Validation of a context analysis method for microRNA data

Stefano Rovetta
Computer and Information Sciences Department, University of Genoa, Italy

Francesco Masulli
Computer and Information Sciences Department, University of Genoa, Italy
and Sbarro Institute for Cancer Research and Molecular Medicine, Temple university, Philadelphia, USA

Giuseppe Russo
Sbarro Institute for Cancer Research and Molecular Medicine, Temple university, Philadelphia, USA

- Knowledge about miRNAs is still incomplete, and prediction is mainly computational.
- A context analysis method was devised to infer possible missing information.
- Experimental (computational) validation was performed to test the method.
- The goal is both to focus lab testing on the most likely miRNA-gene interactions, and to suggest new possible interactions to explore.

Outline

- MicroRNA target prediction and validation
- The value of context in microRNA data analysis
- The data: A Rosetta stone
- The method: Goals and procedures
- Experimental validation
- Conclusions

MicroRNAs

- (miRNA)
- Short, non-coding RNA sequences (22-23 nt)
- Regulate gene expression by inhibiting transcription
- Target genes are specific (but many-to-many relationship)

- Central role in controlling physiological and pathological processes
- Examples:
- Hundreds of miRNAs in the brain, several tissue-specific; Neurogenesis;
- Several types of cancer

miRNA target prediction

- Genes are selectively targeted according to several match/mismatch criteria
- Prediction is mostly a computational task
- Many prediction methods / programs / web repositories
- Lab validation necessary, but indirect

- TargetScan; RNAhybrid; PicTar; miRanda; miRWalk ...
- Prediction programs often do not agree
- Often designed with different selectivity/specificity tradeoffs

Context in miRNA data analysis

- Observation: there are more genes than miRNAs (est. 1000)
- Observation: many genes are target to more than one miRNA
- Observation: many miRNAs target more than one gene

Context in miRNA data analysis

- Hypothesis: miRNAs may work in teams and act on whole pathways or pathway segments
- Evidence of condition-specific signatures in miRNA profiles

- Direct experimental evidence (e.g. Zhang Y et al. "Profiling of 95 microRNAs in pancreatic cancer cell lines and surgical specimens by real-time PCR analysis" World J Surg. 2009 Apr;33(4):698-709)
- Computational methods based on this hypothesis (a simple one in Albertini MC et al. "Predicting microRNA modulation in human prostate cancer using a simple String IDentifier (SID1.0)". J Biomed Inform. 2011 Aug;44(4):615-20).

The data

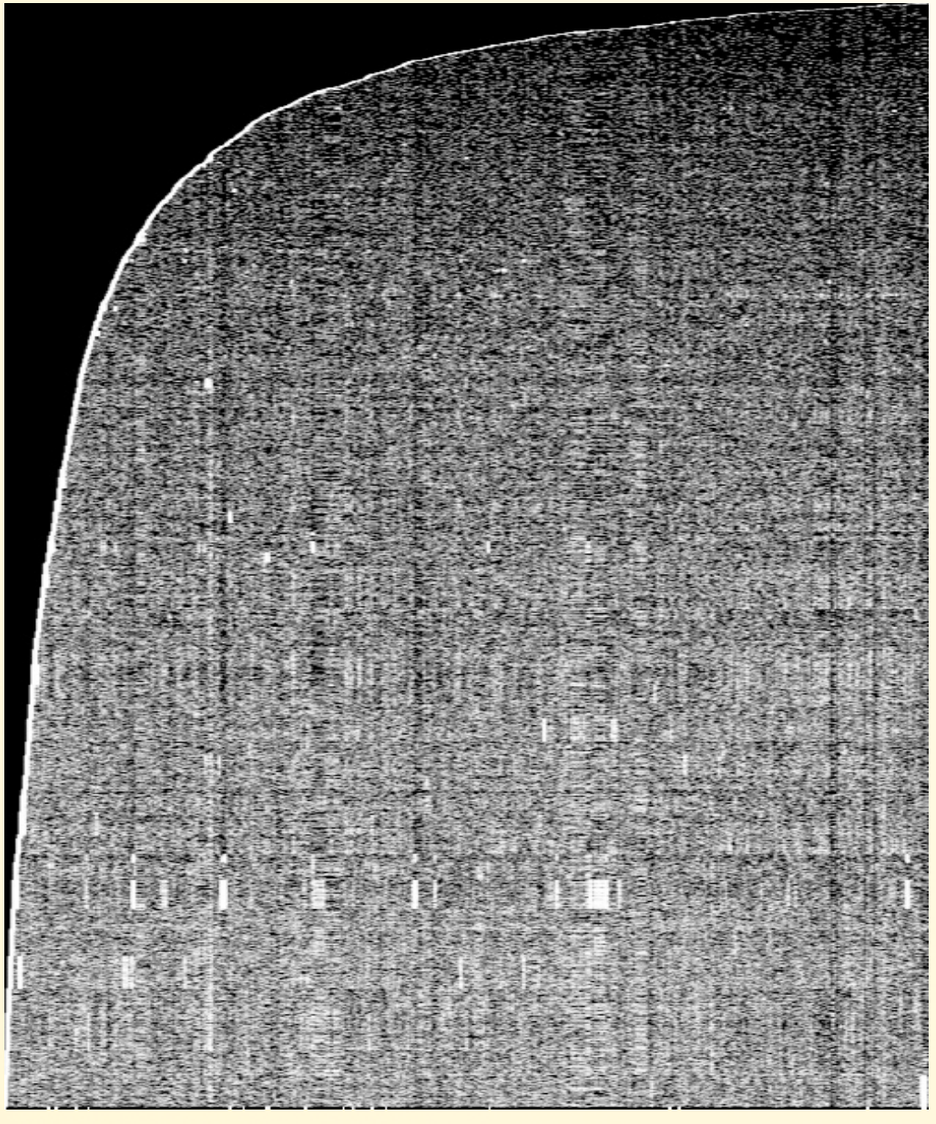
- Several data about miRNAs and their target from several repositories (miRBase, miRWalk, TarBase)
- Basic information: A matrix with

