Predicting MicroRNA Precursors with a Generalized Gaussian Components based Density Estimation Algorithm

Chih-Hung Hsieh, Darby Tien-Hao Chang, Cheng-Hao Hsueh, Chi-Yeh Wu and Yen-Jen Oyang

Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan
Department of Electrical Engineering, National Cheng Kung University, Tainan, Taiwan
Identifying miRNA approaches

- MicroRNAs (miRNAs) play an important role in post-transcriptional regulation of gene expression.
- The earliest computational approaches for discovering pre-miRNAs are mainly based on comparative techniques and can only discover pre-miRNAs that are closely homologous to known miRNAs.
- Alternatively, scientists have been resorted to ab initio approaches to discover pre-miRNAs based on the characteristics of their secondary structures.
- Ab initio approaches do not rely on the existence of homologues.
- The characteristics of their secondary structures are analyzed by machine learning techniques.
Classification algorithms

- Most of people working on this subject have employed the existing kernel based data classification algorithms such as the hidden Markov model (HMM) the support vector machine (SVM) and the kernel density estimator to build the predictors due to the superior prediction performance delivered by these algorithms.

- Many conventional logic based data classification algorithms such as decision trees and decision rules continue to play a major role in some applications due to the interpretability of the logic rules identified by these algorithms.

- The design of the $G^2$DE classifier aims to accommodate the desirable functionalities of both the kernel based and the logic based classification algorithms.
G²DE’s specific

- G²DE can deliver prediction accuracy comparable with the state-of-the-art kernel based machine learning algorithms.
- G²DE provided the user with good interpretability.
- Another useful analysis provided by the learnt model of the G²DE based predictor is sub-class detection.
G$^2$DE decision function

\[ \hat{f}_{\text{svm}}(v) = \sum_{i} y_{i} \alpha_{i} \]

\[ GC(\sigma_{i}, s_{i}) = \frac{1}{\sqrt{\sigma}} \]

\[ \sigma_{i} \quad : \quad \text{Stan} \]

\[ \alpha_{i} \quad : \quad \text{dete} \]

\[ y_{i} \quad : \quad \{ + \} \]

\[ s_{i} \quad : \quad \text{train} \]

\[ v_{i} \quad : \quad \text{quer} \]
Generalized Gaussian Components
A generalized Gaussian component requires many parameter to be determined

- If the user sets this number to $k$ and the total number of features of the data set is $d$, then the learning algorithm of $G^2DE$ needs to figure out the optimal combination of the values of the following $k(d+2)(d+1)/2$ parameters in order to generate one approximate probability density function.

- The optimal combination of parameter values are figured out using the Ranking-based Adaptive Mutation Evolutionary (RAME) algorithm.
RAME

- RAME is an efficient optimization algorithm
Using $G^2$DE to analyze pre-miRNAs

- Convert Dataset into feature vector
- Use $G^2$DE to construct a mixture model
- Group samples into clusters
- Construct classifiers for each clusters
Experiment

- The prediction performance of the employed classification algorithm is evaluated and compared with four classification algorithms.
- A demonstrative analysis is also presented to investigate the interpretability of the employed classification algorithm.
Dataset

miRBase (release 12.0)
human miRNA precursors

CD-HIT

RefSeq & UCSC refGene
8494 pseudo hairpins collected from coding regions (CDSs)

RNAfold

exclude:
1. RNA sequences with <18 base pairs on the stem
2. MFE > -25 kcal/mol
3. multiple loops

692 pre-miRNA

3988 pseudo hairpins
Dataset (Cont.)

692 positive set

3988 negative set

460 (random)

232 (remain)

460 (random)

232 (remain random)

HU920 (training set)

