SciForum
Mol2Net

Machine Learning and Atom-Based Quadratic Indices for Proteasome Inhibition Prediction

Gerardo M. Casañola Martin,[1,2,3]* Huong Le-Thi-Thu,[4] Facundo Perez-Gimenez,[2] and Concepción Abad[1]

[1] Departament de Bioquímica i Biologia Molecular, Universitat de València, E-46100 Burjassot, Spain; emails: gerardo.casanola@uv.es (G.M.C.M) ; concepción.abad@uv.es (C.A)

[2] Unidad de Investigación de Diseño de Fármacos y Conectividad Molecular, Departamento de Química Física, Facultad de Farmacia, Universitat de València, Spain. emails: gerardo.casanola@uv.es (G.M.C.M); facundo.perez@uv.es (F.P.G)

[3] Universidad Estatal Amazónica, Facultad de Ingeniería Ambiental, Paso lateral km 2 1/2 via Napo, Puyo, Ecuador gcasanola@uea.edu.ec (G.M.C.M)

[4] School of Medicine and Pharmacy, Vietnam National University, Hanoi (VNU) 144 Xuan Thuy, Cau Giay, Hanoi, Vietnam ltthuong1017@gmail.com (H.L.T.T)

* Author to whom correspondence should be addressed; E-Mail: gerardo.casanola@uv.es ; Tel.: +34-963543156

Published: 4 December 2015

Abstract: The atom-based quadratic indices are used in this work together with some machine learning techniques that includes: support vector machine, artificial neural network, random forest and k-nearest neighbor. This methodology is used for the development of two quantitative structure-activity relationship (QSAR) studies for the prediction of proteasome inhibition. A first set consisting of active and non-active classes was predicted with model performances above 85% and 80% in training and validation series, respectively. These results provided new approaches on proteasome inhibitor identification encouraged by virtual screenings procedures.

.

Keywords: Atom-based quadratic index, classification and regression model, machine learning, proteasome inhibition, QSAR, TOMOCOMD-CARDD software

Mol2Net YouTube channel*: http://bit.do/mol2net-tube*

1. Introduction

The ubiquitin-proteasome pathway (UPP) is responsible for the selective degradation of the majority of the intracellular proteins in eukaryotic cells and regulates nearly all cellular processes [1]. Disfunction of the ubiquitination machinery or the proteolytic activity of the proteasome is associated with many human diseases [2]. Proteasome inhibitors have been developed being effective for some disorders but sometimes show detrimental effects and resistance. Therefore, efforts are currently directed to the development of new therapeutics with adequated potency and safety properties that target enzyme components of the UPP [3,4].

Ligand-based molecular design and QSAR approaches are promising fields with several applications in drug development, which use a battery of novel molecular descriptors and different classification algorithms for in silico virtual drug screening studies [5,6]. In the present research, we use and compare a set of different machine learning (ML) techniques using the 2D atom-based quadratic indices as attributes with the objective to perform the QSAR modeling of two datasets. The first dataset allows to separate molecules with proteasome inhibitory activity from inactive ones, and the second provides the numerical prediction of the EC_{50}.

2. Results and Discussion

In the case of our classification study, we reduced the inactive subset removing all the cases that fall outside of the applicability domain of our model. Therefore, the dataset remains with 705 chemicals, being 258 active and the rest 447 inactive ones. The first 705 dataset used for classification studies generates 529 in the training set (TS) and 176 compounds in the prediction set (PS). Based on the aspects mentioned above for our case a first step with non-supervised feature reduction filtering was done, by using the Shannon´s entropy as a measure keeping c.a. the 30% of the features (4 143). In a second step a supervised feature reduction filtering was done. In this stage, the process was carried out for the class problem. In this case the features were reduced a 70%, keeping a total of 1248 for the class data. These feature selection processes were carried out with the IMMAN software an “in house” program. Later, in the two-class data the best subset search was done resulting in 43 selected variables. Then wrapper methods associated with the ML techniques were applied to reduce data sets giving different data subsets combinations. Finally, all these subsets were used to generate diverse ML-QSAR models keeping those with the best results for each algorithm. The results for each ML technique used to develop classification QSAR models to predict proteasome inhibitors are shown in Fig. 1.

As it can be observed in Fig. 1 for the TS the fitted models using RF and MLP techniques showed the best accuracies (Ac = 90.17% and Ac = 89.22%) with Mathew´s correlation coefficient (MCC) values of 0.79 and 0.77, respectively. In the case of the PS, the performance of these two QSAR models was of 86.36% (MCC=0.70) and 83.52% (MCC=0.64), respectively. Moreover, can be observed low values of false positive rates, which ensures a good performance at time to perform virtual high-throughput screenings, disminissing the wrong evaluation of predicted positive cases. In the same Fig. 1 can also be noted that RF outperforms other models in most of the quality parameters. Besides, the rest of the models also depicted adequate performances with accuracies values above 85% in the case of the TS and 80 % for the PS.

