Identification of MicroRNA Precursors from Genomic Hairpins Using Kernel Density Estimation

Tien-Hao Chang, Chih-Ching Wang
Outline

- Introduction
- Related Works
- Our Method
- Experimental Results
- Conclusion
Brief Introduction

• MicroRNAs (miRNAs) are ~22 nt RNAs and derive from ~90 nt RNA hairpin structures that play important roles in regulating gene expression in animals and plants.

• The hairpin structure is a necessary feature for the computational classification of novel miRNAs precursor (pre-miRs).
Related Works

- Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine
  - BMC Bioinformatics, 2005 December.
  - 3SVM

- 3SVM is a Support Vector Machine (SVM) for identifying pre-miRs without relying on phylogenetic conservation.
  - Species-specific and non-conserved pre-miRs.
  - 3SVM yielded an overall accuracy of 90.9% for 11 species.
Triplet element: continuous three structures with middle nucleotide. Taking “(” as the same as “(”.

32 triplet element features ---32-dimension vector.
(U(((、U((、U((、U(((、U(((、U((、U(((、U(((

Counting the appearances of the triplet elements.
(12、4、3、1、2、0、0、0、0、10、1、⋯)

32 triplet element features ---32-dimension vector.
(0.1846、0.0615、0.0462、0.0154、0.0308、0、0、0、0、⋯)
Related Works

• De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures.
  – Bioinformatics, 2007 January.
  – miPred

• MiPred is a Support Vector Machine (SVM) for identifying pre-miRs without relying on phylogenetic conservation.
  – Species-specific and non-conserved pre-miRs
  – Validated across 39 species, miPred yields 93.6% accuracy.
Our Method: miR-KDE

• Using a different classifier
  – relaxed variable kernel density estimation (RVKDE)
• Each sequence is summarized as a 33-dimensional feature vector.
  – The first 29 features are derived from miPred.
From the mathematical point of view, the job of the statistical learning algorithm is to identify curves that separate samples with different labels.
With the RKDE based statistical learning algorithm, each training sample is associated with a kernel function with typically a varying width.
In comparison with the decision function of SVM in which only support vectors are associated with a kernel function with a fixed width.
Feature Vector

- Each sequence is summarized as a 33-dimensional feature vector.
  - 17 dinucleotide sequence
  - 7 folding measures
  - 5 normalized feature
  - 4 stem-loop features
Feature Vector: Dinucleotide Sequence

- 16 dinucleotide frequencies $\%XY (X, Y \in \Sigma = [A, C, G, U])$ and 1 aggregate dinucleotide frequency $\%G+C$ ratio.

Counting the Sequence composition variables

$\%G+C = 12/22 = 0.524$ (1 feature)

$\%GA = 3/22 = 0.136$ (16 features)
Feature Vector : Folding Measures

Predicted RNA structure by RNAfold
Minimum Free Energy = -1.50 kcal/mol

- adjusted base pairing propensity, $dP$ : the total number of base pairs divided by the length.
  \[ dP = \frac{6}{22} = 0.2727 \]

- adjusted Minimum Free Energy of folding, $dG$ : the lowest MFE divided by the length
  \[ dG = \frac{-1.50}{22} = -0.0682 \]
Feature Vector : Folding Measures

- Minimum Free Energy of folding index 1, $MFEI_1$ : the ratio of dG to %G+C.
  \[ MFEI_1 = \frac{dG}{%G+C} \]
  \[ = -0.0682/0.524 = -0.1302 \]

- Minimum Free Energy of folding index 2, $MFEI_2$ : the ratio of dG to the number of stem.
  \[ MFEI_2 = \frac{dG}{\text{stem}} \]
  \[ = -0.0682/3 = -0.0227 \]
Feature Vector : Folding Measures

- adjusted shannon entropy, $dQ$:

$$dQ = -\frac{1}{L} \sum_{i<j} P_{ij} \log_2(P_{ij})$$

- adjusted base pair distance, $dD$:

$$dD = \frac{1}{L} \sum_{i<j} P_{ij} (1 - P_{ij})$$

- Here $P_{ij}$ denotes the probability of bases $i$ and $j$ pair.

$$P_{ij} = \sum_{S' \in S(X)} \frac{e^{-E_\alpha/R^*T}}{\sum_{S' \in S(X)} e^{-E_\alpha/R^*T}} \delta_{ij}$$
Feature Vector : Folding Measures

