Active Learning for microRNA Prediction

Mohsen Sheikh Hassani & Dr. James R Green
Carleton University
BIBM 2018, Madrid

December 4th, 2018
MicroRNA (miRNA)

- Short non-coding RNAs
- Typically 18-25 nucleotides
- First miRNA discovered in 1993 (roundworms)
- Next discovery was in 2000
- Today, thousands of known miRNA
Why are miRNA important?

- Through gain- and loss-of-function experiments, evidence shows miRNA regulate the expression of proteins involved in:
  - biological development
  - cell differentiation
  - cell cycle control
  - stress response
  - Related to diseases: cancer, neurological disorders, heart disease

- Predicted to regulate over 60% of transcripts in humans

- May target 60-90% of all mammalian mRNA
Biogenesis

- The biogenesis mechanism plays a key role in miRNA identification
- Either transcribed regions of RNA or introns (pri-miRNA) fold into hairpins
- Cleaved by enzymes called Drosha in nucleus to ~80 nt ds (pre-miRNA)
- Exported to cytoplasm (via Exportin-5 and RanGTP)
- Processed by Dicer (loop cut off) to ~20 bp
- Two strands of mature miRNA:
  - One strand: Incorporated into miRNA-induced silencing complex (miRISC)
  - Other: Released and degraded
Gene regulation

- Exact means of miRNA silencing remains unclear.
- Evidence supports two distinct mechanisms:
  - mRNA degradation: miRNA bind to mRNA and promote degradation
  - Translation inhibition: miRNA bind to mRNA and prevent translation
miRNA identification

- Requires interdisciplinary strategies; integration of experimental approaches with computational methods

- Computational methods are used to predict, experimental methods are used to validate

- Broadly categorized as either de novo miRNA prediction (sequence based) or NGS-based (expression-based)
Computational miRNA prediction

- **De novo**: sequences extracted from genomic data set are classified based on sequence properties
  - Example: look at windows of triplet nts (also single/dinucleotides), how often specific combinations appear
Computational miRNA prediction

- NGS: Predictions made based on patterns of read depth
  - Example: statistics of the read positions and frequencies of the reads
  - Mature sequences are more abundant in the cell → sequenced more frequently
Motivation

- miRNA are critical to our understanding of biological processes
  - Identifying greater numbers = better understanding
  - Inter-disciplinary, identification of miRNA remains a difficult task

- Abundance of unlabeled data, scarcity of labeled examples for many species
  - New NGS methods provide large unlabeled data sets

- Existing methods of miRNA prediction require lots of known samples (supervised)

- We wish to extract the most information from limited labelled and available unlabeled data
Problem Statement

- Explore the application of semi-supervised learning (active learning) to miRNA prediction in order to leverage both labelled and unlabelled data.

- Expected Benefits:
  - Require smaller labelled training sets
  - Applicable to more species
  - More value from wet-lab validation experiments
Active Learning

- A semi-supervised machine learning approach
- Interactively query the user
- Suitable when labeling data is expensive
- Minimizes the overall cost of developing a predictor
Data Set Creation

- NGS expression data
- Known miRNA
- Known functional non-coding RNA
- Genomic data
- Known coding regions
Training Data Preparation

- Candidate pre-miRNA that map to known miRNA from miRbase → True positive
- Candidates not identified as miRNA are aligned to coding region data
- Candidates aligning with at most two mismatches are selected as negative samples + known non-coding RNA

<table>
<thead>
<tr>
<th>Data set</th>
<th># of positive samples</th>
<th># of negative samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>hsa (human)</td>
<td>509</td>
<td>842</td>
</tr>
<tr>
<td>mmu (mouse)</td>
<td>367</td>
<td>844</td>
</tr>
<tr>
<td>dme (fruit-fly)</td>
<td>110</td>
<td>97</td>
</tr>
<tr>
<td>bta (cow)</td>
<td>332</td>
<td>650</td>
</tr>
<tr>
<td>gga (chicken)</td>
<td>193</td>
<td>104</td>
</tr>
<tr>
<td>eca (horse)</td>
<td>364</td>
<td>224</td>
</tr>
</tbody>
</table>
Active Learning Pipeline

- Test/train data split (20%-80%)
- Feature set selection (13-6)
- Initial training set size (10 samples)
- Classifier selection (RF)
- Stopping criterion (11 iterations)

- Query strategy
  - How to spend validation budget?
  - Certainty-based
  - Uncertainty-based
### Results

- **Active Learning**
  - Certainty-based active learning
  - Uncertainty-based active learning

