Combinatorial microRNA target predictions

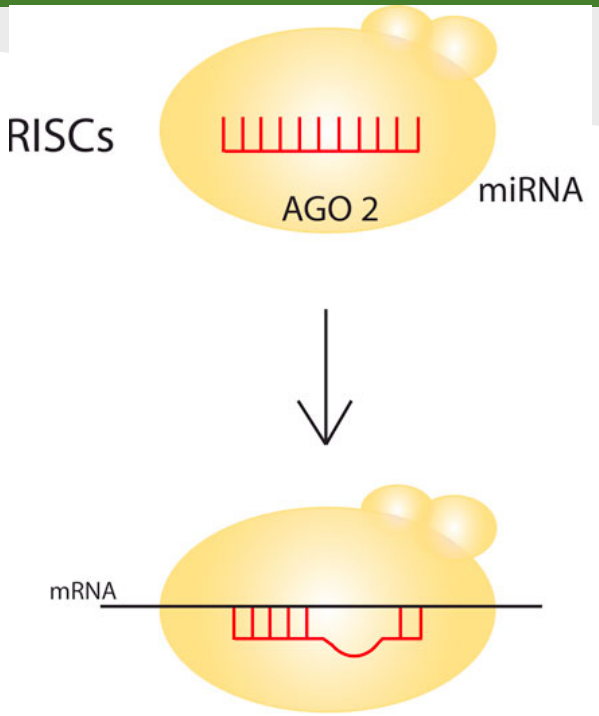
Azra Krek, et al.

Introduction

- microRNA?
  - small noncoding RNAs
    - 18 - 24 nucleotides
  - can bind several different mRNAs
  - bind complementary sites in the 3’ UTR of target genes
    - several different miRNAs can act together
    - multiple target sites

Introduction

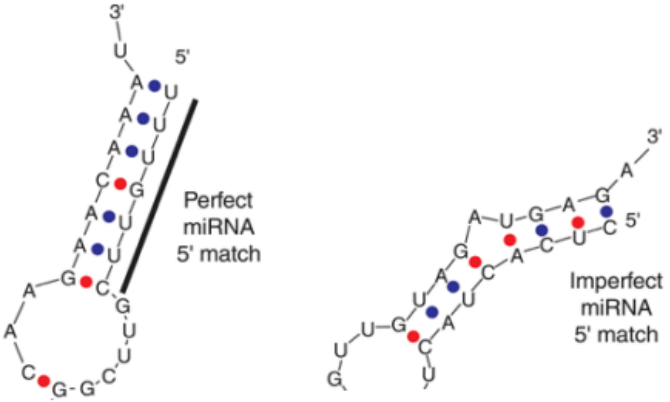
- post-transcriptional regulation of gene expression
  - RISC : gene-knockout/gene-knockdown
- highly conserved binding site
  - target site identification
- problems:
  - don’t know a lot about miRNAs
  - number of confirmed heteroduplexes are small
  - predicted sites can be false positive



source:http://www.biomedcentral.com/content/figures/1741-7007-10-58-1-l.jpg

Introduction

- target site identification
  - perfect nuclei (seed) ~7 nucleotides
    - starting at pos. 1 or 2
  - imperfect nuclei
    - at most one insertion/mutation



source: dissertation "Statistics and Evolution of Functional Genomic Sequence, Dominic Grün"

- problems of over-expression may be avoided by using miRNA repression

Introduction

Paper(2005!):

