MicroRNA prediction with a novel ranking algorithm based on random walks

Yunpen Xu, Xuefeng Zhou, and Weixiong Zhang

Critique (MicroRNA Folding prediction algorithms)
Student: Ilya Minkin

1. Method outline

The paper introduces a novel method of identifying of micro RNA (miRNA) sequences in silico. The method is called miRank. Introduction section describes previously developed methods of predicting miRNA. These methods include:

1) Identifying miRNA that are conserved across set of closely related species. Authors note that this approach is limited, since there are specie-specific or lineage-specific miRNA
2) Supervised classification (usually SVM-based) methods. They require meaningful sets of both positive and negative examples. Since genomes annotations usually do not describe non-coding RNA sequences well, obtaining set of negative examples can be difficult

Proposed method is based on random walks on a weighted graph and takes following as input:

1) A set of known micro RNA sequences, called queries
2) A set of unknown putative candidates, called unknown samples

Each sequence is represented as a vector of real numbers, called features. Since RNA folding plays important role in development properties of an RNA molecule, these features are based on predicted secondary structure of RNA. They include normalized minimum free energy of folding (MFE), number of paired base pairs in the hairpin loop and others (totally 36 features). Each sequence is also assigned a real value f called *belief*, all query sequences are assigned $f_i = 1$ and unknown samples are assigned $f_i \in [0, 1]$. The random walker transits from vertex v_i to v_j and carries portion of f_{v_i} to v_j proportional to weight of the edge between v_i and v_j. More formally, the algorithm works as follows:

1) Construct graph $G = (V, E)$, where vertex set is a set of all input sequences. Two vertices are connected with an edge if they are *close* to each other
2) Assign weights to edges (it is proportional to Euclidean distance between samples)
3) Compute limit values of the belief vector by solving a matrix equation
4) Select n unknown samples with the highest belief – they are predicted miRNA sequences

2. Results

Authors of the paper evaluated their algorithm on two datasets: *H.sapiens* and *A.gambiae*. First they randomly extracted 1000 sequences of fixed length from human genomes that do not overlap with known miRNA precursors and treated them as negative examples. Then they added some known miRNA sequences to this set and ran their program with different number of query samples to determine precision and recall of the method. They obtained precision $> 70\%$ for 1 query sample, and $> 95\%$ for 20 query sequences. They also compared their program with SVM-based model and claim that miRank outperforms SVM with smaller number of positive samples < 5, although for larger number of positive samples performance is very similar.

They also tried to predict novel miRNA sequences in *A.gambiae*. With 38 known query samples they predicted 200 novel miRNA sequences and 78 of them are homologous to some miRNA in other animal genomes.

3. Critique

Biological motivation of the work is well described – it is known that micro RNA plays vital role in gene regulation.

There is an unclear moment in description of the algorithm. Authors say that two vertices in the graph are connected iff they are "close to each other". It raises a natural question – in which case two samples are close to each other? Is this information should be available a priori or two samples are close to each other if weight of the edge between them is larger than some threshold?

It is worth to note the source code is not freely available online, it is available "by request". Also it is written by using commercial MATLAB software with the bioinformatics toolbox, and a recently published paper [2] notes that for this reason they were unable to test miRank.

Instead of comparing miRank with previously developed software, authors implemented SVM-based classifier that uses the same set of features. It is mentioned in the paper that miRank outperforms SVM-based model for in case where number of known positive samples is low. However it is well known that performance of SVM models heavily relies on chosen model parameters [3]. But in the paper authors doesn't describe the strategy of choosing parameters. Is it possible to choose parameters so that SVM outperforms in all cases?

Authors note that miRank doesn't require set of known negative samples and hence doesn't rely on the genome annotation to filter out false miRNA precursors. Of course, it is very important advantage of the algorithm, although it may be not completely true. At it's last stage the algorithm chooses some highest-ranked samples to output. How many samples should be chosen? It is important to choose appropriate cutoff for belief, so that samples with belief lower some constant is filtered out. They used precision-score plots to infer inflection points and selected inflections as cutoff points. Cutoffs were selected separately for each experiment. But to obtain precision-score plot one would need a proper set of known negative samples. So it is not clear how to choose appropriate cutoff without negative samples.

4. Conclusion

The paper introduces novel algorithm for predicting miRNA sequences. It requires only positive samples for it's work and doesn't rely on genome annotation. It was shown that it outperforms SVM-based models when number of known miRNA samples is low (< 5), although in other cases performance is comparable. Hence the method can be successfully used on poorly annotated genomes. But the are some issues that can make it difficult to use – it is not clear how to choose appropriate cutoff without negative samples and it is difficult to access the source code.

References

[1] Yunpen Xu, Xuefeng Zhou, and Weixiong Zhang. MicroRNA prediction with a novel ranking algorithm based on random walks. Bioinformatics (2008) 24 (13): i50-i58.

[2] MiRPara: a SVM-based software tool for prediction of most probable microRNA coding regions in genome scale sequences. Yonggan Wu, Bo Wei, Haizhou Liu, Tianxian Li, and Simon Rayner. BMC Bioinformatics 2011, 12:107.

[3] Parameter selection for support vector machines. Carl Staelin. HPL-2002-354 (R.1) November 10th , 2003.

[4] Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. Chenghai Xue, Fei Li, Tao He, Guo-Ping Liu, Yanda Li, and Xuegong Zhang. BMC Bioinformatics. 2005; 6: 310.