Reliability of putative RRIs

Final presentation of Masterproject
29.07.2021
Sebastian Holler

Supervisors: Florian Eggenhofer, Teresa Müller, Martin Raden
Chair of Bioinformatics
Motivation
Workflow

Data
- CopomuS sequences
- Interaction potential experimentally verified

Feature generation
- Use IntaRNA to predict RRIs
- Generate features from top 1 predictions

Feature selection
- Feature ranking via ReliefF algorithm [1]
- Feature scores indicate most promising ones

Cross-validation
- Repeated k-fold cross-validation
- Support vector machine (SVM) as classifier
- Scoring metrics: f1-score, AUC, sensitivity, specificity

Interaction data

- Challenge: Lack of experimental verified data
- Use data from compensatory mutation experiments

Positive dataset

Negative dataset

Data

ReliefF results

<table>
<thead>
<tr>
<th>Feature name</th>
<th>Description</th>
<th>Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENormGC</td>
<td>E normalized by interGC</td>
<td>0.12430758</td>
</tr>
<tr>
<td>E_hybridNormGC</td>
<td>E_hybrid normalized by interGC</td>
<td>0.1114999</td>
</tr>
<tr>
<td>E</td>
<td>Minimum free energy</td>
<td>0.1091744</td>
</tr>
<tr>
<td>E_hybrid</td>
<td>Hybridization energy</td>
<td>0.0989977</td>
</tr>
<tr>
<td>maxEDNormGC</td>
<td>Max. accessibility energy normalized by interGC</td>
<td>0.04220053</td>
</tr>
<tr>
<td>maxED</td>
<td>Max. accessibility energy</td>
<td>0.03550688</td>
</tr>
<tr>
<td>numBP</td>
<td>Number of base pairs within interaction</td>
<td>0.01767762</td>
</tr>
<tr>
<td>interGC</td>
<td>GC-content within interaction</td>
<td>0.01613493</td>
</tr>
<tr>
<td>diffGC</td>
<td>Difference between interGC and globalGC</td>
<td>0.00770325</td>
</tr>
<tr>
<td>minRange</td>
<td>Min. length of interacting subsequences</td>
<td>0.006607</td>
</tr>
<tr>
<td>maxRange</td>
<td>Max. length of interacting subsequences</td>
<td>0.00314323</td>
</tr>
<tr>
<td>minRangeNorm</td>
<td>minRange normalized by numBP</td>
<td>0.00284907</td>
</tr>
<tr>
<td>distSCodon</td>
<td>Signed distance start codon - interaction</td>
<td>0.00000325</td>
</tr>
<tr>
<td>distSCodonAbs</td>
<td>Absolute distance start codon - interaction</td>
<td>0.00000325</td>
</tr>
<tr>
<td>globalGC</td>
<td>GC-content of whole input sequences</td>
<td>-0.00253847</td>
</tr>
<tr>
<td>minRangeNormInv</td>
<td>Inverse minRange normalized by numBP</td>
<td>-0.00557687</td>
</tr>
<tr>
<td>maxRangeNorm</td>
<td>maxRange normalized by numBP</td>
<td>-0.01102459</td>
</tr>
<tr>
<td>maxRangeNormInv</td>
<td>Inverse maxRange normalized by numBP</td>
<td>-0.01213895</td>
</tr>
</tbody>
</table>

- Energy-based features top ranking
- Features normalized by interGC promising
Cross-validation results
Conclusion

Data
- Data from compensatory mutation experiments
- Interaction potential experimentally verified

Feature generation
- Use IntaRNA to predict RRIs
- Generate features from top 1 predictions

Feature selection
- ReliefF ranking topped by energy features
- Energies normalized by interGC promising

Cross-validation
- Energy alone quite good already
- Hard to get better than baseline
- Top 1-8 feature set performs best