 11]. Notably, not all miRNAs are well expressed in tissues, cell types, and developmental stages that have been sampled [2]. Novel miRNAs tend to be elusive, as they are expressed constitutively in low abundance or they have preferentially restrictive/specific temporal (cell-phase) and spatial (tissue-/cell-type) expression patterns. Furthermore, breakdown products of mRNA transcripts in the background and endogenous ncRNAs (e.g., tRNAs, rRNAs, nat-siRNAs) as well as exogenous siRNAs are dominant players coexisting in the small RNA samples isolated from the cytoplasmic total RNA extracts [12, 2, 3, 11].

Computational strategies have been developed to overcome technical hurdles for experimental methods based on expression screening. First, these methods have been shown to be efficient for finding miRNAs that are expressed constitutively low or tissue-specifically [2]. Moreover, to avoid erroneously designating other nocoding small RNAs and even broken mRNA fragments as novel miRNAs, the flanking genomic sequences of cloned small RNAs are further assessed computationally to check if these cloned small RNAs reside on one arm of the hairpin structures. Only those small RNA sequences occupying the 20nt matched regions on one arm of the hairpin precursors are likely to be miRNAs. However, due to their short length, cloned small RNAs may match to many genome regions that can potentially fold into hairpin structures, and the hairpin structures are not unique to miRNAs exclusively. Thus, genome-wide screening for novel miRNA precursors is technically complicated.

## 1.1 Related works

Up to date, Computational methods have been applied to *C. elegans* [7], *D. melanogaster* [13], *A. thaliana* [5], *Oryza sativa* [5], and *H. sapiens* [4] to identify candidate miRNAs. According their frameworks, these existing methods can be grouped into two categories.

Methods in the first category developed to find close homologs among related miRNAs. In these methods, known miRNA precursors were first folded into typical hairpin structures, local features in the hairpin were extracted, and extreme values of these featured were obtained over all known miRNAs. Then a filter was built to screen the novel hairpined sequences. Finally hairpined sequences passed the filter were further analyzed in related species to assess their evolutionary conservation. Similarity search based and profile-based methods built upon this framework were successful in detection of some miRNAs with conserved homologs in closely related species. In order to identify distant homologs as well as close homologs, a probabilistic co-learning method based on the paired hidden Markov model (HMM) was implemented as a more general miRNA prediction method. It

minimized the false-positive rate to as low as 4.00%, but compromised for a poorer performing sensitivity of only 73.00% [15].

In the second category, supervised machine-learning classification algorithms, e.g., support vector machines (SVMs), were trained on a binary-labeled positive set of genuine miRNA precursors and a negative set of hairpins obtained from exon regions of protein coding genes [9, 16, 18]. Through this inductive learning on their feature vectors, a classifier model and a set of decision rules are devised to discriminate between them. An SVM-based approach, RNAmicro, incorporating sequence and structural information as part of its feature vector, reported incredibly promising efficiencies of 91.16% and 99.47% for sensitivity and specificity, respectively [9].

## 1.2 Motivation and contributions

All methods in two categories discussed above require enough positive samples (known miRNAs). However, for many sequenced genomes, very few miRNAs have been reported. For instance, only 38 of miRNA have been identified in the genome of *A. gambiae*, which is studied in this work. Viruses have been shown to be capable of encoding miRNAs targeting their host genes involved in immune systems [17]. Although identifying potential miRNAs in viral genomes would be valuable for virologists to study virushost interactions, very few of virus miRNAs are reported. known virus miRNA barely share sequence homologies. It is difficult to predict them with available computational tools. Moreover, many sequenced genomes are poorly annotated. Thus, it is even more difficult to prepare negative training sets for the methods in the second category. Furthermore, incorrectly prepared negative training sets may also have harmful impact on the classification model. If the train set and test set are not subject to independent identical distribution (i.i.d), the trained model could be heavily biased with regard to the classification task, which means a poor performance of the model on the test set will be expected [6]. Restricted by all these limitations, available computational tools are not feasible to identify of miRNAs for species whose genome is not well annotated. Since methods in this category cross-species sequence conservation, it is impossible identification of miRNAs , especially for organisms whose closest relatives are partially sequenced or not sequenced yet.

