MicroRNAs

Artemis G. Hatzigeorgiou

Assistant Professor
Department of Genetics, Medical School
Department of Computational and Information Science, SEAS
Center for Bioinformatics
University of Pennsylvania
MicroRNA promoters

Transcription Factor Binding Sites → Pol-II → Primary transcript

A

pri-miR-23a~miR-27a~miR-24-2 (~2.2 kb)

B

Cleavage by Drosha

<table>
<thead>
<tr>
<th>Organism</th>
<th>UCSC Known Genes</th>
<th>Refseq Genes</th>
<th>Genscan Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>A. gambiae</em></td>
<td></td>
<td></td>
<td>16/37 (43.2%)</td>
</tr>
<tr>
<td><em>C. elegans</em></td>
<td></td>
<td>20/116 (17.2%)</td>
<td></td>
</tr>
<tr>
<td><em>C. familiaris</em></td>
<td>1/6 (16.7%)</td>
<td>2/6 (33.3%)</td>
<td></td>
</tr>
<tr>
<td><em>D. melanogaster</em></td>
<td>21/78 (26.9%)</td>
<td>26/78 (33.3%)</td>
<td></td>
</tr>
<tr>
<td><em>G. gallus</em></td>
<td>10/147 (6.8%)</td>
<td>81/147 (55.1%)</td>
<td></td>
</tr>
<tr>
<td><em>H. sapiens</em></td>
<td>166/466 (35.6%)</td>
<td>157/466 (33.7%)</td>
<td>237/466 (50.9%)</td>
</tr>
<tr>
<td><em>M. musculus</em></td>
<td>129/367 (35.1%)</td>
<td>117/367 (31.9%)</td>
<td>203/367 (55.3%)</td>
</tr>
<tr>
<td><em>P. troglodytes</em></td>
<td>21/65 (32.3%)</td>
<td>35/65 (53.8%)</td>
<td></td>
</tr>
<tr>
<td><em>R. norvegicus</em></td>
<td>27/228 (11.8%)</td>
<td>33/228 (14.5%)</td>
<td>117/228 (51.3%)</td>
</tr>
<tr>
<td><em>T. nigroviridis</em></td>
<td></td>
<td></td>
<td>52/143 (36.4%)</td>
</tr>
</tbody>
</table>

Table 1 displays the proportion of miRNAs in UCSC Known Genes, Refseq Genes, and Genscan Genes for each species.
A to I; from A-U to I-U wobble

Figure 1 The consequences of editing of pri-miR-142 by the nuclear editing enzymes, ADAR1p110 or ADAR2. Editing introduces inosines that can affect RNA structure, and this interferes with processing by Drosha. The edited pri-miRNA is then thought to be exported to the cytoplasm, where it is degraded by Tudor-SN.


Table 2 displays the proportion of miRNAs falling into clusters of size two or more for a sample collection of species. Cluster distance is the maximum distance between any two miRNAs considered to be in the same cluster.

<table>
<thead>
<tr>
<th>Organism</th>
<th>Cluster Distance 500 nt</th>
<th>Cluster Distance 1 kb</th>
<th>Cluster Distance 5 kb</th>
<th>Cluster Distance 50 kb</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>H. sapiens</em></td>
<td>99/466 (21.2%)</td>
<td>142/466 (30.5%)</td>
<td>204/466 (43.8%)</td>
<td>230/466 (49.4%)</td>
</tr>
<tr>
<td><em>M. musculus</em></td>
<td>107/367 (29.2%)</td>
<td>133/367 (36.2%)</td>
<td>169/367 (46.0%)</td>
<td>200/367 (54.5%)</td>
</tr>
<tr>
<td><em>R. norvegicus</em></td>
<td>73/228 (32.0%)</td>
<td>91/228 (39.9%)</td>
<td>112/228 (49.1%)</td>
<td>127/228 (55.7%)</td>
</tr>
<tr>
<td><em>G. gallus</em></td>
<td>39/147 (26.5%)</td>
<td>50/147 (34.0%)</td>
<td>65/147 (44.2%)</td>
<td>79/147 (53.7%)</td>
</tr>
</tbody>
</table>
• What is the most biologically meaningful region to search for regulatory factor binding sites?
EST-based approximation of human and mouse miRNA primary transcripts


High-throughput experimental investigation to define full-length transcripts
What is the longest segment that hybridizes to RNA from Drosha depleted cell lines?
The miRNA Promoter Problem

Where do we expect to find regulatory factor binding sites?

