TarMiner

Automatic extraction of miRNA targets from literature

Rodothea-Myrsini Tsoupidi - IMIS, ‘Athena’ RC, Greece

Ilias Kanellos - IMIS, ‘Athena’ RC & ECE, NTUA, Greece

Thanasis Vergoulis - IMIS, ‘Athena’ RC, Greece

Ioannis S. Vlachos - Univ. of Thessaly & UoA, Greece

Artemis G. Hatzigeorgiou - Univ. of Thessaly, Greece

Theodore Dalamagas - IMIS, ‘Athena’ RC, Greece
miRNAs

- Small, non-protein coding RNAs
- Target genes, inhibit protein production
- Related to:
  - Cancer
  - Heart diseases
  - Autoimmune diseases
- Numerous databases for miRNA-gene interactions:
  - mirTarBase
  - tarBase
  - miRecords
Motivation

Issues:
• Curated database maintenance
• Time-consuming process

Goal:
• Automatic miRNA-gene interaction extraction from text

TarMiner exploits:
• Classifier on full texts
• Training on curated data
• Support for many species

I wish I could spend all this time reading in the lab instead...
Workflow
Workflow

miRNAs
• Simple grammar
  • Species prefix
  • <mir/mirna/microRNA/let/lsy/lin>
  • Suffix – identifier
• Found through regular expressions

Genes
• Different ids based on source DB
• We combine different dictionaries
• Tools for synonym identification (mygene, NCBI eUtils)

miRNA names:
- hsa-let-7a-3p
- mmu-mir-1b
- dre-mir-1
Workflow

1. Text processing

“mir-127 post-transcriptionally downregulates Sept7 and suppresses cell growth in hepatocellular carcinoma cells.”

“MIRNA (NN / B-NP) post-transcriptionally (RB / O) downregulate (VBZ / B-VP) GENE (NN /B-NP) suppress (VBZ / B-VP) cell (NN / B-NP) growth (NN / B-NP) in (IN / B-PP) hepatocellular (JJ / B-NP) carcinoma (NN/ I-NP) cell (NN / I-NP).”
Workflow

1. Text processing
2. miRNA-gene pair extraction
Workflow

1. Text processing
2. miRNA-gene pair extraction
3. Optional: pair filtering

“m1 ... g1 ... m2 ... m3 ... g2 ... g3”
“m2 ... g1 ... m3 ... g2”
“m3 ... g2 ... g3”
Workflow

1. Text processing
2. miRNA-gene pair extraction
3. Optional: pair filtering

“m1 ... g1 ... m2 ... m3 ... g2 ... g3”
“m2 ... g1 ... m3 ... g2”
“m3 ... g2 ... g3”
Workflow

1. Text processing
2. miRNA-gene pair extraction
3. Optional: pair filtering
4. Feature Vectors

“In addition, we identified that mir-193 and mir-497 could directly regulate TARBP2 and DICER expression in ACC cells.”

V={ 0,0,1,…,0,1,0,…1,1,0,…,0,1,1,…,0,1,0 }
Workflow

1. Text processing
2. miRNA-gene pair extraction
3. Optional: pair filtering
4. Feature Vectors

“In addition, we identified that mir-193 and mir-497 could directly regulate TARBP2 and DICER expression in ACC cells.”

\[ V = \{ 0,0,1,\ldots,0,1,0,\ldots,1,1,0,\ldots,0,1,1,\ldots,0,1,0 \} \]

Dictionary = \{ target, repress, regulate, ... \}

- Dictionary
- Dependency
- Phrase Chunk
- Location
- Mixed
Workflow

1. **Text processing**
2. **miRNA-gene pair extraction**
3. Optional: pair filtering
4. **Feature Vectors**

“In addition, we identified that mir-193 and mir-497 could directly regulate TARBP2 and DICER expression in ACC cells.”

\[ V = \{ 0,0,1,\ldots,0,1,0,\ldots,1,1,0,\ldots,0,1,1,\ldots,0,1,0 \} \]

NP-following = \{ GENE-repress, GENE-express, \ldots \}
Workflow

1. **Text processing**

2. **miRNA-gene pair extraction**

3. **Optional: pair filtering**

4. **Feature Vectors**

“In addition, we identified that mir-193 and mir-497 could directly regulate TARBP2 and DICER expression in ACC cells.”

