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

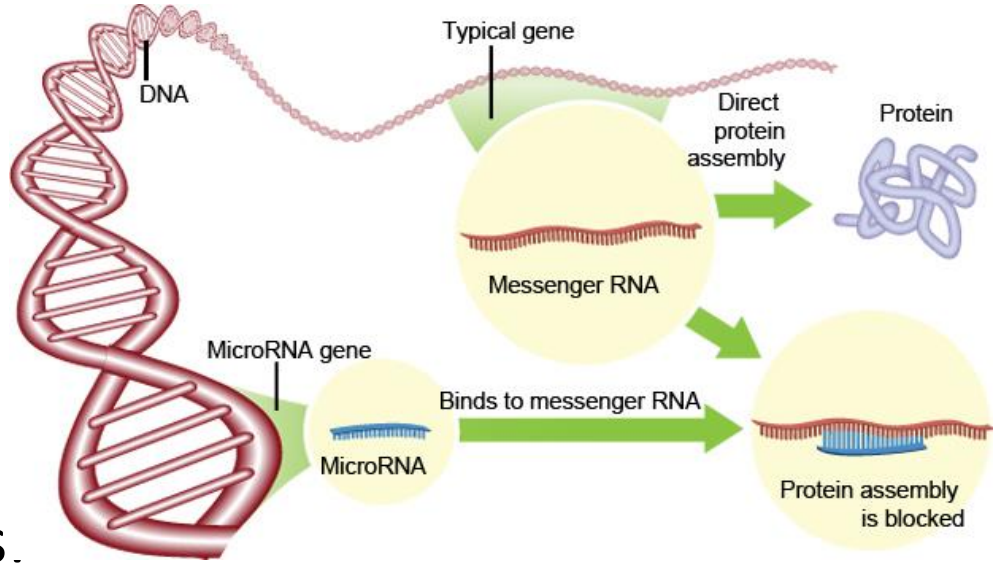
Ioannis S. Vlachos - Univ. of Thessaly & UoA, Greece

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:

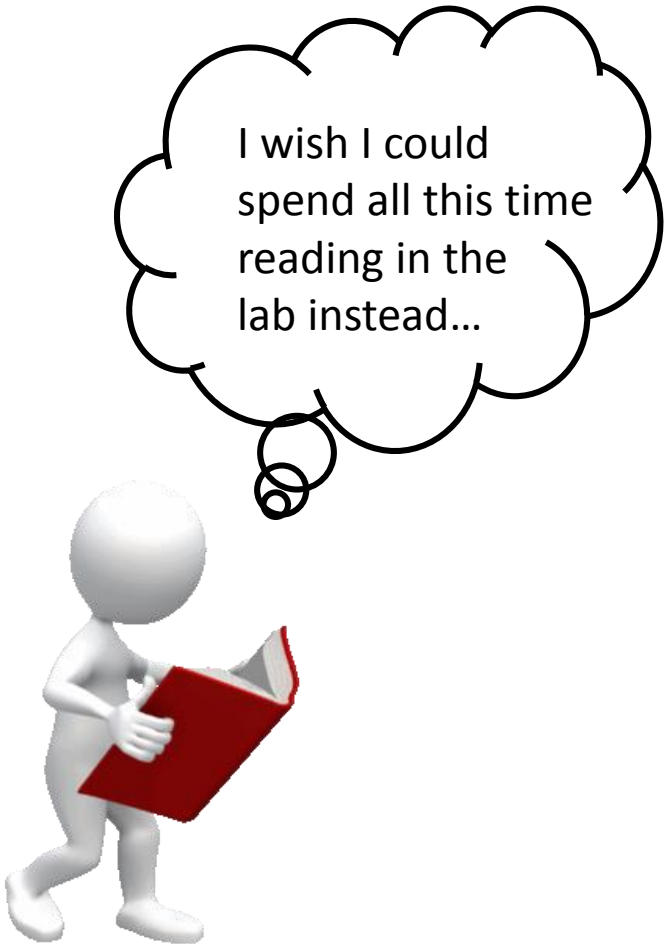
- Curated database maintenance
- Time-cosuming process

Goal:

