A Versatile Combination of Classifiers for Protein Function Prediction

Haneen Altartouri
(haneen.altartouri@ini.rub.de)

Tobias Glasmachers

Institute for Neural Computation
Ruhr University Bochum

27/9/2020
Haneen Altartouri

- PhD student in Institute for Neural Computation, Ruhr University Bochum.
- Member of "Theory of Machine Learning" group lead by Tobias Glasmachers.
- My research is focused on improving the prediction of protein functions.
Contents

1. Introduction
2. The proposed Approach
3. Benchmarks
4. Experiments and results
5. Conclusion
Motivation of protein classification

Computational approaches to solve various protein prediction problems in a faster and more cost-effective manner.
Clustering before the classification

- Classification is often easy if the discriminative features are homogeneous for the whole data set.
- For heterogeneous datasets, we should therefore find homogeneous regions and address them with separate classifiers.
The proposed approach

Evaluating the performance and selecting the best classifier

The best classifier for sub-dataset #1

SDC

RSRC

FDC

Training Classifier

Protein Sequences
AAEVLAYIIVFKQF
NVEKLSTQKSSLRQ
ETVEKIVQYIREPVK
KSRRTLSSFAWLSLD
BYSLHEVDFDPOVL
EYAEQDGKPTILLD
QGSRFLKGMCRPLA
AAEVLAYIIVFKQF
TLYKQGKFGGYG
SFFQDMIFFEALN
VGHGHTFSGTVKLG
TTEVMPVSMKSTVD

Encoded Sequences
0.6121.2565.326
3.835154.8612
6.67851984.122
0.51124.2930.6
0.835124.5612
5.77850984.122
0.711127.2930.6
0.835164.5612
5.77855984.122
0.51124.2930.6
2.836165.5822
3.77850984.122

Encoding Method (PaaLAC)

Training Classifier

SDC

RSRC

FDC

Evaluating the performance and selecting the best classifier

The best classifier for sub-dataset #2
The proposed approach (cont.)

- **Representing Protein Sequences**
  - We used Chou’s Pseudo Amino acid Composition (PseAAC) descriptors.
  - Two sets of Physico-chemical properties (PCPs) were tested:
    1. 3 PCPs (hydrophobicity, hydrophilicity, and side chain mass).
    2. 50 non-redundant PCPs of amino acids.

- **Clustering dataset into sub-datasets**
  - K-means was used.
  - We tuned the number of sub-datasets \((k)\) for each dataset, to study its effect on the proposed approach.
The proposed approach (cont.)

Reducing Feature Vector Dimensionality

- Two reduction techniques were tested:
  1. Recursive Feature Elimination (RFE).
  2. Principal Component Analysis (PCA).

Classifier Selection

- For each sub-dataset we have up to three classifiers available: FDC, SDC, and RSDC.
- We estimate the performance of all three classifiers by means of cross-validation.
- We select the classifier with highest AUC.
## Benchmarks

<table>
<thead>
<tr>
<th>Dataset</th>
<th># of Positives</th>
<th># of Negatives</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA-binding proteins</td>
<td>523 binding proteins</td>
<td>543 non binding proteins</td>
</tr>
<tr>
<td>Antioxidant proteins</td>
<td>250 antioxidant</td>
<td>1547 non-antioxidant</td>
</tr>
<tr>
<td>RNA-binding proteins</td>
<td>2780 binding proteins</td>
<td>7077 non binding proteins</td>
</tr>
<tr>
<td>Antimicrobial peptides (AMP)</td>
<td>869 AMPs</td>
<td>2405 non-AMPS</td>
</tr>
<tr>
<td>Caspase 3 human substrates</td>
<td>247 cleaved peptides</td>
<td>247 non-cleaved peptides</td>
</tr>
<tr>
<td>Major Histocompa. Complex II (MHCII)</td>
<td>3510 binding peptides</td>
<td>1656 non-binding peptides</td>
</tr>
</tbody>
</table>
Selecting the best Classifier

Different classifiers were tested on the full datasets (FDCs): SVM, RF, ANN, and xGBoost.

