A machine learning methodology for enzyme functional classification combining structural and protein sequence descriptors

Afshine Amidi[1], Shervine Amidi[1], Dimitrios Vlachakis[2], Nikos Paragios[1,3], Evangelia I. Zacharaki[1,3]

[1] Center for Visual Computing, Department of Applied Mathematics, École Centrale de Paris, 92295 Châtenay-Malabry, France
[2] Bioinformatics and Medical Informatics Laboratory, Biomedical Research Foundation of the Academy of Athens, Athens, Greece
[3] Equipe GALEN, INRIA Saclay, Île-de-France, Orsay, France

Abstract. The massive expansion of the worldwide Protein Data Bank (PDB) provides new opportunities for computational approaches which can learn from available data and extrapolate the knowledge into new coming instances. The aim of this work is to apply machine learning in order to train prediction models using data acquired by costly experimental procedures and perform enzyme functional classification. Enzymes constitute key pharmacological targets and the knowledge on the chemical reactions they catalyze is very important for the development of potent molecular agents that will either suppress or enhance the function of the given enzyme, thus modulating a pathogenicity, an illness or even the phenotype. Classification is performed on two levels: (i) using structural information into a Support Vector Machines (SVM) classifier and (ii) based on amino acid sequence alignment and Nearest Neighbor (NN) classification. The classification accuracy is increased by fusing the two classifiers and reaches 93.4% on a large dataset of 39,251 proteins from the PDB database. The method is very competitive with respect to accuracy of classification into the 6 enzymatic classes, while at the same time its computational cost during prediction is very small.

Keywords: enzyme classification, protein structure, amino acid sequence alignment, multi-class SVM, PDB database

1 Introduction

Proteins are macromolecules which are made of amino acids. Although many distinct groups of proteins and protein families exist, enzymes constitute key pharmacological targets as their primary role is to catalyze chemical reactions. In contrast to most chemical catalysts which catalyze a wide range of reactions, enzymes are usually highly selective, catalyzing specific reactions only. The latter

are classified into 6 standard categories, Oxidoreductases, Transferases, Hydrolases, Lyases, Isomerases, and Ligases, which are identified by their Enzyme Classification (EC) number.

Knowing the EC number of a given enzyme is necessary for the development of potent molecular agents. Having a large dataset of uniquely annotated (by experimental procedures) enzymes at disposal, the goal of this work is to build classification models that are able to predict the EC number of new enzymes with high precision, repeatability and small computational time.
Previous work has been done on different datasets of enzymes. Dobson and Doig [1] used only structural information and achieved an accuracy of 35% for top-ranked prediction using one-class versus one-class SVM on 498 enzymes from the PDB database. Others used only gene or amino acid sequences and achieved an accuracy stemming from 72.94% on the PDB database using neural network [2] to 96% using neural network on enzyme database [3], but also accuracies between 74% [4] and 88.2% using the Swiss-Prot database [5] [6]. A systematic review on the various approaches used by different research groups, their utility and inference is presented in [7]. The methodologies have been classified according to the type of information used for descriptor generation into bioinformatics approaches and chemoinformatics approaches.
In this paper we present a bioinformatics approach that exploits both structural representation and protein sequence similarity in order to predict *in silico* the EC number of an enzyme using machine learning techniques. The structure is encoded by the torsion angles distribution, whereas the protein sequence is characterized by its alignment error to training sequences in which the class label (EC number) is known. Structural information has been previously used either as validation criteria for newly generated models [8] or during structure calculation to reproduce physically realistic conformational features [9]. In the following we present the method, the results achieved, some discussion and future work.

2 Materials and methods

The outline of the method is shown in Fig. 1. Briefly, two supplementary descriptors are extracted from each protein model based on the structural information (SI) and the amino acid sequence (AA). Each descriptor is introduced into a classifier trained previously on annotated data and then the classifier outputs are fused into a single set of final class probabilities.

