International Journal of
Molecular Sciences

MDPI

Article

Alignment-Free Method to Predict Enzyme Classes and Subclasses

Riccardo Concu * and M. Natália D. S. Cordeiro *

LAQV@REQUIMTE/Department of Chemistry and Biochemistry, Faculty of Sciences, University of Porto, 4169-007 Porto, Portugal

* Correspondence: ric.concu@gmail.com (R.C.); ncordeir@fc.up.pt (M.N.D.S.C.)

Received: 9 September 2019; Accepted: 23 October 2019; Published: 29 October 2019

Abstract: The Enzyme Classification (EC) number is a numerical classification scheme for enzymes, established using the chemical reactions they catalyze. This classification is based on the recommendation of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology. Six enzyme classes were recognised in the first Enzyme Classification and Nomenclature List, reported by the International Union of Biochemistry in 1961. However, a new enzyme group was recently added as the six existing EC classes could not describe enzymes involved in the movement of ions or molecules across membranes. Such enzymes are now classified in the new EC class of translocases (EC 7). Several computational methods have been developed in order to predict the EC number. However, due to this new change, all such methods are now outdated and need updating. In this work, we developed a new multi-task quantitative structure–activity relationship (QSAR) method aimed at predicting all 7 EC classes and subclasses. In so doing, we developed an alignment-free model based on artificial neural networks that proved to be very successful.

Keywords: QSAR; machine learning; artificial neural network; enzyme; enzyme classification; alignment-free

1. Introduction

By the late 1950s, the International Union of Biochemistry and Molecular Biology foresaw the need for unique nomenclature for enzymes. In those years, the number of known enzymes had grown very rapidly and, because of the absence of general guidelines, the nomenclature of the enzymes was getting out of hand. In some cases, enzymes with similar names were catalyzing different reactions, while conversely different names were given to the same or similar enzymes. Due to this, during the third International Congress of Biochemistry in Brussels in August 1955, the General Assembly of the International Union of Biochemistry (IUB) decided to establish an International Commission in charge of developing a nomenclature for enzymes. In 1961, the IUB finally released the first version of the Enzyme Classification (EC) and Nomenclature List. This nomenclature was based on assigning a four number code to enzymes with the following meaning: (i) the first number identifies the main enzyme class; (ii) the second digit indicates the subclass; (iii) the third number denotes the sub-subclass; and (iv) the fourth digit is the serial number of the enzyme in its sub-subclass. Six enzyme classes were identified, with the classification based on the type of reaction catalyzed: oxidoreductases (EC 1), transferases (EC 2), hydrolases (EC 3), lyases (EC 4), isomerases (EC 5) and ligases (EC 6) [1]. Although several revisions have been made to the 1961 version, the six classes identified have not received any change. However, in August 2018, a new class was added. This new class contains the translocases (EC 7), and was added to describe those enzymes catalyzing the movement of ions or molecules across membranes or their separation within membranes. For this reason, some enzymes which had previously been classified in other classes—EC 3.6.3 for example—were now included in the EC 7 class.

Predicting enzyme classes or protein function using bioinformatic tools is still a key goal in bioinformatics and computational biology due to both the prohibitive costs and the time-consuming nature of wet-lab-based functional identification procedures. In point of fact, there are more than four thousand sequences whose function remains unknown so far and this number is still growing [2]. The problem is that our ability to assign a specific function to a sequence is far lower than our ability to isolate and identify sequences. For this reason, significant efforts have been devoted to developing reliable methods able to predict protein function.

Several methodological strategies and tools have been proposed to classify enzymes based on different approaches [3–10]. The Basic Local Alignment Search Tool (BLAST) [11] is likely to be one of the most powerful and used tools which finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates their statistical significance. However, as is the case with all methods, these procedures may fail under certain conditions. In some cases, enzymes with a sequence similarity higher than 90% may belong to different enzyme families and, thus, have different EC annotations [12–14]. On the other hand, some enzymes which share the same first EC number may have a sequence similarity below 30%. Some authors have described this situation well and highlighted the need to develop alignment-free methods, which may be used in a complementary way [15,16]. Other relevant tools based on sequence similarity are the UniProtKB database [17], the Kyoto Encyclopedia of Genes and Genomes (KEGG) [18], the PEDANT protein database [19], DEEPre [20], ECPred [21] and EzyPred [22]. DEEPre is a three-level EC number predictor, which predicts whether an input protein sequence is an enzyme, and its main class and subclass if it is. This method is based on a dataset of 22,198 sequences achieving an overall accuracy of more than 90%. ECPred is another enzymatic function prediction tool based on an ensemble of machine learning classifiers. The creators of this tool developed it using a dataset of approximately 245,000 proteins, achieving score classifications in the 6 EC classes and subclasses like the ones reported by DEEPre. EzyPred is a top-down approach for predicting enzyme classes and subclasses. This model was developed using a 3-layer predictor using the ENZYME [23] dataset (approximately 9800 enzymes when the model was developed), which was able to achieve an overall accuracy above 86%. Other relevant methods with similar classification scores have also been reported [10,15,20,24,25]. All these methods have proved to be robust; however, they are all outdated since they cannot predict the EC 7 classification, and should therefore be updated in accordance with the new EC class.