In this work, we developed a novel ranking algorithm based on random walks to computationally identify novel miRNAs from genomes with very few known miRNAs and/or with poor annotations. Our algorithm needs very few labeled positive samples, and do not require labeled negative data and information of genome annotations. To the best of our knowledge, it is the first such algorithm of its kind. Tested on human data, our approach can identified known miRNAs with relatively high precision and recall. Apply this novel algorithms to *A. gambiae*, we predict many novel miRNA genes, some of which encode mature miRNAs conserved in other species.

## 2 Problem formulation

In this work, we view the miRNA prediction as a problem of information retrieval. Novel miRNAs will be retrieved by the known miRNAs (query samples) from the candidate pool. Specifically, we model this retrieval process by belief propagation on a weighted graph, and propose a novel ranking algorithm based on the random walks method.

For a particular species, the known miRNAs, the putative candidates and their relation can be modeled by a *graph*  $G = \langle V, E \rangle$ . In the graph, each vertex  $v \in V$  represents a datum (a known miRNA precursor or a putative candidate), each edge  $e \in E \subseteq V \times V$  represents the relation of the two vertices linked by it, and the *weight*  $w$  of the edge  $e$  further quantitatively measures the relation.

Generally the edge weights are determined by the pairwise data distances. Therefore, two closely related samples would have an edge with a large weight. The *degree*  $d_i$  of vertex  $v_i$  is  $d_i = \sum_j w_{ij}$ , i.e., the summation of weights of all edges that are connected to  $v_i$ .

We refer known miRNA precursors as query samples and putative candidates as unknown samples. Consider a set of query samples  $X_Q = \{x_{q1}, x_{q2}, \dots, x_{qn}\}$  and a set of unknown samples  $X_U = \{x_{u1}, x_{u2}, \dots, x_{um}\}$ , where  $x_{qi}, x_{uj} \in \mathbb{R}^d$ . We want to rank  $X_U$  with respect to  $X_Q$ . Specifically, we associate each sample  $x_i$  with a relevancy value  $f_i$ , where  $f_i = 1$  for all query samples and  $f_i \in [0, 1]$  for the unknown samples. A higher  $f_i$  value indicates a higher relevancy of  $x_i$  with respect to the queries. Then we sort the relevancy values of all unknown samples and select the top ranked samples as the retrieved samples, which are the predicted miRNA precursors in this work. Therefore, the key of the ranking algorithm is to precisely compute the relevancy value for each unknown sample. In this study, we apply the random walks method to solve the problem ( See sections 3).

There are many works on query by examples in information retrieval and machine learning field [1]. In [19], Zhou *et al* proposed a manifold ranking method, which is the most closely related to our approach. The method ranks the data with respect to the intrinsic manifold structure collectively revealed by the data. Different from their work, our method uses Markov random walks to model the belief propagation process and therefore has a more clear physical interpretation.

## 3 Method

### 3.1 Ranking based on random walks

Random walks is a classic stochastic model on a weighted finite states graph, which exploits the structure of data in a probabilistic way. In the random walks formulation, each sample is treated as a graph vertex, which corresponds to a state on a Markov chain. The one-step transition probability  $p_{ij}$  from  $v_i$  to  $v_j$  is defined by

$$p_{ij} = \frac{w_{ij}}{d_i}, \quad (1)$$

or written in a matrix form

$$P = D^{-1}W, \quad (2)$$

where  $W$  is the weight matrix,  $D$  is a diagonal matrix, whose  $i^{\text{th}}$  diagonal element is  $d_i$ . To facilitate our discussion, we reorder the vertices and arrange them into a vector so that the query samples come first and the unknown samples come last. We then partition the matrix accordingly as

$$P = \begin{pmatrix} P_{QQ} & P_{QU} \\ P_{UQ} & P_{UU} \end{pmatrix}, \quad (3)$$

where  $P_{QQ}$  is an  $n \times n$  transition matrix among the query states,  $P_{QU}$  is an  $n \times m$  transition matrix from the query states to the unknown states,  $P_{UQ}$  is a  $m \times n$  transition matrix from the unknown states to the query states,  $P_{UU}$  is a  $m \times m$  transition matrix among the unknown states. Correspondingly, we partition the weight matrix  $W$  and the degree matrix  $D$  as

$$W = \begin{pmatrix} W_{QQ} & W_{QU} \\ W_{UQ} & W_{UU} \end{pmatrix}, \quad D = \begin{pmatrix} D_{QQ} & O \\ O & D_{UU} \end{pmatrix}, \quad (4)$$

where  $O$  is a matrix with all 0 elements.