Search for regulatory factor binding sites

TSS

Protein-Coding Gene

miRNA primary transcript

3-7KB

Mature miRNA

Where to search

Mature miRNA
miRNA Promoter Element Discovery in Arabidopsis

Data
A published set of Transcription Start Sites (TSSs) for 52 miRNA primary transcripts identified in Arabidopsis via 5’-RACE (Carrington, Aug. 2005).

Project Goal
Are there any known Transcription Factor Binding Sites (TFBSs) which appear in a higher proportion of miRNA promoters than protein-coding gene promoters?
Comparison of binding site frequency

Putative TFBS

TSS

miRNA promoters

Protein-coding gene promoters

Randomly selected Arabidopsis sequences
Histograms of TATA-box binding site locations

miRNA promoter set
protein-coding gene promoter (PGP) set
set of randomly selected Arabidopsis genome sequences.
## Results

<table>
<thead>
<tr>
<th>TF Binding Site Motif</th>
<th>Count</th>
<th>Proportion of Sequences</th>
<th>Posterior Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>miRNA</td>
<td>miRNA</td>
</tr>
<tr>
<td>TATA-box</td>
<td>42</td>
<td>0.81</td>
<td>0.52</td>
</tr>
<tr>
<td>AtMYC2</td>
<td>14</td>
<td>0.27</td>
<td>0.17</td>
</tr>
<tr>
<td>ARF</td>
<td>14</td>
<td>0.27</td>
<td>0.17</td>
</tr>
<tr>
<td>SORLREP3</td>
<td>8</td>
<td>0.15</td>
<td>0.04</td>
</tr>
<tr>
<td>LFY</td>
<td>24</td>
<td>0.46</td>
<td>0.34</td>
</tr>
</tbody>
</table>

- Chart shows the proportion of miRNA promoters, protein-coding gene promoters, and random sequences which contain at least one observation of the given binding site motif.
Two miRNAs with putative ARF binding sites upstream, miR-160 and miR-167, have experimentally supported targets belonging to the ARF gene family.

Promoter Element Discovery in Arabidopsis

predicted secondary structure

comparative analysis
miRNA computational prediction pipeline

- Inverted repeats, composition
- RNA secondary structure prediction
- Energy + structural features
- Cross-species conservation
- SVM
- HMM
- Novel microRNAs: Microarray verification
- 2 851 352 871 bases
Prediction features

1. Stem_Length
2. GC_Content
3. Stem_BPs
4. maxLinHelix
5. MatureCons
6. MatureOppositeCons
7. ArmCons
8. SS_Energy
9. MatureBPs
10. MatureEnergyProfile

=> 10 features for SVM classification
Training the Classifier

Negative data
Hairpins extracted from 3' UTRs of human genes

Positive data
miRBase miRNA hairpin locations

RNA Secondary Structure Prediction

Energy and structural features of hairpins

Train SVM

miRNA hairpin classifier

Classify Test Set

miRNA hairpin SVM classifier performance
Combined miRNA gene prediction: DIANA-microG

Windows of 110 nt: calculate secondary structure

Structural features & conservation features

Classifier (SVM)

0.5

Ranking the results:
- 0.9
- 0.81
- 0.72 known miRNA

8,000 cand., verification with chip experiment.

Data analysis of experimental verification

Expression data from 8184 predicted miRNAs on 6 human tissues

Expression above 98% negative cutoff in any of 6 tissues:

Probe type : #above / total

---------------------------------------------

rRNA : 37 / 44 = 84.1 % } positive control

known ncRNA : 168 / 397 = 42.3 %

true miRNA : 368 / 1782 = 18.0 %

Our predictions : 2834 / 8197 = 33.4 %

Negative random : 10 / 500 = 2.0 % } negative control
microRNA IDENTIFICATION in ENCODE region

Computational approach based on machine learning approach (Support Vector Machine) predicts 8,000 new miRNA candidates.

Printing a chip with these predictions (2 X 30nt probes for each put. miRNA). 30% expressed in 6 tissues (thymus, placenta, lung, ovary, liver and brain)

- ENCODE MicroRNA PREDICTION \rightarrow 271 microRNA candidates
- MICROARRAY TESTING \rightarrow 54 microRNA candidates
- OVERLAP WITH TILING ARRAY EXPRESSION SIGNAL \rightarrow 16 microRNA candidates
Figure 1: Histogram Comparison of Intron Lengths Up to 10kb
These histograms compare the distribution of lengths of UCSC Known Introns containing a miRNA to the distribution of lengths of all UCSC Known Genes, for intron lengths up to 10kb.
miRGen: A database for the study of animal microRNA genomic organization and function

- Where are the miRNAs with respect to UCSC Known Genes?