HU464 (testing set)
## Feature set

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Set 1</strong></td>
<td></td>
</tr>
<tr>
<td>AA, AC, ..., UU</td>
<td>Frequencies of 16 dinucleotide pairs</td>
</tr>
<tr>
<td>%G+C</td>
<td>Percentage of nitrogenous bases which are either G or C</td>
</tr>
<tr>
<td><strong>Set 2</strong></td>
<td></td>
</tr>
<tr>
<td>mfe2</td>
<td>Ratio of $dG$ to the number of stems</td>
</tr>
<tr>
<td>mfe1</td>
<td>Ratio of $dG$ to %G+C</td>
</tr>
<tr>
<td>$dP$</td>
<td>Adjusted base pairing propensity. $dP$ is the number of base pairs observed in the secondary structure divided by the sequence length.</td>
</tr>
<tr>
<td>$dG$</td>
<td>Adjusted minimum free energy of folding. $dG$ is the minimum free energy (MFE) divided by the sequence length.</td>
</tr>
<tr>
<td>$dQ$</td>
<td>Adjusted Shannon entropy. $dQ$ measures the entropy of the base pairing probability distribution (BPPD).</td>
</tr>
<tr>
<td>$dD$</td>
<td>Adjusted base pair distance. $dD$ measures the average distance between all base pairs of structures inferred from the sequence.</td>
</tr>
<tr>
<td>$dF$</td>
<td>Compactness of the tree-graph representation of the sequence.</td>
</tr>
<tr>
<td><strong>Set 3</strong></td>
<td></td>
</tr>
<tr>
<td>$zG$, $zQ$, $zD$, $zP$, $zF$</td>
<td>5 normalized variants of $dP$, $dG$, $dQ$, $dD$ and $dF$</td>
</tr>
<tr>
<td><strong>Set 4</strong></td>
<td></td>
</tr>
<tr>
<td>$lH$</td>
<td>Hairpin length</td>
</tr>
<tr>
<td>$lL$</td>
<td>Loop length</td>
</tr>
<tr>
<td>$lC$</td>
<td>Consecutive base-pairs</td>
</tr>
<tr>
<td>%L</td>
<td>Ratio of loop length to hairpin length</td>
</tr>
</tbody>
</table>

The table shows the order of a feature within the feature set. For example, the fifth feature in the second feature set is $dQ$. 
Prediction performance

- The prediction performance is compared to two kernel based classifiers, SVM and RVKDE, and two logic based classifiers, C4.5 and RIPPER.
- The parameters for each classifier are determined by maximizing the prediction accuracy of ten-fold cross-validation on the HU920.
- A prediction is performed by using the HU920 dataset to predict the HU464 dataset with the selected parameters.
# Prediction performance

<table>
<thead>
<tr>
<th>Feature set</th>
<th>Kernel based classifiers&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Logic based classifiers</th>
<th>Logic based classifiers&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SVM</td>
<td>RVKDE</td>
<td>$G^2$DE</td>
</tr>
<tr>
<td>1</td>
<td>80.17%</td>
<td>77.59%</td>
<td>80.39%</td>
</tr>
<tr>
<td>2</td>
<td><strong>93.32%</strong></td>
<td>92.46%</td>
<td>92.03%</td>
</tr>
<tr>
<td>3</td>
<td>91.60%</td>
<td>91.16%</td>
<td>91.60%</td>
</tr>
<tr>
<td>4</td>
<td>78.66%</td>
<td>79.53%</td>
<td>78.66%</td>
</tr>
<tr>
<td>Average</td>
<td>85.94%</td>
<td>85.18%</td>
<td>85.67%</td>
</tr>
</tbody>
</table>

<sup>a</sup> SVM, RVKDE, $G^2$DE, and $G^2$DE-2 are different kernel-based classifiers.

<sup>b</sup> The number of kernels used for each classifier.
Interpretability

- As the result of previous table suggests that pre-miRNA prediction algorithms have reached the maximum with current feature sets.
- How to interpret the learnt model of machine learning techniques for users is crucial in pre-miRNA prediction.
- Users can obtain more valuable insights among different features by Interpretability.
Example of interpretability

- The first feature of the second feature set (mfe2) and the fifth feature of the second feature set (dQ) used as an example to explain how to interpret the models generated by the G²DE based predictor.

- To analyze parameters of GGC, we compare them to the Pearson product-moment correlations (PMCC).
Interpretability
Conclusions

- G²DE delivered prediction accuracy that is comparable with the state-of-art kernel based machine learning algorithms.
- G²DE provided the user with good interpretability.
- G²DE comprises a small number of generalized Gaussian components and is capable of detecting the sub-clusters or sub-classes of the data set.