Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine

*BMC Bioinformatics* 2005, 6:310

Group: B
Member: 林伯勳，紀永鑫，吳益盛，唐紹祖，許梓亭，白宏益，范文郎
Speaker: 許梓亭
Outline

- Background
- Feature
- Method
- Result
- Conclusion
MicroRNAs (miRNAs) can play important regulatory roles. It will provide a regulatory way to cure cancer.
The past methods for detecting miRNAs.
- MiRscan: for nematodes and human
- MiRseeker: for insects
- MIRcheck & MIRFINDER: for plants

They all utilize comparative genomics information.

All those existing methods cannot be used to predict new miRNAs that have never appear.

For a species that does not have a closely related species sequenced, its miRNAs cannot be found by comparative genomics approaches.
The ability to distinguish pseudo vs. real pre-miRNAs is essential in the computational identification of novel and species-specific miRNAs.

In this study, we focus on the \textit{ab initio} classification of real pre-miRNA from other hairpin sequences with similar stem-loop features (we call them as pseudo pre-miRNAs).
Local Contiguous Structure-Sequence Features (1/4)

- We propose a set of features that combines the hairpin structures with sequence information to characterize the real vs. pseudo pre-miRNAs.
- The features focus on the information of every 3 adjacent nucleotides on the stem, and we name them as triplet elements.
Local Contiguous Structure-Sequence Features (2/4)

- The triplet element is composed of the 3 adjacent nucleotides structures and the middle nucleotide type.
  - paired sub-structure: brackets "(" or ")"
  - unpaired sub-structure: dots ".“
  - nucleotide type: U/A/C/G
Local Contiguous Structure-Sequence Features (3/4)
Local Contiguous Structure-Sequence Features (4/4)

| Triplet elements | Pre-miRNAs | Other hairpins | $F(x_i) = \frac{|\mu_i^+ - \mu_i^-|}{\sigma_i^+ + \sigma_i^-}$ |
|------------------|------------|----------------|-------------------------------------------------|
|                  | $\mu^*$    | $\sigma^*$     | $\mu$              | $\sigma$  |                     |
| A(((            | 0.121      | 0.042          | 0.063            | 0.032     | 0.792              |
| U((            | 0.154      | 0.048          | 0.089            | 0.040     | 0.734              |
| C...            | 0.006      | 0.011          | 0.025            | 0.030     | 0.475              |
| A...            | 0.008      | 0.014          | 0.025            | 0.025     | 0.429              |
| U...            | 0.007      | 0.011          | 0.021            | 0.023     | 0.397              |
| G((            | 0.042      | 0.025          | 0.063            | 0.031     | 0.383              |
| C(              | 0.009      | 0.011          | 0.019            | 0.017     | 0.353              |
| C(              | 0.032      | 0.022          | 0.048            | 0.027     | 0.329              |
| A(              | 0.011      | 0.012          | 0.020            | 0.016     | 0.316              |
| G(              | 0.151      | 0.038          | 0.127            | 0.040     | 0.303              |
| A(              | 0.013      | 0.013          | 0.022            | 0.019     | 0.295              |
| G(              | 0.006      | 0.011          | 0.014            | 0.019     | 0.289              |
| G(              | 0.007      | 0.010          | 0.014            | 0.015     | 0.266              |
| G(              | 0.040      | 0.020          | 0.050            | 0.024     | 0.231              |
| C(((            | 0.119      | 0.030          | 0.105            | 0.034     | 0.230              |
Classification Method of SVM(1/3)

- SVM (Support Vector Machine) has been widely used in classification.

- If we have some data sets of things classified without knowing the rules used for classification, when a new data set comes, SVM can predict which set it should belong to.

(c) Training data and a better classifier  
(d) Applying a better classifier on testing data
Classification Method of SVM(2/3)

- SVM has been widely applied to the prediction and classification of important biology signals.
  - promoters
  - translation initiation sites
  - splicing sites
  - proteins
  - new virus miRNAs
  - functional siRNAs

- SVM is used to classify the real and pseudo pre-miRNAs based on the hairpin structure features.
Classification Method of SVM (3/3)

CUUUUCUACACAGGUUGGAUCGGGUGCAAUUGCUGUGUUCUGUAUGGUAUUGCAUCUGUCCCGGCUGUUGAGUUUGG

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(., U(..., G((, G((., ...)