\[ V = \{ 0, 0, 1, \ldots, 0, 1, 0, \ldots, 1, 1, 0, \ldots, 0, 1, 1, \ldots, 0, 0 \} \]

- **Dictionary**
- **Dependency**
- **Phrase Chunk**
- **Location**
- **Mixed**

Pairs before = \{ [show that MIRNA, identified that MIRNA, directly regulate GENE, found that GENE ... ] \}
Workflow

Trained Classifier:

• Input
  • miRNA-gene pair
  • Pair’s feature vector

• Output
  • True or False
  • Based on calculated probability
Workflow

Training:
• Curated data
• MegaM software trains a max entropy model
  • Calculate optimal feature weights
• Exclude papers with more than \textit{maxInter} interactions
• Large number of interactions not likely in text:
  • Figures
  • Tables
Preliminary Evaluation

Training Data:
• PMC full texts from curated TarBase v7 data
• 5-fold cross validation
• 2/3 training set, 1/3 evaluation set
• Measure:
  • Precision
  • Recall
  • F-measure
• Varying:
  • MinSent
  • MaxInter

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td># Publications</td>
<td>1.236</td>
</tr>
<tr>
<td># verified interactions</td>
<td>2.869</td>
</tr>
</tbody>
</table>
Results

- High precision
- High recall for small maxInter
- Recall drop:
  - More interactions likely not presented in text
  - Author-specific notation used for grouping
- minSent improves precision in some cases

<table>
<thead>
<tr>
<th>minSent/maxInter</th>
<th>1</th>
<th>5</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.8168</td>
<td>0.7901</td>
<td>0.7610</td>
</tr>
<tr>
<td>2</td>
<td>0.8614</td>
<td>0.8196</td>
<td>0.8197</td>
</tr>
<tr>
<td>3</td>
<td>0.8461</td>
<td>0.8151</td>
<td>0.8096</td>
</tr>
<tr>
<td>4</td>
<td>0.8157</td>
<td>0.8443</td>
<td>0.8081</td>
</tr>
</tbody>
</table>

**Precision**

<table>
<thead>
<tr>
<th>minSent/maxInter</th>
<th>1</th>
<th>5</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.7698</td>
<td>0.6103</td>
<td>0.5069</td>
</tr>
<tr>
<td>2</td>
<td>0.7618</td>
<td>0.5784</td>
<td>0.5184</td>
</tr>
<tr>
<td>3</td>
<td>0.7346</td>
<td>0.5941</td>
<td>0.5444</td>
</tr>
<tr>
<td>4</td>
<td>0.7547</td>
<td>0.5991</td>
<td>0.5317</td>
</tr>
</tbody>
</table>

**Recall**

<table>
<thead>
<tr>
<th>minSent/maxInter</th>
<th>1</th>
<th>5</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.7924</td>
<td>0.6883</td>
<td>0.6081</td>
</tr>
<tr>
<td>2</td>
<td>0.8080</td>
<td>0.6780</td>
<td>0.6345</td>
</tr>
<tr>
<td>3</td>
<td>0.7863</td>
<td>0.6871</td>
<td>0.6505</td>
</tr>
<tr>
<td>4</td>
<td>0.7834</td>
<td>0.7004</td>
<td>0.6411</td>
</tr>
</tbody>
</table>

**F-measure**
Conclusion

- TarMiner: automated tool for miRNA-gene interaction extraction
- Applies NLP and classifier
- Real curated data used for training
- 5-fold cross validation
- Future experiments:
  - Evaluate results for interactions of publications not in TarBase
  - Expand extraction to non-textual data (e.g. tables etc.)
Thank you!
miRNA grammar

\[ M = \{ \text{mir}', \text{let}', \text{lsy}', \text{lyn}' \} \]
\[ \text{conn} = \{ \text{and}', \text{or}', ']', ',' \} \]
\[ \text{ext} = \text{known miRNA name extension} \]
Dependency Graph Example
Future Evaluations

- MirPub: miRNA publication search engine
- Tarbase publications for 2013: ~40
  - Most interactions in 3 pubs with table data
  - The rest probably in text
- TarMiner on MirPub 2013:
  - ~3700 articles total
  - ~1800 with miRNA-gene pairs in sentences
  - ~400 pubs with a possible positive classification
- Awaiting evaluation from experts
- Combine TarMiner and mirPub for quicker updates of TarBase