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

Maciej Sikora

Sunday 12th June, 2022

Inspiration and problem overview

- **Data preprocessing for size reduction without losing crucial information.**
- A growing amount of biological data.
- Full Swissprot database: >250 MB
- Full Uniprot database: around 10 GB
- Combining data from other databases would increase those numbers even more.
- The analysis becomes much harder and time-consuming

Inspiration and problem overview

- **Data preprocessing for size reduction without losing crucial information.**
- 0.25% sequences manually annotated in Uniprot
- Automated data annotating
- Need for better models
- More complicated models take longer to train

Biovec

- Biovec [1][2] – a multidisciplinary model combining knowledge from biology, linguistics and statistics.
- Analysis of k-mers in a bigger sequential context.
- Calculating vectors of the probability of k-mers in the given context.
- Using preprocessed metadata for training.

Original Sequence

$$^{(1)}\overrightarrow{M}^{(2)}\overrightarrow{A}^{(3)}\overrightarrow{F}SAEDVLKEYDRRRRMEAL..$$

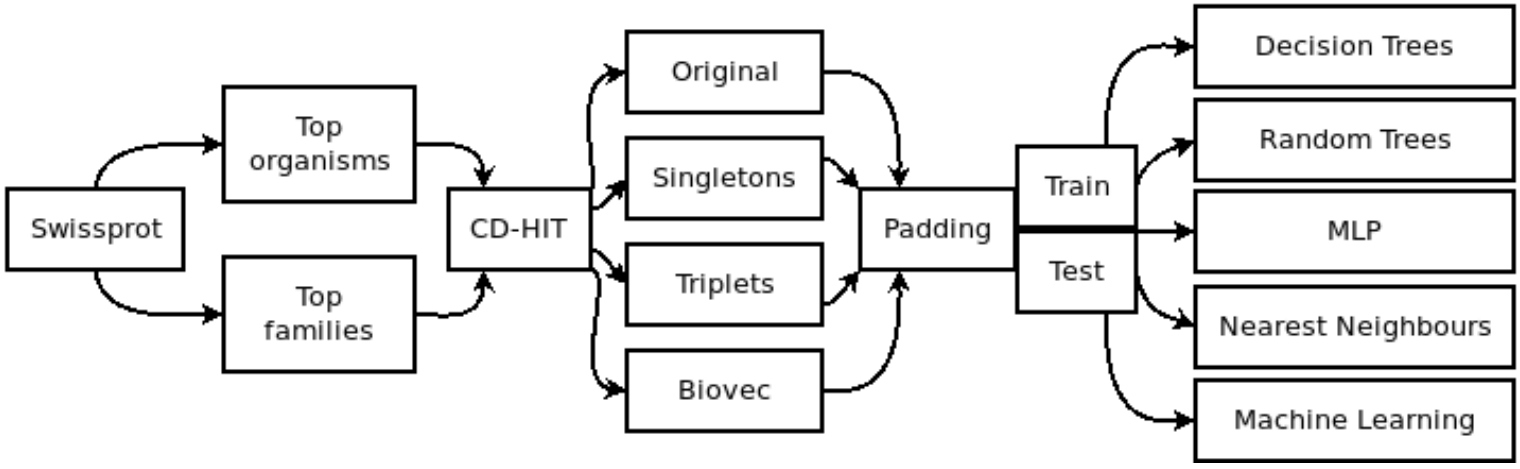
Splittings

$$\begin{cases} 1) & \text{MAF, SAE, DVL, KEY, DRR, RRM, ..} \\ 2) & \text{AFS, AED, VLK, EYD, RRR, RME, ..} \\ 3) & \text{FSA ,EDV, LKE, YDR, RRR, MEA, ..} \end{cases}$$

