Prediction of microRNAs and their targets

- Introduction
  - Brief history
  - miRNA Biogenesis

- Computational Methods
  - Mature and precursor miRNA prediction
  - miRNA target gene prediction

- Summary

Acknowledgments: Eric, Martin, Guy, Grace, Sabah, many Internet sources, and articles
a. DUPLEXES

b. SINGLE STRANDED REGIONS

c. HAIRPINS

HAIRPIN LOOP

HAIRPIN STEM

d. BULGES

BULGE

SINGLE-BASE BULGE

e. INTERNAL LOOPS

f. JUNCTIONS

MISMATCH

SYMMETRIC

INTERNAL LOOP

ASYMMETRIC

INTERNAL LOOP

THREE STEM

FOUR STEM
microRNAs?

- RNA can fold like proteins: possess primary, secondary and tertiary structure
- Secondary hairpin structure crucial to processing of small RNAs
microRNAs?

- ~22nt noncoding small RNAs
- mRNA stability and translation.
- Lin-4 (Lee et al. 1993)
- Let-7 (Reihart et al. 2000)
Junk to Nobel Prize

- 95% “Junk”, 5% proteins

Andrew Z. Fire & Craig C. Mello (2006)
The questions

- Can we predict microRNA genes?
  - ab initio / de novo
  - Homology

- Given a microRNA gene, can we find what genes they regulate, a.k.a targets?
Computational methods to identify miRNA genes: Why?

~500 human miRNAs to date, thousands across species.

However, experimental identification miRNAs is not easy:
  - low expression
  - stability
  - tissue specific
  - Expensive, and long cloning procedure

Predicting miRNAs from genomic sequences provide a valuable alternative/support to cloning.
How do we evaluate these predictions?

TP = True positives
TN = True negatives
FP = False positives
FN = False negatives

\[ TPR(\text{Sensitivity}) = \frac{TP}{TP + FN} \]
\[ FPR(1 - \text{Specificity}) = \frac{FP}{FP + TN} \]
miRNA prediction – Initial methods

MiRscan
find conserved hairpin structures known miRNAs (50) as training set.

Lim et al, Genes and Development 2003

*Human genome (109/109)*
 Align (BLAT annotations),
 Remove protein-coding genes

Mouse genome

Noncoding conserved regions (102/109)

RNAfold

800,000 human stem-loops (102/109)

MiRscan analysis, Retain top 10%

80,000 conserved stem-loops (112/109)

168 miRNA gene candidates with scores > 10.0
(31/109)

MiRscan analysis

16,133 aligned stem-loops (91/109)

Align

Fugu genome

*Lim et al, Genes and Development 2003*
Blue: distribution of MiRscan score of 35,697 sequences
Red: training set
Yellow and purple are verified by cloning or other evidence.
70% Specificity and 50% sensitivity
Comparative Genomics
HMM-based ProMiR

Human microRNA prediction through a probabilistic co-learning model of sequence and structure

Jin-Wu Nam\textsuperscript{1,2}, Ki-Roo Shin\textsuperscript{3}, Jinju Han\textsuperscript{4}, Yoontae Lee\textsuperscript{4}, V. Narry Kim\textsuperscript{4} and Byoung-Tak Zhang\textsuperscript{1,2,3,*}

\textsuperscript{1}Graduate Program in Bioinformatics, \textsuperscript{2}Center for Bioinformation Technology (CBIT), \textsuperscript{3}Biointelligence Laboratory, School of Computer Science and Engineering and \textsuperscript{4}Department of Biological Sciences, Seoul National University, Seoul 151-744, Korea

\[ T_{kl} = P(\pi_i = l \mid \pi_{i-1} = k), \]
\[ E_k(b) = P(x_i = b \mid \pi_i = k). \]
\[ P(x, \pi) = T_0 \prod_{i=1}^{L} E_{\pi_i}(x_i) T_{\pi_i \pi_{i+1}}, \]
\[ \pi^* = \arg \max_{\pi} P(x, \pi). \]
Results of ProMiR

