Detection of miRNA target genes through statistical analysis of DNA motifs in human-mouse 3'-UTR regions.

M. Caselle

Annecy, 10/11/2006

Plan of the talk.

- Introduction to microRNAs:
  - miRNA biology.
  - Role of miRNA in the modern picture of gene regulation.
  - References and databases on miRNA

- Our goal: A general "ab initio" method to identify miRNA targets and (possibly) new miRNA genes based on:
  - conserved overrepresentation of regulatory sequences in the 3'UTR regions of genes
  - strand asymmetry in the 3'UTR regions

- Validation: Identification of several known regulatory sequences in the 3'UTR regions of human genes

- Result: List of new candidate miRNA "seeds"

Working group

miRNA identification:

M. C. and D. Cora' (Theoretical Physics dep. Univ. of Torino)

P. Provero and F. di Cunto (Genetics, Biology and Biochemistry dep. Univ. of Torino)

miRNA regulatory network:

L.Martignetti, I. Molineris, A. Re and G. Sales (Theoretical Physics dep. Univ. of Torino)

References

- D. Cora', C. Herrmann, C. Dieterich, F. Di Cunto, P. Provero and M. Caselle

  "Ab initio identification of putative human transcription factor binding sites by comparative genomics"

  BMC Bioinformatics 2005, 6:110

- D. Cora', F. Di Cunto, M. Caselle and P. Provero "Identification of candidate regulatory sequences in mammalian 3'UTR regions by statistical analysis of oligonucleotide distribution" submited to BMC Bioinformatics

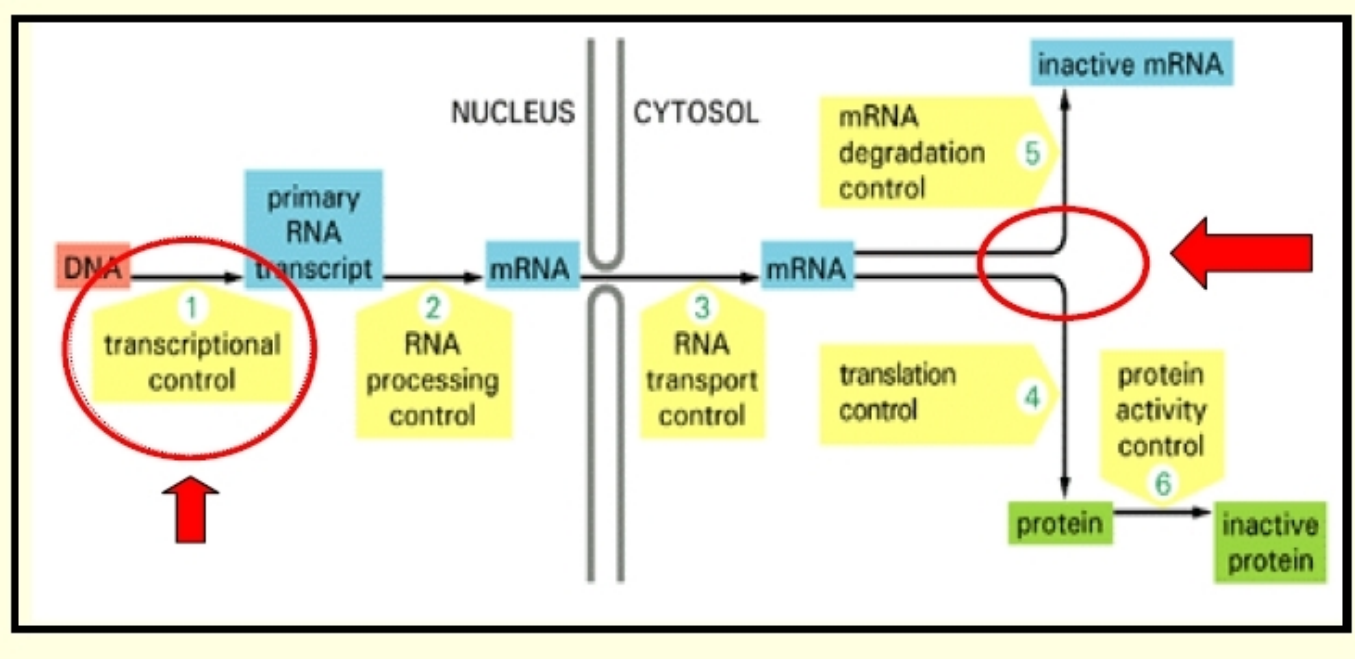
miRNA.

Gene expression can be regulated at many of the steps in the pathway from DNA to RNA and protein.

MicroRNAs (miRNAs) are a family of small RNAs (typically 21 - 25 nucleotide long) that negatively regulate gene expression at the post-transcriptional level.

Members of the miRNA family were initially discovered as small temporal RNAs that regulate developmental transitions in Caenorhabditis Elegans (lin-4). (Chalfieet al. 1981; Lee et al. 1993) but considered only as a peculiarity of worms.

In 2002-2003 it was suddenly realized that miRNA exist in all higher Eukaryotes in several copies and that they play an essential role in development and differentiation of tissues.