In light of what has been referred to so far, the major target of this work was to develop an alignment-free strategy using machine learning (ML) methods to predict the first two digits of the seven EC classes. Previous ML methods have used alignment-free numerical parameters to quantify information about the 2D or 3D structure of proteins [26–29]. Specifically, Graham, Bonchev, Marrero-Ponce, and others [30–34] used Shannon's entropy measures to quantify relevant structural information about molecular systems. In addition, González-Díaz et al. [35–37] introduced so-called Markov–Shannon entropies (θ_k) to codify the structural information of large bio-molecules and complex bio-systems or networks. For comparative purposes, we developed different linear and non-linear models, including a linear discriminant analysis (LDA) and various types of artificial neural networks (ANNs). In addition, we focused our work on performing an efficient feature selection (FS). Nowadays, there are several software packages or tools that may be used to calculate thousands of molecular descriptors (MDs). As a result, a proper FS method is essential to develop robust and reliable quantitative structure–activity relationship (QSAR) models. This is particularly the case when using ANNs, since QSAR models developed with a large set of MDs are really complex, vulnerable to overfitting and difficult to obtain a mechanistic interpretation from [38,39].

2. Results

2.1. LDA Model

As a first step, we used the LDA algorithm implemented in the software STATISTICA® [40] to derive a linear model able to discriminate all of the subclasses of enzymes using a multi-task model, which means that a single model was developed in order to assign each enzyme to a specific class. From the first pool of more than 200 variables, we selected four that clearly had an influence on the model using a supervised forward stepwise analysis. In order to validate the model, we split our dataset, assigning 70% of the entries to the training class and the remaining 30% to the validation class. The latter was used for validation of the model using a cross-validation procedure. The LDA model had the following overall values for specificity: Sp = 99.71%, sensitivity: Sn = 98.16% and accuracy: Acc = 98.66%. In the training series, the model displayed Sp = 99.71%, Sn = 98.13% and Acc = 98.63%, while in the validation series Sp = 99.71, Sn = 98.27, Acc = 98.73. All of these statistics are reported in Table 1.

Table 1. Accuracy for the linear discriminant analysis (LDA) model.

	Training			Validation			Overall		
	All	**−1 = Sn**	**1 = Sp**	**All**	**−1 = Sn**	**1 = Sp**	**All**	**−1 = Sn**	**1 = Sp**
−1	98.13	40,781	778	98.27	13,613	240	98.16	54,394	1018
1	99.7	57	19,498	99.71	19	6498	99.71	76	25,996
Total	98.63	40,838	20,276	98.73	13,632	6738	98.66	54,470	27,014

The linear equation (Equation (1)) for this model is shown below and information regarding its variables is given in Table 5:

$$EC = < Tr3(srn) > * - 0.95 + < Tr5(srn) * -0.80 + DTr5(srn) * -0.80 + Dtr3(srn) * 1.01 - 2.05 \quad (1)$$

Other relevant statistics for the LDA model (both training and validation), such as the Wilk's lambda and Matthews correlation coefficient (MCC), are reported in Table 2.

Table 2. Relevant statistics for the LDA model.

Eigenvalue	**CanonicalR**	**Wilk'sLambda**	**Chi-Sqr.**	**df**	***p*-value**	**MCC**
1.241879	0.744275	0.446054	49334.99	4.000000	0.00	0.97

2.2. ANN models

We then decided to move a step forward and try to develop non-linear models using various neural networks' architectures. We firstly investigated ANN models using either the multi-layer perceptron (MLP) algorithm or the radial basis function (RBF) [41–46]. To do so, we ran a set of 50 ANN-MLP models in order to identify the best topology and architecture. The best model found had an MLP 4-9-2 topology, and was developed using the same four variables used for the LDA model. Additionally, it was able to correctly classify 100% of the cases in both the training and validation series. Table 3 shows the statistical parameters obtained for this model. As can be seen, the MCC value was, as expected, 1.

Table 3. Performance of the best multi-layer perceptron (MLP) model found.

Obs. Sets [a]	Stat. Param. [a]	Pred. Stat. [a]	Predicted sets		
			1	**−1**	**nj**
Training Series					
1	Sp [a]	100	17,500	0	57,039
−1	Sn [a]	100	0	39,539	0
total	Ac [a]	100	17,500	39,539	57,039
Validation Series					
1	Sp [a]	100	8572	0	24,445
−1	Sn [a]	100	0	15,873	0
total	Ac [a]	100	8572	15,873	24,445
Overall					
1	Sp [a]	100	26,072	0	81,484
−1	Sn [a]	100	0	55,412	0
total	Ac [a]	100	26,072	55,412	81,484

[a] Obs. Sets = Observed sets, Stat. Param. = Statistical parameter, Pred. Stat. =Predicted statistics, Sp = Specificity, Sn = Sensitivity, Ac =Accuracy.

For comparative purposes, Table 4 reports the statistics of the 10 best MLP and RBF models found.

Table 4. Resumé of the 10 best MLP and radial basis function (RBF) models.

		Training			Validation			Overall		
Model		**−1 = Sn**	**1 = Sp**	**All**	**−1 = Sn**	**1 = Sp**	**All**	**−1 = Sn**	**1 = Sp**	**All**
BEST MLP: 4-9-2	Total	55,412	26,072	81,484	55,412	26,072	81,484	55,412	26,072	81,484
	Correct	55,412	26,072	81,484	55,412	26,072	81,484	55,412	26,072	81,484
	Incorrect	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Correct (%)	100	100	100	100	100	100	100	100	100
	Incorrect (%)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1.MLP 4-7-2	Total	39,448	17,591	57,039	15,873	8572	24,445	55,412	26,072	81,484
	Correct	39,448	17,567	57,015	15,873	8562	24,435	55,412	26,034	81,446
	Incorrect	0	24	24	0	10	10	0	38	38
	Correct (%)	100	99.86	99.96	100.00	99.88	99.96	100.00	99.85	99.95
	Incorrect (%)	0	0.14	0.04	0.00	0.12	0.04	0.00	0.15	0.05
2.MLP 4-8-2	Total	39,448	17,591	57,039	15,873	8572	24,445	55,412	26,072	81,484
	Correct	39,448	17,565	57,013	15,873	8563	24,436	55,412	26,037	81,449
	Incorrect	0	26	26	0	9	9	0	35	35
	Correct (%)	100	99.85	99.95	100.00	99.90	99.96	100.00	99.87	99.96
	Incorrect (%)	0	0.15	0.05	0.00	0.10	0.04	0.00	0.13	0.04
3.MLP 4-10-2	Total	39,448	17,591	57,039	15,873	8572	24,445	55,412	26,072	81,484
	Correct	39,448	17,565	57,013	15,873	8563	24,436	55,412	26,037	81,449
	Incorrect	0	26	26	0	9	9	0	35	35
	Correct (%)	100	99.85	99.95	100.00	99.90	99.96	100.00	99.87	99.96
	Incorrect (%)	0	0.15	0.05	0.00	0.10	0.04	0.00	0.13	0.04

