A Machine-Learning Approach for the Prediction of Enzymatic Activity of Proteins in Metagenomic Samples

Theodoros Koutsandreas, Eleftherios Pilalis, and Aristotelis Chatziioannou

Metabolic Engineering & Bioinformatics Program, Institute of Biology, Medicinal Chemistry & Biotechnology, National Hellenic Research Foundation, Athens, Greece
th_koutsandreas@hotmail.com, {epilalis,achatzi}@eie.gr

Abstract. In this work, a machine-learning approach was developed, which performs the prediction of the putative enzymatic function of unknown proteins, based on the PFAM protein domain database and the Enzyme Commission (EC) numbers that describe the enzymatic activities. The classifier was trained with well annotated protein datasets from the Uniprot database, in order to define the characteristic domains of each enzymatic sub-category in the class of Hydrolases. As a conclusion, the machine-learning procedure based on Hmmer3 scores against the PFAM database can accurately predict the enzymatic activity of unknown proteins as a part of metagenomic analysis workflows.

Keywords: Machine-learning, Enzymes, Proteins, Metagenomics.

1 Introduction

The emerging field of Metagenomics comprises the collection and analysis of large amounts of DNA that is contained in an environmental niche [1]. Due to the recent advances in high-throughput sequencing, very large amounts of nucleotide sequences can be generated in short time. Because of the increased volume of data, metagenomics is a promising way to identify novel enzymes and protein functions. However, despite the advances in high-throughput sequencing, the development of appropriate analysis tools remains challenging. Here, we developed a classifier for the prediction of protein enzymatic activity in metagenomic samples. Enzymes are proteins that are used in a wide range of applications and industries, such as Biotechnology and Biomedicine. In order to correlate unknown amino acid/nucleotide sequences with enzyme classes the PFAM database and the Enzyme Nomenclature system were used. PFAM is a database of protein families [2]. Each family is represented by a multiple sequence alignment which is generated by Hidden Markov Models (HMMs) with the Hmmer3 [3] program. Proteins consist of one or more functional regions which are called domains, i.e. the existence of a domain in the tertiary structure of a protein, imply a specific function. Thus, proteins of the same family will include identical or similar domains. The PFAM database contains information about protein families,

H. Papadopoulos et al. (Eds.): AIAI 2013, IFIP AICT 412, pp. 81–87, 2013.

their domains and their architecture. Each entry, represented by a PFAM id, corresponds to a single domain. The similarity of an unknown protein with a protein domain may give great information about its function and its phylogenetic relationships.

The Enzyme Nomenclature (EC) is a numerical classification system for enzymes, based on the chemical reaction that they catalyze. It was developed under the auspices of the International Union of Biochemistry and Molecular Biology during the second half of twenty-first century. Each entry of Enzyme Nomenclature is a four-number code, the enzyme commission number (EC number), which is associated with a specific chemical reaction. Thus each enzyme receives the appropriate EC number according to its chemical activity. The first number of code specifies the major category of catalyzed chemical reaction. There are six major categories of catalysed biochemical reactions: Oxidoreductases: 1.-.-.-, Transferases: 2.-.-.-, Hydrolases: 3.-.-.-, Lyases: 4.-.-.-, Isomerases: 5.-.-.-, Ligases: 6.-.-.-. The next two numbers specify the subclasses of major class and the last one states the substrate of the reaction. For instance, the EC number 3.1.3.- refers to the hydrolysis of phosphoric mono-ester bond and 3.1.3.11 refers to the hydrolysis of fructose-bisphosphatase which contains a phosphoric mono-ester bond. The classifier developed in the current study was able to classify unknown amino acid sequences originating from metagenomic analysis to hydrolases classes pursuant to the results of Hidden Markov Model detection. The classifier consisted of separate classification models, where the classification type was binominal, i.e. is EC number or is not EC number. In order to train the classifier we used well-annotated proteins and analyzed them with Hmmer3. The result was the score of similarity between an examined sequence and a protein domain. As a result, the features of the training data were the PFAM ids and the vector of each training example included its scores to the appropriate fields.