  miRNAs as columns

  gene transcripts as rows

  1 at the intersection between a miRNA and a gene transcript listed as a target for that miRNA; 0 otherwise

A visual representation of the data



677 miRNAs x 23683 gene transcripts

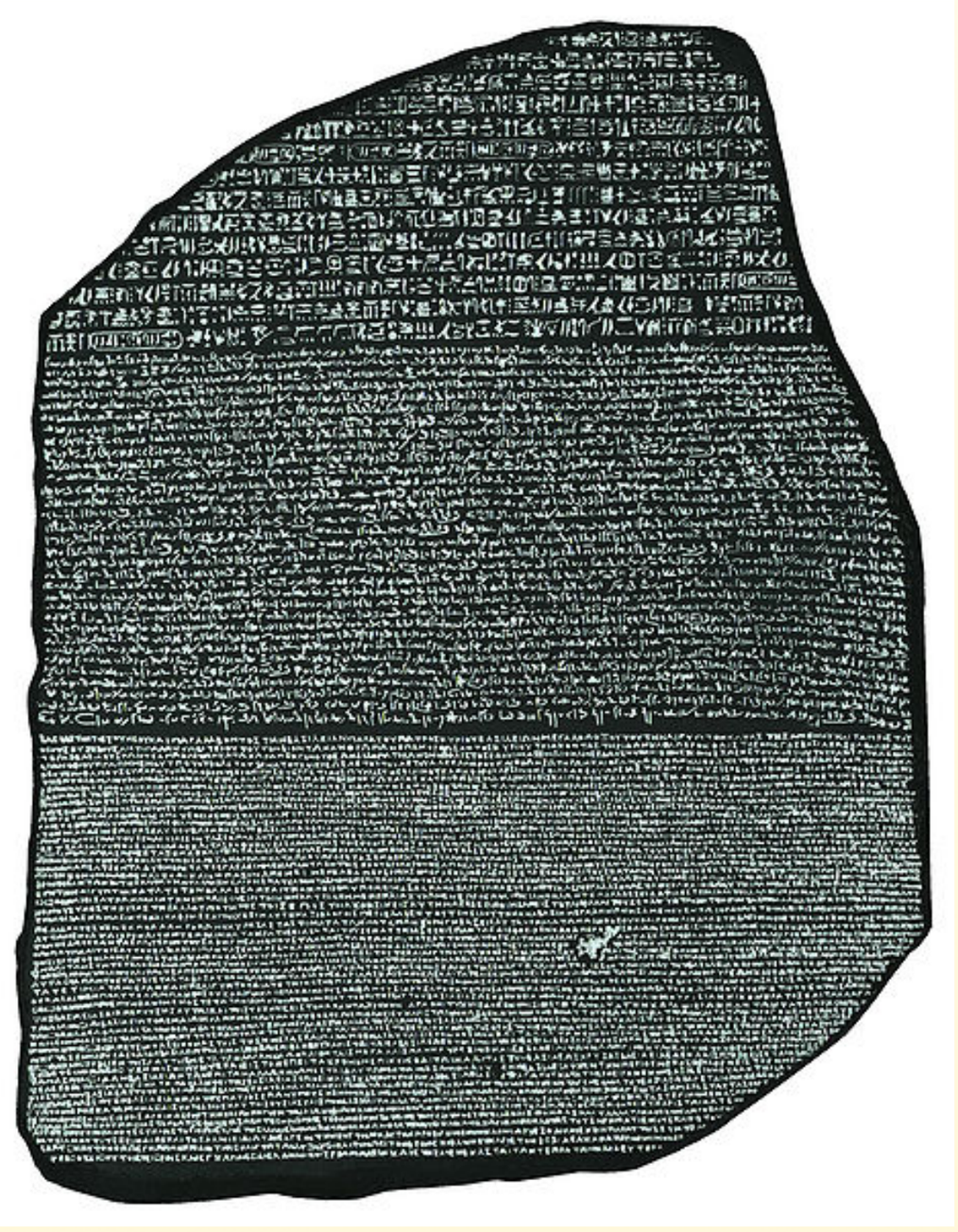
The meaning of 0

- 16 033 391 entries
- 487 409 entries with value 1 (about 3%)
- value 1 means "a match between this miRNA and this transcript has been found, so this gene is a target for this miRNA"
- value 0 means "a match between this miRNA and this transcript HAS NOT been found, so WE DON'T KNOW whether this gene is a target for this miRNA"

- 0 does not mean "no match", but "match not found"
- Either because not validated, or because not even tested

A Rosetta stone

- Hypothesis: patterns in the data may help suggesting when zeroes stand for "possible match but not tested yet"
- Parts of the matrix for which the meaning is known may help in decoding other parts



The method – setting up

Similar in spirit to previous "Rosetta stone" approaches

> Marcotte EM et al. "A combined algorithm for genome-wide prediction of protein function" Nature 1999 402: 83–86.

- Take a set of reference miRNAs known to be involved in a process of interest (e.g., prostate cancer)
- Take a query miRNA to investigate its possible involvement in the same or related processes
- We want to know whether the query miRNA may have targets among the genes targeted by the reference set even if this information is not recorded in our data set

The method – sorting out genes

- Define the set of all genes that are targets of the reference set
- Split this set in two:

  The MATCH subset –

  genes known to be targets of the query miRNA

  The NON-MATCH subset –

  the rest (not known to be targets)

N.B. For the sake of brevity here gene = transcript sequence

The method – decision-making

- Define a suitable similarity between genes
- Compute similarity between each gene in the NON-MATCH subset and each gene in the MATCH subset
- Take NON-MATCH genes with high average similarity as candidate targets for the query miRNA

Distances between genes

- Genes are rows in our data matrix
- Can be considered 677-dimensional vectors
- Similarity between two genes:

$$d(u,v) = u \cdot v$$

- Similarity between a gene and a set of genes (cumulative similarity):

$$cs(u,V) = \sum_{v \in V} \sqrt{d(u,v)}$$

Experimental validation

- 2 tests:
- Recovering matching genes from their own context

  One gene is removed from the list of targets for a miRNA

  Can it be recovered by analyzing the remaining ones?

  (here query = reference)
- Inferring matching genes from external context

  Select query and reference

  One gene is removed from the MATCH subset

  Can it be recovered by analyzing the rest?