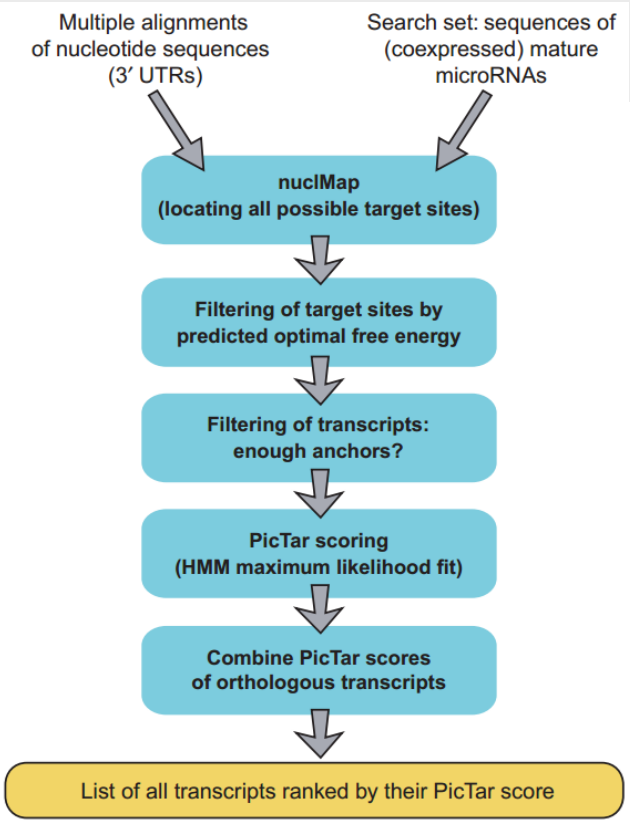
- search for combinations of miRNA binding sites
  - common targets of miRNAs

- PicTar - **p**robabilistic **i**dentification of **c**ombinations of **tar**get sites
  - identification of targets for single miRNAs and combinations of miRNAs

PicTar

- fixed search set of miRNAs and multiple alignments of orthologous nucleotide sequences (3’ UTR)

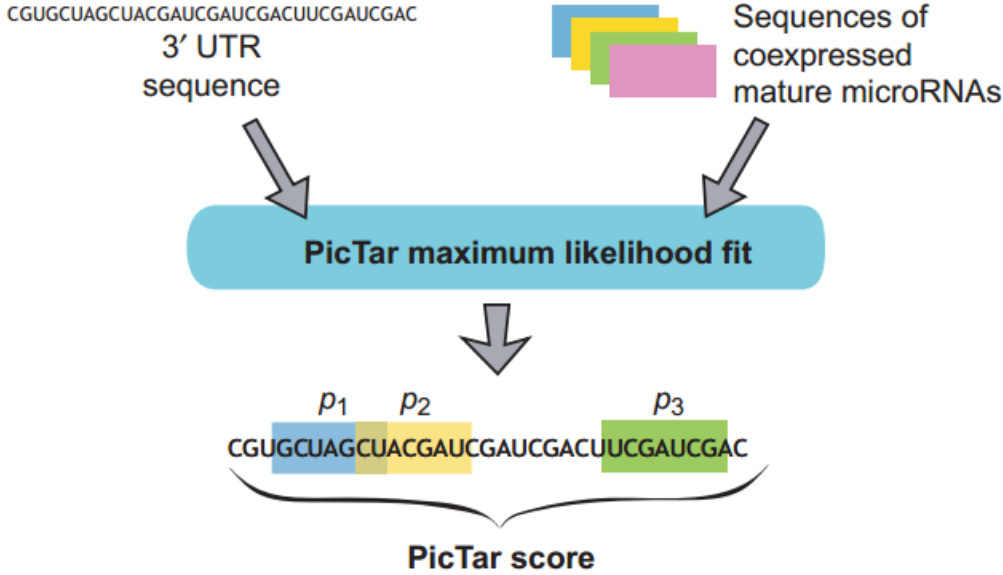
- follows the logic of ahab
    - identification of combinations of transcription factor binding sites



PicTar

- tallies all segmentations of a sequence into binding sites and background
  - computes maximum likelihood score

- probabilities assigned to a single site - modeled in accordance with experimental and computational results