Table 4. *Cont.*

		Training			Validation			Overall		
Model		**−1 = Sn**	**1 = Sp**	**All**	**−1 = Sn**	**1 = Sp**	**All**	**−1 = Sn**	**1 = Sp**	**All**
4.MLP 4-11-2	Total	39,448	17,591	57,039	15,873	8572	24,445	55,412	26,072	81,484
	Correct	39,448	17,566	57,014	15,873	8563	24,436	55,412	26,037	81,449
	Incorrect	0	25	25	0	9	9	0	35	35
	Correct (%)	100	99.86	99.96	100.00	99.90	99.96	100.00	99.87	99.96
	Incorrect (%)	0	0.14	0.04	0.00	0.10	0.04	0.00	0.13	0.04
5.MLP 4-16-2	Total	39,448	17,591	57,039	15,873	8572	24,445	55,321	26,163	81,484
	Correct	39,448	17,567	57,015	15,873	8572	24,445	55,321	26,139	81,460
	Incorrect	0	24	24	0	0	0	0	0	0
	Correct (%)	100	99.86	99.96	100.00	100.00	100.00	100.00	99.91	99.97
	Incorrect (%)	0	0.14	0.04	0.00	0.00	0.00	0.00	0.09	0.03
6.RBF 4-21-2	Total	39,539	17,500	57,039	15,873	8572	24,445	55,412	26,072	81,484
	Correct	39,520	16,426	55,946	15,855	8059	23,914	55,375	24,485	79,860
	Incorrect	19	1074	1093	18	513	531	37	1587	1624
	Correct (%)	99.95	93.86	98.08	99.89	94.02	97.83	99.93	93.91	98.01
	Incorrect (%)	0.05	6.14	1.92	0.11	5.98	2.17	0.07	6.09	1.99
7.RBF 4-29-2	Total	39,539	17,500	57,039	15,873	8572	24,445	55,412	26,072	81,484
	Correct	39,165	17,475	56,640	15,714	8561	24,275	54,879	26,036	80,915
	Incorrect	374	25	399	159	11	170	533	36	569
	Correct (%)	99.05	99.86	99.3	99.00	99.87	99.30	99.04	99.86	99.30
	Incorrect (%)	0.95	0.14	0.7	1.00	0.13	0.70	0.96	0.14	0.70
8.RBF 4-21-2	Total	39,539	17,500	57,039	15,873	8572	24,445	55,412	26,072	81,484
	Correct	39,526	16,138	55,664	15,868	7873	23,741	55,394	24,011	79,405
	Incorrect	13	1362	1375	5	699	704	18	2061	2079
	Correct (%)	99.97	92.22	97.59	99.97	91.85	97.12	99.97	92.09	97.45
	Incorrect (%)	0.03	7.78	2.41	0.03	8.15	2.88	0.03	7.91	2.55
9.RBF 4-28-2	Total	39,539	17,500	57,039	15,197	8571	23,768	53,008	26,060	81,484
	Correct	39,489	16,000	23,489	15,197	8448	23,645	53,008	25,674	78,682
	Incorrect	50	1500	1,450	0	123	123	0	386	386
	Correct (%)	99.87	91.43	95.65	100.00	98.56	99.48	100.00	98.52	99.51
	Incorrect (%)	0.03	7.78	4.35	0.00	1.44	0.52	0.00	1.48	0.49
10.RBF 4-26-2	Total	39,539	17,500	57,039	15,873	8572	24,445	55,412	26,072	81,484
	Correct	11,880	6629	18,509	4748	3170	7918	16,628	9799	26,427
	Incorrect	27659	10871	38530	11125	5402	16527	38784	16273	55057
	Correct (%)	30.05	37.88	32.45	29.91	36.98	32.39	30.01	37.58	32.43
	Incorrect (%)	69.95	62.12	67.55	70.09	63.02	67.61	69.99	62.42	67.57

The results reported in Table 4 clearly indicate that MLP models perform better than RBF ones. Even if the best MLP model was able to achieve 100% overall accuracy, we decided to perform a quantitative analysis to infer whether the MLP models were failing. As can be seen in Table 5, the non-optimal MLP models were particularly problematic in discriminating the EC 6.5 subclass.

Table 5. Quantitative analysis of the non-optimal MLP models.

Model	Class	Fail	Total Class
1. MLP 4-7-2	6.4	1	104
	6.5	34	36
2. MLP 4-8-2	1.6	3	4
	6.4	1	104
	6.5	34	36
3. MLP 4-10-2	1.6	3	4
	6.4	1	104
	6.5	33	36
4. MLP 4-11-2	1.6	3	4
	6.4	1	104
	6.5	32	36
5. MLP 4-16-2	6.4	1	104
	6.5	33	infer 36

Finally, a sensitivity analysis was also performed to assess the influence of the MDs in the model. The results of this analysis are shown in Table 6.