Experiment 1: some details

- Take a miRNA
- Delete one gene from its targets
- Compute cs between the gene and the remaining ones

Experiment 1: results

- All genes have $cs > 0$

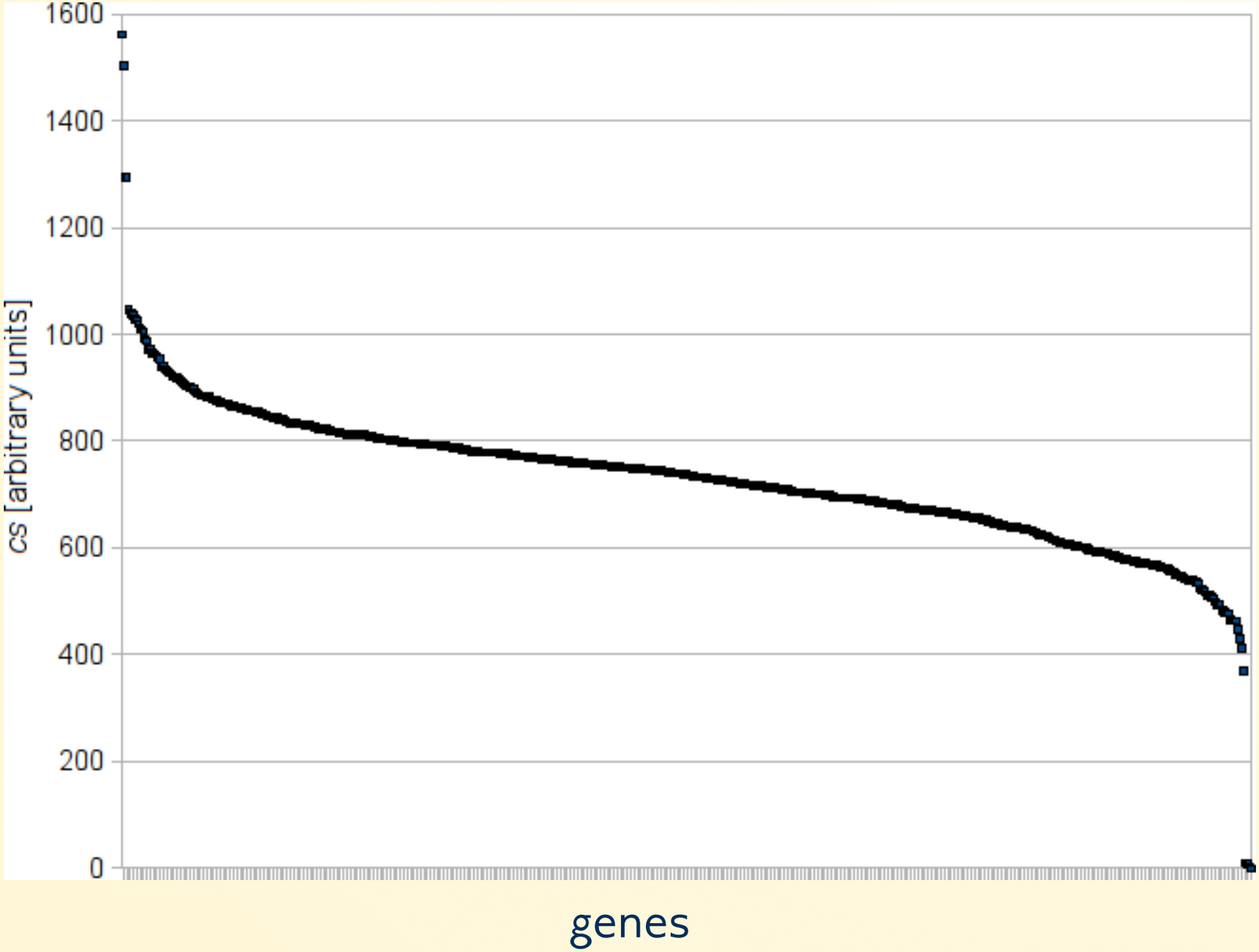
  Their miRNA signature is similar to at least some of the remaining genes