- Second eigenvalue, dF
Feature Vector : normalized feature

- $S(s_n)$ can be dG, dP, dF, dQ, and dD

$$Z(s_n) = \frac{S(s_n) - \mu_n}{\sigma_n}$$

Here, $\mu_n$ and $\sigma_n$ are the sample mean and the standard deviation of the feature $S(s_n)$. 
Feature Vector: Stem-loop Features

Hairpin length = 25 + 8 + 25 = 58
loop length = 8
Length of the longest contiguous base pairs = 8
loop length divided by Hairpin length = 8/58 = 0.138
Positive/Negative Dataset

- Positive set: 1983 pre-miRs
  - From miRBase
  - Without multiple loops

- Negative set: 3988 pseudo hairpins
  - From RefSeq
  - Extracted from the protein-coding regions (CDSs) according to the UCSC refGene annotation dataset
  - MFE lower than -25
### 40 Species

<table>
<thead>
<tr>
<th>Vertebrata</th>
<th>Arthropoda</th>
<th>Viridiplantae</th>
<th>Nematoda</th>
<th>Viruses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homo sapiens</td>
<td>Mus musculus</td>
<td>Apis mellifera</td>
<td>Zea mays</td>
<td></td>
</tr>
<tr>
<td>Ateles geoffroyi</td>
<td>Ovis aries</td>
<td>Drosophila melanogaster</td>
<td>Caenorhabditis briggsae</td>
<td></td>
</tr>
<tr>
<td>Bos taurus</td>
<td>Pan troglodytes</td>
<td>Drosophila pseudoobscura</td>
<td>Caenorhabditis elegans</td>
<td></td>
</tr>
<tr>
<td>Canis familiaris</td>
<td>Rattus norvegicus</td>
<td>Arabidopsis thaliana</td>
<td>Epstein Barr (EBV)</td>
<td></td>
</tr>
<tr>
<td>Danio rerio</td>
<td>Saguinus labiatus</td>
<td>Glycine max</td>
<td>Herpes Simplex (HSV)</td>
<td></td>
</tr>
<tr>
<td>Fugu rubripes</td>
<td>Sus scrofa</td>
<td>Medicago truncatula</td>
<td>Human cytomegalovirus (HCMV)</td>
<td></td>
</tr>
<tr>
<td>Gallus gallus</td>
<td>Tetraodon nigroviridis</td>
<td>Oryza sativa</td>
<td>Kaposi sarcoma-associated herpesvirus (KSHV)</td>
<td></td>
</tr>
<tr>
<td>Lagothrix lagotricha</td>
<td>Xenopus laevis</td>
<td>Physcomitrella patens</td>
<td>Mouse γ-herpesvirus (MGHV68)</td>
<td></td>
</tr>
<tr>
<td>Lemur catta</td>
<td>Xenopus tropicalis</td>
<td>Populus trichocarpa</td>
<td>Rhesus lymphocryptovirus</td>
<td></td>
</tr>
<tr>
<td>Macaca mulatta</td>
<td>Anopheles gambiae</td>
<td>Sorghum bicolor</td>
<td>Simian virus (SV40)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Vertebrata</td>
<td>Arthropoda</td>
<td>Viridiplantae</td>
<td>Nematoda</td>
</tr>
<tr>
<td>----------------</td>
<td>------------</td>
<td>------------</td>
<td>---------------</td>
<td>----------</td>
</tr>
<tr>
<td>308</td>
<td>186</td>
<td>25</td>
<td>52</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>66</td>
<td>70</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>35</td>
<td>107</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>62</td>
<td>71</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>233</td>
<td>1</td>
<td>16</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>64</td>
<td>2</td>
<td>15</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>87</td>
<td>41</td>
<td>115</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>14</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>122</td>
<td>121</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>37</td>
<td>35</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>
Training/Testing Dataset

• Training set
  – HU400, comprises of 200 human pre-miRNAs and 200 pseudo hairpins randomly selected from the positive and negative sets.