2 Dataset

In order to train the classification models, we used well-annotated proteins from the UniProt database [4]. We specifically selected all the reviewed sequences from the reference proteome set of bacteria (taxonomy: 2). The Uniprot database was used because it is a high quality, manually annotated and non-redundant protein database. A total number of 45612 sequences were collected. During the training of the classification models, we selected an amount of known proteins according to their EC numbers. A separate classification model was trained for each EC number, using binominal training data as positive and negative examples. The positive examples were sequences that belonged to a specific EC number (for example 3.1.3.1). In contrast, sequences that belonged to the same EC upper class, but differ in the last digit (i.e. 3.1.1.-), were the negative examples. In this way, we aimed at the detection of differences in features context, which separated a specific EC number protein family from all the other proteins whose EC number differed in the last digit.

3 Methods

3.1 Training of Enzymatic Classification Models

In order to train separate models for each enzymatic category of the Hydrolases class we implemented a procedure which automatically constructed the corresponding training sets (Fig. 1). In the first step, all sequences that are annotated with the specific EC number were selected from the dataset as positive examples. Each EC number has its upper class number (for EC number 3.-.-.- the upper class is all the no 3.-.-.- classes, i.e. all enzymes which are not hydrolases). In the second step the procedure selected, as negative example set, an equal amount of sequences that belonged to the upper class but not to the specific EC number. Consequently, the difference between positive and negative training data examples was the last EC number digit. Note that there were some conditions to be tested, especially for the EC numbers with three or four digits, during the execution of this step. The procedure stopped if the amount of positive examples was less than five or if negative examples were not found. Thereafter these two sets of examples were analyzed by Hmmer3 against the HMM profiles of the PFAM-A database in order to collect the domain scores as training features. For each enzymatic category an HMM profile library was thus constructed, which contained all the PFAM domains having a score against the sequences in the corresponding training set. For models corresponding to three- or four-digits EC numbers, a custom HMM profile was automatically constructed from the positive example set and was added to the HMM profile library as an additional feature representing the whole sequence length. The training was performed with the k-nearest neighbor (k-NN) algorithm using the Euclidean distance of 4-nearest neighbors, in a10-fold cross-validation process (stratified sampling). The choice of the parameter k=4 was made based on the size of the smaller training sets (12-15 examples). Thus, a stable value 4 was given to k representing approximately the one third of the smallest training data.

3.2 Application of the Trained Models to Unlabeled Sequences

The classification procedure automatically performed Hmmer3 analysis on the unlabeled sequences against the PFAM database and collected their scores as features (Fig. 2). Afterwards, the trained models were applied to the unlabeled feature sets starting from the general classification in Hydrolases class (i.e. EC number 3.-.-.-). Then the procedure continued to the subclasses with two, three and four EC number digits. The classification procedure was not hierarchical top-down as there was not any filtering method during the descending in enzyme subclasses, like the exclusion of sequences which are not annotated as hydrolases in the first classification task. Filtering was avoided because we observed that the classification procedure in main hydrolase class (3.-.-.-) and its subclasses (3.1.-.-, 3.2.-.- etc) had some false negatives that were correctly classified during the next steps in more detailed EC number classes with three or four digits. However, not all sequences passed through the classification procedures. As mentioned above, the training set of each EC number contained a particular set of features (HMM profile library). In the unlabeled sequences

classification task, the sequences having no score against this feature set were filtered out as they were considered distant from the upper-level EC number category. The volume of unlabeled sequences to be tested was thus reduced as an execution time optimization of the procedure.