- **Baseline methods**
  - Self-training
  - Passive learning

<table>
<thead>
<tr>
<th>Data set</th>
<th>Self-training average AUPRC</th>
<th>Passive learning average AUPRC</th>
<th>Certainty based average AUPRC</th>
<th>Uncertainty based average AUPRC</th>
</tr>
</thead>
<tbody>
<tr>
<td>hsa</td>
<td>0.788 (+13.1%)</td>
<td>0.789 (+13.2%)</td>
<td>0.797 (+14.4%)</td>
<td>0.875 (+25.7%)</td>
</tr>
<tr>
<td>mmu</td>
<td>0.909 (-0.50%)</td>
<td>0.924 (+1.16%)</td>
<td>0.938 (+2.69%)</td>
<td>0.972 (+6.37%)</td>
</tr>
<tr>
<td>dme</td>
<td>0.896 (-1.68%)</td>
<td>0.914 (+0.30%)</td>
<td>0.917 (+0.66%)</td>
<td>0.924 (+1.44%)</td>
</tr>
<tr>
<td>bta</td>
<td>0.879 (+3.36%)</td>
<td>0.867 (+1.89%)</td>
<td>0.921 (+8.25%)</td>
<td>0.935 (+9.90%)</td>
</tr>
<tr>
<td>gga</td>
<td>0.903 (+1.31%)</td>
<td>0.886 (-0.60%)</td>
<td>0.915 (+2.67%)</td>
<td>0.944 (+6.01%)</td>
</tr>
<tr>
<td>eca</td>
<td>0.956 (+1.39%)</td>
<td>0.954 (+1.17%)</td>
<td>0.968 (+2.67%)</td>
<td>0.971 (+2.95%)</td>
</tr>
<tr>
<td>Avg.</td>
<td>+ 2.83%</td>
<td>+2.86%</td>
<td>+5.23%</td>
<td>+8.72%</td>
</tr>
</tbody>
</table>

![Graphs](15)
Results - continued

<table>
<thead>
<tr>
<th>Data set</th>
<th>Sequence-based average AUPRC</th>
<th>Expression-based average AUPRC</th>
<th>Integrated (miPIE) average AUPRC</th>
<th>miRDeep2 average AUPRC</th>
<th>Active learning average AUPRC</th>
</tr>
</thead>
<tbody>
<tr>
<td>hsa</td>
<td>0.763 (±0.02)</td>
<td>0.789 (±0.01)</td>
<td>0.844 (±0.01)</td>
<td>0.736</td>
<td>0.875 (±0.01)</td>
</tr>
<tr>
<td>mmu</td>
<td>0.907 (±0.01)</td>
<td>0.939 (±0.01)</td>
<td>0.966 (±0.01)</td>
<td>0.915</td>
<td>0.972 (±0.00)</td>
</tr>
<tr>
<td>dme</td>
<td>0.918 (±0.01)</td>
<td>0.893 (±0.01)</td>
<td>0.894 (±0.01)</td>
<td>0.914</td>
<td>0.924 (±0.01)</td>
</tr>
<tr>
<td>bta</td>
<td>0.890 (±0.02)</td>
<td>0.865 (±0.02)</td>
<td>0.905 (±0.02)</td>
<td>0.869</td>
<td>0.935 (±0.01)</td>
</tr>
<tr>
<td>gga</td>
<td>0.886 (±0.02)</td>
<td>0.906 (±0.01)</td>
<td>0.919 (±0.01)</td>
<td>0.923</td>
<td>0.944 (±0.01)</td>
</tr>
<tr>
<td>eca</td>
<td>0.886 (±0.01)</td>
<td>0.906 (±0.01)</td>
<td>0.919 (±0.01)</td>
<td>0.843</td>
<td>0.971 (±0.00)</td>
</tr>
<tr>
<td>Avg.</td>
<td><strong>0.875</strong></td>
<td><strong>0.883</strong></td>
<td><strong>0.908</strong></td>
<td><strong>0.867</strong></td>
<td><strong>0.935</strong></td>
</tr>
</tbody>
</table>

In all plots, the y-axis represents precision while the x-axis is recall.

![Graphs for different species showing precision-recall curves for different methods including sequence-based classifier, expression-based classifier, integrated classifier, miRDeep2, and active learning.]
Conclusions

- Novel active learning approach for the classification of miRNA
- Decreased the number of labeled samples required
- Targeted the problem of limited known data and made use of unlabeled data
- Improved on state-of-the-art performance
Future Work

- Development of high-quality integrated training data sets
  - Pooling multiple NGS datasets to cover multiple conditions

- Experimental validation of predictions
Thank You For Your Attention