In our model, when a random walker transits from  $v_i$  to  $v_j$ , it will transmit the relevancy information of  $v_i$  to  $v_j$ . Specifically, this is a dynamic process of relevancy information propagation: at each

iteration step, a vertex transmits its relevancy information to its neighbors, and simultaneously it also receives the relevancy information from its neighbors. Then, at the end of the iteration step, it updates its own relevancy value by the received information. Note that the query vertices only transmit their relevancy information, but will not update their own relevancy values in the process. Furthermore, since query samples are more important than the unknown samples, they are assigned higher weights for their relevancy information in transmission. This suggests the following updating rule for the relevancy value of  $v_i$ ,

$$f_i^{(k+1)} = \alpha \sum_{x_j \in X_U} p_{ij} f_j^{(k)} + \sum_{x_j \in X_Q} p_{ij} f_j, \quad (5)$$

where  $k$  is the number of iteration,  $\alpha \in (0, 1)$  is the weight of the relevancy information from the unknown samples. We can also write (5) in a matrix form as

$$\mathbf{f}_U^{(k+1)} = \alpha P_{UU} \mathbf{f}_U^{(k)} + P_{UQ} \mathbf{f}_Q. \quad (6)$$

By plugging in (2), we have

$$\mathbf{f}_U^{(k+1)} = \alpha D_{UU}^{-1} W_{UU} \mathbf{f}_U^{(k)} + D_{UU}^{-1} W_{UQ} \mathbf{f}_Q. \quad (7)$$

The convergence of the iteration is guaranteed by the following theorem.

**Theorem 1.** The sequence  $\{\mathbf{f}_U^{(k)}\}$  generated by the updating rule of (7) converges when  $k$  approaches infinity.

*Proof:* For convenience, let  $R_{ij} = \{\alpha D_{UU}^{-1} W_{UU}\}_{ij}$ ,  $\mathbb{D} = [0, 1]^m$ . Let  $\mathbf{b} = D_{UU}^{-1} W_{UQ} \mathbf{f}_Q$ ,  $\mathbf{f}_U^{(k)} = [f_{u1}^{(k)}, f_{u2}^{(k)}, \dots, f_{um}^{(k)}]^T \in \mathbb{D}$ . Then (7) can be written as

$$\mathbf{f}_U^{(k+1)} = R \mathbf{f}_U^{(k)} + \mathbf{b}. \quad (8)$$

We define a map  $T : \mathbb{D} \rightarrow \mathbb{D}$  and the measure  $d$  on it as

$$T(X) = RX + \mathbf{b}, \quad (9)$$

$$d(X_1, X_2) = \max_i |x_{1i} - x_{2i}|, \quad (X_1, X_2 \in \mathbb{D}), \quad (10)$$

where  $x_i$  is the  $i^{\text{th}}$  element of the vector  $X$ . It is easy to show that  $(\mathbb{D}, d)$  is a complete metric space.

According to the *Contractive Mapping Theorem of Banach*, it is suffice to prove that  $T(\mathbf{f})$  is a contractive mapping, which holds if  $R$  satisfies

$$r_i = \sum_j R_{ij} < 1, \quad (i = 1, 2, \dots, m). \quad (11)$$

To prove (11), recall that  $0 < \alpha < 1$ , and

$$D_{UU} = \text{diag}(W_{UQ} \times \mathbf{1}_n + W_{UU} \times \mathbf{1}_m) \quad (12)$$

where  $\mathbf{1}_n = [1, 1, \dots, 1]^T$  is an  $n$ -dimensional vector whose elements are all 1. It follows that

$$\begin{aligned} r_i &= [R_{i1}, R_{i2}, \dots, R_{im}] \times \mathbf{1}_m \\ &< [\dots, (D_{UU})_{ii}^{-1}, \dots] \times W_{UU} \times \mathbf{1}_m \\ &= \frac{1}{(W_{UQ})_{i,\cdot} \times \mathbf{1}_m + (W_{UU})_{i,\cdot} \times \mathbf{1}_m} \cdot (W_{UU})_{i,\cdot} \times \mathbf{1}_m \\ &< 1 \end{aligned} \quad (13)$$