The method has been trained and tested on proteins from the PDB database. Enzymes that were found experimentally to catalyze more than one chemical reaction and were assigned multiple labels in the first level of the Enzyme Classification, were excluded from the analysis due to the uncertainty they introduce in both training and testing phase. Also PDB entries containing amino acids other than the 20 natural ones, were excluded from the AA analysis, as proper physicochemical parameterisation of Selenocysteine (U) and Pyrrolysine (O) was not part of this study. Same goes for ambiguous amino acids that are represented with the letters B, Z, J and X. We concluded that since the X-ray crystallogra-

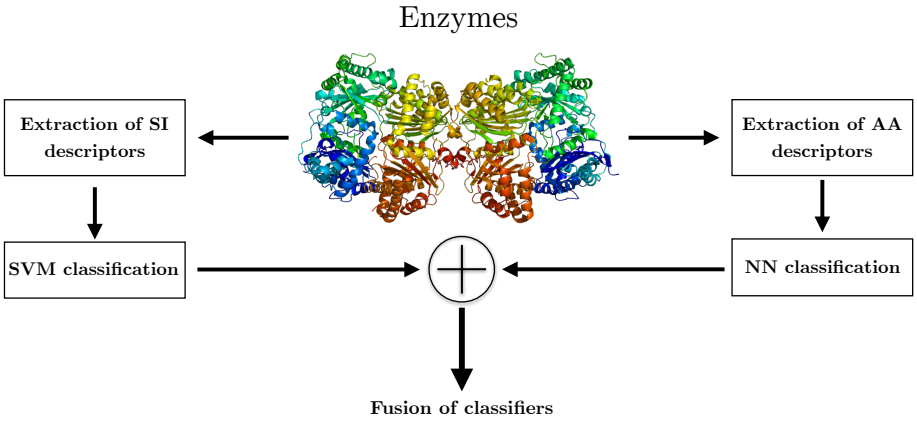


Fig. 1. Overview of the method

phy phases were incapable of giving a clear answer on which amino acid it is and provided that there were thousands of such cases in the full PDB dataset, we would introduce non-quantifiable "noise" to our dataset, which would inevitably sabotage the reliability of our findings.

The functional classes and number of enzymes obtained per class is shown in table 1.

Table 1. Enzyme Classification of the 39,251 enzymes

ID	EC 1	EC 2	EC 3	EC 4	EC 5	EC 6
Name	Oxidoreductase	Transferase	Hydrolase	Lyase	Isomerase	Ligase
Number	7,256	10,665	15,451	2,694	1,642	1,543

2.1 Feature extraction

Proteins are chains of amino acids joined together by peptide bonds. The three-dimensional (3D) configuration of the amino acids chain is a very good predictor of protein function, thus there has been many efforts in extracting an appropriate representation of the 3D structure [10]. Since many conformations of this chain are possible due to the possible rotation of the chain about each carbon (C_α) atom, the use of rotation invariant features is preferred over features based on cartesian coordinates of the atoms. In this study the two torsion angles of the polypeptide chain were used as structural features. The two torsion angles describe the rotation of the polypeptide backbone around the bonds between N-C_α

(angle ϕ) and C_α-C (angle ψ). The probability density of the torsion angles ϕ and $\psi \in [-180°, 180°]$ was estimated by calculating the 2D sample histogram of the angles of all residues in the protein using equal sized bins. When the protein consisted of more than one chain, the torsion angles of all chains were included into the feature vector. Smoothness in the density function was achieved by moving average filtering, i.e. by convoluting the 2D histogram with a uniform kernel. The range of angles was discretized using 19×19 bins centered at $0°$ (with bin size equal to $20°$ for all bins except the 1^{st} and last) and the obtained matrix of structural features was linearized to a 361-dimensional feature vector for each enzyme.