Compared methods of preprocessing

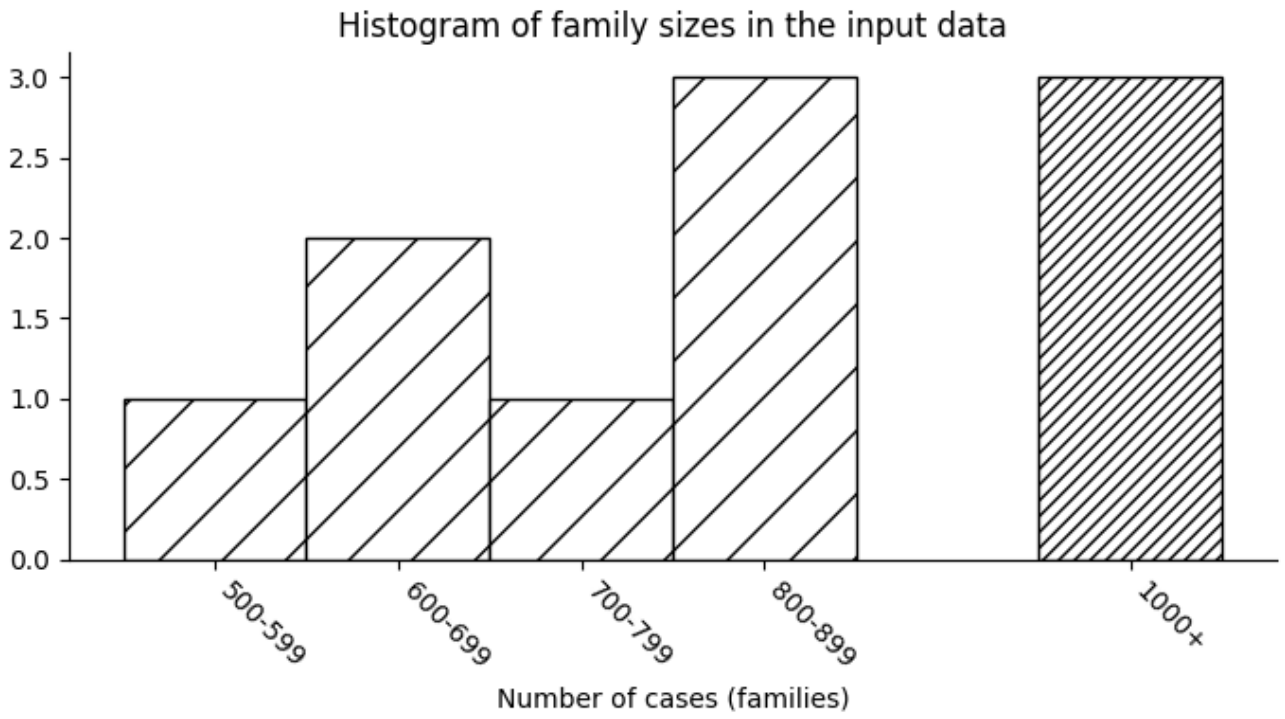
- Data source: Swissprot – manually reviewed part of the Uniprot[3] database
- Filtering by most frequent organisms and families
- CD-HIT[4] – removing very similar protein sequences
- Padding
- Splitting to stratified train and test pools
- Shuffling