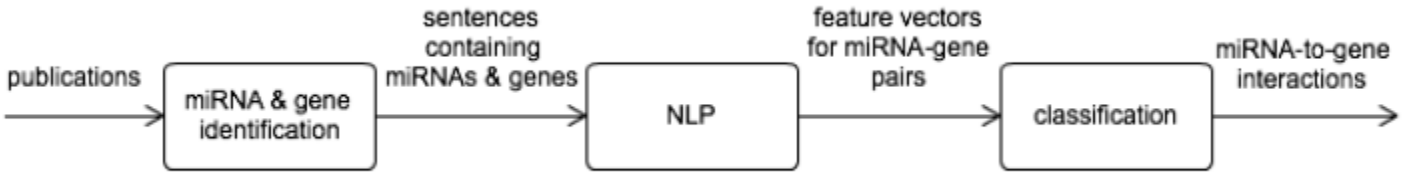
- Automatic miRNA-gene interaction extraction from text

TarMiner exploits:

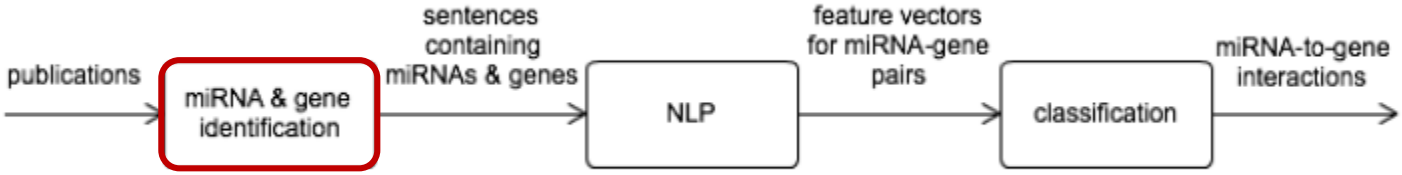
- Classifier on full texts
- Training on curated data
- Support for many species



Workflow


publications
miRNA & gene identification
sentences containing miRNAs & genes
NLP
feature vectors for miRNA-gene pairs
classification
miRNA-to-gene interactions

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

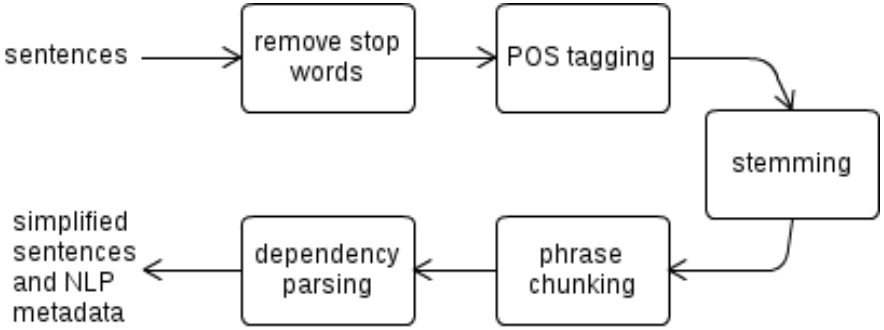
1. Text processing

"mir-127 post-transcriptionally downregulates Sept7 and suppreses 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)."