Therefore,  $T(\mathbf{f})$  is a contractive mapping, and (7) converges to a unique fixed point.  $\square$

Therefore, we can safely substitute the limit of  $\mathbf{f}_U^{(k)}$  and  $\mathbf{f}_U^{(k+1)}$  by  $\mathbf{f}_U$ , which produces

$$\mathbf{f}_U = \alpha D_{UU}^{-1} W_{UU} \mathbf{f}_U + D_{UU}^{-1} W_{UQ} \mathbf{f}_Q. \quad (14)$$

which is equivalent to

$$(I - \alpha D_{UU}^{-1} W_{UU}) \mathbf{f}_U = D_{UU}^{-1} W_{UQ} \mathbf{f}_Q. \quad (15)$$

where  $I$  is an  $m \times m$  identical matrix. Therefore,

$$\mathbf{f}_U = (D_{UU} - \alpha W_{UU})^{-1} W_{UQ} \mathbf{f}_Q. \quad (16)$$

## 3.2 Algorithm

The formula (16) shows that we can compute the relevancy value of each unknown state without actually perform the iteration steps. Therefore, we have the following algorithm for ranking, which runs as follows:

- **Step 1:** Construct the graph  $G = \langle V, E \rangle$ .  
For each pair of  $x_i$  and  $x_j$ , we put an edge between them if they are within  $k$  nearest neighbors of each other.
- **Step 2:** Measure the graph weights  $W$ .  
Here heat kernel is adopted, i.e., if  $x_i$  and  $x_j$  are connected by an edge, then the weight of the edge is defined as

$$W_{ij} = \exp\{-d(x_i, x_j)^2 / \sigma\},$$

where  $d(\cdot, \cdot)$  is a distance measure defined on the graph,  $\sigma$  is the heat kernel parameter.

- **Step 3:** Solve the matrix problem in equation (16).
- **Step 4:** Rank the samples. Sort the relevancy values of all samples, and select the top ranked samples as the retrieved results.

In the miRNA prediction, each sample is represented by a vector of 36 features (See Section 3.3), and distance between two samples is the Euclidean distance of the feature vectors. The extraction of features for miRNA precursors is further discussed following.

## 3.3 Extraction of the global and local sequence-structure features

The most salient characteristics of miRNAs precursors is their hairpined secondary structure. Recent reports have shown that local sequence features are important in miRNAs precursors. In this study, the RNA secondary structures are predicted by RNAfold [10]. We postulate that the entire hairpin-shaped structure of each miRNA precursor can be characterized solely by a feature vector  $x_i$  containing 36 global and local intrinsic attributes at the sequence, structural, and topological levels.

Four important global features at the structural and topological levels are the normalized minimum free energy of folding (MFE), the normalized base-pairing propensity of each arm, the normalized loop length. Here the normalization factor is the length of the precursor sequence.



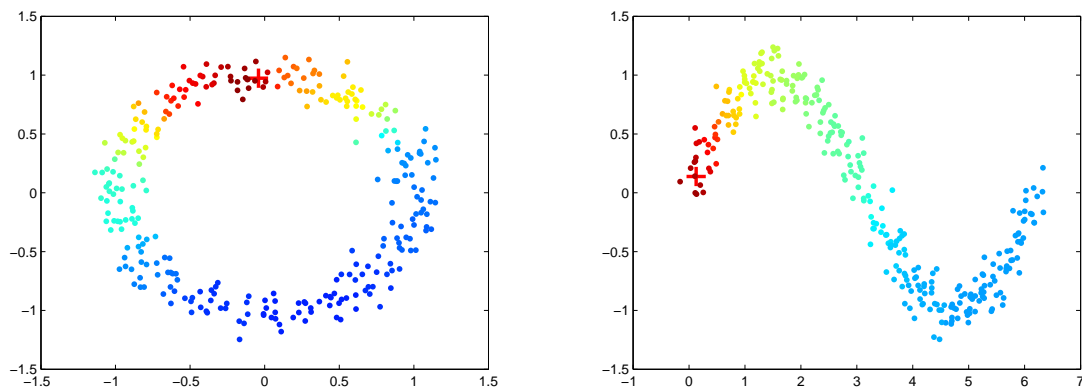


Figure 2: Two toy examples of querying using Markov random walks ranking method.

