IWIMSM-04: Iberoamerican Workshop on Model. and Simulation Methods, Valencia, Spain, 2020

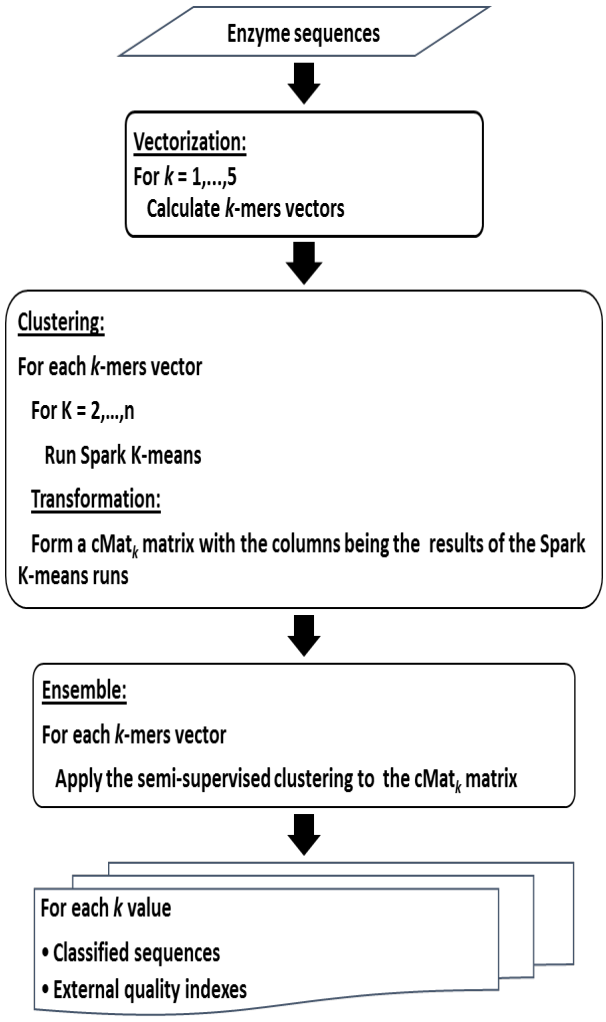
Ensemble K-means for semi-supervised learning in enzymatic activity classification of GH-70 enzymes

Yadelis González Valle [a], Deborah Galpert [a,b]*, Reinaldo Molina-Ruiz [c], Guillermin Aguero-Chapin [d]*

[a] Centro de Investigaciones de Informáticas. Universidad Central ¨Marta Abreu¨ de Las Villas (UCLV), Santa Clara, 54830, Cuba.
[b] Departamento de Ciencia de la Computación. Universidad Central ¨Marta Abreu¨ de Las Villas (UCLV), Santa Clara, 54830, Cuba.
[c] Centro de Bioactivos Químicos (CBQ), Universidad Central ¨Marta Abreu¨ de Las Villas (UCLV), Santa Clara, 54830, Cuba
[d] CIIMAR/CIMAR, Centro Interdisciplinar de Investigação Marinha e Ambiental, Universidade do Porto (UP), Terminal de Cruzeiros do Porto de Leixões, Av. General Norton de Matos s/n 4450-208 Matosinhos, Porto & Departamento de Biologia, Faculdade de Ciências, UP, Rua do Campo Alegre, 4169-007 Porto, Portugal

*Corresponding author DG: deborah@uclv.edu.cu, GACh: gchapin@ciimar.up.pt

Graphical Abstract



Abstract.

The enzymatic activity classification of the GH-70 enzymes is a challenge in Bioinformatics due to the high diversity of these sequences. From the 501 sequences reported when we accessed Cazy.org, just 58 were labeled into 6 EC number classes. In this paper we propose a semi-supervised classification algorithm based on the *k*-mers frequency descriptors with k equals to 2, 3, 4, 5 and 6 as alignment-free measures extracted from the sequences. The high dimensionality of the *k*-mers (c^k) vectors and the increasing number of sequences lead to the application of big data Spark classifiers such as the ones in Apache MLlib. Specifically, the K-means clustering applied in an iterative way yields multiple results that can be ensemble in a semi-supervised second-round clustering step capable of detecting groups of similar sequences including the labeled and the unlabeled ones. Finally, external measures validate the ensemble clustering for the labeled sequences. Further improvements in the clustering and ensemble steps could raise the quality of classification.