The structural description based on the probability density of the torsion angles does not provide any information about the spatial location of the amino acids in the chain, as well as their sequence. The connectivity patterns (protein sequence) reflects the intrinsic composition of the macromolecule and is an important descriptor of spatial composition. Many methods have been developed to quantify the similarity between two protein sequences which are either based on sequence alignment [11] or provide a similarity score without performing prior alignment [1]. A common algorithm that provides a similarity score to each pair of sequences is the Smith-Waterman algorithm [12]. The Smith-Waterman algorithm calculates the optimal local alignment of two sequences by computing a similarity matrix that takes into account matches, mismatches, substitutions, insertions and deletions between the two sequences [13]. Based on this algorithm, a score $\mathcal{S}(i, j)$ between each pair of sequences i and j was obtained and used to calculate the class probability of a sequence based on nearest neighbor classification rule.

2.2 Classification using structural and amino acid sequence information

For a given enzyme, the 361 obtained features representing the protein's structural conformation were introduced into a multi-class SVM to obtain 6 probabilistic output features. A one-versus-all classification scheme was applied in which 6 binary classifications are performed (class i versus not class i) and combined by majority voting rule to decide which of the 6 classes is the most probable. Since the SVM's decision scores are not reflecting probabilities, an additional sigmoid function was fitted to the data in order to map the SVM outputs into pseudo-probabilities [14]. The latter are noted $(p_i^j)_{i \in [\![1,6]\!]}$, with p_i^j reflecting the probability for the enzyme j to belong to the class EC i.
In order to allow the fusion of classifiers, the scores of amino acid sequence alignment (matrix $\mathcal{S}$) were also converted to pseudo-probabilistic output for the second classifier. Specifically, for a given enzyme j, the nearest neighbor of each class was found using as distance measure the alignment scores with the training samples. The alignment scores between the enzyme j and the 6 neighbors were normalized by the sum of their scores:

$$\text{For each enzyme } j, \quad q_i^j = \frac{\max\limits_{\substack{k \in \text{training} \\ k \neq j}} \bigcap \text{EC } i \; \mathcal{S}(k,j)}{\sum\limits_{l=1}^{6} \max\limits_{\substack{k \in \text{training} \\ k \neq j}} \bigcap \text{EC } l \; \mathcal{S}(k,j)}$$

Thus the classifier decision scores for both descriptors were appropriately scaled, such that $\sum_{i=1}^{6} p_i^j = \sum_{i=1}^{6} q_i^j = 1$, and allowed to be combined within a fusion scheme.

2.3 Fusion of classifiers

It has been shown that fusion techniques that combine multiple machine learning methods achieve better predictive performance than any of the constituent methods [15]. In this work we combine the probabilistic output of each classifier, i.e. the SVM based on structural information and the nearest neighbor based on protein sequence alignment. The classification fusion is constructed by fusing the two probabilistic outputs (p_i^j) and (q_i^j) for each class i through a linear combination into a final probability (z_i^j):

$$\text{For each enzyme } j, \quad \forall i \in [\![1\,,6]\!], z_i^j = (1-\alpha)p_i^j + \alpha q_i^j$$

where $\alpha \in [0,1]$ is a weight that regulates the contribution of each classifier. Thus, the problem is to find the optimal parameter α that maximizes the overall classification accuracy.

2.4 Validation scheme

The dataset was randomly split into 80% for training and 20% for testing. In order to determine the optimum α, 20% of the training set has been hold out and used as validation set, while the remaining part was used to train the classifier. Upon selection of α based on the validation set, the classifier was retrained using the whole training set, and the optimum α was used to assess accuracy of the fusion on the testing set. The assessment of the multi-class system was based on the confusion matrix. However, the performance of the system was also evaluated in respect to the cumulative accuracy after i guesses, noted $\text{CA}_i (i \in [\![1,5]\!])$, which represents the classification accuracy after looking at the i highest class probabilities.

3 Results

3.1 Structural information and amino acid sequences

Table 2 shows the cumulative accuracy (CA1 to CA5) for each enzymatic class and each descriptor separately. Classification with amino acid sequences yields

from 12,2% (EC 3) to 49,2% (EC 6) better accuracy than classification with structural information. The accuracy difference between those two approaches might indicate that amino acid sequence content (with 93.4% overall accuracy) is a better predictor of enzyme function than structural information (with 73.5% overall accuracy).