- Which miRNAs fall into clusters at a given inter-miRNA distance?
Figure 1. The sequence database entry for hsa-miR-25. The three sections of the page describe the predicted stem-loop hairpin, mature sequences and primary references. The genomic coordinates and contextual information link to the Ensembl database. Each mature miRNA contains an evidence field, and links are provided to predicted target pages.

http://microrna.sanger.ac.uk/
## Translationally Repressed targets for Human

<table>
<thead>
<tr>
<th>miRNA</th>
<th>Gene</th>
<th>MRB</th>
<th>Single Site Sufficiency</th>
<th>Indirect Support</th>
<th>Direct Support</th>
<th>Paper</th>
<th>Binding</th>
</tr>
</thead>
<tbody>
<tr>
<td>miR-124</td>
<td>M1pa</td>
<td>1 site</td>
<td>Unknown</td>
<td>in vivo overexpression of mRNA</td>
<td>in vitro reporter gene assay (Luciferase) AND immunoblotting</td>
<td>Krek et al, 2005</td>
<td>Binding Pictures</td>
</tr>
<tr>
<td>miR-375</td>
<td>M1pa</td>
<td>1 site</td>
<td>Yes</td>
<td>in vitro 2'-O-methyl inhibition of miRNA AND in vitro overexpression of miRNA</td>
<td>in vitro reporter gene assay (Luciferase) AND immunoblotting</td>
<td>Foy et al, 2004</td>
<td>Binding Pictures</td>
</tr>
<tr>
<td>let-7b</td>
<td>M1pa</td>
<td>Not Given</td>
<td>Unknown</td>
<td>in vivo overexpression of mRNA</td>
<td>in vitro reporter gene assay (Luciferase)</td>
<td>Krek et al, 2005</td>
<td>Binding Pictures</td>
</tr>
<tr>
<td>let-7b</td>
<td>Lin28</td>
<td>1 site</td>
<td>Yes</td>
<td></td>
<td>in vitro reporter gene assay (Luciferase)</td>
<td>Kiriakidou et al, 2004</td>
<td>Binding Pictures</td>
</tr>
<tr>
<td>miR-141</td>
<td>Csk</td>
<td>1 site</td>
<td>Yes</td>
<td></td>
<td>in vitro reporter gene assay (Luciferase)</td>
<td>Kiriakidou et al, 2004</td>
<td>Binding Pictures</td>
</tr>
<tr>
<td>miR-24</td>
<td>MAPK14</td>
<td>1 site</td>
<td>Yes</td>
<td></td>
<td>in vitro reporter gene assay (Luciferase)</td>
<td>Kiriakidou et al, 2004</td>
<td>Binding Pictures</td>
</tr>
<tr>
<td>miR-145</td>
<td>FLJ21308</td>
<td>1 site</td>
<td>Yes</td>
<td></td>
<td>in vitro reporter gene assay (Luciferase)</td>
<td>Kiriakidou et al, 2004</td>
<td>Binding Pictures</td>
</tr>
<tr>
<td>miR-22a</td>
<td>FLJ13158</td>
<td>1 site</td>
<td>Yes</td>
<td></td>
<td>in vitro reporter gene assay (Luciferase)</td>
<td>Kiriakidou et al, 2004</td>
<td>Binding Pictures</td>
</tr>
<tr>
<td>let-7e</td>
<td>SMC1L1</td>
<td>1 site</td>
<td>Yes</td>
<td></td>
<td>in vitro reporter gene assay (Luciferase)</td>
<td>Kiriakidou et al, 2004</td>
<td>Binding Pictures</td>
</tr>
</tbody>
</table>


http://www.diana.pcbi.upenn.edu/tarbase.html
Current mammalian target prediction programs

• Widely used mammalian target prediction programs

  1. TargetScan (MIT, late 2003)
  2. DIANA-microT (UPENN, early 2004)
  3. MiRanda (Sloan-Kettering, 2004)
  4. TargetScanS (MIT, 2005)
  5. PicTar (NYU, 2005)

• Each program applies a slightly different set of “rules” that are thought to govern miRNA:target interactions

• How do these programs compare? What are the relative advantages of each? What are the limitations?

Evaluate mammalian target prediction programs

• Sensitivity
  • ~85 human/mouse miRNA:gene interactions that have direct experimental support for at least one target site (includes 32 different miRNAs)

• Specificity
  • How many total predictions are made by each program?
Minimum free energy alignment with a miRNA sequence

An algorithm based on dynamic programming is calculating the optimal path between each window and the miRNA.

