A Self-training Approach for Functional Annotation of UniProtKB Proteins

Maryam Abdollahyan[1,2], Rabie Saidi[2,*], Fabrizio Smeraldi[1], Maria J. Martin[2]
1 Queen Mary University of London, Mile End Road, London E1 4NS, UK
2 European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridgeshire CB10 1SD, UK
*To whom correspondence should be addressed: rsaidi@ebi.ac.uk

1. INTRODUCTION

Automatic annotation systems are essential to reduce the gap between the amount of protein sequence data and functional information in public databases such as UniProtKB (1). These systems rely on manually annotated (also called labelled) data to learn rules for predicting annotations. Manually labelled data are, however, often scarce or time consuming to obtain as they have to be reviewed by expert human curators. On the other hand, unlabelled data are abundant and comparatively easy to gather. In this work, we present a self-training (2) automatic annotation approach that utilises unlabelled data in order to improve the accuracy of predictions. We evaluated our system on a set of entries in UniProtKB/Swiss-Prot. The results show improvement in different performance metrics when self-training is used. The generated model was then used to predict metabolic pathway involvement of UniProtKB/TrEMBL proteins. As a result, it covered 86% of the proteins currently annotated by UniProt pipelines, but also could annotate 6.7 million proteins that lacked any previous pathway annotations.

2. MOTIVATION AND METHODS

In a previous work, we introduced the Association-Rule-Based Annotator (ARBA), a multiclass annotation system for automatic classification and annotation of UniProtKB proteins. The system was evaluated on UniProtKB/Swiss-Prot prokaryotic data where it achieved very promising results (3). However, 89.7% Of UniProtKB/Swiss-Prot entries have been automatically annotated using HAMAP annotation rules (4). This leads to the following question: can ARBA still predict annotations in the absence of annotations provided by HAMAP and using only manual annotations? To answer this question, we considered the task of predicting metabolic pathways in UniProtKB bacterial data using two different approaches: first, the original ARBA as defined in (3) and second, ARBA combined with self-training as introduced below.

In order to deal with small amount of labelled data, ARBA was self-trained on UniProtKB/Swiss-Prot data. Proteins that contain pathway annotations constitute the labelled dataset while those that do not contain any pathway annotation constitute the unlabelled dataset. The system performs self-training in two main steps. In the first step, annotations are propagated from the labelled to the unlabelled set based on their similarity. The similarity criterion is defined as a Boolean that is true if the two proteins have the same attributes; e.g., signatures. In the second step, ARBA iteratively learns from these data, retrains itself and adds to the labelled instances until a desired performance level is reached. The output of self-training is the final learning dataset using which the annotation model was built. This model was validated in two 2-fold cross-validation runs. The results, averaged over the two runs, are shown in Table 1. Finally, we applied this model to predict metabolic pathways in UniProtKB/TrEMBL bacterial data which are poorly covered, currently 3.5%. A comparison between the annotation coverage of the system before and after self-training is shown in Figure 1.

3. RESULTS AND DISCUSSION

Table 1 shows the evaluation metrics obtained by ARBA without HAMAP's annotations. Results indicate that while the system performs well in its original form, its performance is improved with self-training, as illustrated by noticeable increases in recall and AUC.

Table 1. Evaluation metrics obtained by ARBA before and after self-training

Metric	Without Self-training	With Self-training
Precision	98.4%	99.7%
Recall	71.2%	89.4%
AUC	86.2%	95.8%

Figure 1 provides statistics from the UniProtKB/TrEMBL proteins annotated by ARBA. Without self-training, ARBA covered 4,564,250 entries, where 3,083,501 proteins (denoted by *New*) lacked any previous pathway annotations, resulting in an increase in pathway coverage from 3.5% to 7.4%, and 1,480,749 proteins (denoted by *Overlap*) had been previously annotated by UniProt pipelines. 1,224,264 proteins (denoted by *Missing*) that had been previously annotated were not covered by ARBA. In comparison, with self-training, the number of proteins exclusively annotated by ARBA was increased to 6,687,267, resulting in a coverage of 12.1%. The number of missed entries was notably reduced to 394,205. These results demonstrate the benefits of using self-training algorithms to provide functional annotations for UniProtKB where manually curated annotations are rare.

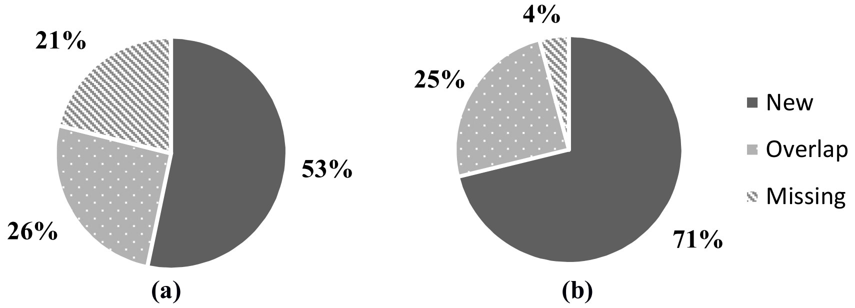

Figure 1. Pathway annotation coverage for UniProtKB/TrEMBL by ARBA, (a) without self-training and (b) with self-training

4. AVAILABILITY

Models for pathway prediction, generated with and without self-training based on release 2017_02 along with a Java Archive (JAR) package for applying them to UniProtKB bacterial data are available at http://www.ebi.ac.uk/~rsaidi/arba/self-training/.

5. REFERENCES

1. The UniProt Consortium. 2017. UniProt: the universal protein knowledgebase. *Nucleic Acids Research* 45(D1): D158-D169.
2. Zhu X. 2005. Semi-supervised learning literature survey (Report No. 1530). Computer Sciences, University of Wisconsin-Madison.
3. Boudellioua I. et al. 2016. Prediction of Metabolic Pathway Involvement in Prokaryotic UniProtKB Data by Association Rule Mining. *PLOS ONE* 11(7): e0158896.
4. Pedruzzi I. et al. 2015. HAMAP in 2015: updates to the protein family classification and annotation system. *Nucleic Acids Research* 43(D1): D1064-D1070.