## 4.2 Datasets for miRNA predictions

All reported miRNAs precursor sequences of *H. sapiens* and *A. gambiae* (533 and 38, respectively) were downloaded from the miRBase (<http://microrna.sanger.ac.uk/sequences/>) [8] as of September 1, 2007.

Genome sequences of *H. sapiens* and *A. gambiae* were downloaded from from UCSC Genome Browser (<http://genome.ucsc.edu/>).

For the *H. sapiens* genome, We randomly extracted non-overlapping fragments of 90nt long from *H. sapiens* genome. If the fragments do not have any overlaps with known *H. sapiens* miRNA precursors, we further predicted their secondary structures using RNAfold [10]. The criteria for selecting the fragments with hairpin structures are as follows: minimum of 18 base pairings on the stem of the hairpin structure, maximum of -12 kcal/mol free energy of the secondary structure, and no multiple loops. The thresholds 18 and -12 are the lowest number of base pairings and the highest free energy among all the genuine animal miRNA precursors. 1,000 of such fragments were used in this study as negative data.

Each chromosome of *A. gambiae* was fragmented, from 5'-end to 3'-end, using a sliding window of 90nt, with an increment of 45nt. The secondary structures of these fragments were predicted by RNAfold [10], and hairpined fragments were selected the same criteria described above. The the selected hairpin sequences form our candidate pool. In the fragmentation, some putative candidates may be cut into two pieces, and lose the hairpin structures, hence are missed in our candidate pool. To avoid this, we further fragmented, with the same sliding window and increment, the sequences between each pair of fragments that are next to each other. The secondary structures of the new set of fragments were predicted and selected by the same tool and criteria. This process was iterated until no more hairpined fragments was found. Eventually we totally obtained 22,297 hairpin sequences, which included the 38 known miRNAs precursors.

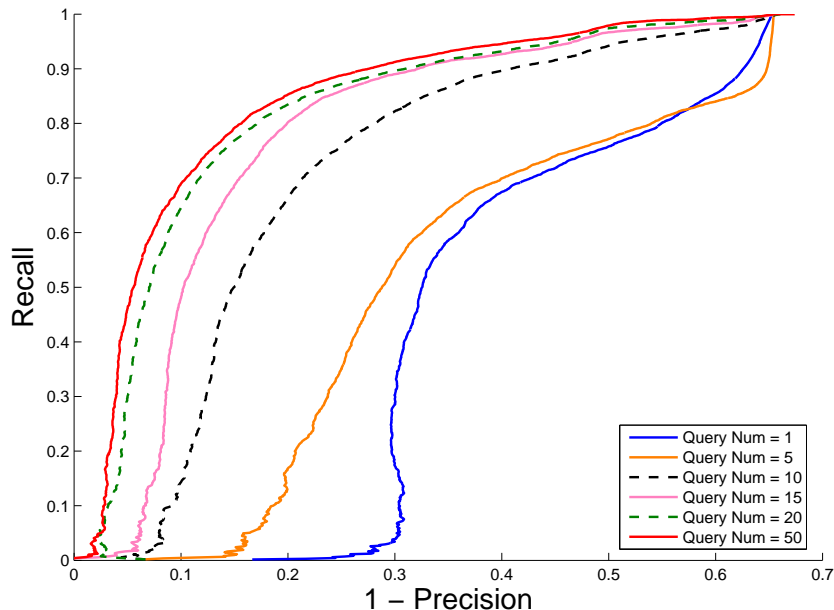


Figure 3: Precision-recall curve obtained by setting different number of queries.