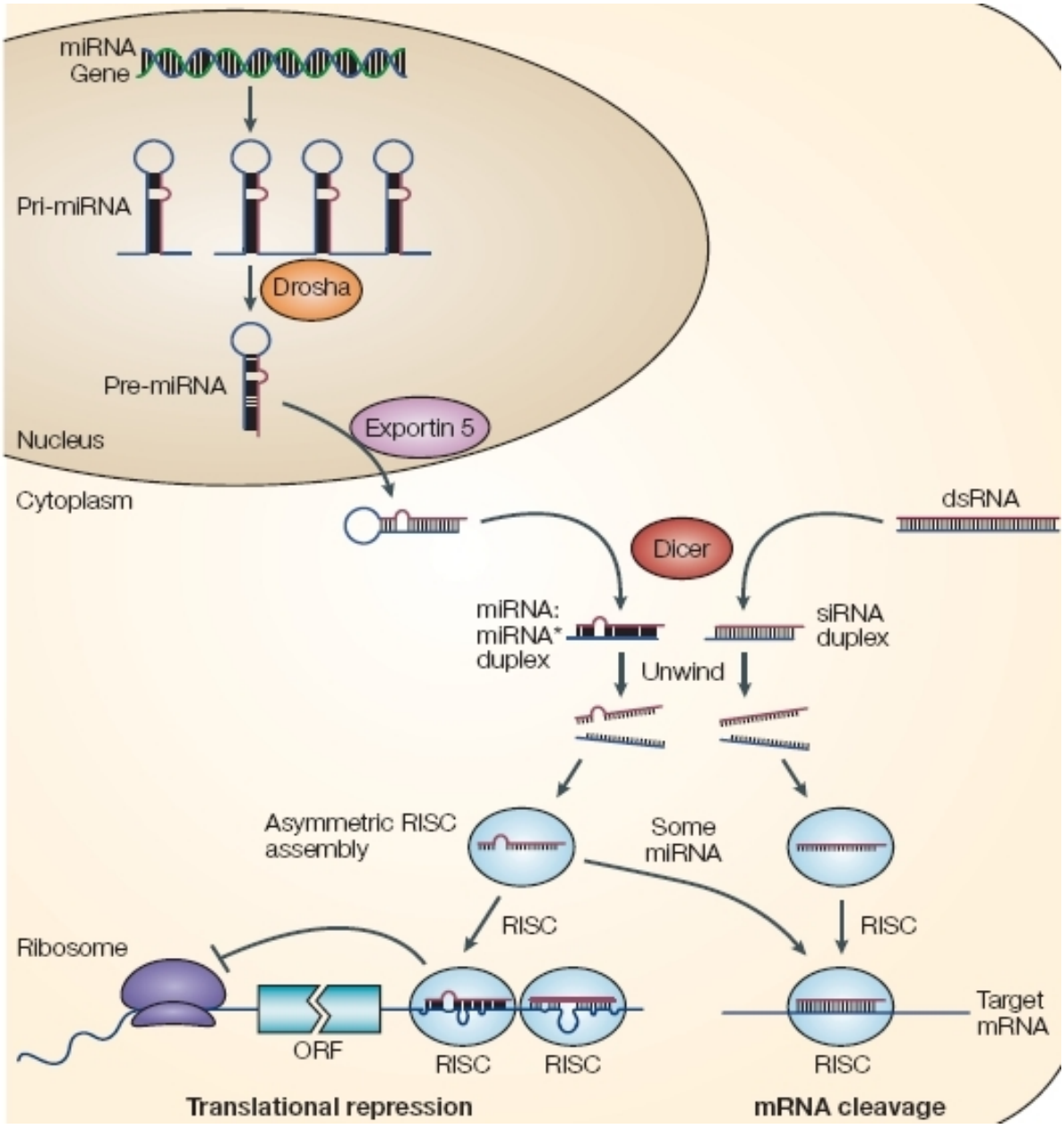

NUCLEUS
CYTOSOL
inactive mRNA
mRNA degradation control
5
primary RNA transcript
DNA
mRNA
mRNA
1
transcriptional control
2
RNA processing control
3
RNA transport control
translation control
4
protein activity control
6
protein
inactive protein

miRNA biology

miRNA derive from larger precursors that form imperfect stem-loop structures.

- the nascent miRNA transcripts (pri-miRNA) are processed into 70 nucleotide precursors (pre-miRNA).
- The precursor is cleaved to generate 21 - 25 nucleotide mature miRNAs in cytoplasm.
- The miRNA gene regulation mechanism requires the coupling with a special protein complex called RNA-Induced Silencing Complex (RISC).
- Even if the precise mechanism of action of the miRNA / RISC complex is not very well understood, the current paradigm is that miRNAs are able to negatively affect the expression of a "target" gene via mRNA cleavage or translational repression,

