miRNA Discovery & Prediction Algorithms

Sergei Lebedev

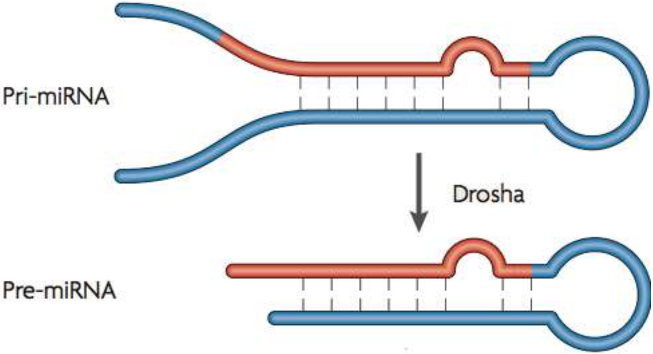
October 13, 2012

What is miRNA?

- microRNA or miRNA, ≈ 22 nucleotide-long **non-coding** RNA;
- mostly expressed in a tissue-specific manner and play crucial roles in cell proliferation, apoptosis and differentiation during cell development;
- thought to be involved in post-transcriptional control in plants and animals;
- linked to disease[1], for example *hsa-miR-126* is associated with retinoblastoma, breast cancer, lung cancer, kidney cancer, asthma etc.

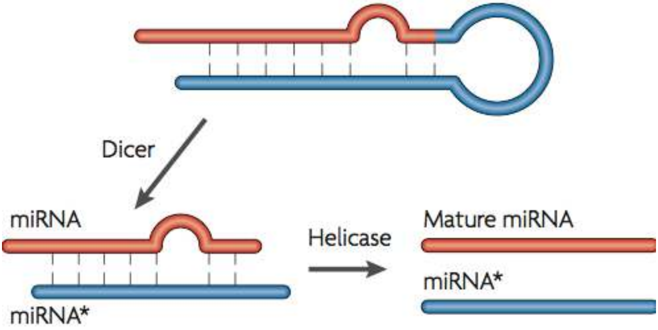
[1] See http://www.mir2disease.org for details.

miRNA in action: nucleus [1]



- **pri-miRNA** is transcribed by RNA polymerase II and seem to possess promoter and enchancer regions, similar to protein coding genes;
- pri-miRNA is then cleaved into (possibly multiple) **pre-miRNA** by an enzyme complex *Drosha*.

miRNA in action: cytoplasm [1]



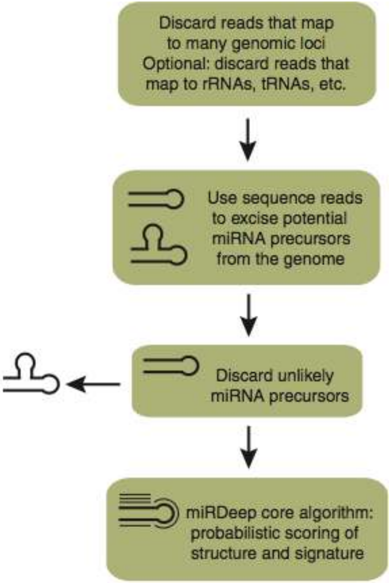
- *Dicer* removes the stem-loop, leaving two complementary sequences: miRNA and miRNA*, the latter is not known to have any regulatory function.
- Mature miRNA base-pairs with 3' UTR of target mRNAs and blocks protein syntesis or causes mRNA degradation.

miRNA identification

- Biological methods: northern blots, qRT-PCR[2], micro arrays, RNA-seq or miRNA-seq.
- Bioinformatics to the rescue! the usual strategy: first sequence everything, RNA-seq in this case, then try to make sense of whatever the result is.
- In this talk: miRDeep [2], MiRAlign [3], MiRank [4].
- A lot of existing tools out of scope, most can be described with a one liner: *"We've developed a novel method for miRNA identification, based on machine learning approach, SVM, HMM!"*.

[2] RT for reverse transcription, not real-time.

mirDeep



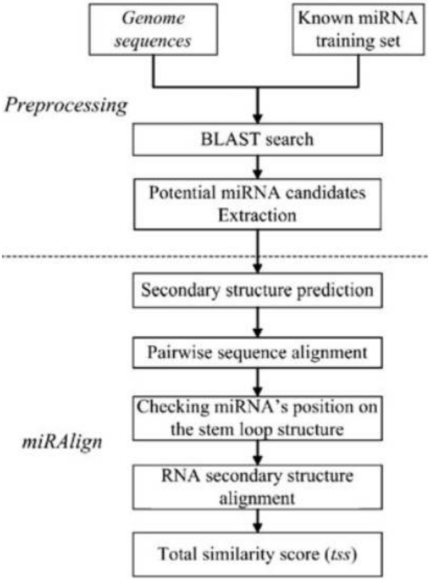
Discard reads that map to many genomic loci Optional: discard reads that map to rRNAs, tRNAs, etc.

