Classification of Terpene Synthases Using a Transformer-based Language Model

Klara Alicja Kotkowska, Mads Cort Nielsen and Niels Jakob Larsen - *DTU Health Tech & DTU Biosustain, Technical University of Denmark*

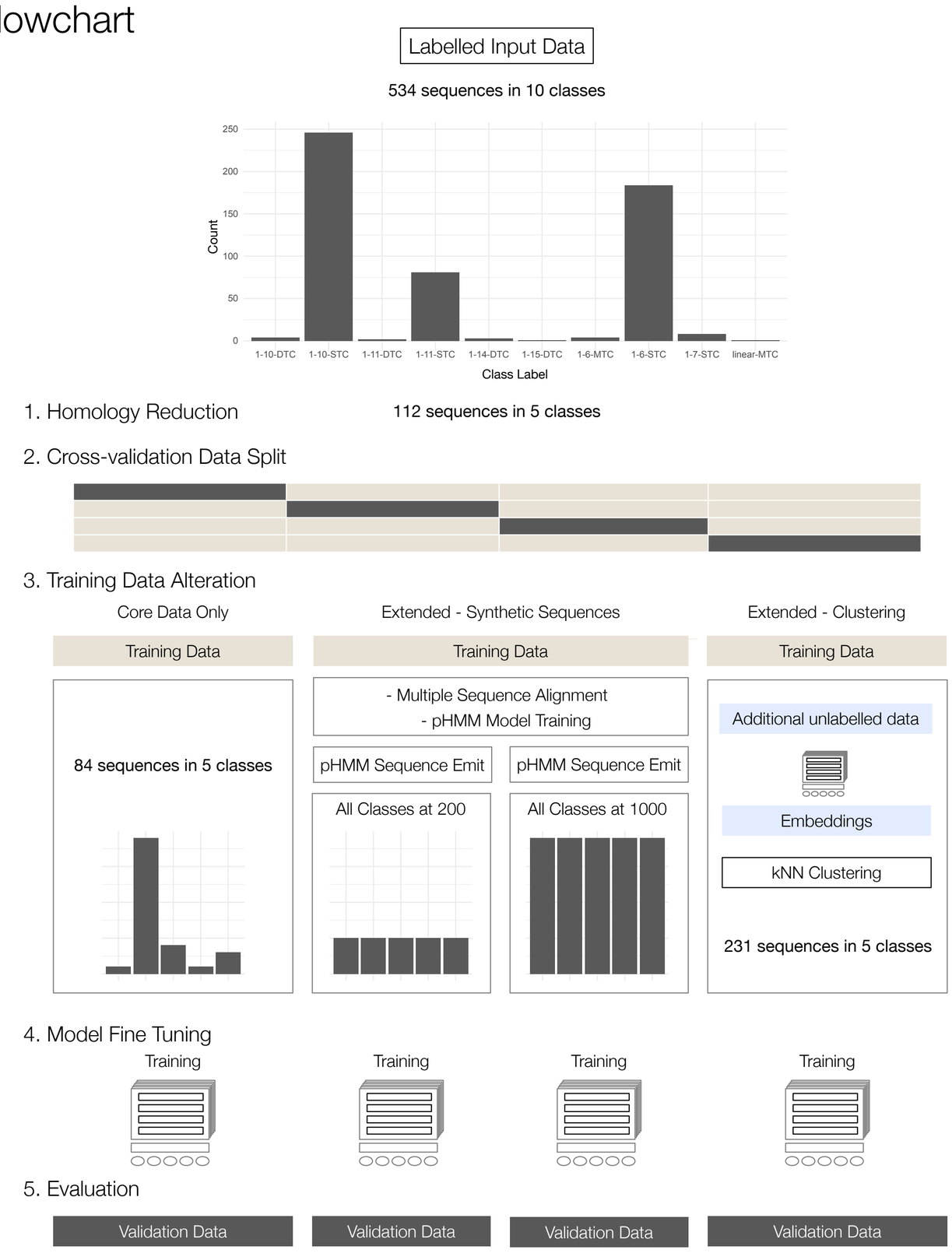
Introduction

Terpenoids or isoprenoid molecules are a very large class known for natural products. With estimates of 30.000 to 50.000 different molecules in the class (Yamadaa et al.). All the different terpenoid molecules have the same five-carbon isopentenyl diphosphate (isoprene) precursor. The isoprene units are added together to form a linear branched carbon chain. The linear branched carbon chain must undergo at least one cyclisation to form a terpenoid molecule. The cyclisation process is performed by the protein terpenoid cyclase/synthase (Rudolf et al.).

There are 534 biologically classified terpenoid cyclases. They have been clustered into classes based on which atoms are cyclised by the terpenoid cyclase. This project focused on fine-tuning the transformer ESM-2 (Lin et al.), to make a multi-class classifier to reliably classify unknown terpenoid cyclase sequences.

The recent big leaps in natural language modelling and transformers have enabled the processing and prediction of biological sequences more feasible.

Flowchart



Materials and Methods

The labelled input data was clustered using CD-hit (v. 4.8.1) for homology reduction. The core data was then split in four folds, with stratification (scikit-learn v. 1.1.3) for cross validation. For each fold, the training data was then used for the model as is or extended with two different methods. First method was aligning the data within each class (Clustalo v. 1.2.4), training a hidden Markov on the alignment data, and using this to extend the data by emitting synthetic sequences (hmmer v. 3.3.2) up to either 200 or 1000 sequences within each class. The second method was to include additional unlabelled synthase sequences in the data, extracting embeddings for all sequences from the ESM-2, and using this to cluster the additional sequences to the classes using a k-nearest neighbours algorithm (scikit-learn v. 1.1.3).