ProMiR: 96% Specificity, 73% sensitivity
miRScan: 70% Specificity and 50% sensitivity
BayesMiRNAfind

(1) # of base pairs.
(2) # of bulges.
(3) # of loops
(4) # of asymmetric loops.
(5) # of bulges of various lengths
(6) # of asymmetric features
(7) Distance of miRNA from foot & loop
(8) Nucleotide sequence ‘words’ with lengths 4–9 are extracted from the candidate 21 nt sequence
C. Elegans (nematode worm)                          Mouse

(1) # of base pairs.  
(2) # of bulges.  
(3) # of loops  
(4) # of asymmetric loops.  
(5)# of bulges of various lengths  
(6) # of asymmetric features  
(7) Distance of miRNA from foot & loop  
(8) Nucleotide sequence ‘words’ with lengths 4–9 are extracted from the candidate 21 nt sequence
MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features

Peng Jiang, Haonan Wu, Wenkai Wang, Wei Ma, Xiao Sun and Zuhong Lu

State Key Laboratory of Bioelectronics, Department of Biological Science and Medical Engineering, Southeast University, Nanjing, 210096, P. R. China

**miPred:** 93.21% Specificity, 89.4 sensitivity  
**ProMiR:** 96% Specificity, 73% sensitivity  
**miRScan:** 70% Specificity, 50% sensitivity
Training
Random Forest

N cases in training set, M input variables

- Sample N cases at random, with replacement, from the original data. This sample will be the training set for growing the tree.
- At each node, m variables (m << M) are selected at random out of the M and the best split on these m is used to split the node. The value of m is held constant during the forest growing.
- Each tree is grown to the largest extent possible. There is no pruning.
miPred-- Random Forest

- Trained on RFAM data set of 60 cloned miRNAs and random negative set (250 putative miRNA hairpins) with a variety of features
- Independently construct 500 trees
- MFE—Minimum free energy value
- P—Randomized sequences to evaluate MFE
- Structure composition

<table>
<thead>
<tr>
<th>Rank</th>
<th>Features</th>
<th>Mean decrease accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>P-value</td>
<td>15.80</td>
</tr>
<tr>
<td>2</td>
<td>MFE</td>
<td>5.48</td>
</tr>
<tr>
<td>3</td>
<td>C ...</td>
<td>2.04</td>
</tr>
<tr>
<td>4</td>
<td>U</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>A ...</td>
<td>0.83</td>
</tr>
<tr>
<td>7</td>
<td>G ...</td>
<td>0.76</td>
</tr>
<tr>
<td>8</td>
<td>U..</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>G.</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>A</td>
<td>..</td>
</tr>
</tbody>
</table>

**Diagram:**

```
precursor

<table>
<thead>
<tr>
<th>G</th>
<th>G</th>
<th>GAGG</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>C</td>
<td>UUCU</td>
</tr>
<tr>
<td>C</td>
<td>U</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GUAGGUGUAUAGUU</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>UGGGGCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>UCCCCG</td>
</tr>
</tbody>
</table>

| UAGGGUAUC |
```
Other strategies

- RNA22 – pattern based
- miRank – Random walks
- RNAmicro – Support Vector Machine
- Reverse motif search w conservation analysis – PMID: 15735639 (Xie et al) and others
- miRNAminer – Sequence similarity based
miRNA target prediction = ?