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

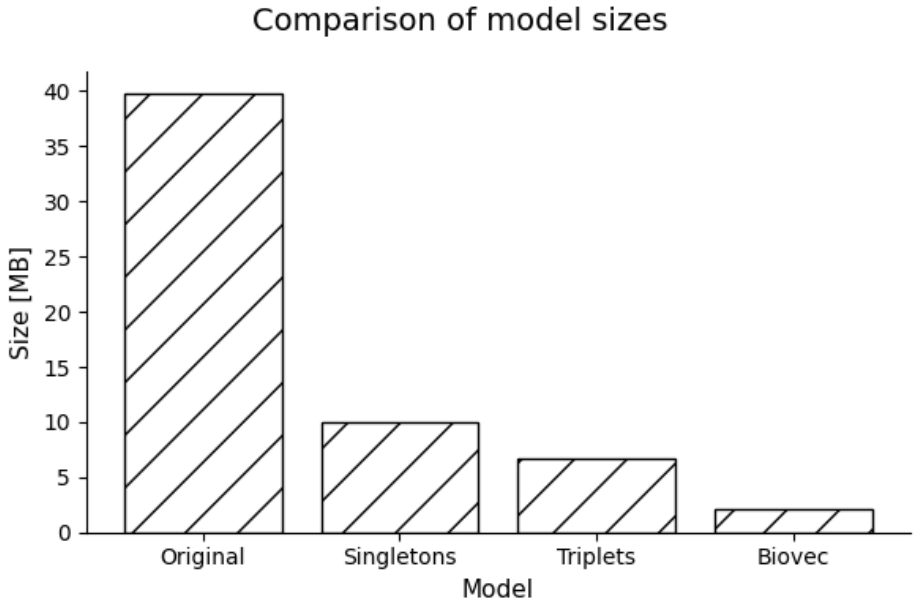


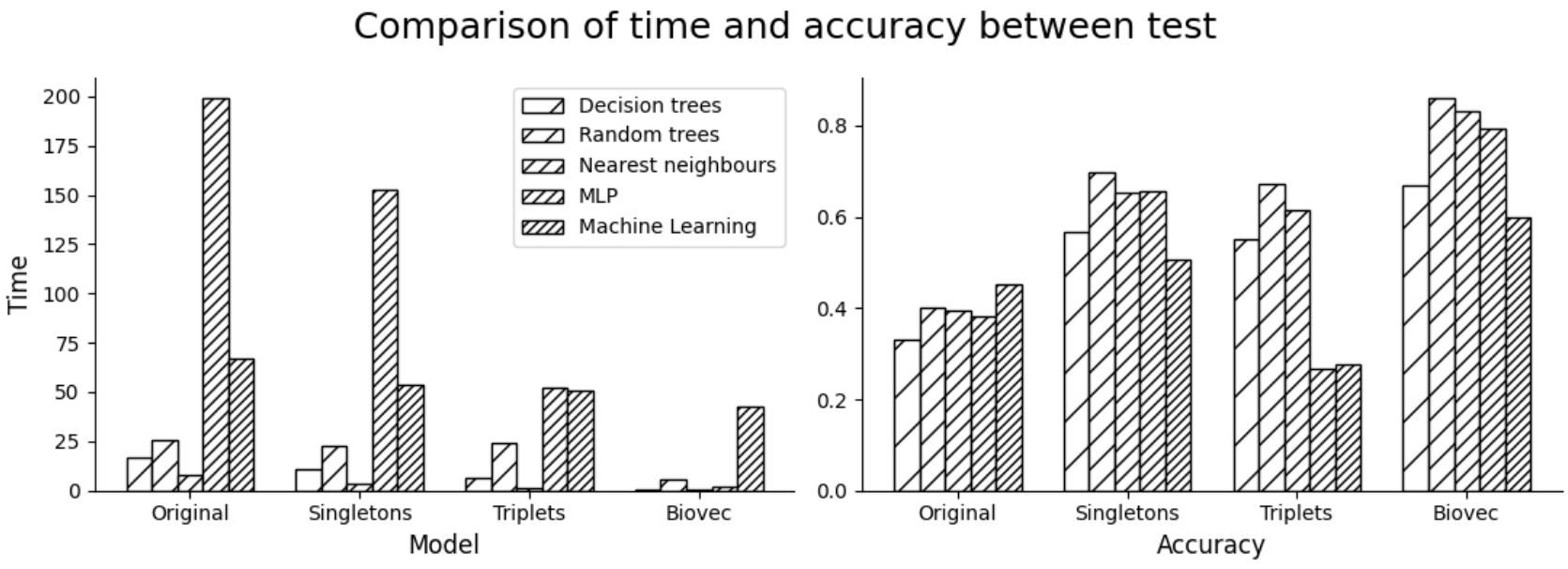
Tested models

- Decision trees
- Random trees
- MLP
- Nearest neighbours
- Machine Learning (Simple Dense model)
- Grid Search for optimal parameters
- Cross-validation



Results


Comparison of model sizes
Size [MB]
0
5
10
15
20
25
30
35
40
Original
Singletons
Triplets
Biovec
Model


Comparison of time and accuracy between test
Time
0
25
50
75
100
125
150
175
200
Original
Singletons
Triplets
Biovec
Model
Decision trees
Random trees
Nearest neighbours
MLP
Machine Learning
0.0
0.2
0.4
0.6
0.8
Original
Singletons
Triplets
Biovec
Accuracy

Summary

- Treating protein data as string objects might have a negative impact on the training process.
- Simple conversion to numerical data improves both accuracy and runtime.
- Further numerical categorization for triplets while speeding up the learning process is losing some of the information leading to decreased accuracy.
- The neighbourhood of amino acids can be analysed using a more sophisticated biovec model.
- Biovec yields both better results and up to 100 times faster training process (MLP) compared to the original model.
- Biovec model data weighs over 20 times less than the original one.

References

1. Article Source: Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics
2. https://gitlab.com/victiln/protvec Asgari E, Mofrad MRK (2015) Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics. PLOS ONE 10(11): e0141287. https://doi.org/10.1371/journal.pone.0141287
3. The UniProt Consortium, UniProt: the universal protein knowledgebase in 2021, Nucleic Acids Research, Volume 49, Issue D1, 8 January 2021, Pages D480–D489, https://doi.org/10.1093/nar/gkaa1100
4. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006 Jul 1;22(13):1658-9. doi: 10.1093/bioinformatics/btl158. Epub 2006 May 26. PMID: 16731699.