• Testing set
  – HU216, comprises of the remaining 108 human pre-miRNAs and randomly selected 108 pseudo hairpins.
  – NH3350, comprises of the remaining 1675 non-human pre-miRNAs and randomly selected 1675 pseudo hairpins.
Experiment design

• The proposed miR-KDE is evaluated by three experiments:
  – a five-fold cross validation on the human pre-miRNA set HU400.
  – using the model trained by the first experiment to predict another human pre-miRNA set HU216.
  – using the model trained by the first experiment to predict the non-human pre-miRNA set NH3350.
Prediction System Assessment

- **TP**: the number of real pre-miRs detected.
- **FN**: the number of real pre-miRs missed.
- **TN**: the number of pseudo hairpins correctly classified.
- **FP**: the number of pseudo hairpins incorrectly classified as pre-miRs.
# Prediction System Assessment

<table>
<thead>
<tr>
<th>Measure</th>
<th>Abbreviation</th>
<th>Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>%SE</td>
<td>$TP/(TP + FN)$</td>
</tr>
<tr>
<td>Specificity</td>
<td>%SP</td>
<td>$TN/(TN + FP)$</td>
</tr>
<tr>
<td>Accuracy</td>
<td>%ACC</td>
<td>$(TP + TN)/(TP + TN + FP + FN)$</td>
</tr>
<tr>
<td>F-measure</td>
<td>%Fm</td>
<td>$2TP/(2TP + FP + FN)$</td>
</tr>
<tr>
<td>Matthews’ correlation</td>
<td>%MCC</td>
<td>$(TP \times TN - FP \times FN)$ [sqrt((TP + FP) \times (TN + FN) \times (TP + FN) \times (TN + FP)))]</td>
</tr>
</tbody>
</table>

**Notes:**
- TP: True Positive
- TN: True Negative
- FP: False Positive
- FN: False Negative
## Experimental results on human pre-miRNAs

<table>
<thead>
<tr>
<th></th>
<th>%SE</th>
<th>%SP</th>
<th>%ACC</th>
<th>%Fm</th>
<th>%MCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Five-fold cross-validation on HU400</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3SVM</td>
<td>86.5%</td>
<td>91.5%</td>
<td>89.0%</td>
<td>88.7%</td>
<td>78.1%</td>
</tr>
<tr>
<td>miPred</td>
<td>87.5%</td>
<td>98.0%</td>
<td>92.8%</td>
<td>92.3%</td>
<td>86.0%</td>
</tr>
<tr>
<td>miR-KDE</td>
<td>90.5%</td>
<td>97.5%</td>
<td>94.0%</td>
<td>93.8%</td>
<td>88.2%</td>
</tr>
<tr>
<td>Using HU400 to predict HU216</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3SVM</td>
<td>83.3%</td>
<td>86.1%</td>
<td>84.7%</td>
<td>84.5%</td>
<td>69.5%</td>
</tr>
<tr>
<td>miPred</td>
<td>88.0%</td>
<td>88.0%</td>
<td>88.0%</td>
<td>88.0%</td>
<td>75.9%</td>
</tr>
<tr>
<td>miR-KDE</td>
<td>88.9%</td>
<td>92.6%</td>
<td>90.7%</td>
<td>90.6%</td>
<td>81.5%</td>
</tr>
</tbody>
</table>
Experimental results on non-human pre-miRNAs

<table>
<thead>
<tr>
<th></th>
<th>%SE</th>
<th>%SP</th>
<th>%ACC</th>
<th>%Fm</th>
<th>%MCC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>3SVM</strong></td>
<td>91.5%</td>
<td>88.7%</td>
<td>90.1%</td>
<td>90.2%</td>
<td>80.2%</td>
</tr>
<tr>
<td><strong>miPred</strong></td>
<td>96.7%</td>
<td>90.4%</td>
<td>93.6%</td>
<td>93.7%</td>
<td>87.3%</td>
</tr>
<tr>
<td><strong>miR-KDE</strong></td>
<td>95.8%</td>
<td>93.5%</td>
<td>94.7%</td>
<td>94.7%</td>
<td>89.3%</td>
</tr>
<tr>
<td>with <em>miPred’s %SP</em></td>
<td>97.4%</td>
<td>90.4%</td>
<td>93.9%</td>
<td>94.1%</td>
<td>88.1%</td>
</tr>
</tbody>
</table>
## Cross-Genus