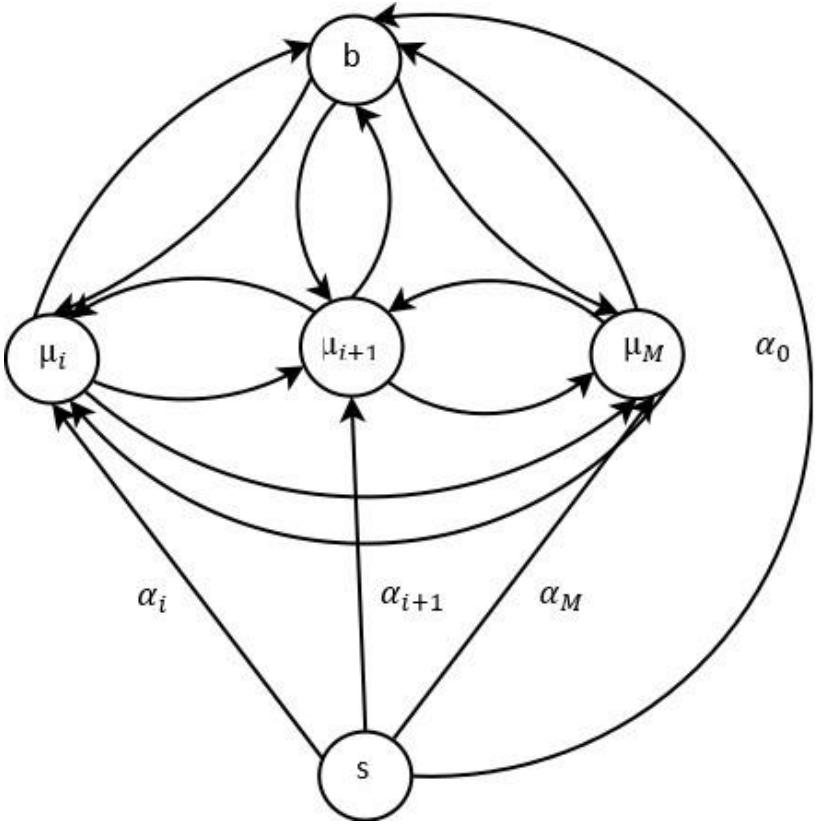
PicTar

- cross-species comparisons
    - filtering out false positives by using sequence alignment to eight vertebrates

- candidate genes:
    - UTRs with a minimal number of evolutionarily conserved binding sites

- PicTar scores the candidate sequences for each species separately
    - scores are combined to obtain the final PicTar score

PicTar - HMM



p = transition probability from $i = 1$ *to* M $\forall_{p_i}$
(prior probabilities of binding)
p_0= transition probability to the background
arrows from one state to another state,
have a transition probability p

M = # of different miRNAs
μ_i = microRNA
α = initial probability
b = background
s = start

selfloops are <u>not</u> displayed

PicTar - HMM

$$P(S \mid \theta) = \sum_{T} P(S \mid \theta, T)$$

$$P(S \mid \theta, T) = \prod_{i=1}^{N(T)} \rho_i \cdot m_i(s)$$

Material & Methods

Data sets of known and randomized mature miRNA

- miRNA sequences from Rfam
  - added 9 miRNAs
- extracted a subset of miRNAs conserved between human, chimpanzee, mouse, rat, dog and chicken
- constructed a set of unique miRNAs (lumping together seq. with same 1-7 or 2-8 base pairs)
  - 58 unique sequences
- generated cohorts of unique randomized miRNAs

Material & Methods
Vertebrate 3’ UTR sequences and alignments

- extracted genome-wide multiple alignments of 8 vertebrates

- define multiple alignments of 3’ UTRs
  - cover human, chimpanzee, mouse, rat and dog for 90% of human 3’ UTR sequences

- restricted human 3’ UTR sequences
  - unique
  - masking repeats

	Human	Chimpanzee	Mouse	Rat	Dog	Chicken	Pufferfish	Zebrafish
1	19,253,481	18,720,159	15,610,779	15,071,221	17,356,774	5,485,265	1,334,211	1,688,879
2	14,575,934	14,224,691	13,144,375	12,699,682	13,873,555	4,398,114	1,136,336	1,430,061

Material & Methods
Identification of single miRNA target sites

- used experimental results to:
  - define probabilities for a mRNA sequence to be a binding site
  - insertions/deletions allowed if free energy does not increase