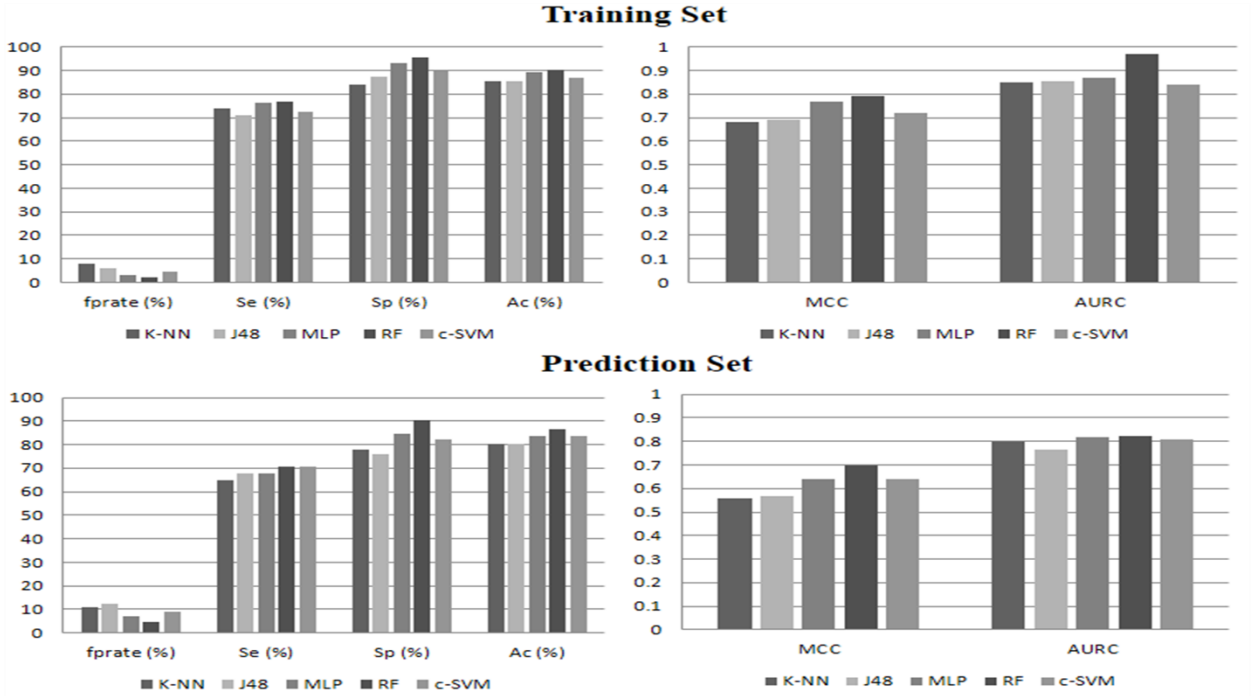

Figure 1. Performance of the ML-based QSAR classifiers

3. Materials and Methods

In this study the molecular descriptors atom-based quadratic indices were calculated using the TOMOCOMD software version 1.0 [7]. We also attempt the different feature selection methods implemented in the IMMAN software [8]. Moreover, the attribute selection method based on BestSubset Search (BSS) of LDA discriminant analysis was used [9]. Later, the wrapper and ranker methods of Waikato environment for knowledge analysis (WEKA) [10] were considered. As a final stage, the parameter tuning optimization for each ML technique was performed to find the best ML-QSAR models.

A dataset derived from a luminescent cell-based dose titration retest counterscreen assay to identify inhibitors of the proteasome pathway was selected from PubChem BioAssay (AID 2486) where the name, structures, compound identifier (CID), and activities can be found. First, a curation process on the database was assessed removing salts, and inorganic compounds. The main difficulty of the ML approaches is to select attributes from a large list of candidates to describe the data. This is because the complete set of molecular descriptors is not needed for the description of the proteasome inhibition. In this sense, the addition of non-relevant attributes can cause noise to the ML systems [10]. Therefore, the feature selection approaches are very suitable to deal with this kind of problem. In this work, different schemes of attribute selection including filter and wrapper approaches implemented in WEKA [10] are examined to select the best attribute subset for each ML technique. Some details, advantages and drawbacks of the two approaches can be reviewed in many works dealing with this subject [11-13].

The machine learning methods shows impressive performances a wide diversity of studies involving automated, text classification and drug design [14-16]. Based on this the machine learning approaches selected were: support vector machine, artificial neural network and k-nearest neighbor also included in the list of

the top ten algorithms used in data mining [17]. Besides the random forest technique was included because is fast and robust approach with recent succesfull application into many problems [18-20]. For each ML method applied in this study, various schemes of selecting attributes were examined and for each selected subset, various models were developed and checked out.

4. Conclusions

In this work, a QSAR study on a diverse and enlarged proteasome inhibitor database collected from the PubChem Bioassay is shown for the first time. The random forest algorithm demonstrates to be the best technique for the modeling of the proteasome inhibitory activity with high accuracies values in the training and test set. The low false positive rates observed validates the presented workflow based on ML-QSAR for the prediction of active proteasome inhibitors compounds from inactive ones.

