Protein function annotation from sequence: prediction of residues interacting with RNA

R. V. Spriggs, Y. Murakami, H. Nakamura and S. Jones

April 23, 2009

BIOINFORMATICS

Department of Chemistry and Biochemistry, School of Life Sciences, John Maynard-Smith Building, University of Sussex, Falmer, Brighton, BN1 9QG, UK
INTRODUCTION

- All eukaryotic proteomes are characterized by a significant percentage (18–40%) of proteins of unknown function (PUFs).

- As more divergent genomes are sequenced the number of PUFs increases, computational methods that predict function from sequence are essential.
Computational function annotation of protein sequences in the field of RNA binding has addressed two problems:

(i) Predicting RNA-binding residues (RBRs) in protein sequences known to bind RNA.

(ii) Predicting RNA-binding function for complete proteins of unknown function.
INTRODUCTION

- PiRaNhA is the method to predict RBR within protein sequences, and use these predicted RBR to predict RNA-binding function at the protein level.

- SVM have been widely used for prediction using either PSSMs or physicochemical properties as sequence features.

- PiRaNhA is the first method to combine both of these feature types.
Datasets of protein–RNA complexes

- From the Protein Data Bank
- Defined an RBR as any residue $\leq 3.9\text{Å}$ distance from the RNA.
- $\geq 5$ protein–RNA contacts,
- No restriction was made on which organism.
A non-redundant set of RNA-binding proteins derived from the 86-protein set used by Kumar et al. (2008).

Four homologous and one have-no-RBR sequences were removed.

Including 41 rRNA-binding, 15 tRNA-binding, 7 mRNA-binding and other RNA-type-binding proteins.
Testing (RNAtestset42)


- Remove all homologous proteins with RNAset81.

- Including 24 rRNA-binding, 9 tRNA-binding, 2 mRNA-binding and other RNA-type-binding proteins.
Sequence feature vectors

Four properties of the residues are used as features in training the SVM:

1. PSSM: the evolutionary conservation of the residue positions.
2. Interface Propensity (IP): how likely a residue of a specific type in an RNA binding site.
3. Predicted accessibility (pA): predicted RSA.
4. Hydrophobicity (H)
Windows

- to enable the sequence environment of the residue to be taken into account.

- the properties of all the residues in the sub-sequence (window) are used to describe the residue in the centre.

- windows of various lengths have been tested, covering 5–25 residues.
Performance measures

- TP, FP, TN, FN
- Precision: TP/(TP+FP)
- Sensitivity (SN): TP/(TP+FN)
- Specificity (SP): TN/(TN+FP)
- Accuracy (ACC): (TP+TN)/(TP+FN+TN+FP)
- MCC: \(\frac{(TP \times TN) - (FP \times FN)}{((TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN))^{1/2}}\)
- \(F_{0.5}\)-measure: \(F_\beta = \frac{(1 + \beta^2) \times R \times P}{(\beta^2 \times P) + R}\),
  \((R = \text{recall}, P = \text{precision}, \beta = 0.5)\)
Performance of SVM trained on the RNAset81 dataset

- positive: 2938; negative: 16175

Table 1. The 5-fold cross-validation of all possible combinations of sequence features using RNAset81

<table>
<thead>
<tr>
<th>SVM</th>
<th>Feature vector</th>
<th>Window size</th>
<th>SN (%)</th>
<th>SP (%)</th>
<th>ACC (%)</th>
<th>MCC</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIII</td>
<td>PSSM + IP + pA + H</td>
<td>23</td>
<td>56.3</td>
<td>92.8</td>
<td>87.2</td>
<td>0.499</td>
<td>0.860</td>
</tr>
<tr>
<td>III</td>
<td>PSSM + pA</td>
<td>25</td>
<td>52.7</td>
<td>93.9</td>
<td>87.6</td>
<td>0.497</td>
<td>0.860</td>
</tr>
<tr>
<td>VI</td>
<td>PSSM + IP + H</td>
<td>25</td>
<td>58.4</td>
<td>92.3</td>
<td>87.1</td>
<td>0.506</td>
<td>0.859</td>
</tr>
<tr>
<td>VII</td>
<td>PSSM + pA + H</td>
<td>23</td>
<td>58.5</td>
<td>92.1</td>
<td>87.0</td>
<td>0.504</td>
<td>0.859</td>
</tr>
<tr>
<td>V</td>
<td>PSSM + IP + pA</td>
<td>23</td>
<td>56.9</td>
<td>92.6</td>
<td>87.1</td>
<td>0.500</td>
<td>0.858</td>
</tr>
<tr>
<td>II</td>
<td>PSSM + IP</td>
<td>25</td>
<td>53.9</td>
<td>93.7</td>
<td>87.5</td>
<td>0.499</td>
<td>0.859</td>
</tr>
<tr>
<td>IV</td>
<td>PSSM + H</td>
<td>23</td>
<td>58.3</td>
<td>91.7</td>
<td>86.6</td>
<td>0.492</td>
<td>0.855</td>
</tr>
<tr>
<td>I</td>
<td>PSSM</td>
<td>23</td>
<td>60.1</td>
<td>90.8</td>
<td>86.1</td>
<td>0.490</td>
<td>0.855</td>
</tr>
</tbody>
</table>

