Improving the State of the Art in Machine Learning Methods for Protein-RNA Interface Prediction

Rasna Walia
Cornelia Caragea, Drena Dobbs & Vasant Honavar

Bioinformatics & Computational Biology
Computer Science
Genetics Development & Cell Biology
Iowa State University
Importance of Protein-RNA Interactions

- Essential roles in many biological processes
  - Transcription
  - Translation
  - Viral infectivity

Source: [http://www.youtube.com/watch?v=NsVEx1nFHJc&feature=player_embedded](http://www.youtube.com/watch?v=NsVEx1nFHJc&feature=player_embedded)
Protein-RNA Interface Prediction

- Experimental methods:
  - Too time and labor intensive

- Computational methods

- Accurate prediction of protein-RNA interactions can contribute to:
  - New molecular tools for modifying gene expression
  - Novel therapies for infectious & genetic diseases
  - Detailed understanding of molecular mechanisms involved in protein-RNA recognition
State of the Art in RNA Binding Site Prediction?

- **Problem**: difficult to compare results due to different datasets, experimental design (cross-validation procedures) & evaluation procedures
- **Our goal**: systematic comparison of methods and features
Feature Representations

- Sequence
- Sequence PSSMs
- Smoothed Sequence PSSMs
- Structure
- Structure PSSMs

MPVGSLEKQVIFARFPGD
Interface Definition

- Interface residue = an amino acid whose atom is within 5 Å of an atom in bound RNA
## Datasets

<table>
<thead>
<tr>
<th>Description</th>
<th>Sequence identity</th>
<th>Resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RB106</strong> 106 non-redundant protein chains from the PDB</td>
<td>≤ 30%</td>
<td>≤ 3.5 Å</td>
</tr>
<tr>
<td><strong>RB144</strong> 144 non-redundant protein chains from the PDB</td>
<td>≤ 30%</td>
<td>≤ 3.5 Å</td>
</tr>
<tr>
<td><strong>RB199</strong> 199 non-redundant protein chains from the PDB</td>
<td>≤ 30%</td>
<td>≤ 3.5 Å</td>
</tr>
</tbody>
</table>
### RB144 Evaluation

<table>
<thead>
<tr>
<th>Naïve Bayes</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>F-Measure</th>
<th>AUC of ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence</td>
<td>83.06</td>
<td>0.26</td>
<td>0.37</td>
<td>0.74</td>
</tr>
<tr>
<td>Sequence PSSM</td>
<td>71.50</td>
<td>0.62</td>
<td>0.45</td>
<td>0.75</td>
</tr>
<tr>
<td>Smoothed Seq PSSM</td>
<td>69.76</td>
<td>0.66</td>
<td>0.45</td>
<td>0.76</td>
</tr>
<tr>
<td>Structure</td>
<td>83.70</td>
<td>0.34</td>
<td>0.44</td>
<td>0.77</td>
</tr>
<tr>
<td>Structure PSSM</td>
<td>73.65</td>
<td>0.63</td>
<td>0.48</td>
<td>0.78</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Support Vector Machines</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>F-Measure</th>
<th>AUC of ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence</td>
<td>81.94</td>
<td>0.18</td>
<td>0.29</td>
<td>0.72</td>
</tr>
<tr>
<td>Sequence PSSM</td>
<td>84.05</td>
<td>0.36</td>
<td>0.46</td>
<td>0.79</td>
</tr>
<tr>
<td>Smoothed Seq PSSM</td>
<td>83.75</td>
<td>0.31</td>
<td>0.43</td>
<td>0.78</td>
</tr>
<tr>
<td>Structure</td>
<td>83.76</td>
<td>0.31</td>
<td>0.42</td>
<td>0.77</td>
</tr>
<tr>
<td>Structure PSSM</td>
<td>84.73</td>
<td>0.38</td>
<td>0.49</td>
<td>0.81</td>
</tr>
</tbody>
</table>

**Naïve Bayes**
- Best accuracy: Structure
- Best AUC of ROC: Structure PSSM

**SVMs**
- Best accuracy: Structure PSSM
- Best AUC of ROC: Structure PSSM

*Unbalanced data*
Ensemble of Classifiers

PDB ID: 1DI2 Chain: A

True Positives: Red
False Positives: Blue
Conclusions

• Both evolutionary information (PSSMs) and structural information improve prediction of RNA-binding residues in proteins
• Need to evaluate SVM modifications that optimize metrics other than accuracy
• Evaluate additional representations and new machine learning algorithms
Acknowledgments

• **Honavar Lab**
  – Dr. Vasant Honavar
  – Dr. Yasser El-Manzalawy
  – Fadi Towfic
  – Cornelia Caragea

• **Dobbs Lab**
  – Dr. Drena Dobbs
  – Ben Lewis

http://bindr.gdcb.iastate.edu/PRIDB
http://bindr.gdcb.iastate.edu/RNABindR
New Directions

• So far, “protein-centric” approach used to predict interface residues

• Use *homology-based* methods & *partner-specific* predictions:
  – Given a query protein-RNA complex, find homologs of interacting protein & RNA partners; transfer interface information to query
  – Approach has been successful for protein-protein interface prediction
    • (Li X., Dobbs, D. & Honavar, V., in press)

• Collect features from **RNA side** of complexes as input for interface residue predictors
  – Challenge: RNA forms myriad tertiary structures
PRIDB: Protein-RNA Interface DataBase

- Comprehensive database: 926 protein-RNA complexes from the PDB (Nov 2010)
- Provides
  - atomic level information regarding interfacial contacts
  - pre-calculated benchmark datasets of protein-RNA complexes for evaluating the performance of interface prediction methods

- Lewis, Walia et al. (2010)
  Nucleic Acids Research

http://bindr.gdcb.iastate.edu/PRIDB
References

Evaluation Metrics

\[
\text{specificity} = \frac{\text{number of True Negatives}}{\text{number of True Negatives} + \text{number of False Positives}}
\]

\[
\text{sensitivity} = \frac{\text{number of True Positives}}{\text{number of True Positives} + \text{number of False Negatives}}
\]

\[
\text{Precision} = \frac{tp}{tp + fp}
\]

\[
\text{Recall} = \frac{tp}{tp + fn}
\]

\[
\text{True Negative Rate} = \frac{tn}{tn + fp}
\]

\[
\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}
\]

\[
F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}
\]