Predicting which genes are regulated by which miRNA(s)?
Common strategy
Method: TargetScan

1. Use 7 nt segment of the miRNA as the ‘microRNA seed’ to find the perfect complementary motifs in the UTR regions.
2. Extend each seeds to find the best energy
3. Assign a score, Z.
4. Rank Give a rank \((R_i)\) according to that species.
5. Repeat above process.
6. Keep those genes for which \(Z_i > Z_c\) and \(R_i < R_c\).
TargetScan

- Signal to noise ratio
  - Don’t have a large training set
  - Estimate of false positive
RNAhybrid

Fast and effective prediction of microRNA/target duplexes

MARC REHMSMEIER, PETER STEFFEN, MATTHIAS HÖCHSMANN, and ROBERT GIEGERICH

1Universität Bielefeld, International NRW Graduate School in Bioinformatics and Genome Research, 33501 Bielefeld, Germany
2Universität Bielefeld, Technische Fakultät, Praktische Informatik, 33501 Bielefeld, Germany

Cross-hybridization Good Cross-hybridization Good
RNAhybrid MFE statistics

- MFE is an optimized score
  - EVD – applicable to max/min of independent random variables
  \[
  Z = -\frac{\text{MFE}}{\log(mn)}
  \]
  \[
  P(Z \leq t) = \exp(-e^{-(t-\varepsilon)/\theta})
  \Rightarrow t = \varepsilon - \theta \log(-\log(P))
  \]
- Regression analysis yields $\varepsilon$ and $\theta$
- Statistical significance $1-P$
## Resources (miRNA target prediction)

<table>
<thead>
<tr>
<th>Method</th>
<th>Organism</th>
<th>Website</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precomputed predictions on searchable websites</td>
<td></td>
<td></td>
</tr>
<tr>
<td>miRNA target predictions at EMBL</td>
<td>Flies</td>
<td><a href="http://www.russell.embl-heidelberg.de/miRNAs/">http://www.russell.embl-heidelberg.de/miRNAs/</a></td>
</tr>
<tr>
<td>miRanda</td>
<td>Flies, vertebrates</td>
<td><a href="http://www.microrna.org/miranda.html">http://www.microrna.org/miranda.html</a></td>
</tr>
<tr>
<td>mirBase</td>
<td>Vertebrates, insects, nematodes</td>
<td><a href="http://microrna.sanger.ac.uk/targets/v2/">http://microrna.sanger.ac.uk/targets/v2/</a></td>
</tr>
<tr>
<td>PicTar</td>
<td>Vertebrates, flies, nematodes</td>
<td><a href="http://pictar.bio.nyu.edu">http://pictar.bio.nyu.edu</a></td>
</tr>
<tr>
<td>TargetScan, TargetScanS</td>
<td>Vertebrates</td>
<td><a href="http://genes.mit.edu/targetscan">http://genes.mit.edu/targetscan</a></td>
</tr>
<tr>
<td>Ref. 27</td>
<td>Flies, nematodes</td>
<td><a href="http://tavaziolab.princeton.edu/mirnas/">http://tavaziolab.princeton.edu/mirnas/</a></td>
</tr>
<tr>
<td>RNA hybrid</td>
<td>Flies</td>
<td><a href="http://www.techfak.uni-bielefeld.de/persons/marc/mirna/targets/drosophila">http://www.techfak.uni-bielefeld.de/persons/marc/mirna/targets/drosophila</a></td>
</tr>
</tbody>
</table>

### Tools for locating miRNA targets

- RNAhybrid
- DIANA-MicroT
- RNA22

### Databases of targets with experimental support

- Tarbase
- Argonaute
- miRNAMAP

For other published miRNA target predictions, see ref. 28 (nematodes), ref. 47 (*D. melanogaster*) and ref. 20 (vertebrates).
Additional Resources for miRNAs

- miRBase data base of microRNAs
- Ensembl database – UTRs, gene regions, etc
- UCSD genome browser – genomes, conservation
- GEO – Gene expression Omnibus
miRNA target recognition

- One target, multiple miRNA (Cooperative interaction)
- One miRNA, multiple targets (Multiplicity/Promiscuity)
- 5’ end sequence is important
- Structural accessibility is important
- Lots of predicted targets, which ones are important remains a question