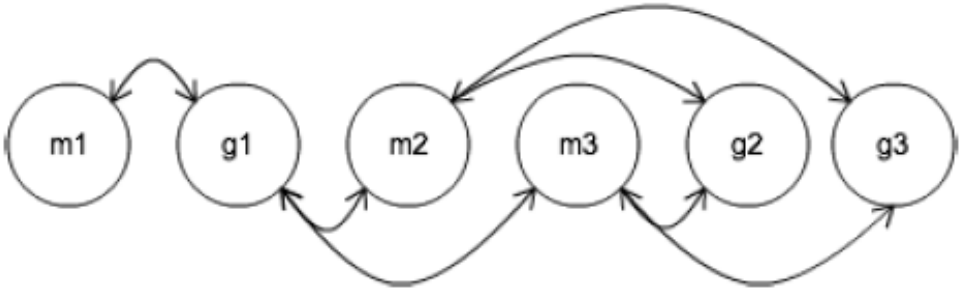
Noun, singular

Intermediate of noun phrase



Workflow

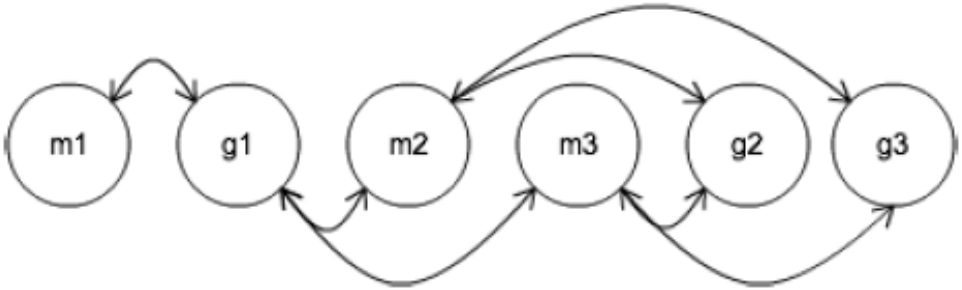
1. Text processing
2. miRNA-gene pair extraction



“m1 … g1 … m2 … m3 … g2 … g3”

Workflow

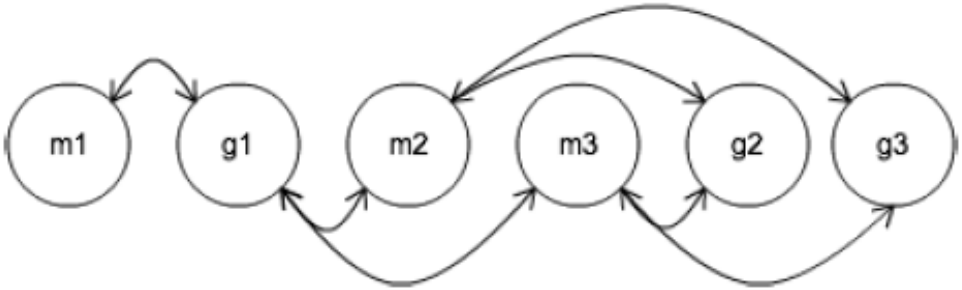
1. Text processing
2. miRNA-gene pair extraction
3. Optional: pair filtering



“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

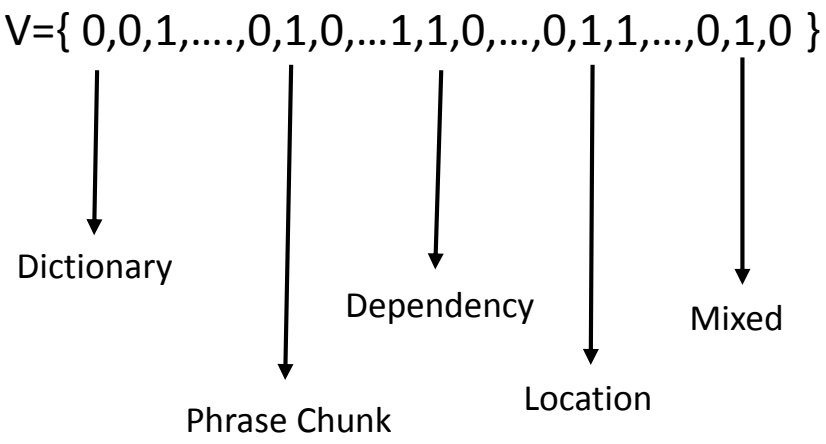


“m1 … g1 … m2 … m3 … g2 … g3”
“m2 … g1 … m3 … g2”
“m3 … g2 … g3”

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."



Workflow

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

Dictionary = { target, repress, regulate, … }

"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 }

Dictionary

Phrase Chunk

Dependency

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,….,0,1,0,…1,1,0,…,0,1,1,…,0,1,0 }