Introduction

The GH-70 enzyme family contains low similarity sequences also known as divergent sequences so it is a problem of great importance in Bioinformatics to increase the efficiency of the functional classification of these enzymes. This is why the use of various descriptors free of enzyme alignment appears as a trend in this type of classification [1].

Alignment-free descriptors are methods of extracting sequence similarity based on their modeling using molecular descriptors. An example of these descriptors are word frequency based methods [2][3] which are based on functions of the form $D: x \longrightarrow \mathbb{R}^r$, where a sequence x of length n is converted to a vector of length r. In this way, the methods based on *k*-tuples, *k*-words or *k*-mers, with $k \leq n$, make a correspondence of the sequence with a vector $\pi_x^k = \left(\frac{N_{k,1}}{n-k+1}, \frac{N_{k,2}}{n-k+1}, \ldots, \frac{N_{k,c^k}}{n-k+1}\right)$, whose components $N_{k,1i}, i = 1,..,c^k$ represent the frequency of subsequences of length k, where c^k is the total of all possible *k*-mers of the finite alphabet A of c characters, here representing the twenty amino acids.

The comparison performed using similarity or dissimilarity measures between vectors allows the partition of the vector dataset through the application of an automated learning technique as clustering. When clustering is based on the similarity of the objects, you want the objects belonging to the same group to be as similar as possible and the objects belonging to different groups to be as different as

possible [4][5]. On the other hand, when comparing sequences, one can take into account the previous classification information, that is, the labels of some objects, as in the enzyme classification problem, the previous classification of some sequences into their enzyme activity classes (EC numbers). With this purpose, it is convenient to use semi-supervised learning when there are few labeled sequences. Semi-supervised learning is a branch of machine learning that results from combining supervised and unsupervised learning [6][7].

Materials and Methods

The ensemble clustering transformation algorithm in [8] lies in running a clustering algorithm (or multiple algorithms) several times (say, n times) with different parameter values where each run produce a categorical feature of the new categorical dataset representing the original vector dataset. Applying the transformation on an object $x \in \mathbb{R}^r$ will create a new object $x^* \in$ cMat with categorical values representing the cluster number of x in each run of K-means. The dimension of x_i^* is n−1 where (n-1 $\ll r$).

In this work we use 501 enzyme sequences available on the site cazy.org[1] where the enzymatic activity classification into six EC numbers is known for 58 sequences available in cazy.org/GH70_characterized[2]. Then, we exclude three enzymes from the labeled group: the enzyme "CDX66820.1" with a double classification which means that it has double enzymatic activity, also, “P08987” and “P49331” that were not among the 501 sequences.

For the 501 enzymes, 2-mers, 3-mers, 4-mers, 5-mers, 6-mers were calculated. Spark's K-means for each calculated k-mers was run for K values from 2 to 50. With the results, a $cMat_k$ matrix was formed for each k-mers dataset of vectors and a semi-supervised classification algorithm was applied that used the available information on the previously classified enzymes. In this incremental clustering algorithm, for each query categorical vector q_i^*, the vector matrix a $cMat_k$ is searched for any equal vector x_j^* corresponding to an enzyme belonging to the class C_l. Then, the q_i^* is assigned this class C_l. In case the algorithm does not find equality with any of the enzymes previously classified, q_i^* will then be allocated in a seventh class.

Results and Discussion

The results of the ensemble clustering method are shown in Table 1 and Table 2. From the 55 labeled enzymes, 46 were correctly classified with the use of 2-mers, 50 with 3-mers, 49 with 4-mers and 47 with 5-mers.