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Genus</th>
<th>%SE</th>
<th>%SP</th>
<th>%ACC</th>
<th>%Fm</th>
<th>%MCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>miPred</td>
<td>Vertebrata</td>
<td>95.3%</td>
<td>88.8%</td>
<td>92.1%</td>
<td>92.3%</td>
<td>84.3%</td>
</tr>
<tr>
<td></td>
<td>Arthropoda</td>
<td>98.8%</td>
<td>89.0%</td>
<td>93.9%</td>
<td>94.2%</td>
<td>88.2%</td>
</tr>
<tr>
<td></td>
<td>Viridiplantae</td>
<td>98.2%</td>
<td>93.6%</td>
<td>95.9%</td>
<td>96.0%</td>
<td>91.9%</td>
</tr>
<tr>
<td></td>
<td>Nematoda</td>
<td>97.2%</td>
<td>90.4%</td>
<td>93.8%</td>
<td>94.0%</td>
<td>87.8%</td>
</tr>
<tr>
<td></td>
<td>Viruses</td>
<td>97.2%</td>
<td>93.1%</td>
<td>95.1%</td>
<td>95.2%</td>
<td>90.4%</td>
</tr>
<tr>
<td>miR-KDE</td>
<td>Vertebrata</td>
<td>93.4%</td>
<td>92.8%</td>
<td>93.1%</td>
<td>93.2%</td>
<td>86.3%</td>
</tr>
<tr>
<td></td>
<td>Arthropoda</td>
<td>100%</td>
<td>92.0%</td>
<td>96.0%</td>
<td>96.2%</td>
<td>92.3%</td>
</tr>
<tr>
<td></td>
<td>Viridiplantae</td>
<td>98.4%</td>
<td>95.0%</td>
<td>96.7%</td>
<td>96.8%</td>
<td>93.4%</td>
</tr>
<tr>
<td></td>
<td>Nematoda</td>
<td>97.2%</td>
<td>92.7%</td>
<td>94.9%</td>
<td>95.0%</td>
<td>89.9%</td>
</tr>
<tr>
<td></td>
<td>Viruses</td>
<td>94.4%</td>
<td>97.2%</td>
<td>95.8%</td>
<td>95.8%</td>
<td>91.7%</td>
</tr>
</tbody>
</table>
Cross-Genus

- Viruses are speculated to lack miRNA processing proteins such as Drosha, Dicer and RISC.
- Viral miRNAs utilize such processing proteins from their hosts to regulate viral expression after infecting. Thus, viral-encoded pre-miRNAs are likely to have very similar characteristics to those pre-miRNAs from the host (i.e., human).
Comparison of the feature set and the classification mechanism

- We next investigate the effect of using RVKDE by separating the two differences of miR-KDE to miPred:
  - introducing the four stem-loop features.
  - using RVKDE instead of SVM.
## Comparison of the feature set and the classification mechanism

<table>
<thead>
<tr>
<th></th>
<th>29 miPred’s features</th>
<th>29 miPred’s features + 4 stem-loop features</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>%SE</td>
<td>%SP</td>
</tr>
<tr>
<td><strong>HU216</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>SVM</strong></td>
<td>88.0%</td>
<td>88.0%</td>
</tr>
<tr>
<td><strong>RVKDE</strong></td>
<td>85.2%</td>
<td>90.7%</td>
</tr>
<tr>
<td><strong>NH3350</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>SVM</strong></td>
<td>96.7%</td>
<td>90.4%</td>
</tr>
<tr>
<td><strong>RVKDE</strong></td>
<td>94.8%</td>
<td>93.4%</td>
</tr>
</tbody>
</table>
Decision boundaries of SVM and RVKDE

• To explain the characteristic of RVKDE in miRNA prediction, four cases are selected to demonstrate its difference to SVM from the view of decision boundary.

• miR-KDE adopts only 29 features derived from miPred.
Decision boundaries of SVM and RVKDE

- A pre-miRNA *Caenorhabditis elegans* miR-260, classified correctly by miR-KDE but incorrectly by miPred.
Conclusion

- The 4 stem-loop features help around 0.6%~0.9% for accuracy, sensitivity and specificity
- The RVKDE helps more ~0.5% for accuracy and ~2.5% for specificity, but loses around 1.5%~1.9% sensitivity
- The miR-KDE yields a good overall accuracy of 94.7%, and has advantages over two compared \textit{ab initio} approaches for species-specific pre-miRNAs.
Appendix

- Probability of a structure \( S_\alpha \in S(X) \)

\[
P(S_\alpha) = \frac{e^{-E_\alpha/RT}}{Z}
\]

where \( Z = \sum_{S_\alpha \in S(X)} e^{-E_\alpha/RT} \)

\( E_\alpha \) is the free energy of \( S_\alpha \)