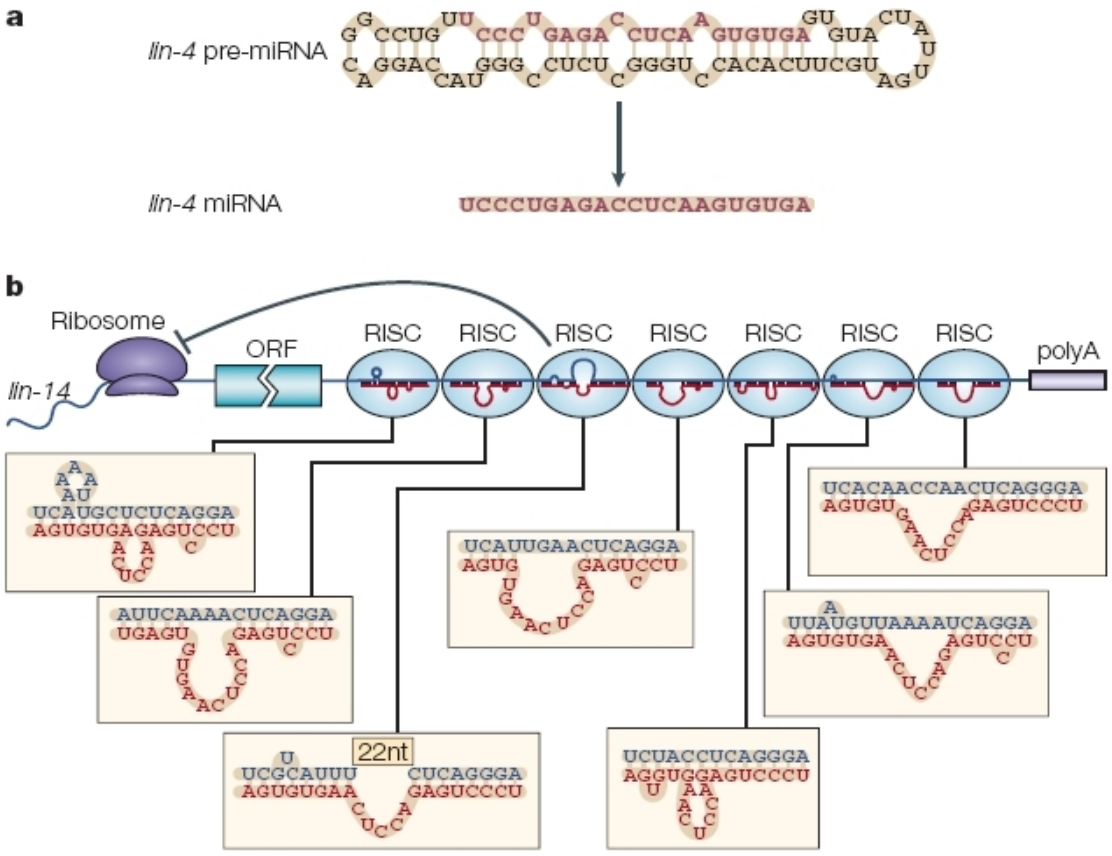
after antisense complementary base-pair matching to specific target sequences in the 3'-utr of the regulated genes.



- The functions in which miRNAs are involved are extremely wide and, in animals, they include: developmental timing, pattern formation and embryogenesis, differentiation and organogenesis, growth control and cell death, with putative involvement in human diseases in the case of *H. Sapiens*.

- As for the regulatory mechanism, it is also clear that the miRNAs control is a one-to-many process, meaning that each miRNA is expected to control from one to hundreds of targets. Moreover, each specific miRNA binding site is also often overrepresented in a given 3'-utr sequence. There are also evidences of a combinatorial mechanism, meaning that a certain mRNA can be under control of many different miRNAs

- The above observations tell us that Transcription Factors and miRNA share a very similar behaviour. The main difference between the two is that while TF act as a sort of on/off switch, miRNA role is to fine tune the gene expression.


a
lin-4 pre-miRNA
lin-4 miRNA
UCCCUGAGACCUCAAGUGUGA
b
Ribosome
ORF
lin-14
RISC
RISC
RISC
RISC
RISC
RISC
RISC
polyA
22nt

- MiRNAs also show interesting evolutionary properties between different species. Up to one third of the miRNAs discovered in *C. elegans* have and orthologous in human. On the other hand, specie-specific miRNAs exist and, in particular, it is established that primates have an own class of miRNA genes.

- Tracing back this evolutionary pattern it is possible to guess that miRNA appaeared as a new regulatory mechanism about 500 Myears ago. It is interesting to observe that this time scale almost coincides with the impressive explosion of new species in the Cambrian age and with the almost simulatenous appearance of retrotransposons in Eukaryotes.

miRNA References

- He and Hannon "MicroRNA: small RNAs with a big role in gene regulation."
  Nat Rev. Genet. 2004 Jul;5(7):522-31.
- Bartel, D "MicroRNAs: Genomics, Biogenesis, Mechanism, and Function."
  Cell. 2004 Jan 23;116(2):281-97.
- John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS "Human MicroRNA targets."
  PLoS Biol. 2004 Nov;2(11):e363. Epub 2004 Oct 05.
- S. Yoon and G. De Micheli "Computational identification of MicroRnas and their targets."
  Birth Defect Research (Part c) 78:118-128 (2006).
- N. Rajewsky. "microRna target predictions in animals."
  Nature Genetics Supplement Vol.38 s8-s13 (2006).
- Griffiths-Jones S. "The microRNA registry."
  Nucleic Acids Res. 2004 Jan 1;32 Database issue:D109-11.
- `http://www.microrna.org/`

miRNA databases

Despite all these results, our knowledge on miRNA is still largely unsatisfactory.

