PETcofold: predicting conserved interactions and structures of two multiple alignments of RNA sequences

CS681 class presentation by Dogan Altunbay
Outline

- Introduction
  - Motivation
  - Related Study
- Algorithm
- Performance Evaluation
- Conclusion
Introduction

- Motivation
  - ncRNAs are getting more interest, since they modulate the activity of other types of RNA.
  - Predicting RNA-RNA interactions is one method for assigning functions to ncRNAs.
  - Existing methods consider interactions between of a pair of single sequences.
  - Motivation of this study is to combine evolutionary and thermodynamic information of RNA secondary structures to construct a probabilistic hierarchical model using multiple alignments of RNA sequences.
Introduction

- Related Study
  - Methods that evaluate only the base pairs involved in duplex formation: Rehmsmeier et al., Tafer et al.
  - Methods that predict a joint secondary structure of two RNAs by folding their concatenated sequences: Andronescu et al., Bernhart et al., Dirks et al.
  - Methods that predict a joint secondary structure without restriction on interaction types: Bompfünewerer et al., Busch et al., Mückstein et al., Richter et al.
  - Methods that handle more complex joint structures: Alkan et al., Chitsaz et al., Huang et al., Pervouchine et al. Salari et al.
PETcofold Algorithm

- Hierarchical folding of two RNA alignments, A1 and A2, in two steps
  - Step 1:
    - Search for highly reliable base pairs by applying maximum expected accuracy approach on A1 and A2 to calculate unified reliabilities from evolutionary reliabilities and thermodynamic probabilities.
    - Find the partial structures $\sigma_1^p$ and $\sigma_2^p$, such that unified reliability is greater than $\delta$, and base pair probability is greater than $\gamma$.
  - Step 2:
    - Concatenate A1 and A2, and search for conserved interactions and structures using $\sigma_1^p$ and $\sigma_2^p$, reliabilities of which should exceed both $\sigma_1^p$ and $\sigma_2^p$.
    - Compute a joint probability value for each base pair with respect to thermodynamic probabilities computed by RNAcofold and evolutionary reliabilities computed by Pfold.
    - Find a combined structure using the joint probability information.
PETcofold Algorithm (cont.)

Input
Alignments $A_1$ and $A_2$

Step 1
Structure prediction of $A_1$ and $A_2$ using PETfold model

Select basepairs with high reliability ($\geq \delta$)

$p$-partial structures $\sigma^p_1$ and $\sigma^p_2$

Probable ($\geq \gamma$)

YES

Increase threshold $\delta$

NO

Step 2
Structure prediction of combined alignment constrained to $\sigma^p_1$ and $\sigma^p_2$ using adapted PETfold model

$pseudoknots$

$\sigma^p_1 \cup \sigma^p_2 \cup \sigma_{int}$

Output
Combined structure $\sigma^p_1 \cup \sigma^p_2 \cup \sigma_{int}$
Test datasets

- **Literature data**
  - Interactions between bacterial sRNAs and target mRNAs
  - Final dataset contains 13 different RNAs and 32 interactions from *Escherichia coli* K12 (*E.coli*), *Salmonella typhimurium* LT2 and *Staphylococcus aureus* N315

- **Simulated data**
  - Artificial test data with high covariance is generated to avoid the following aspects:
    - The orthologs predicted during dataset preparation from literature data may include false positives.
    - Weak covariance at interaction sites limits the full potential of algorithm.
Performance Evaluation

- **Evaluation metric**: correlation of the predictions to their literature structures.

- **Matthews correlation coefficient**:
  - TP: True Positive
  - TN: True Negative
  - FP: False Positive
  - FN: False Negative

\[
    MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}
\]

- The algorithm is also compared with single sequence based methods: inteRNA, Pairfold, RactIP, RNAcofold.
Results

- Performance of parameters $\delta$ and $\gamma$ is investigated with varying values.
  - $\delta$: maximal intra-molecular base pair reliability
  - $\gamma$: minimal partial structure probability
  - $\delta=0.9$ and $\gamma=0.9$ gives highest MCC
Results

- Performance on simulated data
  - Covariance at interaction sites is increased by scaling the branch lengths of phylogenetic trees
## Results

- **Comparison with other RNA joint secondary structure methods**

<table>
<thead>
<tr>
<th>sRNA-target pair</th>
<th>MCC</th>
<th>Run time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>PETcofold</strong></td>
<td><strong>interRNA</strong></td>
</tr>
<tr>
<td></td>
<td><strong>-extstem</strong></td>
<td>80% 100%</td>
</tr>
<tr>
<td>MicA-ompA</td>
<td>0.87 0.83 0.49 0.51 0.86 0.74 0.57 0.57 0.80 0.67</td>
<td>28.7 28.4 69493.1 3.2 3.0 0.2</td>
</tr>
<tr>
<td>OxyS-fhlA</td>
<td>0.80 0.82 0.64 0.64 0.61 0.61 0.48 0.48 0.61 0.61</td>
<td>20.6 19.3 129636.7 1.9 2.0 0.2</td>
</tr>
<tr>
<td>RyhB-uof-fur</td>
<td>0.13 0.13 0.12 0.00 0.21 0.21 0.19 0.00 0.21 0.21</td>
<td>26.4 25.3 65599.2 2.6 2.7 0.2</td>
</tr>
<tr>
<td>RyhB-sodB</td>
<td>0.67 0.71 0.70 0.68 0.65 0.51 0.65 0.59 0.65 0.63</td>
<td>15.4 15.2 23579.3 1.7 2.0 0.1</td>
</tr>
<tr>
<td>Average</td>
<td>0.62 0.62 0.49 0.46 0.58 0.52 0.47 0.41 0.57 0.53</td>
<td>22.8 22.1 72077.1 2.4 2.5 0.2</td>
</tr>
</tbody>
</table>
Results

- Example prediction
Conclusion

- The paper presents PETcofold method for the prediction of a joint secondary structure of two interacting RNAs.
- Covariance information improves the performance of prediction.
- Comparisons showed that evolutionary information from multiple sequences also improve the performance of prediction.
Questions