Table 2. CA using structural information and amino-acid sequences separately

	Structural information					**Amino acid sequence**				
Category	**CA1**	**CA2**	**CA3**	**CA4**	**CA5**	**CA1**	**CA2**	**CA3**	**CA4**	**CA5**
EC 1	0.746	0.884	0.943	0.977	0.992	0.967	0.988	0.996	0.999	1.000
EC 2	0.762	0.920	0.970	0.990	0.997	0.937	0.964	0.977	0.996	1.000
EC 3	0.798	0.909	0.961	0.982	0.994	0.920	0.958	0.975	0.992	0.999
EC 4	0.596	0.685	0.790	0.900	0.950	0.892	0.922	0.944	0.983	0.993
EC 5	0.547	0.635	0.730	0.827	0.939	0.900	0.954	0.970	0.976	0.979
EC 6	0.304	0.476	0.605	0.822	0.926	0.796	0.822	0.841	0.871	0.964
Overall	0.735	0.864	0.924	0.965	0.986	0.934	0.966	0.979	0.993	0.998

3.2 Fusion of both information

Table 3 shows the most often predicted class by each method after each guess. It can be seen that the order of predicted classes varies when different features (SI and AA) are used for classification denoting that each descriptor captures different aspects of enzymatic function.

The fusion of information relies on the study of the function $\alpha \mapsto \mathrm{Acc}(\alpha)$, where Acc is the accuracy on the validation set obtained by weighting the probabilities obtained by the SI-based and AA-based classifiers. Fig. 2 illustrates the classification accuracy in the interval $\alpha \in [0.8, 1]$. The value that maximizes the accuracy is $\alpha = 0.94$ and was chosen for the assessment of the method on the testing dataset. As expected, the optimum α gives a more significant weight to amino acid sequences descriptors than to structural information descriptors.

The total cumulative accuracy (for multiple guesses) is shown in Fig. 3 for each of the three methods, while Fig. 4 illustrates the cumulative accuracy for each enzymatic class separately. For Transferases and Hydrolases the fusion of features induced an average increase of accuracy by 1% on each guess compared to the accuracy obtained by amino acid sequences only. For the other classes, the accuracy of the fusion is comparable to the accuracy by AA indicating that structural information does not additionaly contribute to classification.

Table 3. Most representative guesses for each true class

		Guesses					
True class	**Method**	1^{st}	2^{nd}	3^{rd}	4^{th}	5^{th}	6^{th}
EC 1	SI	EC 1	EC 2	EC 2	EC 4	EC 5	EC 6
	AA	EC 1	EC 2	EC 3	EC 4	EC 6	EC 5
EC 2	SI	EC 2	EC 3	EC 1	EC 6	EC 5	EC 5
	AA	EC 2	EC 3	EC 1	EC 1	EC 5	EC 5
EC 3	SI	EC 3	EC 2	EC 1	EC 6	EC 5	EC 5
	AA	EC 3	EC 2	EC 1	EC 1	EC 6	EC 5
EC 4	SI	EC 4	EC 2	EC 1	EC 1	EC 5	EC 5
	AA	EC 4	EC 3	EC 2	EC 2	EC 1	EC 5
EC 5	SI	EC 5	EC 3	EC 2	EC 1	EC 6	EC 4
	AA	EC 5	EC 2	EC 1	EC 3	EC 6	EC 6
EC 6	SI	EC 2	EC 2	EC 1	EC 1	EC 5	EC 5
	AA	EC 6	EC 1	EC 2	EC 1	EC 4	EC 5