There is a catalogue of about 580 putative human miRNA genes obtained with in silico analysis, but only 131 of them are experimentally validated. The total content of of miRNA genes in human is expected to be much larger

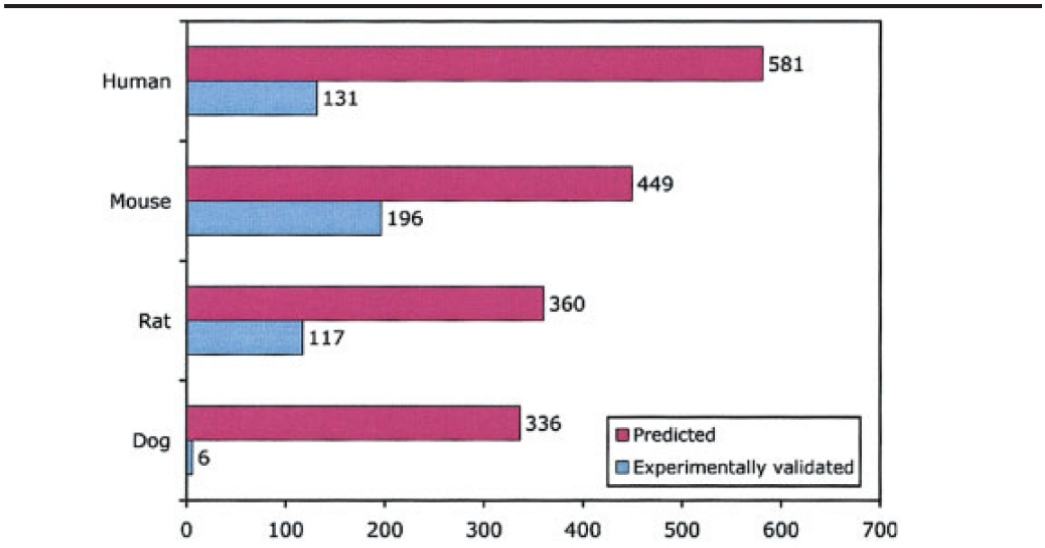


Figure 1. The total numbers of predicted and experimentally validated miRNAs for four different species (http://mirnamap.mbc.nctu.edu.tw/php/statistics.php).

It is commonly believed that each miRNA could target up to hundreds of genes, however the identification of this targets turned out to be a very difficult task. Existing in silico studies starting from the known or putative miRNAs lead to contradictory results and to huge number of false positives. This is a real problem since the functional characterization of miRNAs heavily relies on the identification of miRNA target genes.

There are a few "reference" databases on miRNAs. In particluar miRBase, miRanda and miRNAMap.

TABLE 1. Online Resources for miRNA Research

Name	URL	Main feature	References
miRNA registry/ miRBase	http://microrna.sanger.ac.uk	miRNA sequences, annotations, and predicted targets	Griffiths-Jones (2004, 2006)
miRNAMap	http://mirnamap.mbc.nctu.edu.tw	Genomic maps for miRNA genes and targets	Hsu et al. (2006)
MiRscan	http://genes.mit.edu/mirscan	miRNA gene scan	Lim et al. (2003a, b); Ohler et al. (2004)
RNA regulatory networks	http://www.mirz.unibas.ch	Putative miRNA gene and target scan	Sewer et al. (2005)
TargetScan/ TargetScanS	http://genes.mit.edu/targetscan	Prediction of miRNA targets	Lewis et al. (2005, 2003)
PicTar	http://pictar.bio.nyu.edu	miRNA target prediction for vertebrates and flies	Grun et al. (2005); Krek et al. (2005)
miRanda	http://www.microma.org	Human, flies, and zebrafish miRNA target search	Enright et al. (2003); John et al. (2004)
DIANA-microT	http://www.diana.pcbi.upenn.edu/cgi-bin/micro_t.cgi	Human, mouse, rat miRNA target scan	Kiriakidou et al. (2004)
RNAhybrid	http://bibiserv.techfak.uni-bielefeld.de/rnahybrid	Prediction of miRNA binding sites	Rehmsmeier et al. (2004)
Tarbase	http://www.diana.pcbi.upenn.edu	List of experimentally supported miRNA targets	Sethupathy et al. (2006)
miRU	http://bioinfo3.noble.org/miRNA/miRU.htm	Plant miRNA target finder	Zhang (2005)
TargetBoost	https://demo1.interagon.com/demo	miRNA-target binding characterization	Saetrom et al. (2005)
Vienna package	http://www.tbi.univie.ac.at/~ivo/RNA	RNA secondary structure prediction and comparison	Hofacker (2003)
Mfold package	http://www.bioinfo.rpi.edu/~zukerm/rna	RNA folding and hybridization prediction	Mathews et al. (1999); Zuker (2003)

Our project

The goal of our research project is to find these targets (and possibly also new candidate miRNA genes) by directly looking at the genomic sequence and thus with a "ab initio" unbiased method.

Contrary to the existing approaches we do not start with the known miRNAs but use them at the end of analysis to validate our results.

The problem

Summary of the most relevant features of miRNA regulatory interaction in mammalians

- Targets are located in the 3'-UTRs.

