Benchmark comparison of protein sequence preprocessing effect on learning task for Pfam family classification

Maciej Sikora

Saturday 11th June, 2022

Inspiration and problem overiew

- **Data preprocessing for size reduction without loosing crucial information.**
- Growing amount of biological data
- Uniprot

Inspiration and problem overiew

- **Data preprocessing for size reduction without loosing crucial information.**
- Most data is unreviewed
- Automated data annotating
- Need for better models
- More complicated models take longer to train

Compared methods of preprocessing

- Data source: Swissprot – manually reviewed part of the Uniprot database
- Filtering by most frequent organisms and families
- CD-HIT – removing very simmilar protein sequences
- Padding
- Splitting to stratified train and test pools
- Shuffling

- Original
- Singletons – dtype int8
- Triplets – dtype int16
- Biovec

Tested models

- Decision trees
- Random trees
- MLP
- Nearest neighbours
- Machine Learnig (Simple Dense model)

- Grid Search for optimal parameters
- Cross-validation

Results

Summary

- Treating protein data as string objects might have a negative impact on the training process.
- Simple conversion to numerical data improves both quality and runtime.
- Further numerical categorization for triplets while speeding up the learning process is loosing some of the information leading to decreased accuracy.
- Neighbourhood of aminoacids can be analysed using more sophisticated biovec model.
- Biovec yields both better results and over 500% faster time compared to the original model.
- Biovec model data weights over 20 times less than the original one.