### 4.3 Experiment results on a set of human data

We first evaluate the performance of our algorithm on *H. sapiens* data (see Section 4.2). The prediction performance is assessed by the recall and the precision, which are, respectively, defined as:

$$recall = \frac{TP}{TP + FN} \quad (17)$$

$$precision = \frac{TP}{TP + FP} \quad (18)$$

The number of query samples is the most important parameter in the algorithm. With the parameter  $k$  fixed to 12,  $\alpha$  fixed to 0.995 and  $\sigma$  fixed to 0.15, we perform six experiments on *H. sapiens* data with 1, 5, 10, 15, 20 and 50 known miRNA precursors as query samples, and combine the rest known ones with the 1,000 hairpin sequences randomly extracted from the genome to form the candidate pool. We treat all the 1,000 hairpin sequences as negative samples although it is possible that some of them could be the real miRNA precursors. In each experiments, we sequential choose  $n$  topmost ranked samples, then compute the precision and recall of the retrieval samples. The results are shown in Figure 3.

In the extreme case of querying with only one sample, we still can approach a more than 70% precision in a few of the topmost retrieval samples. This result shows that our algorithm is very promising to predict miRNAs for some virus species with only one or two identified miRNAs. When we query with more than 20 samples, we can obtain a high precision of over 95%. This is important in miRNA prediction for some genomes with very few reported miRNAs and poor annotation information. In the research of miRNA, a low false positive rate is much more desirable. It worths pointing out that the precisions in these experiments is still under-estimated due to the possible true miRNA precursors in the 1,000 hairpin sequences extracted from *H. sapiens* genome.

Obviously, samples with higher ranks are most likely to be the true miRNAs, which can be observed in Figure 3. These samples are of most confidence for further experimental validation.

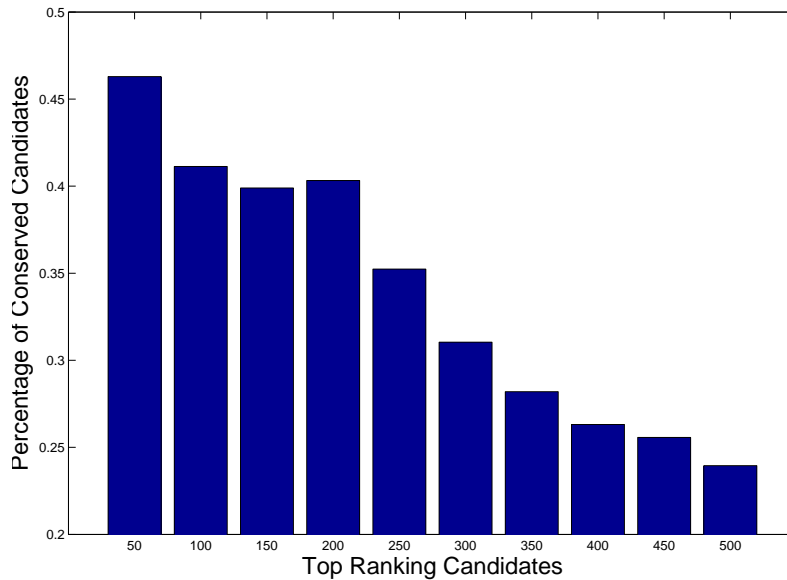


Figure 4: Percentage of conserved miRNAs in different sets of putative candidates. X axis shows the sets of putative candidates with different ranks. For instance, '200' means that candidates in this set rank from 151 to 200.

#### 4.4 Novel miRNA genes identified in *A. gambiae*

*A. gambiae* is one of the most important vectors of malaria in Africa, and one of the most efficient malaria vectors in the world. Identification of miRNAs is very important to study the development of *A. gambiae*, hence may broaden our perspectives on the control of the prevalence of malaria [14]. Up to date, only 38 miRNAs are reported and curated in miRBase [8]. As we have discussed in Section 1.2, available tools are not feasible to predict miRNAs in large scale for *A. gambiae*.

We take the 38 miRNAs curated in miRBase [8] as the query samples, and apply our algorithm to 22,259 *A. gambiae* candidate sequences with hairpined secondary structures (See Section 4.2). Obviously, samples with higher ranks are most likely to be the true miRNAs, which can be observed in Figure 3 from the experiments on *H. sapiens* data. These samples are of most confidence for further experimental validation. We take the 200 with the topmost ranks as our predictions. Mature miRNAs of 76 predicted miRNA precursors are conserved in at least one of the animal species.