- Perfect complementarity or G-U pairing (usually at most one) between the target 3'-UTR and the first seven nucleotides 1-7 or 2-8 of miRNA (in few cases also 3-9).

- Evolutionary conservation of miRNA target sites.

hence the goal is to find words of 7 nucleotides which are evolutionary conserved and show an uneven statistical distribution in the 3'UTR of human genes

However unfortunately the standard tools of sequence datamining are uneffective for this problem:

- the word's distribution in the human 3'UTR is surprisingly uneven: almost all the words are underrepresented or overrepresented! (reasons: CpG islands and A/T bias)

- Conseved sequences between men and mouse are not enriched of known validated miRNA targets! (reason: Blast is unable to identify conserved words if they are too short (less than 11)!)

Our proposal to overcome these problems

- "Conserved overrepresentation"

- Breaking of the strand symmmetry

Conserved overrepresentation.

We organized the whole human gene content in sets as follows:

- For each word of 7 bp's (there are 16384 of them) we computed the frequency in the 3'UTR sequences of the whole genome considered as a single sample: these were our reference frequencies.

- Then we counted occurrences of the word in the 3'UTR region of each gene separately.

- If the number of occurencies of the word in the 3'UTR region of gene G was statistically significant (compared to a binomial distribution based on the above reference frequencies), then we included gene G in the set the set $S(\text{word})$.

Then we performed the same analysis in mouse. By using the orthology map (almost each human gene has an orthologous in mouse) we can superimpose the two collections of sets and look for possible intersections.

It is possible to evaluate the non-randomness of all these intersections using the hypergeometric distribution plus a suitable Bonferroni correction.

We term the sets (and hence the words) which survive this analysis as "conserved overrepresented".

This is our first list of candidates (miRna seeds / miRNA targets)

Strand asymmetry.

A completely different approach to obtain putative miRNA targets is based on the observation that they should break the strand symmetry which is otherwise always conserved in the genome.

Strand asymmetry can be evaluated by looking in the 3'UTR of a given gene to the difference between the frequency of a word and of its reverse complement

A careful treatment of the background distribution (the "null hypothesis") is mandatory due to A/T bias. Solution: Markov model

Table: - Nucleotide composition of 3' UTR regions. The base frequencies of 3' UTR regions in human and mouse, excluding the masked repeats.

	Human	Mouse
A	0.2683	0.2638
C	0.2199	0.2237
G	0.2210	0.2254
T	0.2908	0.2871

We generate using a MM artificial sequences which have the "correct" (i.e. the one found in real biological sequences) short scale distribution of bases (up to 4 nucleotides) and look for signatures of strand asymmetry at a larger scale (7 bp):

On the sequences produced by the Markov chain we computed, for each oligo w, the mean $\mu(w)$ and standard deviation $\sigma(w)$ of the quantity