Table 6. Sensitivity analysis for the artificial neural network (ANN) model.

Input Variable	Variable Sensitivity	Variable Name/Details
<Tr5(srn)>	15,896,991	Expected value of Trace of order 5 of the srn for the sequence
D Tr5(srn)	1,288,626	Deviation of Trace of order 5 of the srn with respect to the mean value of the class
<Tr3(srn)>	591,331.9	Expected value of Trace of order 3 of the srn for the sequence
D Tr3(srn)	108.7591	Deviation of Trace of order 3 of the srn with respect to the mean value of the class

Sensitivity analysis refers to the assessment of the importance of predictors in a developed model, with higher values of sensitivity being assigned to the most important predictors. As seen, the high sensitivity values found for some of the parameters suggest that the model's performance can drastically fall if the parameters used in the model are removed. On the other hand, parameters with lower values of sensitivity may be discarded since they are not relevant to the performance of the model and may lead to an overfitted model. Regarding the variables presented in Table 6, they are traces of the *n* connectivity matrices of the amino acid sequences. The terms 3 and 5 represent the order of the matrix used in the calculation. The terms within brackets ("< >") represent the mean value of each subclass, while "D" stands for the difference (or distance) between each amino acid sequence and the mean value of its subclass. This basically means that the model, in order to correctly predict each sequence as an enzyme and then input it into the specific subclass, is calculating the distance between each input and the mean of its subclass. This is in fact how a multi-target model works.

3. Discussion

The main aim of this study was to develop a new QSAR-ML model able to predict enzyme subclasses considering the new and recently introduced EC class 7. We retrieved from the Protein Data Bank (PDB) more than 26,000 enzyme and 55,000 non-enzyme sequences in order to build up our dataset. All of the enzyme sequences belonged to one of the 7 main classes and 65 subclasses. The EC 7 class was introduced just few months ago and, due to this, all of the current models do not include

this new enzyme class. As a result, the classification or prediction such models are performing may be misleading. Hence, the development of new models which are capable of predicting all enzyme classes and subclasses—including the EC 7 class—are of utmost importance. In view of this, we developed a new machine learning model able to discriminate between enzymes and non-enzymes. In addition, the model was capable of assigning enzymes to a specific enzyme subclass. We generated linear and non-linear models using alignment-free variables to find the best model to predict EC classes and subclasses. The results of the linear model were impressive since with only four MDs the model could discriminate between enzymes and non-enzymes, as well as assign a specific EC class and subclass to each enzyme sequence. We checked the accuracy and robustness of the model and the results clearly indicate that the model is reliable. Regarding the validation, we performed a classical cross-validation procedure using 30% of the dataset. This led to almost the same results for the training and validation sets, indicating once more the robustness of the model and approach.

Although the accuracy of the derived LDA model was near 100%, we decided to further test our approach by developing some neural network models, which usually improve LDA results. To the best of our knowledge, an MLP is generally considered the best ANN algorithm and, in this case, had the potential to improve our linear model. As previously reported, the MLP was able to perfectly discriminate between enzymes and non-enzymes, in addition to assigning each enzyme sequence to a specific subclass. It is also remarkable that the best model only needed nine neurons in the hidden layer. This low number of neurons, considering the number of sequences and variables, suggest that the model is not suffering from an overfitting problem. Mechanistic interpretation of ANN models is always a challenging task since these models do not lead to simple linear equations. A sensitivity analysis may then be used to analyze the influence of each MD on the model. For the ANN model, we carried out such an analysis to evaluate the weight of each variable in the model. This analysis is also useful for identifying redundant variables in models, assisting in their eliminatation to avoid an unlikely overfitting problem. In the case of the ANN model, we identified that the same four variables used in the LDA model were able to perfectly discriminate between enzymes and non-enzymes and assign each enzyme sequence to a specific subclass.

Finally, we also tested RBF models, which afforded results that were worse than the MLP models. In fact, the general accuracy was lower when compared to the MLP models, which usually need less neurons to achieve greater accuracy.

4. Materials and Methods

4.1. Dataset

From the PDB, we retrieved a total of 81,486 protein FASTA sequences. Of those sequences, 26,073 were enzymes, while 55,413 were non-enzymes (α-proteins, β-proteins, membrane proteins, and so forth). Each of the 26,073 enzyme sequences belonged to one of the 65 enzyme subclasses. In order to avoid redundant sequences, we selected the enzymes using the specific EC classification query module of the PDB and then double-checked the dataset, eliminating double entries. Regarding the non-enzyme sequences, we randomly downloaded protein sequences belonging to different classes, such as membrane proteins, multi-domains, alfas and betas. The complete list of EC subclasses is reported in Supplemental Material S1, while Table 7 reports the number of entries for each one of the subclasses.

Table 7. Number of entries for each subclass.

