Benchmarking of sequence preprocessing step for family classification task

Sikora Maciej[1]

[1] Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Poland

Abstract

With the growing amount of protein data, there is a need for improving automated prediction models. However, more sophisticated ones are increasingly resource-heavy and slower, thus requiring preprocessing steps to yield satisfying results in finite time. Because successful compression of biological sequences should speed up the learning process without losing crucial information, standard string compression might not fit here. Instead, protein-dedicated models are required, taking advantage of the smaller amino acid alphabet as well as factoring wider biological context behind one-letter abbreviations. Context analysing and learning models aren't a problem specific to protein data and are commonly used in natural language text analysis. Multiple such preprocessing models can be formulated as well as multiple algorithms can be used for the learning process depending on the research goal. Here I show a step-by-step comparison between several preprocessed protein datasets for the family classification tasks, evaluating with most common multi-class classifier algorithms with runtime and accuracy as metrics. Results show how important the sequence preparation step is, significantly improving model quality. We also highlight, that in pursuit of size reduction we shouldn't forget about accuracy assessment.

Introduction

For the last 5 years unannotated part of the Uniprot database – Swissprot, increased over 3 times from 73,711,881 (release 2017_01) to 230,328,648 entries (2022_01) [XXXXXX].

Methods and materials

Results and discussion

References