Dictionary

Phrase Chunk

Dependency

Location

Mixed

NP-following = { GENE-repress, GENE-express, … }

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 }

Dictionary

Phrase Chunk

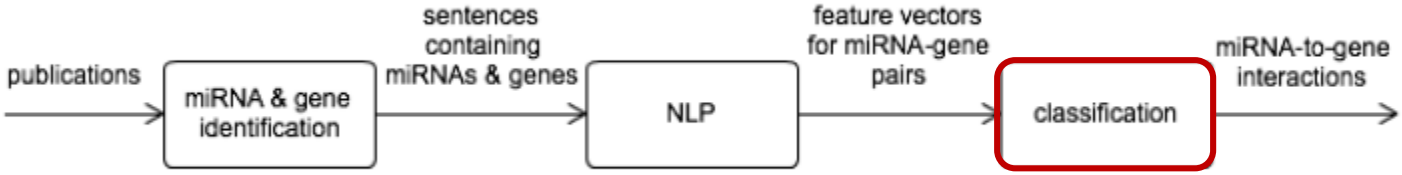
Dependency

Location

Mixed

Pairs before = { [show that MIRNA, identified that MIRNA, directly regulate GENE, found that GENE … }

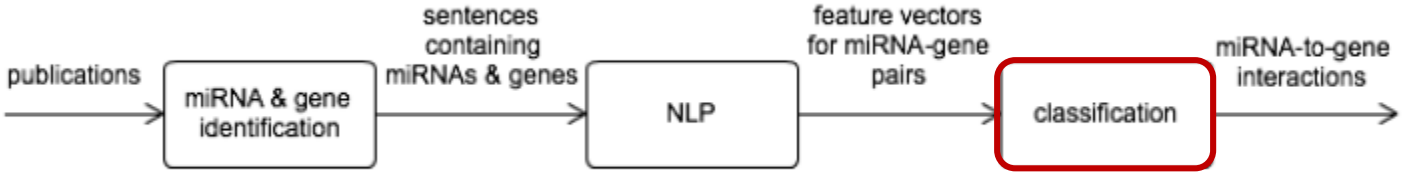
Workflow



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 *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

Characteristic	Value
# Publications	1.236
# verified interactions	2.869

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

minSent/maxInter	1	5	10
1	0.8168	0.7901	0.7610
2	0.8614	0.8196	0.8197
3	0.8461	0.8151	0.8096
4	0.8157	0.8443	0.8081

Precision

minSent/maxInter	1	5	10
1	0.7698	0.6103	0.5069
2	0.7618	0.5784	0.5184
3	0.7346	0.5941	0.5444
4	0.7547	0.5991	0.5317

Recall

minSent/maxInter	1	5	10
1	0.7924	0.6883	0.6081
2	0.8080	0.6780	0.6345
3	0.7863	0.6871	0.6505
4	0.7834	0.7004	0.6411