- free energy(using RNAhybrid) below a cutoff value
  - perfect nuclei: 33% of optimal free energy ~ 5% discarded
  - imperfect nuclei: 66%

- perfect nucleus assigned a probability p to be a binding site

- imperfect nucleus: $\frac{1-p}{\#imperfect\ nuclei}$

Material & Methods
Scoring combinations of target sites

- computes a maximum likelihood score
  - RNA sequence targeted or not

- 5 implementation details
  - 1. sets the length of miRNA binding sites to the length of the corresponding nuclei
  - 2. short 3’ UTR sequences cannot be used
  - 3. Baum-Welch algorithm to compute the maximum likelihoods
  - 4. optimized prior for background
  - 5. order of the model for background sequence is set to 0

Material & Methods

Genome-wide picTar runs and cross-species comparison

- precomputed positions of all possible miRNA nuclei in all UTR seq.
- nuclei fall into overlapping alignment for all species?
  - check: optimal free energy of miRNA:mRNA duplex
    - anchors
- filtering improbable target sites
  - compute score for each UTR
- final score: averaged the scores for all species that defined anchor sites

Test of PicTar

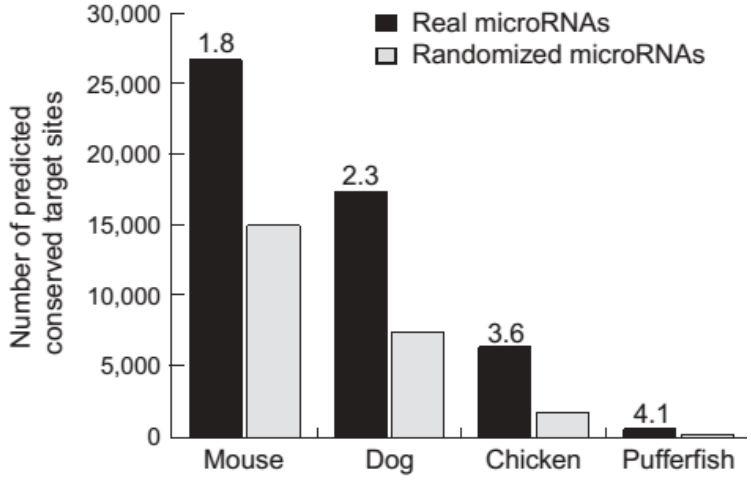
- search in genome-wide set of 10,607 *C. elegans* and *C. briggsae* 3’ UTR sequences
  - miRNAs: *lin-4* or *let-7*
  - known targets *lin-14*, *hbl-1*, *daf-12* and *lin-28*
    - ranked first, second, fourth and seventh

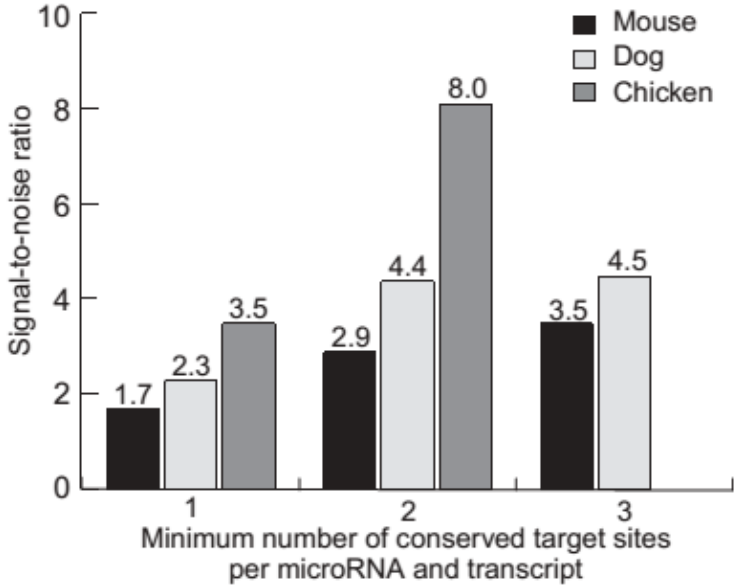
- in vertebrates:
  - construction of a multiple alignment of human annotated 3’ UTRs from 7 other vertebrates
    - to reduce false positive rate