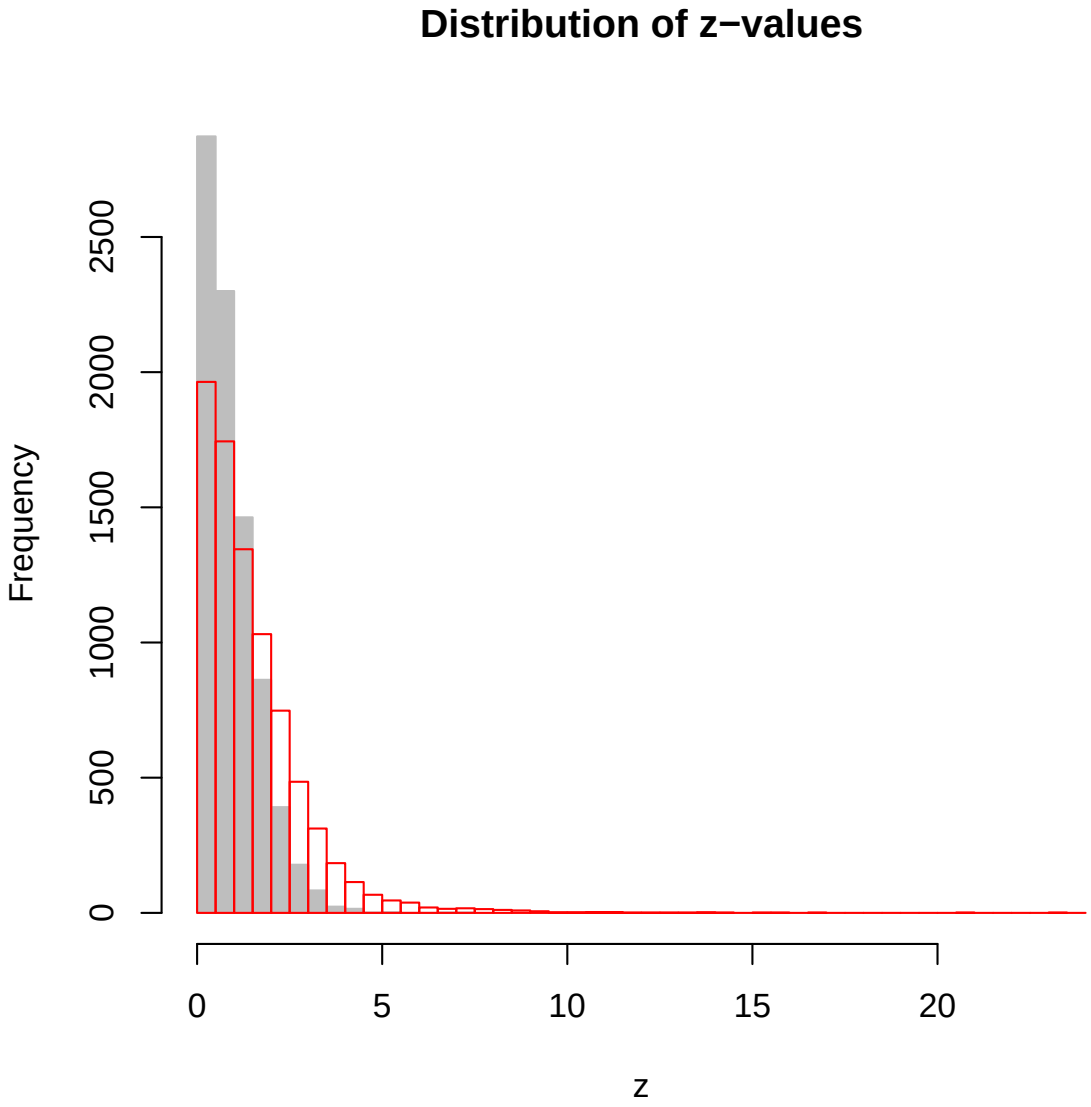
$$a(w) = n(w) - n(\bar{w})$$

where $n(w)$ is the number of occurrences of w. $a(w)$ was then computed for the actual sequences, and a z value was constructed as

$$z(w) = \frac{a(w) - \mu(w)}{\sigma(w)}$$

where $a(w)$ refers now to the actual sequence. A P-value was finally associated to each oligo w assuming a standard normal distribution of the z-values.

Selecting the words with (Bonferroni corrected) $p < 0.01$ we obtained a list of 214 words in human and139 in mouse. Out of these 113 were in common. This is our second list of candidate miRnas.


Distribution of z−values
Frequency
0
500
1000
1500
2000
2500
0
5
10
15
20
z

Results 1

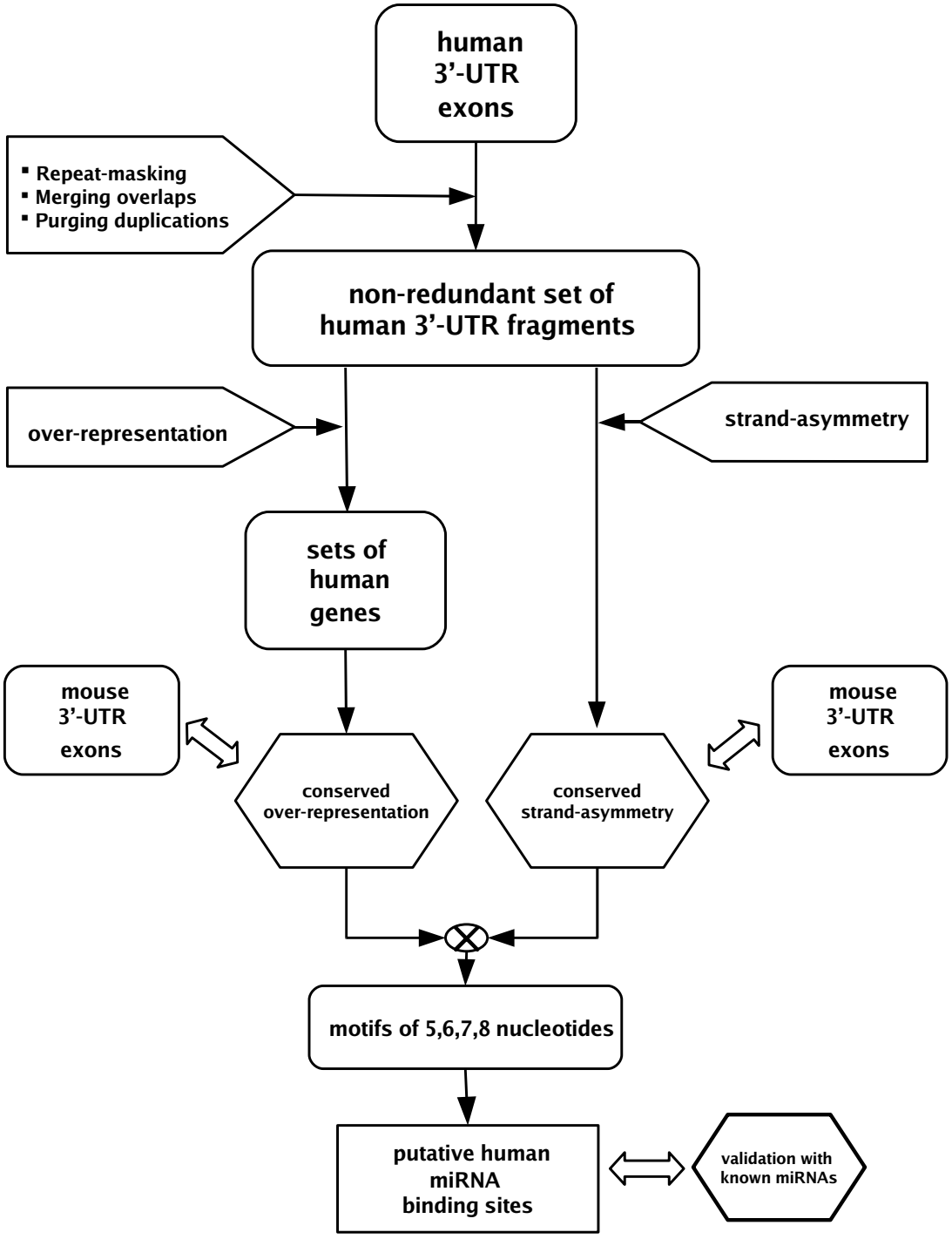
- From the "conserved overrepresentation" analysis we end up with 465 different words of 7 letters which fulfill all the requirements.