EC Subclass	Number of Sequences	EC Subclass	Number of Sequences	EC Subclass	Number of Sequences
1.1	555	2.3	722	4.6	120
1.2	250	2.4	424	4.99	95
1.3	172	2.5	291	5.1	176
1.4	108	2.6	19	5.2	74
1.5	5	2.7	3112	5.3	247
1.6	4	2.8	71	5.4	160
1.7	91	2.9	10	5.5	115
1.8	165	3.1	1559	5.6	159
1.9	73	3.11	7	5.99	3
1.10	555	3.13	3	6.1	277
1.11	136	3.2	700	6.2	38
1.12	32	3.3	164	6.3	291
1.13	123	3.4	1481	6.4	104
1.14	244	3.5	561	6.5	36
1.15	162	3.6	417	7.1	8827
1.16	173	3.7	69	7.2	927
1.17	121	3.8	77	7.4	189
1.18	45	3.9	3	7.5	187
1.20	250	4.1	486	7.6	197
1.21	28	4.2	460		
1.23	3	4.3	97		
2.1	522	4.4	39		
2.2	107	4.5	25		

4.2. Molecular Descriptor Calculation

The software S2SNet [47] was used to transform each protein sequence into one sequence recurrence network (SRN). The SRN of a protein sequence can be constructed starting from one of two directions: (1) from a sequence graph with linear topology by adding amino acid recurrence information, or (2) from a protein representation graph with star graph (SG) topology by adding sequence information [48–52]. Note that, in both of these SRN representations of a protein sequence, the amino acids are the nodes and are paired (na and nb) in the network (being connected by a link, αab = 1) if they are adjacent and/or neighbour recurrent nodes. This means that αab = 1 if the topological distance between na and nb is d = 1 (chemically bonded amino acids), or if they are the nearest neighbour amino acid of the same type (A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V, X) with minimal topological distance, dab = min(dab), between them. The first node in the sequence (centre of the star graph) is a bias or a dummy non-residue vertex.

Secondly, we needed to transform the SRN of each sequence into one stochastic matrix $^{1}\Pi$. The elements of $^{1}\Pi$ were found by considering the probability (p_{ab}) of reaching an amino acid (node nj) by walking from another amino acid (node ni) through a walk of length dij = 1 (Equation (2)):

$$p_{ab} = \frac{\alpha_{ab}}{\sum_{n=1}^{n=L} \alpha_{ab}} \quad (2)$$

Note that the number of amino acids in the sequences was equal to the number of nodes (n) in the SRN graph, and was also equal to the number of rows and columns in $^{1}\Pi$, the length of the sequence (L), and the maximal topological distance in the sequence max(dab). In this work, we quantified the information content of a peptide using the Shannon entropy values (θ_k) of the k-th natural powers of

the Markov matrix $^{1}\Pi$. The same procedure was used to quantify the information of the q-seqs ($^{q}\theta_k$) and r-seqs ($^{r}\theta_k$). The formula for the Markov–Shannon entropy $^{q}\theta_k$ is as follows (Equation (3)):

$$^{q}\theta_k(seq) = -\sum_{a=0}^{a=L} {}^{k}p_a \cdot \log\left({}^{k}p_a\right) \quad (3)$$

where $^{k}p_a$ represents the absolute probability of reaching a node moving throughout a walk of length k with respect to any node in the spectral graph. Further details of this formula can be seen in previous works [35–37].

In the Supplemental Material S2, we report the complete list of sequence entries with the respective value of the MD used to develop the models.

4.3. Multi-Target Linear model

The LDA model was developed using the General Discriminant tool implemented in the software STATISTICA [40]. The model is based on a multi-task approach, meaning it is able to predict if a sequence belongs to one out of the seven EC classes. It starts by identifying the presence of enzyme activity εq(ci) = 1 of subclass ci (or the absence of this activity εq(ci) = 0) for a query protein with a known amino acid sequence. The linear model is based on a linear equation, which directly correlates the dependent variable (enzyme or not) with the independent variable (MD). The multi-target LDA model was developed as follows. Once the MD were calculated, we computed the mean value of each subclass and then the difference between each sequence and the mean value of its subclass. Due to the model's incorporation of the mean value of each subclass and the difference between each sequence, as well as the mean value of its subclass, the model is able to achieve a multi-target prediction. For further information regarding this statistical technique, please refer to the bibliography [53–55]. This same procedure was used also for the development of the multi-target ANN model. The validation of the model was performed using the cross-validation module implemented in the software. This procedure is aimed at assessing the predictive accuracy of a model. The test split the dataset into a training set and a validation set, ensuring that if an entry was included in the test set it could not be used in the validation test. In so doing, the model was developed using the cases in the training or learning sample, which, in our study, was 70% of the dataset. The predictive accuracy was then assessed using the remaining 30% of the dataset [56,57]. Standard statistics, such as the specificity (Sp), sensitivity (Sn), probability of error (p), cross-validation, and the Matthews correlation coefficient (MCC) [58], were used to assess the discriminatory power of the model.

4.4. Non-Linear Models

The non-linear models were developed using the neural network tool implemented in the software STATISTICA. In order to identify the best topology and architecture, we ran a large set of 50 models with various topologies. This step is crucial to avoid an (albeit unlikely) overfitting problem. We examined RBF and MLP networks since these usually perform better than other algorithms. The discriminatory power of the models was assessed using the cross-validation method. The models were validated using the cross-validation tool implemented in the ANN module of the STATISTICA software. In this validation procedure, the software automatically assigns 70% of the dataset to training the model. Once the model is trained, the remaining 30% of the inputs are used for validation. It is important to note that if an entry is used in the training set it cannot be used for the validation series.

5. Conclusions

Developing new, reliable, and robust methods for predicting protein function and enzyme class and subclasses is a key goal for theoreticians, especially in light of the recently introduced EC 7 class. In this work, we developed linear and non-linear models using an alignment-free approach to discriminate between enzymes and non-enzymes, as well as assign each enzyme sequence to a specific EC class. The best LDA model showed an overall accuracy of 98.63%, which is considered a remarkable result. However, we decided to explore further and develop some non-linear models using two different algorithms: MLP and RBF. While the latter was unable to improve the results of the LDA model, the MLP model was able to achieve an overall accuracy of 100%. This means that it was able to perfectly discriminate between enzymes and non-enzymes and identify the EC class of each enzyme.

Supplementary Materials: Supplementary materials can be found at http://www.mdpi.com/1422-0067/20/21/5389/s1.