Table 1. Results obtained by the ensemble clustering with *2*-mers and 3-mers.

Classes	Classification by classes with 2-mers			Classification by classes with 3-mers		
	Its enzymatic classification unknown	Matches previous enzymatic classification	Total	Its enzymatic classification unknown	Matches previous enzymatic classification	Total
C_1	152	37	189	182	41	223
C_2	6	1	7	6	1	7
C_3	0	0	0	1	1	2
C_4	36	5	41	10	4	14
C_5	4	2	6	1	2	3
C_6	8	1	9	8	1	9
C_7	249	0	249	243	0	243
Total	455	**46**	**501**	**451**	**50**	**501**

[1] http://www.cazy.org

[2] http://www.cazy.org/GH70_characterized.html

Table 2. Results obtained by the ensemble clustering with 4-mers and 5-mers.

Classes	Classification by classes with 4-mers			Classification by classes with 5-mers		
	Its enzymatic classification unknown	Matches previous enzymatic classification	Total	Its enzymatic classification unknown	Matches previous enzymatic classification	Total
C_1	134	38	172	157	37	194
C_2	5	2	7	6	1	7
C_3	1	1	2	1	1	2
C_4	47	5	52	31	5	36
C_5	3	2	5	2	2	4
C_6	0	1	1	8	1	9
C_7	262	0	262	249	0	249
Total	452	**49**	**501**	**454**	**47**	**501**

To validate the grouping made by the proposed method, a confusion matrix was calculated for each of the six groups of enzyme activity and for each of the *k*-mers that were used. The values obtained from the external evaluation metrics with the labeled sequences can be seen in Table 3. In bold-face are highlighted the highest results.

Table 3. Values of the external evaluation metrics of the classifications.

External Measure	Values obtained with 2-mers	Values obtained with 3-mers	Values obtained with 4-mers	Values obtained with 5-mers	Values obtained with 6-mers
Accuracy	0.95	0.95	0.95	**0.96**	0.95
Precision	0.73	0.78	**0.81**	0.79	**0.81**
Recall	0.61	0.86	0.72	0.85	**0.93**
F1 Measure	0.74	0.76	0.72	0.76	**0.83**

Conclusions

The proposed procedure allows grouping in a semi-supervised manner, by including available classification information in the dataset. The *k*-mers from 2 to 6 were used to group sequences with the K-means algorithm implemented in Apache Spark, based on the Euclidean distance to measure the dissimilarity between pairs of sequences. When validating the clustering, values of 0.95 were obtained with 6-mers in the Accuracy measure, 0.81 in Precision, 0.93 in Recall, and 0.83 in F1 Measure. For further studies we may use other *k*-mers values, with $k > 6$ and different parameter values of the ensemble.

References

[1] G. J. Davies and M. L. Sinnott, "The sequence-based classifications of carbohydrate-active enzymes. Sorting the diverse," *Regul. Biochem. J. Class. Pap.*, pp. 27–32, 2008.

[2] P. Melsted and J. k. Pritchard, "Efficient counting of k-mers in DNA sequences using a bloom filter," *BMC Bioinformatics*, vol. 12, no. 333, pp. 1–7, 2011.

[3] U. Gunasinghe, D. Alahakoon, and S. Bedingfield, "Extraction of high quality k-words for alignment-free sequence comparison," *J. Theor. Biol.*, vol. 358, pp. 31–51, 2014, doi: 10.1016/j.jtbi.2014.05.016.

[4] R. Kruse, C. Döring, and M. Lesot, *Fundamentals of Fuzzy Clustering, in Advances in Fuzzy Clustering and its Applications*, J. Valente. 2007.

[5] F. Höppner, F. Klawonn, R. Kruse, and T. Runkler, *Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition*. 1999.

[6] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-Supervised Learning*. 2006.

[7] Xiaojin Zhu, *Semi-Supervised Learning Literature Survey*. 2005.
[8] L. Abdallah and M. Yousef, "GrpClassifierEC: a novel classification approach based on the ensemble clustering space," 2020.