Use sequence reads to excise potential miRNA precursors from the genome

Discard unlikely miRNA precursors

miRDeep core algorithm: probabilistic scoring of structure and signature

MiRAlign


Genome sequences
Known miRNA training set
Preprocessing
BLAST search
Potential miRNA candidates Extraction
Secondary structure prediction
Pairwise sequence alignment
Checking miRNA's position on the stem loop structure
miRAlign
RNA secondary structure alignment
Total similarity score (tss)

miRank: overview

- Treat miRNA identification problem as a problem of information retrieval, where novel miRNAs are to be retrieved from a set of candidates by the known query samples – “true” miRNAs.
- More formally, given a set of known pre-miRNAs X_Q as *query samples* and a set of putative candidates X_U as *unknown samples*, rank X_U with respect to X_Q.
- To do so, compute the relevancy values $f_i \in [0, 1]$ for all unknown samples, assuming $f_i = 1$ for query samples.
- After that, simply select n ranked samples, which constitute to predicted pre-miRNA.
- Makes sense, right?

miRank: how does it work?

- miRank models belief propagation process by doing Markov random walks on a graph, where each vertex corresponds to either known pre-miRNA or a putative candidate and two vertices are connected by an edge if the two vertices are *"close to each other"*.
- Each edge on the graph is assigned a weight w_{ij}, proportional to the Euclidean distance between the samples v_i and v_j (see next slide on how samples are represented).
- When a random walker transits from v_i to v_j it transmits the relevancy information of v_i to v_j by the following update rule:

$$f_i^{(k+1)} = \alpha \sum_{x_j \in X_U} p_{ij} f_j^{(k)} + \sum_{x_j \in X_Q} p_{ij} f_j \qquad p_{ij} = \frac{w_{ij}}{deg(v_{ij})}$$

miRank: features

Global

- normalized minimum free energy of folding (MFE);
- normalized no. of paired nucleotides on both arms;
- normalized loop length.

Local – RNAFold

```
GUAGCACUAAAGUGCUUAUAGUGCAGGUAGUGUUUAGUUAUCUACUGCAUUAUGAGCACUUAAAGUACUGC
((((.(((.((((((((((((((((.(((((......)).))))))))))))))))))))..))).))))
```

- Each nucleotide is either paired, denoted by a bracket (– 5' arm,)– 3' arm, or unpaired – .;
- Each local feature is a "word" of length 3, further distinguished by the nucleotide in the middle position, examples: **((.**, .((.

miRank: good parts, bad parts & magic

- The method doesn't require any genomic annotations, except for the set of query samples.
- $\approx 75\%$ precision and $\approx 70\%$ recall even with **very** few query samples (1, 5) – hard to validate, because the source code was never released.
- The notion of *similarity* between query samples, which defines the graph structure is unclear, even though it looks critical for algorithm performance.
- Two user-specified parameters, n – number of predicted samples and α – the weight of unknown samples in the relevancy value. How do they affect precision-recall and how to choose them?
- Overall, it seems like miRank isn't used much by biologists[3].

[3] http://www.ncbi.nlm.nih.gov/pubmed?linkname=pubmed_pubmed_citedin&from_uid=18586744

References

K. Chen and N. Rajewsky.
The evolution of gene regulation by transcription factors and microRNAs.
Nat. Rev. Genet., 8(2):93–103, Feb 2007.

M. R. Friedlander, W. Chen, C. Adamidi, J. Maaskola, R. Einspanier, S. Knespel, and N. Rajewsky.
Discovering microRNAs from deep sequencing data using miRDeep.
Nat. Biotechnol., 26(4):407–415, Apr 2008.

X. Wang, J. Zhang, F. Li, J. Gu, T. He, X. Zhang, and Y. Li.
MicroRNA identification based on sequence and structure alignment.
Bioinformatics, 21(18):3610–3614, Sep 2005.

Y. Xu, X. Zhou, and W. Zhang.
MicroRNA prediction with a novel ranking algorithm based on random walks.
Bioinformatics, 24(13):i50–58, Jul 2008.