Author Contributions: Conceptualization, R.C.; Data curation, R.C.; Investigation, R.C.; Methodology, R.C.; Software, R.C.; Supervision, M.N.D.S.C.; Validation, M.N.D.S.C.; Writing—original draft, R.C.; Writing—review & editing, M.N.D.S.C.

Funding: This work was supported by UID/QUI/50006/2019 with funding from FCT/MCTES through national funds.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Nomenclature, E. Enzyme nomenclature: Recommendations (1972) of the international union of pure and applied chemistry and the international union of biochemistry: Supplement i: Corrections & additions (1975). *Biochim. Et Biophys. Acta (BBA) Enzymol.* **1976**, *429*, 1–45.
2. Rose, P.W.; Prlić, A.; Altunkaya, A.; Bi, C.; Bradley, A.R.; Christie, C.H.; Di Costanzo, L.; Duarte, J.M.; Dutta, S.; Feng, Z.; et al. The RCSB protein data bank: Integrative view of protein, gene and 3D structural information. *Nucleic Acids Res.* **2016**, *45*, D271–D281. [PubMed]
3. Jensen, L.J.; Gupta, R.; Blom, N.S.; Devos, D.; Tamames, J.; Kesmir, C.; Nielsen, H.; Stærfeldt, H.; Rapacki, K.; Workman, C.; et al. Prediction of Human Protein Function from Post-translational Modifications and Localization Features. *J. Mol. Biol.* **2002**, *319*, 1257–1265. [CrossRef]
4. Davidson, N.J.; Wang, X. Non-Alignment Features based Enzyme/Non-Enzyme Classification Using an Ensemble Method. In Proceedings of the Ninth International Conference on Machine Learning and Applications, Washington, DC, USA, 12–14 December 2010; pp. 546–551.
5. Wang, Y.C.; Wang, X.B.; Yang, Z.X.; Deng, N.Y. Prediction of enzyme subfamily class via pseudo amino acid composition by incorporating the conjoint triad feature. *Protein Pept. Lett.* **2010**, *17*, 1441–1449. [CrossRef] [PubMed]
6. Concu, R.; Dias Soeiro Cordeiro, M.; Munteanu, C.R.; Gonzalez-Diaz, H. Ptml model of enzyme subclasses for mining the proteome of bio-fuel producing microorganisms. *J. Proteome Res.* **2019**, *18*, 2735–2746. [CrossRef]
7. Dobson, P.D.; Doig, A.J. Distinguishing enzyme structures from non-enzymes without alignments. *J. Mol. Biol.* **2003**, *330*, 771–783. [CrossRef]
8. Che, Y.; Ju, Y.; Xuan, P.; Long, R.; Xing, F. Identification of Multi-Functional Enzyme with Multi-Label Classifier. *PLoS ONE* **2016**, *11*, e0153503. [CrossRef]
9. Amidi, A.; Amidi, S.; Vlachakis, D.; Megalooikonomou, V.; Paragios, N.; Zacharaki, E.I. EnzyNet: Enzyme classification using 3D convolutional neural networks on spatial representation. *PeerJ* **2018**, *6*, e4750. [CrossRef]
10. Hu, Q.N.; Zhu, H.; Li, X.; Zhang, M.; Deng, Z.; Yang, X.; Deng, Z. Assignment of EC Numbers to Enzymatic Reactions with Reaction Difference Fingerprints. *PLoS ONE* **2012**, *7*, e52901. [CrossRef]
11. Cock, P.J.A.; Chilton, J.M.; Grüning, B.; Johnson, J.E.; Soranzo, N. Ncbi blast integrated into galaxy. *Gigascience* **2015**, *4*, 39. [CrossRef]
12. Todd, A.E.; Orengo, C.A.; Thornton, J.M. Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.* **2001**, *307*, 1113–1143. [CrossRef] [PubMed]

