Analysis of the Signal Peptide dataset

November 28, 2019
Signal Peptide

- A short peptide (typically 15-30 residues long), destined towards the secretory pathway
- Cleaved during translocation across membrane existing in all 3 kingdoms of life
Our dataset

- FASTA format is a text-based format for representing either nucleotide sequences or peptide sequences, in which base pairs or amino acids are represented using single-letter codes. A sequence in FASTA format begins with a single-line description, followed by lines of sequence data. The description line is distinguished from the sequence data by a greater-than (">"), symbol in the first column. It is recommended that all lines of text be shorter than 80 characters in length.
Our dataset

- The FASTA file contains for each protein (in order):
  - Header (e.g. ">Q8TF40|EUKARYA|NO_SP|0")
  - Protein sequence (first 70 residues only)
  - Residue annotation

```plaintext
>P07093|EUKARYA|SP|3
MNWHLPLLASVTLPSCSHFNPLSLEELGSNTGIQVFNQIVKSRPHDNIVISPHGASVLGLMLQLGAD
SSSSSSSSSSSSSSSSSSSS0000000000000000000000000000000000000000
```
Our dataset

The header contains information about:

● The protein ID (e.g. "Q8TF40")
● The kingdom of life the organism (that contains the protein) belongs to (e.g. "EUKARYA")
● The type of signal peptide the protein contains (e.g. "NO_SP")
● The data set split the protein belongs to (e.g. "0")
Our dataset

- 20,758 proteins
- 4 types of signal peptides
- 6 residue types
- 20% sequence similarity
Our dataset

- 5 splits for **cross-validation** with similar residue distribution

- **Cross-validation** is a resampling procedure used to evaluate machine learning models on a limited data sample. The procedure has a single parameter called $k$ that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called $k$-fold cross-validation.
Class distributions

- Strong dataset imbalance: most proteins don’t contain Signal Peptides

<table>
<thead>
<tr>
<th>Class</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SP</td>
<td>Signal Peptide</td>
</tr>
<tr>
<td>LIPO</td>
<td>Lipoprotein Signal Peptide</td>
</tr>
<tr>
<td>TAT</td>
<td>Tat Signal Peptide</td>
</tr>
<tr>
<td>NO_SP</td>
<td>No Signal Peptide</td>
</tr>
</tbody>
</table>
Residue annotations

S  Sec/SPI signal peptide
T  Tat/SPI signal peptide
L  Sec/SPII signal peptide
I  Cytoplasm
M  Transmembrane
O  Extracellular
Prediction Baseline

Binary Classification

- SP: 8.8%
- NON-SP: 91.3%
Dealing with class imbalance

- Undersampling (majority classes)
- Oversampling (minority classes)
- Class weights
- SMOTE (synthetic samples)
ELMo Embeddings

- ELMo Embeddings: Embedded Language Models
- Used in Natural Language Processing
- In our case, embeddings represent the context of each residue
- Either 64 dim or 1024 dim per residue
Learning from high-dimensional data

- Reduce the dimensions
- t-SNE
- Techniques for dimensionality reduction and clustering that preserve the proportionality of the objects

-> Visualization of high dimensionality datasets
PCA vs t-SNE
Results of t-SNE for the 64 dim embeddings
Results of t-SNE for the 64 dim embeddings for L signal peptides
Results of t-SNE for the 64 dim embeddings for S signal peptides
Results of t-SNE for the 64 dim embeddings for T signal peptides
Notes

- Results are based on the perplexity = 30
- Not a lot of information
- 1024 dimensional embeddings can be more helpful
References

- https://zhanglab.ccmb.med.umich.edu/FASTA/
- https://machinelearningmastery.com/k-fold-cross-validation/
Thank you very much!