# Prediction of Enzymatic Activity of Proteins Based on Structural and Functional Domains

Theodoros G. Koutsandreas, Eleftherios D. Pilalis, and Aristotelis A. Chatziioannou, *Member, IEEE*

***Abstract*—The prediction of the putative enzymatic function of uncharacterized proteins is a major problem in the field of metagenomic research, where large amounts of sequences can be rapidly determined. In this work a machine-learning approach was developed, that attempts the prediction of enzymatic activity based on three protein domain databases, PFAM, CATH and SCOP, which contain functional and structural information of proteins as Hidden Markov Models. Separate and combined classifiers were trained by well-annotated data and their performance was assessed in order to compare the predictive power of different attribute sets corresponding to the three protein domain databases. All classifiers performed well, with an average accuracy of ~96% and an average AUC score of 0.84. As a conclusion, the classification procedure can be integrated to more extended metagenomic analysis workflows.**

## I. INTRODUCTION

DUE to the recent advances in high-throughput sequencing, very large amounts of nucleotide sequences can be rapidly generated[1]. The functional characterization of these sequences, and in particular of the corresponding encoded proteins, is a challenging task that demands the development of novel data mining tools. In this work, we developed a multi-classifier for the prediction of enzymatic activity of novel protein sequences. Enzymes are proteins that are used in a wide range of applications and industries, such as Biotechnology and Biomedicine[2].

In order to classify unknown sequences into enzyme categories, the PFAM [3], CATH [4], SCOP [5] databases and the Enzyme Nomenclature system were used as attributes. The Enzyme Nomenclature (EC) [6] is a numerical classification system for enzymes, based on the chemical reaction that they catalyze. Each entry of Enzyme Nomenclature is a four-field 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 field specifies the major category of catalyzed chemical reaction (Oxidoreductases: 1.-.-.-, Transferases: 2.-.-.-, Hydrolases: 3.-.-.-, Lyases: 4.-.-.-, Isomerases: 5.-.-.-, Ligases: 6.-.-.-). The next two fields specify the subclasses of the 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.

Proteins consist of one or more functional and/or structural regions which are called domains and imply specific functions. PFAM is a database of protein families that contains functional information on conserved protein domains through evolution. Each family is represented by a multiple sequence alignment, which serves as the basis for the generation of a Hidden Markov Model (HMM). CATH and SCOP are structural domain databases that contain information on different protein folds and adopt a top-down hierarchy system in order to classify protein domains in several levels according to sequence similarity, secondary structure, architecture, function similarity etc. They also represent each protein domain with an HMM profile and are partially, but not completely, overlapping. The exploitation of all aforementioned databases is an appropriate means for constructing protein attribute spaces, because the conserved protein domains reflect the modular nature of proteins structure and function. In Table I are shown examples of HMM databases features.

TABLE I
HMM PROFILE EXAMPLE FOR EACH DATABASE

| Database type | ID | Participating function |
|---|---|---|
| PFAM | PF00122 | ATP hydrolysis (proton as ligand) |
| CATH | 3kvnA01 | Cell outer membrane, biofilm formation, lipid biosynthetic process, cell motility |
| SCOP | 0048053 | Carboxylic ester bond hydrolysis |

PFAM, CATH and SCOP Hidden Markov Model id examples. Each model is generated by the multiple sequence alignment of sequences which are characterized by a specific functionality.

The multi-classifier presented in this work was intended to predict EC numbers by using as attributes the domain

Manuscript received July 30, 2013. This work was supported by the "Cooperation" program 09SYN-11-675 (DAMP), O.P. Competitiveness & Entrepreneurship (EPAN II).

T. G. Koutsandreas is with the Metabolic Engineering and Bioinformatics Program, Institute of Biology, Medicinal Chemistry and Biotechnology, National Hellenic Research Foundation, Athens, Greece (e-mail: th_koutsandreas@hotmail.com).

E. D. Pilalis is with the Metabolic Engineering and Bioinformatics Program, Institute of Biology, Medicinal Chemistry and Biotechnology, National Hellenic Research Foundation, Athens, Greece (e-mail: epilalis@eie.gr).

A. A. Chatziioannou is with the Metabolic Engineering and Bioinformatics Program, Institute of Biology, Medicinal Chemistry and Biotechnology, National Hellenic Research Foundation, Athens, Greece (corresponding author to provide phone: +302107273751, e-mail: achatzi@eie.gr).