- Similarly from the "strand asymmetry" analysis we end up with 214 different words of 7 letters which fulfill all the requirements.

- The two sets have 78 words in common. The probability that two sets of size 465 and 214 chosen at random from a pool of 16384 words ($16384 = 4^7$) have an intersection of 78 word is $p = 5.6\ 10^{-66}$. Since the two analyses are completely independent this is a strong indication that the targets that we find are biologically relevant

Results 2

The most reliable list of predictions is obtained by intersecting the 113 words which show strand asymmetry in both human and mouse (a condition which we name conserved strand asymmetry) *and* the 465 words which fulfill conserved overrepresentation. In this way we select only 59 candidate regulatory sequences. Out of these 14 correspond to known putative miRNA binding sequences.


human
3'-UTR
exons
Repeat-masking
Merging overlaps
Purging duplications
non-redundant set of
human 3'-UTR fragments
over-representation
strand-asymmetry
sets of
human
genes
mouse
3'-UTR
exons
conserved
over-representation
conserved
strand-asymmetry
mouse
3'-UTR
exons
motifs of 5,6,7,8 nucleotides
putative human
miRNA
binding sites
validation with
known miRNAs

Validation 1

It is interesting to compare the results that we find with the MiRBase database of putative miRNAs. Out of the 580 existing putative miRNA we find (keeping into account the three possible frames and the existing degeneration among miRNAs) 1017 different putative "target words".

- In the set of the 465 conserved overrepresented words there are 74 of the above putative targets. The probability of this to be a random intersection is $p = 5\ 10^{-14}$

- In the set of the 214 strand asymmetric words there are 41 of the above putative targets: $p = 8\ 10^{-11}$

Example: miR-29

```
GGTGCTA hsa-miR-29c TAGCACCatttgaaatcggt
GGTGCTA hsa-miR-29b TAGCACCatttgaaatcagtgtt
GGTGCTA hsa-miR-29a TAGCACCatctgaaatcggtt

TGGTGCT hsa-miR-29c tAGCACCAtttgaaatcggt
TGGTGCT hsa-miR-29b tAGCACCAtttgaaatcagtgtt
TGGTGCT hsa-miR-29a tAGCACCAtctgaaatcggtt


GO Function: extracellular matrix
             structural constituent
GO:0005201    p = 7.40828e-07


GO Cellular component: collagen
GO:0005581        p = 8.67893e-09
```

Table 2 - Comparison between the results of our computational approach and binding sites of known miRNAs in miRBase The rows indicate the computational methods as described in the text: CO=conserved overrepresentation; SA= strand asymmetry; CSA=conserved strand asymmetry (oligos displaying strand asymmetry in both human and mouse); $CO \cap CSA$ oligos identified by both CO and CSA. The columns are : N: Number of oligos identified computationally; M: Number of these that coincide with the binding sites of experimentally validated human miRNAS; E: number of such matches expected by chance; N_m: number of different experimentally validated human miRNAs binding the N oligos (many 7-oligos can bind more than one miRNA) P: P-value from exact Fisher test, taking into account that there are 16384 possible 7-mers 1017 of which are binding sites of known miRNAs.

Method	N	M	E	N_m	P
CO	465	74	28.9	116	$5.0 \cdot 10^{-14}$
SA	214	41	13.3	101	$7.9 \cdot 10^{-11}$
CSA	113	19	7.01	54	$6.3 \cdot 10^{-5}$
$CO \cap CSA$	59	14	3.66	46	$1.1 \cdot 10^{-5}$
$CO \cup SA$	601	94	37.3	150	$4.8 \cdot 10^{-17}$

A similar comparison can be performed with the list of putative miRNAs contained in miRNAMap. In this case the total number of 7-mers which are putative miRNA binding sites is 939.

Method	N	M	E	N_m	P
CO	465	76	26.7	101	$6.7 \cdot 10^{-17}$
SA	214	46	12.3	79	$3.9 \cdot 10^{-15}$
CSA	113	23	6.48	42	$8.8 \cdot 10^{-8}$
$CO \cap CSA$	59	15	3.38	31	$7.8 \cdot 10^{-7}$
$CO \cup SA$	601	98	34.4	135	$1.8 \cdot 10^{-21}$

Validation 2

Besides miRNAa binding sites there are other well known regulatory sequences in 3'UTR. We expect to find also them among our candidates.