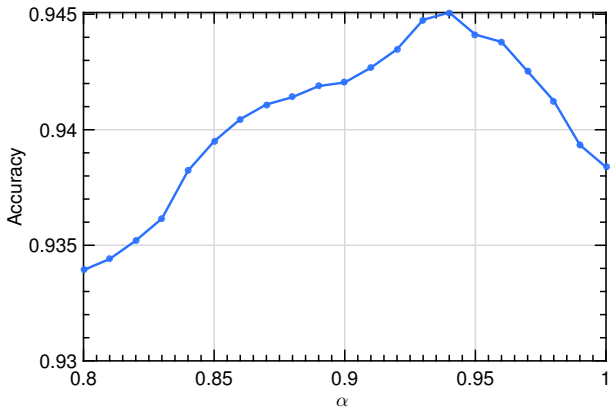


Fig. 2. Classification accuracy on the validation dataset as a function of the weight α

Fig. 5 shows the confusion matrix of the classification fusion. The confusion matrix allows to see whether an enzyme class is more or less predictable and shows how the classification error is distributed among classes. Each row of the confusion matrix has been normalized so that coefficient (i,j) represents the proportion of enzymes from class EC i that are predicted as belonging to EC j.

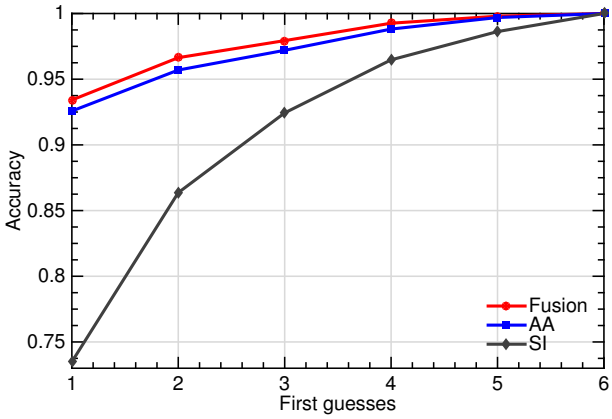
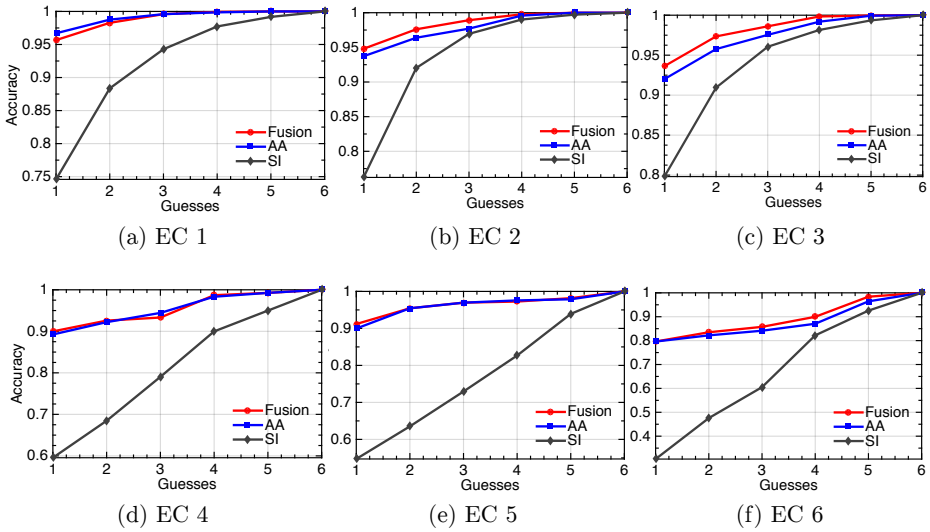


Fig. 3. Comparison of the overall CA between the 3 methods



(a) EC 1 (b) EC 2 (c) EC 3

(d) EC 4 (e) EC 5 (f) EC 6

Fig. 4. Comparison of the CA for each true class between the 3 methods

The highest misclassification rates are observed for Ligases with 10.4% of them classified as Transferases and 6.5% as Hydrolases. Conversely, Oxidoreductases, Transferases and Hydrolases have very small misclassification rates.