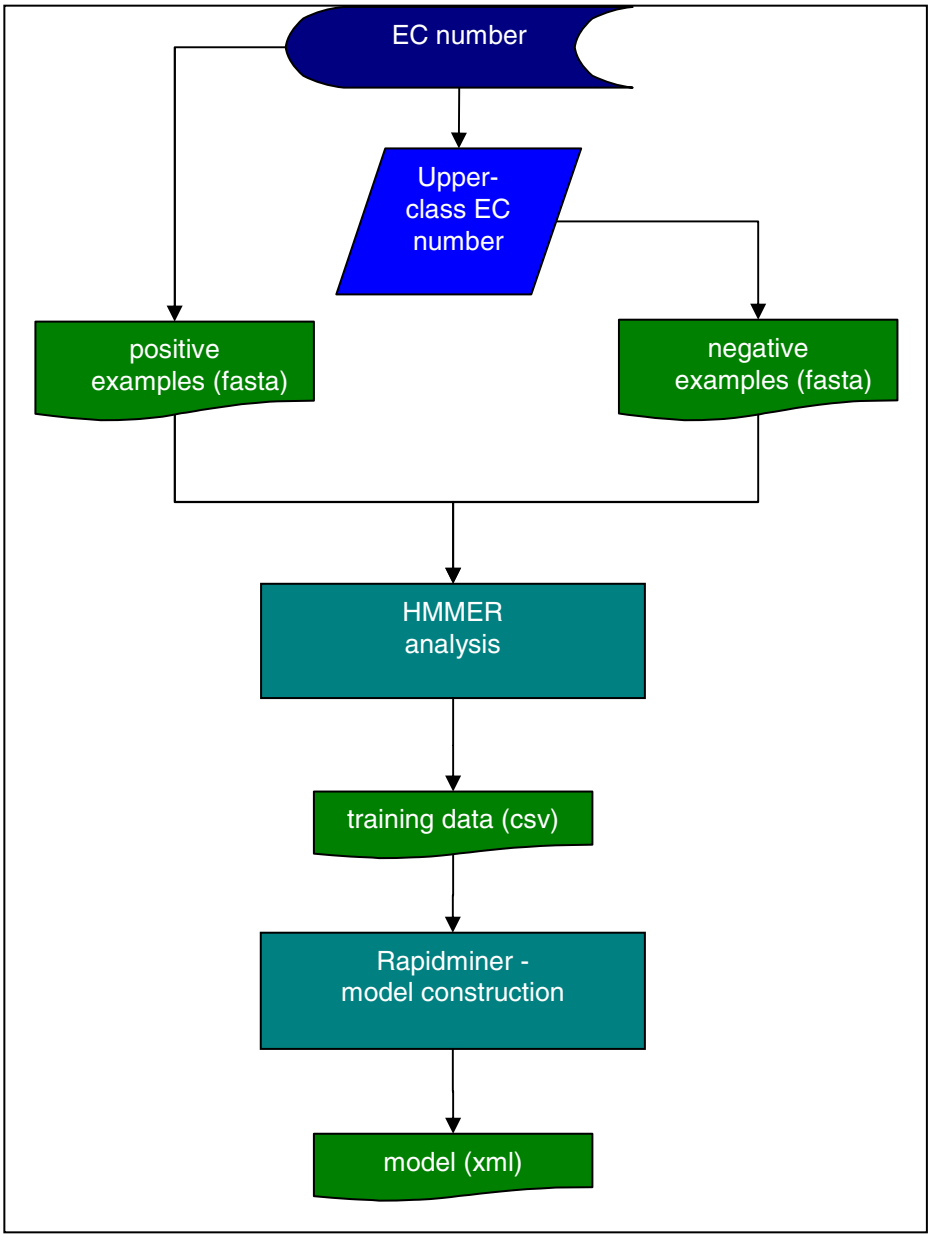

Fig. 1. Procedure of training of the EC number classification models

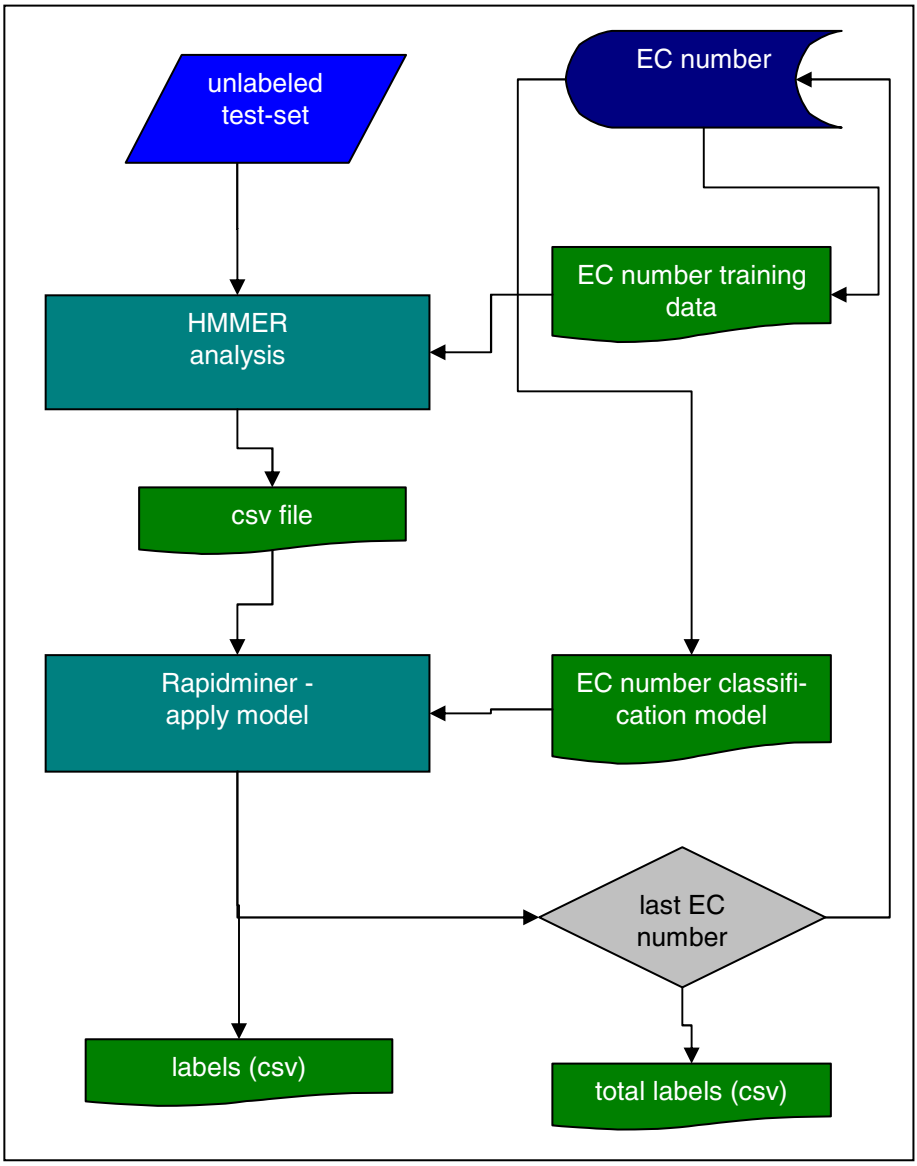

Fig. 2. Procedure of classification of unlabeled sequences

3.3 Software Tools

The aforementioned procedures were implemented in Python scripts. All datasets and results were stored and queried in a MySQL database. The training and the application of the models was performed using RapidMiner (http://rapid-i.com/).