\( R = \) gas constant

\( T \) is the temperature
Appendix

- **Base-pair distance**

\[
d_{BP}(S_{\alpha}, S_{\beta}) = |S_{\alpha} \cup S_{\beta}| - |S_{\alpha} \cap S_{\beta}| = |S_{\alpha}| + |S_{\beta}| - 2|S_{\alpha} \cap S_{\beta}|
\]

\[
= \sum_{i<j} (\delta_{ij}^{\alpha} + \delta_{ij}^{\beta} - 2\delta_{ij}^{\alpha}\delta_{ij}^{\beta})
\]

- **Average Base-pair distance**

\[
< d_{BP} > = \frac{1}{2} \sum_{S_{\alpha} \subseteq S(\mathcal{X})} \sum_{S_{\beta} \subseteq S(\mathcal{X})} [P(S_{\alpha})P(S_{\beta}) \sum_{i<j} \delta_{ij}^{\alpha} + \delta_{ij}^{\beta} - 2\delta_{ij}^{\alpha}\delta_{ij}^{\beta}]
\]

\[
= \frac{1}{2} \sum_{i<j} \left[ \sum_{S_{\alpha} \subseteq S(\mathcal{X})} P(S_{\alpha}) \delta_{ij}^{\alpha} \sum_{S_{\beta} \subseteq S(\mathcal{X})} P(S_{\beta}) + \sum_{S_{\alpha} \subseteq S(\mathcal{X})} P(S_{\alpha}) \sum_{S_{\beta} \subseteq S(\mathcal{X})} P(S_{\beta}) \delta_{ij}^{\beta} - 2 \sum_{S_{\alpha} \subseteq S(\mathcal{X})} P(S_{\alpha}) \delta_{ij}^{\alpha} \sum_{S_{\beta} \subseteq S(\mathcal{X})} P(S_{\beta}) \delta_{ij}^{\beta} \right]
\]

\[
= \frac{1}{2} \sum_{i<j} \left[ P_{ij} + P_{ij} - 2 P_{ij}P_{ij} \right]
\]

\[
= \sum_{i<j} P_{ij} - P_{ij}^2
\]
Alternative Algorithm 1: KNN
Decision Function of the KNN Algorithm

\[ f(\mathbf{v}) = \sum_{i=1}^{k} \text{sgn}(s_i), \]

where \( s_1, s_2, ..., s_k \) are the \( k \) nearest samples of query vector \( \mathbf{v} \).

The input query located at \( \mathbf{v} \) is predicted to be a positive instance if \( f(\mathbf{v}) \geq 0 \); otherwise it is predicted to be a negative instance.
Decision Function of the RVKDE

\[ f_{RVKDE}(\mathbf{v}) = \sum_{s_i} y_i \cdot \frac{1}{\sigma_i} \cdot \exp \left( -\frac{||\mathbf{v} - \mathbf{s}_i||^2}{2\sigma_i^2} \right) \]

- \( \mathbf{v} \): testing sample
- \( y_i \): class value of a training sample \( s_i \)
- \( \sigma_i \): is local density of the proximity of \( s_i \)
Decision Function of the SVM

\[ f_{\text{SVM}}(v) = \sum_{s_i} y_i \cdot \alpha_i \cdot \exp(-\gamma \|v - s_i\|^2) \]

- \( v \) : testing sample
- \( y_i \) : class value of a training sample \( s_i \)
- \( \alpha_i \) : is determined by a constrained quadratic optimization.
- \( \gamma \) : user-specified parameter
Decision boundaries of SVM and RVKDE

- A pre-miRNA *Caenorhabditis elegans* miR-260, classified correctly by miR-KDE but incorrectly by miPred.
- X-axis is “UU”, and y-axis is “dP”.

Decision boundaries of SVM and RVKDE

- A pseudo hairpin classified incorrectly by miPred and correctly by miR-KDE.
- X-axis is “CC”, and y-axis is “GG”.

![Graphical representation of decision boundaries.](https://via.placeholder.com/150)
Decision boundaries of SVM and RVKDE

- A pre-miRNA, *Zea mays* miR168a, classified correctly by miPred but incorrectly by miR-KDE
- X-axis is “CG”, and y-axis is “dG”. 
Decision boundaries of SVM and RVKDE

- A pseudo hairpin correctly classified by miPred but incorrectly by miR-KDE.
- X-axis is “GG”, and y-axis is “dG”.