- Poly-A signal AATAAA. We found 12 instances of this signal among the 59 putative sequenecs which satisfy all our selections

- ARE (AU rich elements) ATTTA. This sequence seems to trigger mRNA instability. We found 3 instances of this signal among our 59 entries.

- PUF (Pumilio-FBF protein family elements) TGTANATA. Similarly to miRNA this class of proteins seem to downregulate the target genes. We found 9 instances of this signal among our 59 entries.

Validation 3

Among the (small number of) experimentally validated miRna targets, a particularly interesting role is played by the human miR-124 .

Recently in Wang and Wang, NAR 34 1646-1652 (2006) a list of 8 genes downregulated after overexpression of miR-124 appeared. All the three words potentially associated to the binding site of this miRNA were identified by our procedure. (GTGCCTT and TGCCTTA by all methods while GCCTTAA by conserved overrepresentation only). We checked that 5 out of these 8 genes indeed appear in the sets associated to the above words, thus supporting the reliability of our method.

New putative regulatory sequences.

Once we eliminate from the list of the 59 words which satisfy all our selections the 14 already known miRNA seeds and the 24 non-miRNA related regulatory sequences discussed above we end up with a list of 21 new putative regulatory sequences. Work is in progress to characterize these sequences and, possibly, to validate them.

Conclusions.

We propose a new method to identify miRNA targets and possibly new putative miRNA genes. The method is based on:

- Analysis of the statistical distribution of oligonucleotides in the 3'UTR region of the genes.
- A careful treatment of the evolutionary conservation of these targets
- Analysis of the strand symmetry breaking

We studied its performances in the human case. We found some already known miRNA genes and miRNA targets which we used as a validation test of the method. We also found several previuosly unknown putative seeds and target sites, which we expect to be of biological relevance.

7-mer word	know regulators
AAACTTG	
AATAAAC	PolyA 6
AATAAAG	PolyA 6
AATCATG	
AGCACAA	hsa-miR-218
ATAAAAG	PolyA 5
ATAAAGG	PolyA 5
ATAAAGT	PolyA 5
ATAAATG	PolyA 5
ATATTTT	
ATTAAAG	
ATTGTAA	PUF 5
ATTTAAG	ARE
ATTTATA	ARE
CAATAAA	PolyA 6
CCAATAA	PolyA 5
CTAATAA	PolyA 5
CTATTTT	
CTTTGTA	hsa-miR-524* hsa-miR-520d*
GAATAAA	PolyA 6
GACCAAA	
GCAATAA	PolyA 5

(continued)

(continued)

7-mer word	know regulators
GCACTTT	hsa-miR-520d hsa-miR-93 hsa-miR-106a hsa-miR-520h hsa-miR-520a hsa-miR-520e hsa-miR-519b hsa-miR-20a hsa-miR-106b hsa-miR-372 hsa-miR-520b hsa-miR-17-5p hsa-miR-520g hsa-miR-519c hsa-miR-519d hsa-miR-20b hsa-miR-519a hsa-miR-520c hsa-miR-526b* hsa-miR-519e
GGTGCTA	hsa-miR-29c hsa-miR-29b hsa-miR-29a
GTAAATA	PUF 6
GTACATA	PUF 6
GTATTTT	
GTGCAAT	hsa-miR-92b hsa-miR-367 hsa-miR-92 hsa-miR-363 hsa-miR-32 hsa-miR-25
GTGCCTT	hsa-miR-506 hsa-miR-124a
GTTATTT	
GTTTACA	hsa-miR-30a-5p hsa-miR-30b hsa-miR-30d hsa-miR-30e-5p hsa-miR-30c
TACTGTA	hsa-miR-101 hsa-miR-199a* hsa-miR-144
TATATGT	
TATTTAT	ARE
TATTTTG	

(continued)

(continued)

7-mer word	know regulators
TATTTTT	
TCAATAA	PolyA 5
TGCAATA	hsa-miR-92b hsa-miR-92 hsa-miR-32
TGCCTTA	hsa-miR-506 hsa-miR-124a
TGTAAAT	PUF 7
TGTACAT	PUF 7
TGTATAT	PUF 7
TGTGAAT	
TGTTTAC	hsa-miR-30a-5p hsa-miR-30b hsa-miR-30d hsa-miR-30e-5p hsa-miR-30c
TTATATT	hsa-miR-410
TTGCCTT	
TTGTAAA	PUF 6
TTGTATA	hsa-miR-381
TTGTATT	
TTTATAA	
TTTATAT	
TTTGCAC	hsa-miR-19b hsa-miR-19a
TTTGTAA	PUF 5
TTTGTAT	PUF 5
TTTTATA	
TTTTGTA	
TTTTGTT	
TTTTTAA	
TTTTTAT	

7-mer word
AAACTTG
AATCATG
ATATTTT
ATTAAAG
CTATTTT
GACCAAA
GTATTTT
GTTATTT
TATATGT
TATTTTG
TATTTTT
TGTGAAT
TTGCCTT
TTGTATT
TTTATAA
TTTATAT
TTTTATA
TTTTGTA
TTTTGTT
TTTTTAA
TTTTTAT