Acknowledgments

Casañola-Martin. G.M. and Castillo-Garit. J.A thank the program 'Estades Temporals per a Investigadors Convidats' for a fellowship to research University (2013-2014) at Valencia. Marrero-Ponce, Y. thanks to the program 'International Professor' for a fellowship to work at Cartagena University in 2013-2014. Le-Thi-Thu, H. gratefully acknowledge support from the National Vietnam National University, Hanoi

Conflicts of Interest

"The authors declare no conflict of interest".

References and Notes

1. Varshavsky, A. The ubiquitin system, an immense realm. *Annu. Rev. Biochem* **2012**, *81*, 167-176.
2. Rastogi, N.; Mishra, D.P. Therapeutic targeting of cancer cell cycle using proteasome inhibitors. *Cell Division* **2012**, *7*, 26.
3. de Bettignies, G.; Coux, O. Proteasome inhibitors: Dozens of molecules and still counting. *Biochimie* **2010**, *92*, 1530-1545.
4. Pevzner, Y.; Metcalf, R.; Kantor, M.; Sagaro, D.; Daniel, K. Recent advances in proteasome inhibitor discovery. *Expert Opinion on Drug Discovery* **2013**, *8*, 537-568.
5. Rescigno, A.; Casañola-Martin, G.M.; Sanjust, E.; Zucca, P.; Marrero-Ponce, Y. Vanilloid derivatives as tyrosinase inhibitors driven by virtual screening-based qsar models. *Drug Test Anal* **2011**, *3*, 176-181.
6. Kumar, D.; Kapoor, A.; Thangadurai, A.; Kumar, P.; Narasimhan, B. Synthesis, antimicrobial evaluation and qsar studies of 3-ethoxy-4-hydroxybenzylidene/4-nitrobenzylidene hydrazides. *Chin. Chem. Lett* **2011**, *22*, 1293-1296.
7. Marrero-Ponce, Y.; Valdés-Martini, J.R.; García Jacas, C.R. *Tomocomd-cardd qubils software qubils-mas. Version 1.0*, CAMD-BIR Unit, Universidad Central "Marta Abreu" de Las Villas, 2012.
8. Barigye, S.J.; Pino Urias, R.W.; Marrero-Ponce, Y. *Imman (information theory based chemometric analysis) version 1.0.*, 2011.
9. *Statistica (data analysis software system) vs 6.0*, StatSoft Inc: Tulsa,OK:, 2001.

10. Witten, I.H.; Frank, E. *Data mining: Practical machine learning tools and techniques*. 2nd ed. ed.; Morgan Kaufmann: Burlington, MA, 2005.
11. Ben Meskina, S. In *On the effect of data reduction on classification accuracy*, 2013.
12. Shahlaei, M. Descriptor selection methods in quantitative structure-activity relationship studies: A review study. *Chemical Reviews* **2013**, *113*, 8093-8103.
13. Inza, I.; Larrañaga, P.; Blanco, R.; Cerrolaza, A.J. Filter versus wrapper gene selection approaches in DNA microarray domains. *Artificial Intelligence in Medicine* **2004**, *31*, 91-103.
14. Baumes, L.A.; Ranilla, J. A study on factors affecting the reproducibility of a chemical tongue analysis responding to amino acids. *Combinatorial Chemistry and High Throughput Screening* **2013**, *16*, 572-583.
15. Gertrudes, J.C.; Maltarollo, V.G.; Silva, R.A.; Oliveira, P.R.; Honório, K.M.; Da Silva, A.B.F. Machine learning techniques and drug design. *Current Medicinal Chemistry* **2012**, *19*, 4289-4297.
16. Le-Thi-Thu, H.; Marrero-Ponce, Y.; Casañola-Martin, G.M.; Cardoso, G.C.; Chávez, M.D.C.; Garcia, M.M.; Morell, C.; Torrens, F.; Abad, C. A comparative study of nonlinear machine learning for the "in silico" depiction of tyrosinase inhibitory activity from molecular structure. *Molecular Informatics* **2011**, *30*, 527-537.
17. Wu, X.; Kumar, V.; Ross, Q.J.; Ghosh, J.; Yang, Q.; Motoda, H.; McLachlan, G.J.; Ng, A.; Liu, B.; Yu, P.S.*, et al.* Top 10 algorithms in data mining. *Knowledge and Information Systems* **2008**, *14*, 1-37.
18. Ziegler, A.; König, I.R. Mining data with random forests: Current options for real-world applications. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **2014**, *4*, 55-63.
19. Verikas, A.; Gelzinis, A.; Bacauskiene, M. Mining data with random forests: A survey and results of new tests. *Pattern Recognition* **2011**, *44*, 330-349.
20. Chen, X.; Ishwaran, H. Random forests for genomic data analysis. *Genomics* **2012**, *99*, 323-329.