Test of PicTar

false positive rate

- estimate false positive rate
  - recorded all perfectly binding conserved target sites (“anchors”) and for randomized DNA

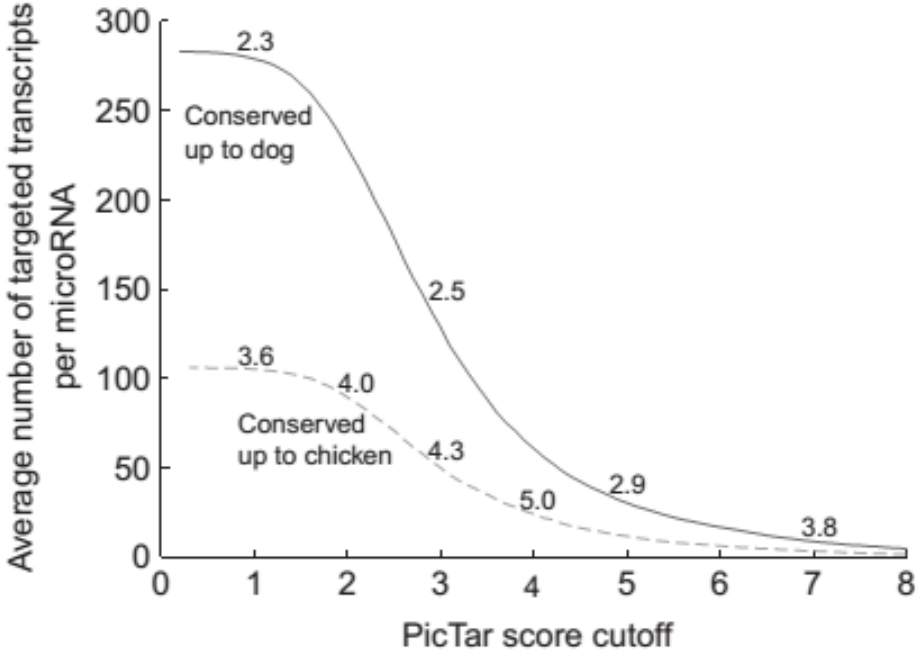




Test of PicTar

false positive rate

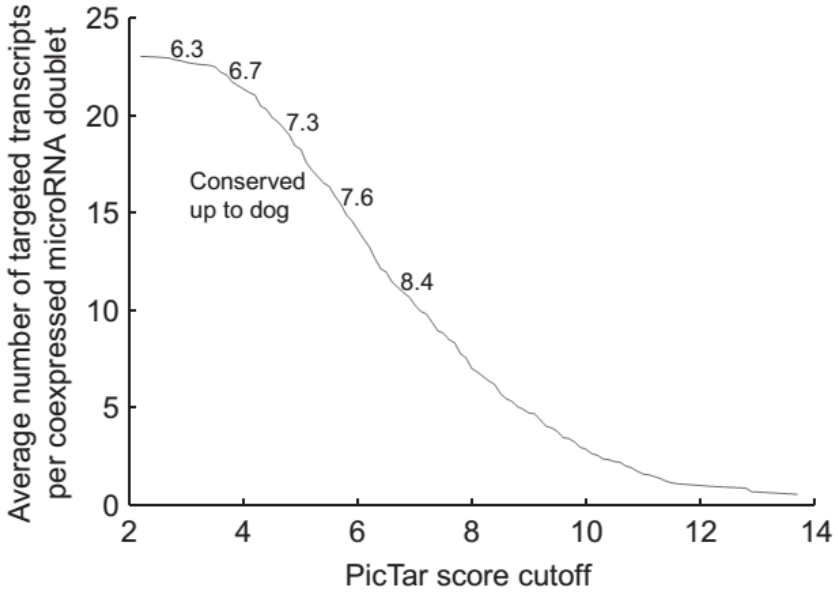
- ranked target prediction for all available, conserved miRNAs