4 Results

The procedure comprised 163 classifiers. The accuracies and F-scores in 10-fold cross validation, of a representative classifiers part, are listed in Table 1. The vast majority of classifiers had an accuracy of above 95%. Table 2 indicates the mean value and the standard deviation of accuracies for each level of enzyme class. Classifiers with two digits (i.e. 3.1.-.-, 3.2.-.-, 3.4.-.-, 3.5.-.-, 3.6.-.-) had a mean accuracy of 94.56%, those with three digits had a mean accuracy of 97.71% and those with four digits had a mean accuracy of 99.39%. The high performances of the classifiers show that the PFAM domain scores were able to separate the training examples accurately. This state indicates a high vector profile difference between two classes and the lack of significant noise.

Table 1. Classifiers accuracies in 10-fold cross-validation

EC number	Accuracy (%)	F_score (%)	EC number	Accuracy (%)	F_score (%)
3.-.-.-	90.25	88.40	3.1.1.45	100.00	100.00
3.1.-.-	94.26	94.42	3.1.1.61	100.00	100.00
3.5.-.-	93.44	93.76	3.1.1.85	100.00	100.00
3.-6.-.-	94.76	94.33	3.1.21.2	100.00	100.00
3.1.1.-	96.03	95.89	3.1.26.3	100.00	100.00
3.1.13.-	100.00	100.00	3.1.2.6	100.00	100.00
3.1.2.-	96.12	96.00	3.1.3.1	95.45	90.91
3.1.21.-	94.74	94.12	3.1.3.5	99.00	98.88
3.1.22.-	99.20	99.13	3.1.4.17	100.00	100.00
3.1.3.-	97.21	97.18	3.2.1.52	100.00	100.00
3.2.1.-	98.39	98.57	3.2.2.27	100.00	100.00
3.4.11.-	95.29	95.12	3.4.11.9	100.00	100.00
3.5.3.-	98.64	98.53	3.4.13.9	100.00	100.00
3.6.1.-	99.31	99.33	3.4.16.4	93.33	94.12
3.6.5.-	99.29	99.23	3.4.21.107	100.00	100.00
3.1.11.2	100.00	100.00	3.5.1.1	100.00	100.00
3.1.11.5	100.00	100.00	3.5.2.5	100.00	100.00
3.1.11.6	98.29	98.86	3.5.3.6	100.00	100.00
3.1.13.1	100.00	100.00	3.5.4.4	97.14	96.30
3.1.1.1	95.83	93.33	3.6.1.7	100.00	100.00
3.1.1.29	99.31	99.25	3.6.3.12	100.00	100.00
3.1.1.3	87.50	82.35	3.6.3.30	100.00	100.00
3.1.1.31	100.00	100.00	3.6.4.12	99.21	99.43

Table 2. Mean and standard deviation in function with the amount of digits in EC numbers

amount of digits	mean value	standard deviation
2	94.86	0.45
3	97.71	0.19
4	99.39	0.09

5 Conclusion

In conclusion, the machine-learning procedure based on Hmmer3 scores against the PFAM database performed well and accurately predicted the enzymatic activity of unknown proteins. Future developments will include the use of other protein motifs databases, like CATH and SCOP and the development of more efficient data mining algorithms. The procedure will also be extended to other enzymatic classes (here we focused on the group of Hydrolases) and will be run-time optimized for its application on very large datasets. Finally, it will be implemented as an independent tool and it will be integrated in more extended metagenomic analysis workflows.

Acknowledgements. The presented work in this paper has been funded by the "Co-operation" program 09SYN-11-675 (DAMP), O.P. Competitiveness & Entrepreneurship (EPAN II).

References

1. Lorenz, P., Eck, J.: Metagenomics and industrial applications. Nat. Rev. Microbiol. 3(6), 510–516 (2005)
2. Finn, R.D., et al.: The PFAM protein families database. Nucleic Acids Res. 36(Database issue), D281–D288 (2008)
3. Finn, R.D., Clements, J., Eddy, S.R.: HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 39(Web Server issue), W29–W37 (2011)
4. Apweiler, R., et al.: UniProt: The Universal Protein knowledgebase. Nucleic Acids Res. 32(Database issue), D115–D119 (2004)