- Some genes have low cs, but not many

  The less similar gene for each miRNA has $cs < 10$ only in 5 cases

  All others have cs in the range $[\ 369, 1561\ ]$

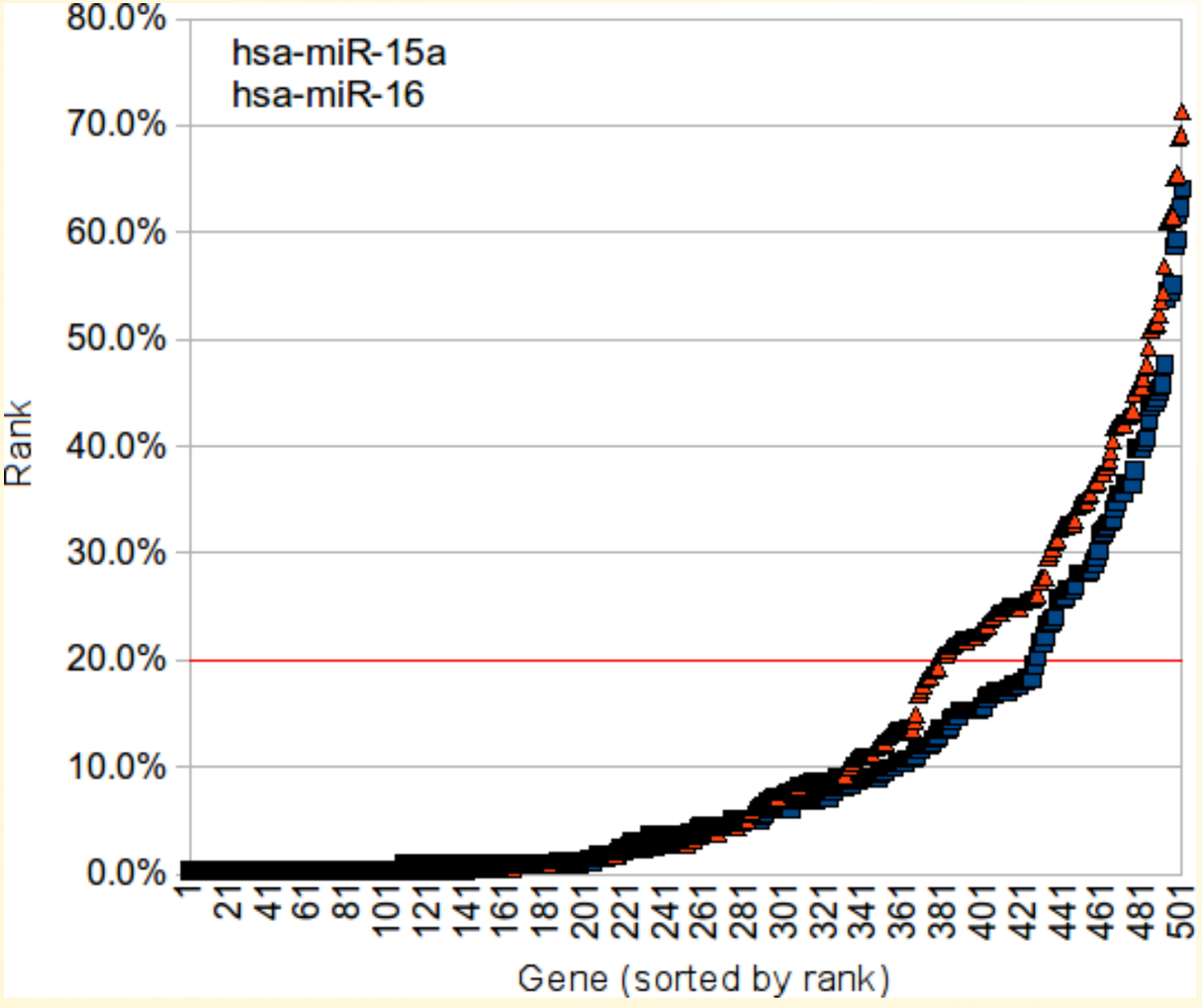
Experiment 1: results



Experiment 2: some details

- Take all pairs of miRNAs (677 x 676 = 457652 pairs)
- For each pair, use one miRNA as a query and the other one as the reference
- Remove each gene from the MATCH subset (from tens to thousands, depending on the miRNA) and place in the NON-MATCH subset
- Compute *cs* for each gene in NON-MATCH and rank genes
- Check the rank of al NON-MATCH genes: is the removed gene among the first ones?
- Swap query and reference, then repeat

Experiment 2: results (first example)