Interspecies conservation has been widely applied to predict of miRNA genes in both animal and plant species. For many reported miRNAs, there are three major observations on their conservations: First, the mature miRNA sequences are conserved, whereas the rest of the precursor sequences have diverged. Second, the ability of the precursor sequences to form a hairpined secondary structure is conserved, although they have diverged. Third, the conserved mature miRNA homologs are mostly located on the same arm of the haripined secondary structures. We have applied to the hairpin structure as the most significant filter to extract the 22,259 candidates. We further analyzed the conservation of mature miRNAs of our top candidates. Figure 4 shows the distributions of conserved mature miRNAs in the top 500 candidates. Evidently, the set of candidates with the higher ranks include more conserved miRNAs. Figure 5 shows two novel putative miRNAs that we predict. According to the IDs of their homologs in other species, these two miRNAs are named as aga-mir-135, and aga-mir-49, respectively. Note that all of the homologs are located on the same arms of the hairpins in corresponding species.

```

mmu-mir-135b CGCUCUGCUGUGGCCUAUGGCUUUUCAUUCUUAUGUGAUUGCUGCUCGG
AACUCAUGUAGGGCUAAAAGCCAUGGGCUACAGUGAGGGGCAAGCUCC
rno-mir-135b CGCUCUGCUGUGGCCUAUGGCUUUUCAUUCUUAUGUGAUUGCUGUUCGG
AACUCAUGUAGGGCUAAAAGCCAUGGGCUACAGUGAGGGGCAAGCUCC
mdo-mir-135b AGCUCUCUGCUGUGGCCUAUGGCUUUUCAUUCUUAUGUGAUUGCUGUUCGG
AACUCAUGUAGGGCUAAAAGCCAUGGGCUACAGGGAGGGGAGAGCCUCC
hsa-mir-135b CACUCUGCUGUGGCCUAUGGCUUUUCAUUCUUAUGUGAUUGCUGUCCAA
ACUCAUGUAGGGCUAAAAGCCAUGGGCUACAGUGAGGGGCGAGCUCC
aga-mir-135 GGAGUAUGGCUUUUCAUUCUUAUGUGUAAACUCAUCAUA
UGAUAAAGGUGAUGAUGGUCGGUUUGCCAGCGAUUCAAGUGGACAACAUUACUUC

cel-mir-49 UUUUGAAAAAGACCACCGUCCGCAGUUUGUUGUGAUGUGCUCCAAGCAAUCAUGA
GUCUGAAGCACCACGAGAAGCUGCAGAUUGGAGGUUCUGAUUU
cbr-mir-49 ACCGAAACCAUUUGCCAUCGCGAGUUUUUGUAGUGUGCUCCGCGCCAUCUUGA
GCCCCAAGCACCACGAGAAGCUGCAGAUUGAAGUUUUGGUU
aga-mir-49 UCUGCAGUGUGUGUUUGUGUGUGAAAAGCUAACAUUAGUCAUUUGGUCU
UUUUGCCAACGAAGCACCACGAGAAGCUGCAGA

```

Figure 5: Two examples of novel miRNA identified in *A. gambiae* and conserved in more than one other animal species. According to the IDs of their homologs in other species, we name them aga-mir-135, adn aga-mir-49, respectively.

## 5 Conclusions and final remarks

In this study, we treat the miRNA prediction as a problem of information retrieval—retrieve novel miRNAs by the known miRNAs (query samples) from genome-scale candidate pools. We model the novel miRNA retrieval process by belief propagation on a weighted graph constructed from known miRNAs and all items in the candidate pool. We then propose a novel ranking algorithm based on the random walks method.

Our method has the following striking advantage. First it does not require any negative sample and annotation information for the genome of interest. Secondly, it does not rely on cross-species conservation. Moreover, it can approach very high prediction accuracy by using very few known miRNAs, for instance, one known miRNA in the extreme case.