The free energies of dinucleotide pairs are used as scoring matrix. Canonical base pairing and G-U wobbles are allowed. Loops and bulges have extra penalty.
and some heuristics …

• Compute for each miRNA a list of targets sorted after their minimum free energy

• Identification of targets conserved in human / mouse orthologs.

• Selection of 13 targets.
Performance Spectrum

- Union of all programs has 100% sensitivity, but predicts a total of ~30,000 miRNA:target interactions.
- Intersection of TSS and PT has 57% sensitivity, and predicts a total of ~7,000 miRNA:target interactions.
- Intersection of all programs has 0% sensitivity, but only predicts a total of 2 miRNA:target interactions.

Wide spectrum is largely due to differences in methodology for 3’-compensatory predictions.

miRNA:target interaction categories

• Perfect base pairing to at least 7 nucleotides starting from the first or second nucleotide at the 5’-end of the miRNA

• Binding to the 3’-end of the miRNA is currently considered irrelevant, but TarBase indicates that it is often extensive.

• Imperfect or shorter stretch of base pairing to at least 7 nucleotides starting from the first or second nucleotide at the 5’-end of the miRNA

• Extensive binding to the 3’-end of the miRNA in order to compensate for the weaker binding to the miRNA 5’-end
For a query-able interface to pre-compiled Diana-microT 2.0 target predictions for the human genome, please visit [here](http://diana.pcbi.upenn.edu/diana-microT2).
Resolving the molecular mechanisms of some polymorphic disease associations

- SNPs that occur in functional miRNA target sites could affect miRNA binding
- Map all annotated SNPs from dbSNP onto all experimentally supported target sites from TarBase
- 2 of the 5 SNPs occur in a region that disrupts the 5’-dominant binding
- 1 of these 2 SNPs is genotyped according to ALFRED (ALlele FREquency Database)
- Does this SNP impair miR-155 binding and silencing of AGTR1?
AGTR1: 1166A: 1166C: AGCATTAG... AGCCTTAG...

miR-155: TGA...TCGTAATT

Wild-type AGTR1 levels

Transl. repression

Elevated AGTR1 levels

No transl. repression
A newly identified role for miR-155 in hypertension

Experimental validation

• In vitro luciferase assay to test the prediction

qRT-PCR for mature miR-155 expression in fibroblast cells from monozygotic twins discordant for trisomy 21.

$P = 7.8 \times 10^{-5}$
Effect of other types of sequence variation

- SNPs do not map onto any known miRNAs
- But do miRNAs undergo A → I RNA editing?

pre-miR-376a

A → I editing can almost completely alter miRNA targeting activity (I pairs with G)

miRNAs and Cancer Specimens Studied

Human miRNA genes (207) identified from miRNA registry
http://www.sanger.ac.uk/Software/Rfam/mirna/index.html

A total of 253 human cancer specimens examined:

134 ovarian cancer specimens (107 primary tumors and 27 cell lines)
73 breast cancer specimens (55 primary tumors and 18 cell lines)
46 melanoma cell lines
miRNA DNA Copy Number Alterations in Human Cancer

miRNAs with Copy Number Changes Shared by 3 Cancer Types

22 miRNA genes with copy number gains and 7 with losses were shared by all three types of cancer.

38 miRNA genes had no copy number change in any of those cancer samples.
DIANA Lab

Molly Megraw, PhD cand.
Praveen Sethupathy, PhD cand.

And. Kouranov, PhD.
P. Fitziev, res. fellow
Benoit Corda, res. Fellow

Support
NSF Career Award Grant
NIH pre-doctoral training grant

Collaborators
Shane Jensen, Dep. of Statistics, Upenn
Krit. Kalandidis, IMBB, FORTH

Martin Reczko, ICS, FORTH
Timothy Hughes, University of Toronto
Aris Economidis, Regeneron Pharmaceutical

St. Antonarakis, University of Geneva
Ch. Borel, University of Geneva

Zis. Mourelatos, Dep. of Pathology, Upenn
Mar. Kiriakidou, Dep. of Pathology, Upenn
Pet. Nielson, Dep. of Pathology, UPenn

Qih. Huang, Wistar Institute, Philadelphia

George Coukos, Dep. of Gyn. Upenn

F. Pereira , A. Bernal
http://microrna.sanger.ac.uk/
http://www.diana.pcbi.upenn.edu/
http://pictar.bio.nyu.edu
http://genes.mit.edu/targetscan/
http://www.microrna.org/