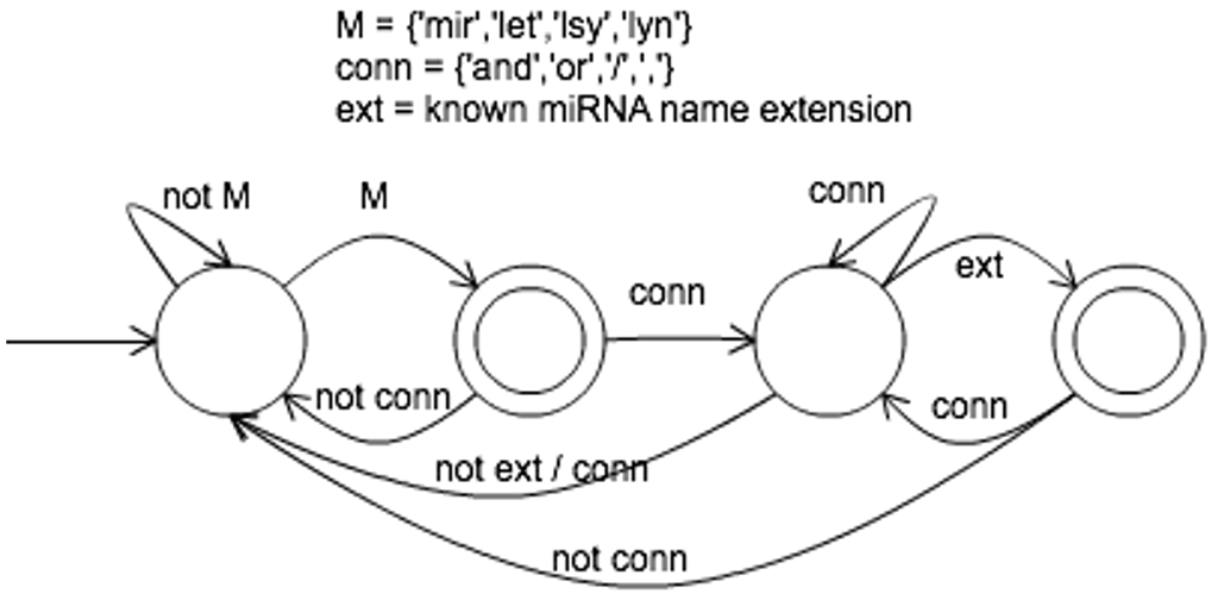
F-measure

Conclusion

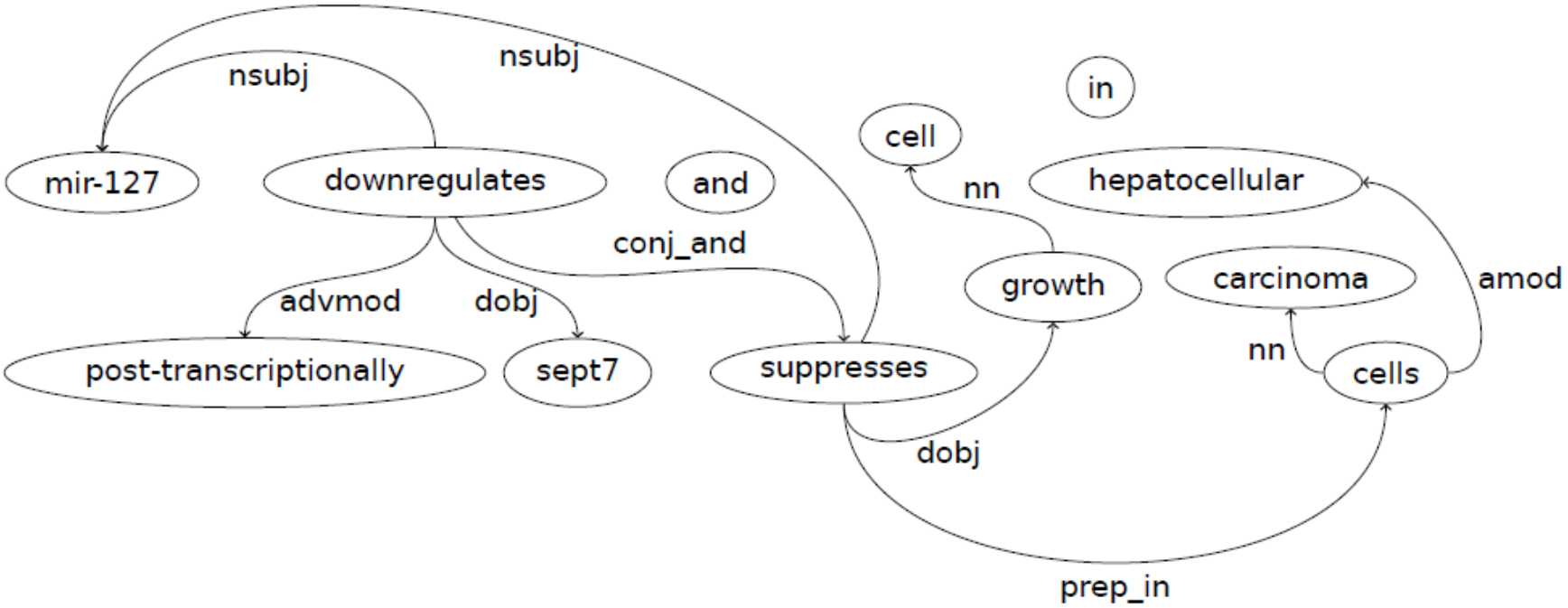
- TarMiner: automated tool for miRNA-gene interaction extraction
- Applies NLP and classifier
- Real curated data used for training
- 5-fold cross validation
- Future experiments:
  - Evaluate results for interactions of publications not in TarBase
  - Expand extraction to non-textual data (e.g. tables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

Thank you!

miRNA grammar



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