The results showed that:
1. SVM is the best choice for most datasets using 50 PCPs.
2. RF is the best choice when using 3 PCPs.
Applying feature reduction on sub-datasets

- The importance of the features differs not only between the two sub-datasets, but also from the full dataset.
- Therefore, applying feature reduction on a per-cluster basis has the potential to improve overall performance.
Experiments and Results (cont.)

DNA-binding proteins (SVM)

- FDC only (the baseline)
- SDC [FDC(2), SDC(2)]
- RSDC [FDC(1), SDC(0), RSDC (1(RFE))]

DNA-binding proteins (RF)

- FDC only (the baseline)
- SDC [FDC(2), SDC(1)]
- RSDC [FDC(0), SDC(0), RSDC (2(PCA))]

DNA-binding proteins (SVM)

- FDC only (the baseline)
- SDC [FDC(2), SDC(2)]
- RSDC [FDC(1), SDC(0), RSDC (1(RFE))]

DNA-binding proteins (RF)

- FDC only (the baseline)
- SDC [FDC(2), SDC(2)]
- RSDC [FDC(1), SDC(0), RSDC (1(RFE))]

Experiments and Results (cont.)

Antioxidant proteins (SVM)

- FDC only (the baseline)
- SDC [FDC(2), SDC(2)]
- RSDC [FDC(2), SDC(1), RSDC (1(RFE))]

Antioxidant proteins (SVM)

- FDC only (the baseline)
- SDC [FDC(2), SDC(2)]
- RSDC [FDC(2), SDC(1), RSDC (1(RFE))]

Antioxidant proteins (RF)

- FDC only (the baseline)
- SDC [FDC(3), SDC(1)]

Antioxidant proteins (RF)

- FDC only (the baseline)
- SDC [FDC(3), SDC(1)]
Experiments and Results (cont.)
Experiments and Results (cont.)

### AMP peptides (SVM)

<table>
<thead>
<tr>
<th>Method</th>
<th>AUC (%)</th>
<th>MCC (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FDC only</td>
<td>0.97</td>
<td>0.72</td>
</tr>
<tr>
<td>[FDC(3), SDC(2)]</td>
<td>0.96</td>
<td>0.75</td>
</tr>
<tr>
<td>RSDC</td>
<td>0.98</td>
<td>0.80</td>
</tr>
</tbody>
</table>

### AMP peptides (RF)

<table>
<thead>
<tr>
<th>Method</th>
<th>AUC (%)</th>
<th>MCC (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FDC only</td>
<td>0.97</td>
<td>0.72</td>
</tr>
<tr>
<td>[FDC(21), SDC(4)]</td>
<td>0.96</td>
<td>0.75</td>
</tr>
<tr>
<td>RSDC</td>
<td>0.98</td>
<td>0.80</td>
</tr>
</tbody>
</table>
Experiments and Results (cont.)

• In most cases, RFE shows that the frequencies of amino acids play an important role in classifying the sequences inside the clusters, while the sequence order has a higher impact on classifying the full dataset.
• For datasets containing long protein sequences, RFE shows that the optimal sets of features for clusters contain only a bit more than 50% of all available descriptors.
Conclusion

• We have studied the effect of exploiting homogeneous sub-datasets inside protein sequence data by training multiple classifiers on sub-datasets.

• The proposed approach handles each sub-dataset as a separate classification problem that requires tuning the hyper-parameters and finding the best features separately.

• We have evaluated the performance of SVM and RF classifiers inside the sub-datasets, and RFE and PCA are tested as a reduction feature algorithms.

• SVM and SVM-RFE achieved good performance for most datasets.
• The performance of the proposed approach depends on the number of sub-datasets, the encoding method, and for each cluster the classifier with its hyperparameters and the feature reduction method applied.

• The results indicate that the proposed approach improved the overall performance of function prediction of protein sequences in the most cases.

• Results indicate that many protein sequence datasets suffer from heterogeneity.
Thank you