13. Tian, W.; Skolnick, J. How Well is Enzyme Function Conserved as a Function of Pairwise Sequence Identity? *J. Mol. Biol.* **2003**, *333*, 863–882. [CrossRef] [PubMed]
14. Rost, B.; Liu, J.; Nair, R.; Wrzeszczynski, K.O.; Ofran, Y. Automatic prediction of protein function. *Cell. Mol. Life Sci. CMLS* **2003**, *60*, 2637–2650. [PubMed]
15. Nagao, C.; Nagano, N.; Mizuguchi, K. Prediction of Detailed Enzyme Functions and Identification of Specificity Determining Residues by Random Forests. *PLoS ONE* **2014**, *9*, 84623. [CrossRef] [PubMed]
16. Quester, S.; Schomburg, D. EnzymeDetector: An integrated enzyme function prediction tool and database. *BMC Bioinform.* **2011**, *12*, 376. [CrossRef]
17. The UniProt, C. Ongoing and future developments at the universal protein resource. *Nucleic Acids Res.* **2011**, *39*, D214–D219. [CrossRef]
18. Kanehisa, M. From genomics to chemical genomics: New developments in KEGG. *Nucleic Acids Res.* **2006**, *34*, D354–D357. [CrossRef]
19. Frishman, D.; Mokrejs, M.; Kosykh, D.; Kastenmüller, G.; Kolesov, G.; Zubrzycki, I.; Gruber, C.; Geier, B.; Kaps, A.; Albermann, K.; et al. The pedant genome database. *Nucleic Acids Res.* **2003**, *31*, 207–211. [CrossRef]
20. Li, Y.; Wang, S.; Umarov, R.; Xie, B.; Fan, M.; Li, L.; Gao, X. Deepre: Sequence-based enzyme ec number prediction by deep learning. *Bioinformatics* **2018**, *34*, 760–769. [CrossRef]
21. Dalkiran, A.; Rifaioglu, A.S.; Martin, M.J.; Cetin-Atalay, R.; Atalay, V.; Doğan, T. ECPred: A tool for the prediction of the enzymatic functions of protein sequences based on the EC nomenclature. *BMC Bioinform.* **2018**, *19*, 334. [CrossRef]
22. Shen, H.B.; Chou, K.C. EzyPred: A top–down approach for predicting enzyme functional classes and subclasses. *Biochem. Biophys. Res. Commun.* **2007**, *364*, 53–59. [CrossRef] [PubMed]
23. Bairoch, A. The enzyme data bank. *Nucleic Acids Res.* **1993**, *21*, 3155–3156. [CrossRef] [PubMed]
24. Kumar, C.; Choudhary, A. A top-down approach to classify enzyme functional classes and sub-classes using random forest. *EURASIP J. Bioinform. Syst. Biol.* **2012**, *2012*, 1. [CrossRef] [PubMed]
25. Matsuta, Y.; Ito, M.; Tohsato, Y. Ecoh: An enzyme commission number predictor using mutual information and a support vector machine. *Bioinformatics* **2013**, *29*, 365–372. [CrossRef]
26. Agüero-Chapin, G.; González-Díaz, H.; Molina, R.; Varona-Santos, J.; Uriarte, E.; González-Díaz, Y. Novel 2D maps and coupling numbers for protein sequences. The first QSAR study of polygalacturonases; isolation and prediction of a novel sequence fromPsidium guajavaL. *FEBS Lett.* **2006**, *580*, 723–730. [CrossRef]
27. Concu, R.; Dea-Ayuela, M.; Pérez-Montoto, L.G.; Prado-Prado, F.J.; Uriarte, E.; Fernandez, F.B.; Podda, G.; Pazos, A.; Munteanu, C.-R.; Ubeira, F.; et al. 3D entropy and moments prediction of enzyme classes and experimental-theoretic study of peptide fingerprints in Leishmania parasites. *Biochim. Biophys. Acta (BBA) Proteins Proteom.* **2009**, *1794*, 1784–1794. [CrossRef]
28. Concu, R.; Dea-Ayuela, M.A.; Pérez-Montoto, L.G.; Bolas-Fernández, F.; Prado-Prado, F.J.; Podda, G.; Uriarte, E.; Ubeira, F.M.; González-Díaz, H. Prediction of Enzyme Classes from 3D Structure: A General Model and Examples of Experimental-Theoretic Scoring of Peptide Mass Fingerprints ofLeishmaniaProteins. *J. Proteome Res.* **2009**, *8*, 4372–4382.
29. Bernardes, J.S.; E Pedreira, C. A review of protein function prediction under machine learning perspective. *Recent Pat. Biotechnol.* **2013**, *7*, 122–141. [CrossRef]
30. Barigye, S.J.; Marrero-Ponce, Y.; Pérez-Giménez, F.; Bonchev, D. Trends in information theory-based chemical structure codification. *Mol. Divers.* **2014**, *18*, 673–686. [CrossRef]
31. Graham, D.J.; Malarkey, C.; Schulmerich, M.V. Information Content in Organic Molecules: Quantification and Statistical Structure via Brownian Processing. *J. Chem. Inf. Comput. Sci.* **2004**, *35*, 44.
32. Graham, D.J.; Schacht, D. Base information content in organic molecular formulae. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 942. [CrossRef] [PubMed]
33. Graham, D.J. Information content and organic molecules: Aggregation states and solvent effects. *J. Chem. Inf. Modeling* **2005**, *45*, 1223–1236. [CrossRef] [PubMed]
34. Graham, D.J. Information Content in Organic Molecules: Brownian Processing at Low Levels. *J. Chem. Inf. Modeling* **2007**, *38*, 376–389. [CrossRef] [PubMed]
35. González-Díaz, H.; Molina, R.; Uriarte, E. Markov entropy backbone electrostatic descriptors for predicting proteins biological activity. *Bioorganic Med. Chem. Lett.* **2004**, *14*, 4691–4695.
36. González-Díaz, H.; Saíz-Urra, L.; Molina, R.; Santana, L.; Uriarte, E. A Model for the Recognition of Protein Kinases Based on the Entropy of 3D van der Waals Interactions. *J. Proteome Res.* **2007**, *6*, 904–908. [CrossRef]