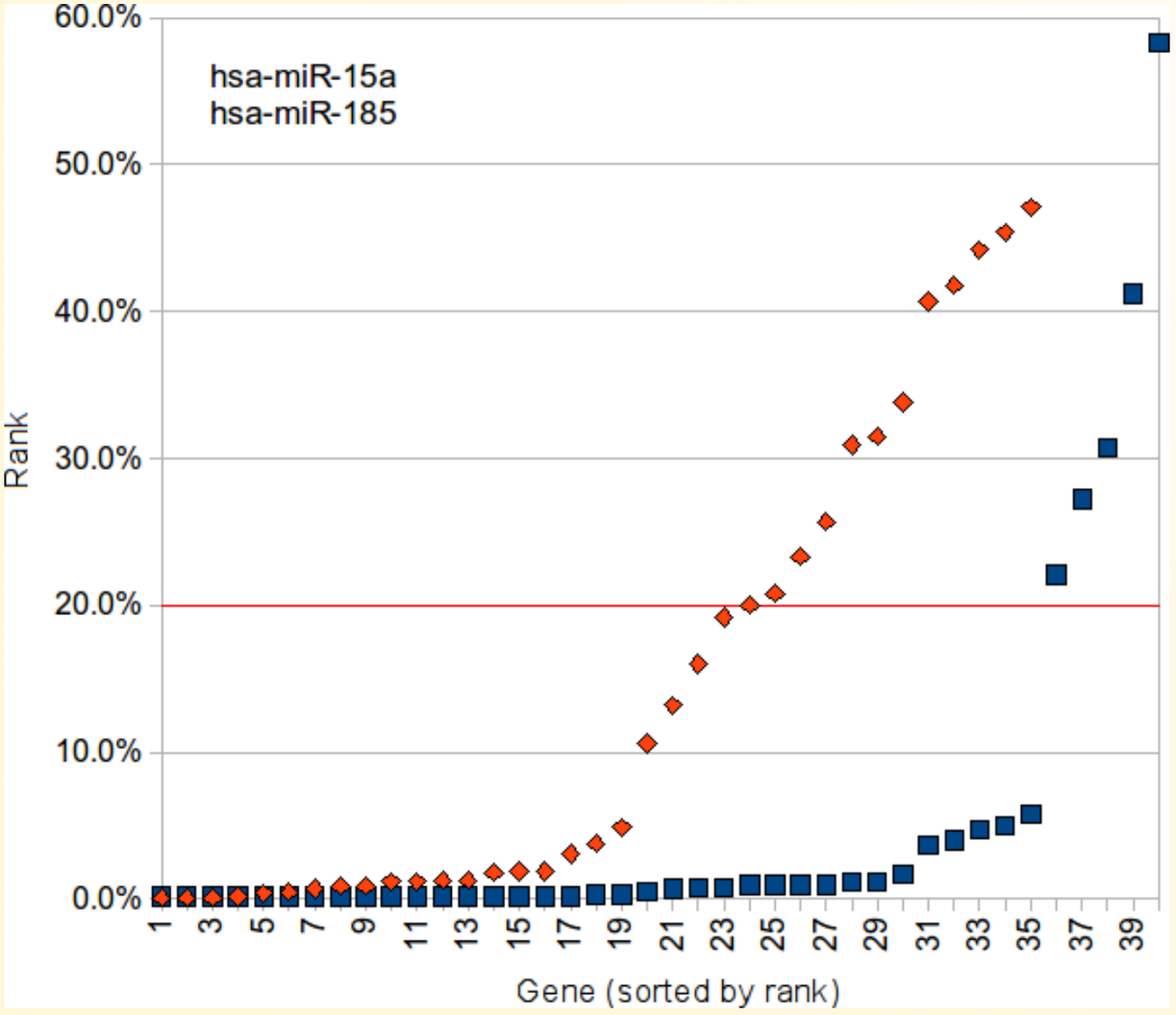


- query hsa-miR-15a, reference hsa-miR-16
- query hsa-miR-16, reference hsa-miR-15a

Experiment 2: results (first example)

- MiRNAs are related (belong to the same cluster)
- Genes involved: 506
- About 70% of the removed genes come up among the top 20% as candidate targets
- No big difference between using one or the other miRNA as query – They are related, and the graph confirms it

Experiment 2: results (second example)



- query hsa-miR-185, reference hsa-miR-15a
- query hsa-miR-15a, reference hsa-miR-185

Experiment 2: results (second example)

- MiRNAs are not related
- Genes involved: 40
- Notable difference between using one or the other miRNA as query.

Conclusion

- A method for suggesting possible target genes for miRNAs
- Can also be used to detect false positives
- Tested according to two experimental strategies

- Theoretically not limited to the specific problem of miRNA target prediction

– Thanks for listening to the end