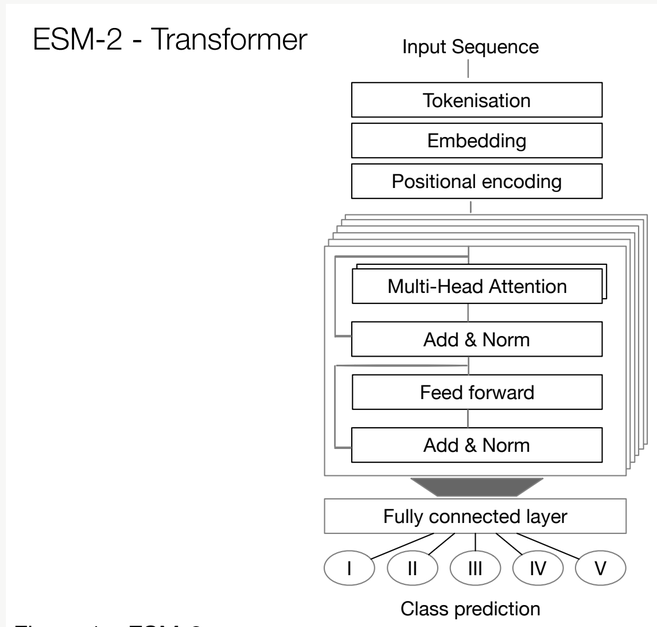


Figure 1 – ESM-2

ESM-2 Transformer

The ESM-2 model (Lin et al.) is a variation on roBERTa (Robustly Optimized BERT Pretraining Approach) which is based on BERT - **B**idirectional **E**ncoder **R**epresentations from **T**ransformers. The model is unsupervised pretrained on Uniref50 database by predicting masked amino acids. The specific model used here is a reduced version with 6 ESM transformer layers, a total of 8M parameters and an embedding size of 320. The output is connected to a dense layer with an output node for each terpene class. In this project we fine tuned it on each of the different training data splits, by training it for for 5 - 50 epochs with a batch size of 5. Each of the differently trained models was then evaluated on the same validation split.

Discussion

Small, homologous and imbalanced dataset was a major impediment to an acceptable fine-tuning of the model, as well as avoiding overfitting. The model performed better when trained on artificially extended datasets. As the model performed well only for the largest class, the metrics taking into account the balance of the data show fair performance.

Principal component analysis of sequences' embeddings (not shown) has not demonstrated clear separation of most of the classes. It might be that natural language models are not the most suitable for amino-acid sequence representation or that the protein structures of synthases are very similar.

For further development of the tool, it has to be taken into consideration, that there are also less common classes, that do not have a typical structure.

Results

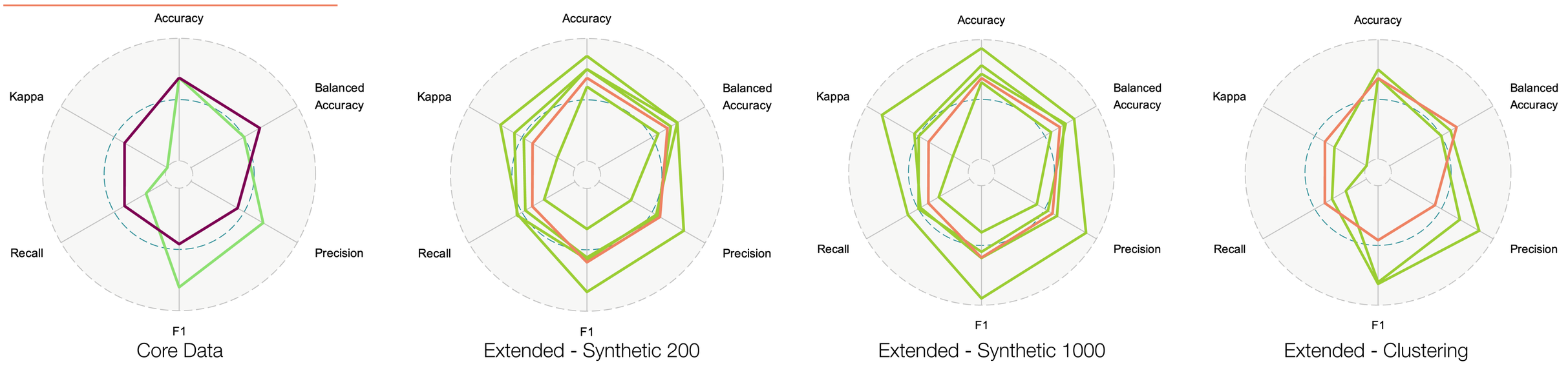


Figure 2 – Results of training ESM-2 on 3 altered training sets of terpene synthase sequences.

References

Evolutionary-scale prediction of atomic level protein structure with a language model. Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, Alexander Rives bioRxiv 2022.07.20.500902; doi: https://doi.org/10.1101/2022.07.20.500902

Terpene synthases are widely distributed in bacteria. Yuuki Yamadaa, Tomohisa Kuzuyamab, Mamoru Komatsua, Kazuo Shin-yac, Satoshi Omurad, David E. Canee, and Haruo Iked; doi: https://doi.org/10.1073/pnas.1422108112

Terpene synthases in disguise: enzymology, structure, and opportunities of non-canonical terpene synthases. Rudolf JD, Chang CY. Nat Prod Rep. 2020 Mar 25;37(3):425-463. doi: 10.1039/c9np00051h. PMID: 31650156; PMCID: PMC7101268.