37. Riera-Fernandez, P.; Munteanu, C.-R.; Escobar, M.; Prado-Prado, F.J.; Martín-Romalde, R.; Pereira, D.; Villalba, K.; Duardo-Sánchez, A.; González-Díaz, H. New Markov–Shannon Entropy models to assess connectivity quality in complex networks: From molecular to cellular pathway, Parasite–Host, Neural, Industry, and Legal–Social networks. *J. Theor. Biol.* **2012**, *293*, 174–188. [CrossRef]
38. Cherkasov, A.; Muratov, E.N.; Fourches, D.; Varnek, A.; Baskin, I.I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y.C.; Todeschini, R.; et al. QSAR Modeling: Where Have You Been? Where Are You Going To? *J. Med. Chem.* **2014**, *57*, 4977–5010. [CrossRef]
39. Basak, S.C.; Natarajan, R.; Mills, D.; Hawkins, D.M.; Kraker, J.J. Quantitative Structure—Activity Relationship Modeling of Juvenile Hormone Mimetic Compounds for Culex pipiens Larvae, with a Discussion of Descriptor-Thinning Methods. *J. Chem. Inf. Modeling* **2006**, *37*, 65–77. [CrossRef]
40. Hill, T.; Lewicki, P. Statistics Methods and Applications. In *A Comprehensive Reference for Science, Industry and Data Mining*; StatSoft: Tulsa, OK, USA, 2006; Volume 1, p. 813.
41. Shahsavari, S.; Bagheri, G.; Mahjub, R.; Bagheri, R.; Radmehr, M.; Rafiee-Tehrani, M.; Dorkoosh, F.A. Application of artificial neural networks for optimization of preparation of insulin nanoparticles composed of quaternized aromatic derivatives of chitosan. *Drug Res.* **2014**, *64*, 151–158. [CrossRef]
42. Tenorio-Borroto, E.; Rivas, C.G.P.; Chagoyán, J.C.V.; Castañedo, N.; Prado-Prado, F.J.; Garcia-Mera, X.; González-Díaz, H. ANN multiplexing model of drugs effect on macrophages; theoretical and flow cytometry study on the cytotoxicity of the anti-microbial drug G1 in spleen. *Bioorganic Med. Chem.* **2012**, *20*, 6181–6194. [CrossRef]
43. Honório, K.M.; De Lima, E.F.; Quiles, M.G.; Romero, R.A.F.; Molfetta, F.A.; Da Silva, A.B.F.; Da Silva, A.B.F. Artificial Neural Networks and the Study of the Psychoactivity of Cannabinoid Compounds. *Chem. Biol. Drug Des.* **2010**, *75*, 632–640. [CrossRef] [PubMed]
44. Jung, E.; Choi, S.H.; Lee, N.K.; Kang, S.K.; Choi, Y.J.; Shin, J.M.; Choi, K.; Jung, D.H. Machine learning study for the prediction of transdermal peptide. *J. Comput. Mol. Des.* **2011**, *25*, 339–347. [CrossRef] [PubMed]
45. Erol, R.; Ogulata, S.N.; Sahin, C.; Alparslan, Z.N.; Erol, R. A Radial Basis Function Neural Network (RBFNN) Approach for Structural Classification of Thyroid Diseases. *J. Med Syst.* **2008**, *32*, 215–220. [CrossRef] [PubMed]
46. Bezerianos, A.; Papadimitriou, S.; Alexopoulos, D. Radial basis function neural networks for the characterization of heart rate variability dynamics. *Artif. Intell. Med.* **1999**, *15*, 215–234. [CrossRef]
47. Munteanu, C.-R.; Magalhaes, A.; Duardo-Sánchez, A.; Pazos, A.; González-Díaz, H. S2SNet: A Tool for Transforming Characters and Numeric Sequences into Star Network Topological Indices in Chemoinformatics, Bioinformatics, Biomedical, and Social-Legal Sciences. *Curr. Bioinform.* **2013**, *8*, 429–437. [CrossRef]
48. Vazquez, J.; Aguiar, V.; Seoane, J.A.; Freire, A.; Serantes, J.; Dorado, J.; Pazos, A.; Munteanu, C.-R. Star Graphs of Protein Sequences and Proteome Mass Spectra in Cancer Prediction. *Curr. Proteom.* **2009**, *6*, 275–288. [CrossRef]
49. Randić, M.; Zupan, J.; Vikic-Topic, D. On representation of proteins by star-like graphs. *J. Mol. Graph. Model.* **2007**, *26*, 290–305. [CrossRef]
50. Fernández-Blanco, E.; Aguiar-Pulido, V.; Munteanu, C.R.; Dorado, J. Random Forest classification based on star graph topological indices for antioxidant proteins. *J. Theor. Biol.* **2013**, *317*, 331–337. [CrossRef]
51. Fernandez-Lozano, C.; Cuiñas, R.F.; Seoane, J.A.; Fernández-Blanco, E.; Dorado, J.; Munteanu, C.-R. Classification of signaling proteins based on molecular star graph descriptors using Machine Learning models. *J. Theor. Biol.* **2015**, *384*, 50–58. [CrossRef]
52. Munteanu, C.R.; González-Díaz, H.; Magalhães, A.L. Enzymes/non-enzymes classification model complexity based on composition, sequence, 3D and topological indices. *J. Theor. Biol.* **2008**, *254*, 476–482. [CrossRef]
53. Wang, H.; Yan, L.; Huang, H.; Ding, C. From Protein Sequence to Protein Function via Multi-Label Linear Discriminant Analysis. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2017**, *14*, 503–513. [CrossRef] [PubMed]
54. Hendryli, J.; Fanany, M.I. Classifying Abnormal Activities in Exam using Multi-Class Markov Chain LDA Based on MODEC Features. In Proceedings of the 4th International Conference on Information and Communication Technology (ICoICT), Bandung, Indonesia, 25–27 May 2016; pp. 1–6.
55. Safo, S.E.; Ahn, J. General sparse multi-class linear discriminant analysis. *Comput. Stat. Data Anal.* **2016**, *99*, 81–90. [CrossRef]
56. Beleites, C.; Salzer, R. Assessing and improving the stability of chemometric models in small sample size situations. *Anal. Bioanal. Chem.* **2008**, *390*, 1261–1271. [CrossRef] [PubMed]

57. Ion-Mărgineanu, A.; Kocevar, G.; Stamile, C.; Sima, D.M.; Durand-Dubief, F.; Van Huffel, S.; Sappey-Marinier, D. Machine Learning Approach for Classifying Multiple Sclerosis Courses by Combining Clinical Data with Lesion Loads and Magnetic Resonance Metabolic Features. *Front. Mol. Neurosci.* **2017**, *11*, 398. [CrossRef] [PubMed]
58. Boughorbel, S.; Jarray, F.; El-Anbari, M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLoS ONE* **2017**, *12*, 0177678. [CrossRef]

© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).