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

Normalizing the triplet element count vector:
( 0.1846, 0.0615, 0.0462, 0.0154, 0.0308, 0, 0, 0, 0.1538, 0.0154, ... )
The distribution of miRNAs on human genome

- Intrinsic miRNA (~1/3)
  - synthesis by polymerase II
- Intergenic miRNA (~2/3)
  - synthesis by polymerase III
- miRNA gene in alternatively spliced exons (rare)
  - synthesis by polymerase II
  - splice out by alternative splicing

- Total experimentally validated miRNA sequences (474)

- Reference: miRBase
  - [http://microrna.sanger.ac.uk/sequences/](http://microrna.sanger.ac.uk/sequences/)

- Reference: miRNAMAP
  - *Nucleic Acids Res.*, Jan 2006; 34: D135 - D139
# Results by SVM Classification of real vs. pseudo pre-miRNAs

<table>
<thead>
<tr>
<th>Test set</th>
<th>Type</th>
<th>Size</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TE-C</td>
<td>Real(^1)</td>
<td>30</td>
<td>93.3 (28/30)</td>
</tr>
<tr>
<td></td>
<td>Pseudo(^2)</td>
<td>1000</td>
<td>88.1 (881/1000)</td>
</tr>
<tr>
<td>CONSERVED-HAIRPIN</td>
<td>Pseudo(^2)</td>
<td>2444</td>
<td>89.0 (2174/2444)</td>
</tr>
<tr>
<td>UPDATED</td>
<td>Real(^1)</td>
<td>39</td>
<td>92.3 (36/39)</td>
</tr>
</tbody>
</table>

\(^1\)Real: real human pre-miRNAs.

\(^2\)Pseudo: pseudo pre-miRNA hairpins.
## Results by SVM Classification of real vs. pseudo pre-miRNAs (2/2)

<table>
<thead>
<tr>
<th>Species</th>
<th># of pre-miRNAs</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mus musculusi</td>
<td>36</td>
<td>94.4</td>
</tr>
<tr>
<td>Rattus norvegicus</td>
<td>25</td>
<td>80</td>
</tr>
<tr>
<td>Callus gallus</td>
<td>13</td>
<td>84.6</td>
</tr>
<tr>
<td>Dnio rerio</td>
<td>6</td>
<td>66.7</td>
</tr>
<tr>
<td>Caenorhabditis briggsae</td>
<td>73</td>
<td>95.9</td>
</tr>
<tr>
<td>Caenorhabditis elegans</td>
<td>110</td>
<td>86.4</td>
</tr>
<tr>
<td>Drosophila pseudoobscura</td>
<td>71</td>
<td>90.1</td>
</tr>
<tr>
<td>Drosophila melanogaster</td>
<td>71</td>
<td>91.5</td>
</tr>
<tr>
<td>Oryza sativa</td>
<td>96</td>
<td>94.8</td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
<td>75</td>
<td>92</td>
</tr>
<tr>
<td>Epstein Barr Virus</td>
<td>5</td>
<td>100</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>581</strong></td>
<td><strong>90.9</strong></td>
</tr>
</tbody>
</table>
Conclusion

- It was largely limited to classifying miRNA sequences that fold into secondary structures without multiple loops.

- It might be necessary to reconsider what we previously regard as new miRNAs features.

- The method SVM might be used in other fields of biology.
Reference

Thanks for your listening.
Additional information about SVM

Support Vector Machines

- **Training** vectors: \( x_i, i = 1, \ldots, l \)

- Consider a simple case with **two classes**:
  
  Define a vector \( y \)

  \[
  y_i = \begin{cases} 
  1 & \text{if } x_i \text{ in class 1} \\
  -1 & \text{if } x_i \text{ in class 2},
  \end{cases}
  \]

- A hyperplane which separates all data
Additional information about SVM

- A separating hyperplane: $w^T x + b = 0$
  
  \[(w^T x_i) + b > 0 \quad \text{if } y_i = 1\]
  
  \[(w^T x_i) + b < 0 \quad \text{if } y_i = -1\]

- Decision function $f(x) = \text{sign}(w^T x + b)$, $x$: test data

Variables: $w$ and $b$: Need to know coefficients of a plane

Many possible choices of $w$ and $b$
Additional information about SVM

- Select $w, b$ with the maximal margin.
  
  Maximal distance between $w^T x + b = \pm 1$  
  
  Vapnik’s statistical learning theory. (will be discussed later)

  $$(w^T x_i) + b \geq 1 \quad \text{if } y_i = 1$$
  $$(w^T x_i) + b \leq -1 \quad \text{if } y_i = -1$$

  \hspace{1cm} (1)

- Distance between $w^T x + b = 1$ and $-1$:

  $$2/\|w\| = 2/\sqrt{w^T w}.$$ 

- $\max 2/\|w\| \equiv \min_{w,b} w^T w / 2$

  $$\min_{w,b} \frac{1}{2} w^T w$$
  
  $y_i ((w^T x_i) + b) \geq 1, \quad \text{from (1)}$
  
  $i = 1, \ldots, l.$