The code was written in Matlab language and was performed using High-Performance Computing (HPC) resources from the "mesocentre" computing center of École Centrale de Paris (`http://www.mesocentre.ecp.fr`) supported by CNRS. Running on 2 Intel X5650 with 4GB RAM, the computational time

Actual \ Predicted	EC 1	EC 2	EC 3	EC 4	EC 5	EC 6
EC 1	**0.957**	0.012	0.027	0.002	0.001	0.001
EC 2	0.011	**0.948**	0.033	0.002	0.001	0.004
EC 3	0.017	0.041	**0.936**	0.003	0	0.001
EC 4	0	0.059	0.032	**0.900**	0.006	0.004
EC 5	0.040	0.015	0.018	0.015	**0.912**	0
EC 6	0.026	0.104	0.065	0.003	0.007	**0.796**

Fig. 5. Confusion matrix

required for predicting the enzymatic class of a new protein was less than a minute.

4 Conclusions

The bioinformatics bottleneck is a major problem in the fields of Next Generation Sequencing (NGS) and Drug Design (DD). Data accumulation is becoming a challenge on its own as the handling and analysis of huge datasets is not a trivial task. Therefore there is dear need in the field of bioinformatics to establish and optimize new methodologies that are faster and much more efficient for routine tasks in the aforementioned fields. A major drawback in protein classification is the fact that protein similarity is still judged upon primary sequence comparisons rather than combining structural information too. Using structural information was quite challenging a while ago as significant computing power was required. However, nowadays it is not too difficult to use optimized workstations and supercomputers to analyze all structural information available. Remarkably, the RCSB PDB database has now more than 114,000 entries. In 1990 there were only 507 structures available and in 2000 only 13,591 structures. Quick and reliable classification of NGS generated sequences is of paramount importance in the field of DD, as the problem of pharmacological target identification still holds. Classifying quickly and reliably, the several thousands of proteins that are being identified and sequenced on a daily basis would lead to an optimized pipeline

that could yield results of great interest to protein science, DD and NGS related fields such as anticancer and antiviral therapy.

The proposed approach is fully automated, reproducible and computationally very efficient during prediction phase. Also the achieved accuracy (73.5%) by the SVM classifier based on structural information is significantly higher than the reported accuracy (35%) of the SVM classifier by Dobson and Doig [1] based also on protein structure. We used the torsion angles distribution to represent structure, whereas in [1] 55 structural attributes were originally extracted, including residue fractions, surface fractions, secondary structure content, cofactors, metals and general properties. Then attribute subset selection was applied based on backwards elimination.

Furthermore, the weighted combination of classifiers in the proposed scheme led to a significant increase of accuracy (27.1% for the first guess) relatively to the classification performed using only structural information, while the accuracy increase relatively to the AA-based classification was only marginal (0.9%). The trend is similar for the cumulative accuracy with multiple guesses. As can be seen from Fig. 3, the classification fusion performs better in respect to the cumulative accuracy with multiple guesses than the single classifiers.

The two classes that showed the best results after fusion (Transferases and Hydrolases) are the ones which have the largest number of samples in the dataset. This could be related to the selection of α through maximization of overall accuracy which is biased towards the majority classes. It is well known that most machine learning techniques rely on balanced datasets in which each class includes the same number of samples. In the future, we plan to investigate bootstrapping and data aggregation techniques [16] which provide efficient solutions to unbalanced datasets.

Moreover, in the current study, we investigated two common supervised classification techniques, the nearest neighbor and SVM. The nearest neighbor is a very simple classifier often used as baseline to assess the necessity of more advanced tools. SVM is a very popular algorithm due to its ability to capture complex relationships between the datapoints and its high predictive power. Also the number of required hyper-parameters to be optimized in SVM is small (usually 2, i.e. the parameters controlling the misclassification penalty and smoothness), and thus quite easy to tune even by manual search or grid-search. Ongoing work includes the investigation of different classifiers, in terms of accuracy and performance.

5 Acknowledgment

This research was partially supported by European Research Council Grant Diocles (ERC-STG-259112).