Test of PicTar

false positive rate

- PicTar score dependent sensitivity and specificity of target site predictions
  - 4 sets of coexpressed miRNAs + corresponding sets of randomized miRNAs - require 2 anchor sites



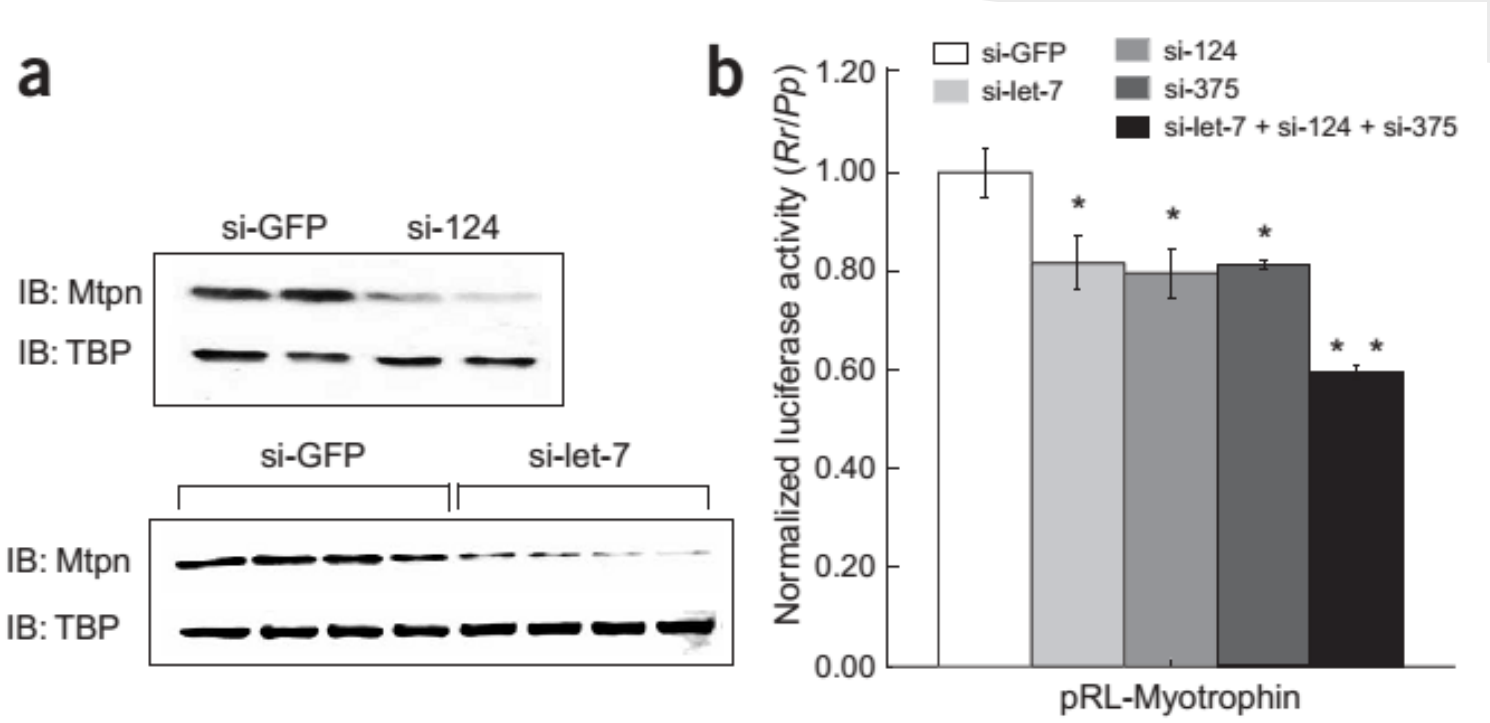
PicTar - common binding site

- hypothesized 3 miRNAs in the murine pancreatic cell line MIN6
  - *miR-124*, *miR-375* and *let-7b* may act together