Genomes of many species have been sequenced, but their annotation is far from complete and genomes of their close relative species have not been sequenced yet. Thus large number of false positive candidates with hairpinned secondary structures can not be filtered out based on genome annotation or by phylogenetic conservation. Moreover, many species including viruses have only very few reported miRNAs. Existing methods are not feasible in all these situations. Taking all the advantages of our approach, we can apply it to solve all these problems. In the tests on human data, we can obtain a prediction accuracy of more than 95% relying on very few known miRNAs as labeled samples. We further conduct a study on *A. gambiae* with our approach. 200 novel miRNA precursors are predicted. 78 of the putative miRNA precursors encode mature miRNAs that are conserved in at least one animal species.

## References

- [1] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, May 1999.
- [2] D.P. Bartel. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, 116(2):281–97, Jan 2004.
- [3] E. Berezikov, E. Cuppen, and R.H. Plasterk. Approaches to microRNA discovery. *Nat Genet*, 38 Suppl:S2–7, Jun 2006.
- [4] E. Berezikov, V. Guryev, J. van de Belt, E. Wienholds, R.H. Plasterk, and E. Cuppen. Phylogenetic shadowing and computational identification of human microRNA genes. *Cell*, 120(1):21–4, Jan 2005.
- [5] E. Bonnet, J. Wuyts, P. Rouze, and Y. Van de Peer. Detection of 91 potential conserved plant microRNAs in *Arabidopsis thaliana* and *Oryza sativa* identifies important target genes. *Proc Natl Acad Sci U S A*, 101(31):11511–6, Aug 2004.
- [6] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2000.
- [7] Y. Grad, J. Aach, G.D. Hayes, B.J. Reinhart, G.M. Church, G. Ruvkun, and J. Kim. Computational and experimental identification of *C. elegans* microRNAs. *Mol Cell*, 11(5):1253–63, May 2003.
- [8] S. Griffiths-Jones, R.J. Grocock, S. van Dongen, A. Bateman, and A.J. Enright. miR-Base: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res*, 34(Database issue):D140–4, Jan 2006.
- [9] J. Hertel and P.F. Stadler. Hairpins in a Haystack: recognizing microRNA precursors in comparative genomics data. *Bioinformatics*, 22(14):e197–202, Jul 2006.
- [10] I.L. Hofacker. Vienna RNA secondary structure server. *Nucleic Acids Res*, 31(13):3429–31, Jul 2003.
- [11] M.W. Jones-Rhoades, D.P. Bartel, and B. Bartel. MicroRNAs and their regulatory roles in plants. *Annu Rev Plant Biol*, 57:19–53, 2006.
- [12] M. Lagos-Quintana, R. Rauhut, W. Lendeckel, and T. Tuschl. Identification of novel genes coding for small expressed RNAs. *Science*, 294(5543):853–8, Oct 2001.
- [13] E.C. Lai, P. Tomancak, R.W. Williams, and G.M. Rubin. Computational identification of *Drosophila* microRNA genes. *Genome Biol*, 4(7):R42, 2003.
- [14] A. Moffett, N. Shackelford, and S. Sarkar. Malaria in Africa: vector species’ niche models and relative risk maps. *PLoS ONE*, 2(9):e824, 2007.
- [15] J.W. Nam, K.R. Shin, J. Han, Y. Lee, V.N. Kim, and B.T. Zhang. Human microRNA prediction through a probabilistic co-learning model of sequence and structure. *Nucleic Acids Res*, 33(11):3570–81, 2005.

- [16] K.L. Ng and S.K. Mishra. De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. *Bioinformatics*, 23(11):1321–30, Jun 2007.
- [17] N. Stern-Ginossar, N. Elefant, A. Zimmermann, D.G. Wolf, N. Saleh, M. Biton, E. Horwitz, Z. Prokocimer, M. Prichard, G. Hahn, D. Goldman-Wohl, C. Greenfield, S. Yagel, H. Hengel, Y. Altuvia, H. Margalit, and O. Mandelboim. Host immune system gene targeting by a viral miRNA. *Science*, 317(5836):376–81, Jul 2007.
- [18] C. Xue, F. Li, T. He, G.P. Liu, Y. Li, and X. Zhang. Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics*, 6:310, 2005.
- [19] Dengyong Zhou, Jason Weston, Arthur Gretton, Olivier Bousquet, and Bernhard Schölkopf. Ranking on data manifolds.