Models are ordered by AUC value. AUC shown to three decimal places.
compared with the published performance measures

Table 2. Comparison of reported performance measures for five other prediction methods and the best model from the current work based on the RNAset81 dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>RBR definition (Å)</th>
<th>Cross-validation</th>
<th>SN (%)</th>
<th>SP (%)</th>
<th>ACC (%)</th>
<th>AUC</th>
<th>MCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>PiRaNhA (SVM VIH)</td>
<td>3.9 (HBPLUS)</td>
<td>5 fold</td>
<td>56.3</td>
<td>92.8</td>
<td>87.2</td>
<td>0.86</td>
<td>0.50</td>
</tr>
<tr>
<td>PPRInt (Kumar et al., 2008)</td>
<td>6</td>
<td>5-fold</td>
<td>53.1</td>
<td>89.6</td>
<td>81.2</td>
<td>–</td>
<td>0.45</td>
</tr>
<tr>
<td>PRINTr (Wang et al., 2008)</td>
<td>ENTANGLE (Allers and Shamoo, 2001)</td>
<td>7-fold</td>
<td>55.9</td>
<td>–</td>
<td>87.1</td>
<td>0.83</td>
<td>0.43</td>
</tr>
<tr>
<td>NN-based (Jeong and Miyano, 2006; Jeong et al., 2004)</td>
<td>6</td>
<td>10-fold</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>0.77</td>
<td>0.41</td>
</tr>
<tr>
<td>RNABindR (Terribilini et al., 2007)</td>
<td>5</td>
<td>leave-1-out</td>
<td>33</td>
<td>95</td>
<td>83</td>
<td>–</td>
<td>0.36</td>
</tr>
<tr>
<td>BindN (Wang and Brown, 2006)</td>
<td>3.5</td>
<td>5-fold</td>
<td>66.3</td>
<td>69.8</td>
<td>69.3</td>
<td>0.73</td>
<td>0.27</td>
</tr>
</tbody>
</table>

Methods are ordered by MCC value.
Predicting binding residues in the RNA_testset42 dataset

Table 3. RBR predictions for the RNA_testset42 dataset using PiRaNhA and using RNABindR, PPRInt and BindN

<table>
<thead>
<tr>
<th>Method</th>
<th>SN (%)</th>
<th>SP (%)</th>
<th>ACC (%)</th>
<th>MCC</th>
<th>$F_{0.5}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>PiRaNhA</td>
<td>53.0</td>
<td>90.0</td>
<td>84.5</td>
<td>0.41</td>
<td>0.49</td>
</tr>
<tr>
<td>RNABindR (Terribilini et al., 2007)</td>
<td>37.4</td>
<td>93.8</td>
<td>85.5</td>
<td>0.36</td>
<td>0.48</td>
</tr>
<tr>
<td>PPRInt (Kumar et al., 2008)</td>
<td>70.4</td>
<td>73.9</td>
<td>73.4</td>
<td>0.34</td>
<td>0.36</td>
</tr>
<tr>
<td>BindN (Wang and Brown, 2006)</td>
<td>55.0</td>
<td>80.2</td>
<td>76.4</td>
<td>0.29</td>
<td>0.35</td>
</tr>
</tbody>
</table>

Methods are ordered by MCC value.
Predicting RNA-binding function at the protein level

- The 134 positives were created from UniProtKB/Swiss-Prot using strict keywords (RNA-, rRNA- and tRNA-binding), discard proteins in RNAset81.

- The 4849 negatives were created in the same way, but did not feature any keywords that could imply an RNA-binding function, and random select 134 proteins.
Predicting RNA-binding function at the protein level

- The PiRaNhA server was used to make predictions for all the residues in the 268 proteins.

- A series of statistical features were then evaluated as features in a second SVM to predict RNA-binding function at the protein level.

- The features tested for this second stage SVM included the minimum, median, mean, variance, skew, range and kurtosis.
Table 4. Performance measures for the five-fold cross-validation of the 2-stage SVM to predict RNA-binding function at the protein level for a non-redundant dataset of 268 proteins.

<table>
<thead>
<tr>
<th>Feature vector</th>
<th>SN (%)</th>
<th>SP (%)</th>
<th>ACC (%)</th>
<th>AUC</th>
<th>MCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum, maximum, mean, range, third quantile, skew, kurtosis of decision values from PiRaNhA</td>
<td>80.0</td>
<td>72.3</td>
<td>76.1</td>
<td>0.80</td>
<td>0.53</td>
</tr>
</tbody>
</table>
The method is based on an SVM that is the first to integrate PSSMs, amino acid interface propensities, predicted residue accessibility and hydrophobicity in the feature vector.

Predictions for the RNAtestset42 dataset using PiRaNhA and three other prediction servers showed that PiRaNhA is the best performing method.

While the two-stage SVM method requires further testing, table4 show potential to accurately predict if a protein binds RNA, as well as where.