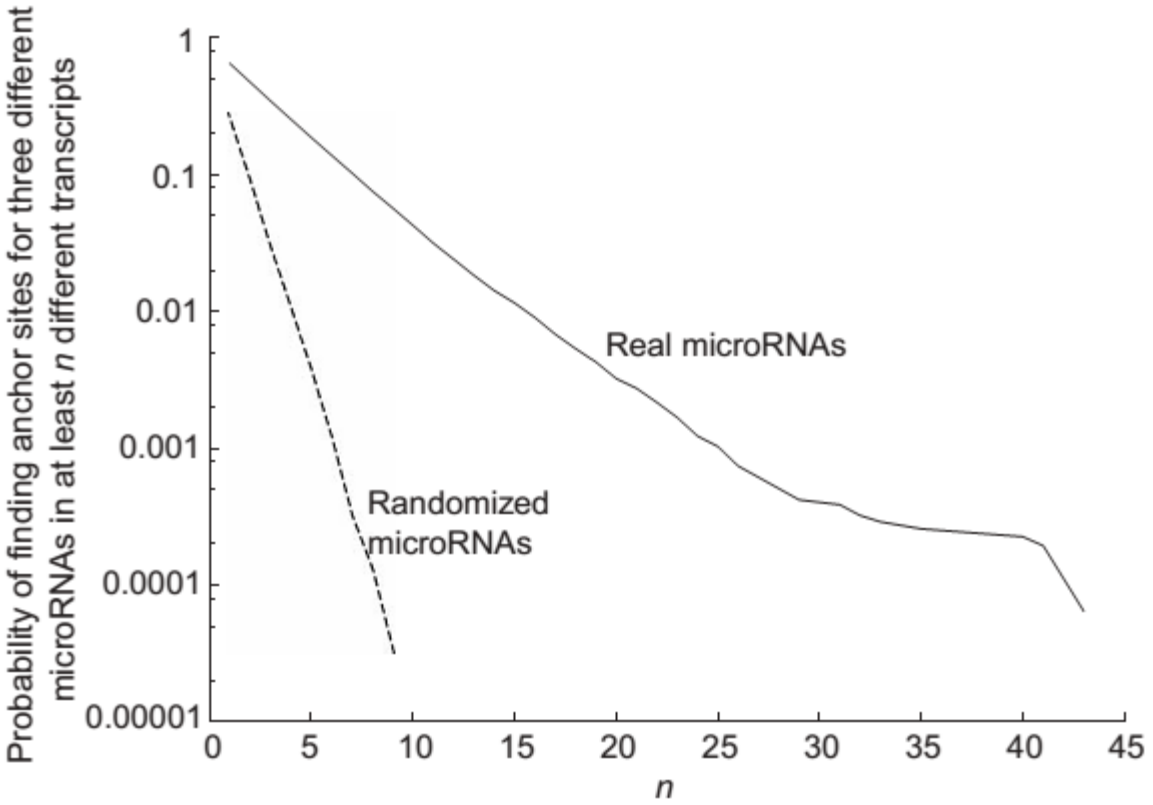
- examined results for *Mtpn* (known target of *miR-375*)
  - with *miR-375*: rank 102
  - with *miR-375* and *miR-124*: rank 15
  - with *miR-375*, *miR-124* and let-7b: rank 4

PicTar - common binding site

Validation



Estimate of miRNAs that may regulate target genes



Conclusions

- developed a computational approach that identifies targets that are likely regulated by miRNAs in common pathways

- showed that seq. comparison using genome-wide alignments across 8 vertebrates reduce the FP-rate of miRNA target prediction
  - predict on average ~200 targeted transcripts per miRNA

- let to experimental validation of Mtpn
  - regulated coordinately by 3 miRNAs

Conclusions

- PicTar is from 2005!
  - good prediction

- problem: based on data that is over 10 years out of date
  - “today” more:
    - databases/sequences for 3’ UTRs
    - knowledge about the biology from miRNAs

- based on free energy calculations (filtering)
  - my not be complete or accurate
  - quality of data used in free energy calculation

Thank you for the attention