978-1-4799-3163-7/13/$31.00 ©2013 IEEE

matching scores of protein sequences in HMM analyses. For each enzymatic category in the level of four fields, four different models were trained, one for each domain database and one with all databases combined, in the scope to compare the classification performance of the three different attribute spaces. In addition, two different training modes were employed and compared, one that comprised structurally but not functionally similar sequences in the negative examples, and one with inversely, functionally, but not necessarily structurally, related sequences.

## II. METHODS

### A. Protein Dataset

The enzymatic class of Hydrolases was selected due to their high biotechnological and industrial interest [7-8]. In order to train the classification models protein data were retreived from the UniProt/SwissProt database [9], which contains high quality, manually annotated protein sequences. All the reviewed sequences whose EC number was belonged to 3.-.-.- class (i.e. Hydrolases), were collected, which corresponds to a total amount of 42767 sequences. Also, we collected 383225 reviewed sequences from complete proteome sets, which were not Hydrolases (i.e. they belonged to other enzyme classes or they were not enzymes) and their annotation fields did not contain phrases like "hypothetical", "putative", "probable", "by similarity", "by homology", "enzyme", "virus" and "fragment", in order to reject incompletely annotated sequences, viral sequences and protein fragments

### B. Training of enzyme 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. For each four-field EC number, two different training sets were constructed. In both sets the sequences for the positive examples were protein sequences annotated with the specific four-field EC number. In regard to the negative examples, in the first training set (Training 1) the procedure selected an approximate equal amount of random sequences and sequences that belonged to the upper EC number class but not to the specific EC number (i.e. they catalyze the same chemical reaction with different substrate). In the second training set (Training 2), the negative protein set was constructed by submitting the positive examples to a three-iteration PSI-BLAST [10] procedure (e-value threshold 0.01) in order to collect similar sequences, which did not belong to the two-level upper EC number class or they are not enzymes. The USEARCH clustering tool [11] was used to reduce the redundancy of all sets. The minimum amount of sequences representing an enzyme category was set to 5.

In order to construct the attribute space, the initial protein sequence sets were analyzed by HMMER3 [12] against the HMM profiles of the PFAM-A, the CATH and the SCOP databases (e-value cut-off <0.01), in order to collect the domain matching scores as training features. For each enzymatic category an HMM profile library was thus constructed, which contained all the PFAM, CATH and SCOP domains having a matching score against the sequences in the corresponding training set.

For each enzymatic category, the procedure constructed a classification model corresponding to the three domain types (PFAM, CATH, SCOP) and a model with all domain features together, in order to compare the performances of the three different attributes separately and combined. Training data dimensionalities differed among these three databases. An average amount of PFAM training features was ~ 90, while CATH and SCOP yielded an average amount of ~250 and ~400 features respectively.

The training was performed with the k-nearest neighbor (k-NN) algorithm, embedded in a 10-fold cross-validation process (stratified sampling). For training sets with less than 50 examples, defined as "small training sets", we used a bootstrapping operator (sampling with replacement) in order to reach the amount of 50 examples. Also for each training set, we calculated the information gain ratio, in order to filter-out features with small information content and to reduce data dimensionality. Additionally, a pre-training optimization procedure was performed, in order to find the best k-NN parameters (k amount, numerical measure). We tested k-NN performance for k=[1,10] for "small training sets" and k=[1,20] for larger data sets in function with the following distances: Euclidean Distance, Chebychev Distance, Correlation Similarity, Dice Similarity, Inner Product Similarity, Jaccard Similarity, Manhattan Distance, Max Product Similarity, Overlap Similarity. In order to optimize the parameter set for each classifier, the Accuracy and the Area Under Curve (AUC) scores were calculated. The double of k and the distance metric that maximized the mean value of accuracy and AUC sum was determined as the best parameter set for each classifier independently. These two measures represent the overall classifier performance and its ability to distinguish the two classes. The AUC is a measure of the probability that the classifier will correctly classify a randomly selected example, thus indicating the performance of the model in regard to overfitting. Additionally, the F-score, which is an estimation of the accuracy based on both the precision and the recall, was calculated as following:

$$Fscore = 2 \times \frac{precision \times recall}{precision + recall}$$

### C. 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 [13].

## III. RESULTS

The multi-classifier performed well in predicting EC numbers by using as attributes the domain matching scores of protein sequences in HMM analyses. The final multi-classifier consisted of 408 models derived from Training 1 datasets and 225 models derived from Training 2 datasets. The majority of the classification models were optimized with the Euclidean Distance as the k-NN distance metric, while the optimized value of k was depended on each training data. The average values of the performance scores are shown In Table II and Table III.