References

1. Dobson, P.D., Doig, A.J.: *Predicting Enzyme Class From Protein Structure Without Alignments.* J Mol Biol (January 2005)

2. Osman, M.H., Choong-Yeun Liong, I.H.: *Hybrid Learning Algorithm in Neural Network System for Enzyme Classification.* ICSRS **2**(ISSN 2074-8523) (July 2010)
3. Volpato, V., Adelfio, A., Pollastri, G.: *Accurate prediction of protein enzymatic class by N-to-1 Neural Networks.* BMC Bioinformatics (doi: 10.1186/1471-2105-14-S1-S11) (January 2013)
4. desJardins, M., Karp, P.D., Krummenacker, M., Lee, T.J., Ouzounis, C.A.: *Prediction of Enzyme Classification from Protein Sequence without the use of Sequence Similarity.* ISMB (1997)
5. Kumar, C., Choudhary, A.: *A top-down approach to classify enzyme functional class and sub-classes using random forest.* In: EURASIP J Bioinform Syst Biol. (February 2012)
6. Lee, B.J., Lee, H.G., Lee, J.Y., Ryu, K.H.: *Classification of Enzyme Function from Protein Sequence based of Feature Representation.* In: Bioinformatics and Bioengineering, 2007. BIBE 2007. Proceedings of the 7th IEEE International Conference. (2007) 741–747
7. Sharma, M., Garg, P.: *Computational Approaches for Enzyme Functional Class Prediction: A Review.* Current Proteomics **11**(1) (2014) 17–22
8. Read, R., Adams, P., 3rd Arendall, W., Brunger, A., Emsley, P., Joosten, R., Keyweft, G., Krissinel, E., Lütteke, T., Otwinowski, Z., Perrakis, A., Richardson, J., Sheffler, W., Smith, J., Tickle, I., Vriend, G., Zwart, P.: *A new generation of crystallographic validation tools for the protein data bank.* PubMed (doi: 10.1016/j.str.2011.08.006) (October 2011)
9. Bermejo, G., Clore, G., Schwieters, C.: *Smooth statistical torsion angle potential derived from a large conformational database via adaptive kernel density estimation improves the quality of NMR protein structures.* Proteine Science (doi: 10.1002/pro.2163) (December 2012)
10. Lie, J., Koehl, P.: *3D representations of amino acids—applications to protein sequence comparison and classification.* Computational and Structural Biotechnology Journal **11**(doi:10.1016/j.csbj.2014.09.001) (August 2014) 47–58
11. Sharif, M.M., Thrwat, A., Amin, I.I., Ella, A., Hefeny, H.A.: *Enzyme Function Classification Based on Sequence Alignment.* Volume 340 of Advances in Intelligent Systems and Computing. Springer India (January 2015)
12. Jensen, L., Gupta, R., Blom, N., Devos, D., Tamames, J., Kesmir, C., Nielsen, H., Stærfeldt, H., Rapacki, K., Workman, C., Andersen, C., Knudsen, S., Krogh, A., A. Valencia, S.B.: *Prediction of Human Protein Function from Post-translational Modifications and Localization Features.* JMB (June 2002)
13. Smith, T., Waterman, M.: *Identification of common molecular subsequences.* Journal of Molecular Biology **147**(doi:10.1016/0022-2836(81)90087-5) (March 1981) 195–197
14. Platt, J.C.: *Probabilistic Outputs for Support Vector Machines and Comparison to Regularized Likelihood Methods.* In: Advances in large margin classifiers, MIT Press (1999) 61–74
15. Mohammed, A., Guda, C.: *Application of a hierarchical enzyme classification method reveals the role of gut microbiome in human metabolism.* BMC Genomics **16**(doi:10.1186/1471-2164-16-S7-S16) (June 2015)
16. Chawla, N.V.: 40. Number doi: 10.1007/0-387-25465-X40 in Data mining and knwoledge discovery handbook. In: *Data mining for imbalanced datasets: an overview.* Springer US (2000) 853–867