TABLE II
AVERAGE PERFORMANCE OF TRAINING 1 MODELS

| Database type | Accuracy (%) | AUC | F-Score (%) |
|---|---|---|---|
| CATH | 95.94 +/- 7.48 | 0.83 +/- 0.14 | 94.80 +/- 11.25 |
| PFAM | 97.41 +/- 5.59 | 0.84 +/- 0.13 | 97.23 +/- 5.53 |
| SCOP | 96.42 +/- 7.05 | 0.84 +/- 0.14 | 95.54 +/- 10.29 |
| TOTAL | 97.30 +/- 5.45 | 0.84 +/- 0.13 | 96.86 +/- 6.81 |

Average values of accuracy, Area Under Curve and F-score in Training 1, where the negative example set comprised functionally-related proteins (an upper EC number level).

The classifiers based on either CATH, PFAM or SCOP domains showed high accuracies and were able to effectively separate the enzymatic classes. The combination of the three types of features in one classifier did not seem to offer any significant advantage.

TABLE III
AVERAGE PERFORMANCE OF TRAINING 2 MODELS

| Database type | Accuracy (%) | AUC | F-Score (%) |
|---|---|---|---|
| CATH | 94.13 +/- 8.62 | 0.83 +/- 0.15 | 92.53 +/- 13.42 |
| PFAM | 94.21 +/- 7.04 | 0.83 +/- 0.14 | 93.02 +/- 10.61 |
| SCOP | 94.97 +/- 5.92 | 0.84 +/- 0.14 | 94.04 +/- 8.21 |
| TOTAL | 95.46 +/- 5.92 | 0.84 +/- 0.13 | 94.76 +/- 7.22 |

Average values of accuracy, Area Under Curve and F-score in Training 2, where the negative example set comprised structurally-related proteins (obtained by PSI-BLAST).

## IV. DISCUSSION

Overall, with respect to the particular enzymatic class of Hydrolases, both training methods and all three attribute spaces yielded roughly the same performance. In future work the performance of the procedure in all enzymatic classes will be investigated and the best attribute spaces will be included in an integrative enzyme classifier. Additionally, the exploitation of domain attributes may be evaluated for protein function classification problems other than the enzymatic activity.

Moreover, in future development of this approach, the dimensionality reduction of the combined feature space by Principal Component Analysis may reduce the risk of over-fitting in regard to novel, uncharacterized sequences.

Finally, the classification procedure will be integrated in metagenomic analysis workflows given that the dimensionality of the final attribute spaces for each enzyme category is not prohibitive for high throughput environments.

## REFERENCES


[1] Logares, R., et al., Environmental microbiology through the lens of high-throughput DNA sequencing: Synopsis of current platforms and bioinformatics approaches, Journal of Microbiological Methods, 2012. 91(1): p.106-13.

[2] Kirk O. et al., Industrial enzyme applications, Current Opinion in Biotechnology, 2002. 13: p. 345-51.

[3] Finn, R.D., et al., *The PFAM protein families database.* Nucleic Acids Res, 2008. **36**(Database issue): p. D281-8.

[4] Sillitoe, I., et al., New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. Nucleic Acids Res, 2013. 41(Database issue): p. D490-8.

[5] Lo Conte, L., et al., SCOP: a structural classification of proteins database. Nucleic Acids Res, 2000. 28(1): p. 257-9.

[6] Barrett AJ (1992) Enzyme Nomenclature. Academic Press, San Diego, California.

[7] Delgado-G M., et al. Halophilic hydrolases as a new tool for the biotechnological industries, J Sci Food Agric, 2012. 92: p. 2575-80.

[8] Luen-Luen L., et al. Bioprospecting metagenomes: glycosyl hydrolases for converting biomass, Biotechnology for Biofuels 2009. 2(10)

[9] Apweiler, R., et al., *UniProt: the Universal Protein knowledgebase.* Nucleic Acids Res, 2004. **32**(Database issue): p. D115-9.

[10] Altschul S.F., et al., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Research, 1997. 25(17): p. 3389-402.

[11] Edgar, R.C., Search and clustering orders of magnitude faster than BLAST. Bioinformatics, 2010. 26(19): p. 2460-1.

[12] Finn, R.D., J. Clements, and S.R. Eddy, *HMMER web server: interactive sequence similarity searching.* Nucleic Acids Res, 2011. **39**(Web Server issue): p. W29-37.

[13] http://rapid-i.com/