QSAR Modelling for Drug Discovery: Predicting the Activity of LRRK2 inhibitors for Parkinson's Disease using Cheminformatics Approaches

Víctor Sebastián-Pérez[1], María J. Martínez[2], Carmen Gil[1], Nuria E. Campillo[1], Ana Martínez[1], and Ignacio Ponzoni[2,*]

[1] Centro de Investigaciones Biológicas (CIB, CSIC), Ramiro de Maeztu 9, 28040 Madrid, España
[2] Instituto de Ciencias e Ingeniería de la Computación (UNS–CONICET), Departamento de Ciencias e Ingeniería de la Computación, Universidad Nacional del Sur (UNS), Bahía Blanca, Argentina

*E-mail: ip@cs.uns.edu.ar

Abstract. Parkinson's disease is one of the most common neurodegenerative disorders in elder people and the leucine-rich repeat kinase 2 (LRRK2) is a promising target for its pharmacological treatment. In this paper, QSAR models for identification of potential inhibitors of LRRK2 protein are designed by using an in house chemical library and several machine learning methods. The applied methodology works in two steps: first, several alternative subsets of molecular descriptors relevant for characterizing LRRK2 inhibitors are identified by a feature selection software tool; secondly, QSAR models are inferred by using these subsets and three different methods for supervised learning. The performance of all these QSAR models are assessed by traditional metrics and the best models are analyzed in statistical and physicochemical terms.

Keywords: Cheminformatics, QSAR, Machine Learning, Parkinson's disease, LRRK2.

1. Introduction

Nowadays, the search of effective treatments for neurodegenerative diseases is one of the urgent clinical and social needs. Number of people affected by those pathologies, including Alzheimer's and Parkinson's diseases, increase every year, mainly in developed countries, directly associated to the longer life expectancy. Parkinson's Disease (PD) is the second most common human neurodegenerative disorder in people over 60 years of age, affecting 1 in 100 people and increasing to that affects 2–3% of the population ≥65 years of age. It is associated with Lewy bodies, abnormal aggregates of α-synuclein protein, and loss of dopaminergic neurons in the substantia nigra. Although clinical diagnosis is based on the existence of bradykinesia and other cardinal

motor characteristics, Parkinson disease is associated with many non-motor symptoms that add to overall disability.

Epidemiological and genetic studies carried on several families in Asia, the United States, and Europe led to discover in 2004 that mutations in a new gene, known as leucine-rich repeat kinase 2 (LRRK2), are a major genetic risk factor for familiar and sporadic PD [1]. Today, LRRK2 is one of the most pursuing and promising targets for the future pharmacological treatment of PD. In this sense, big efforts are being done both from academia and pharmaceutical industry with the goal of developing selective and brain-permeable LRRK2 inhibitors as a strategy for PD [2, 3]. LRRK2 is an unusual large protein (2527 amino acids) classified as a member of the ROCO superfamily. It presents a leucine-rich repeat (LRR) domain, a kinase domain, a DFG-like motif, a RAS domain, a GTPase domain, a MLK-like domain, and a WD40 domain. The protein is present mainly in the cytoplasm, although it is also related to the mitochondrial outer membrane. The physiological role of LRRK2 is poorly understood and many of its substrates remain unclear. However, it has been proposed to be beneficial for preventing neurodegeneration [4,5] and several LRRK2 inhibitors are being developed as neuroprotective agents for PD. Some studies revealed that LRRK mutations increases aggregation of α-synuclein in dopaminergic neurons that are exposed to α-synuclein fibrils [6].

Quantitative structure–activity/property relationship (QSAR/QSPR) modeling has been established itself as one of the major computational molecular modeling methodologies, playing a central role in drug identification or optimization. QSAR/QSPR models allow to identify relationships between structural information of chemical compounds (molecular descriptors) and a physicochemical or biological property under study. Now, these techniques are widely used as a surrogate for experimental studies to predict the activity of the molecules from their structure. In particular, machine learning methods had become extensively used in this field during the last decades [7].

Regarding literature, very few QSAR studies in LRRK2 have been published. The works that have been reported presented a very limited predictive activity for the external validation datasets [8,9]. In this paper, novel QSAR models for predicting putative inhibitors of LRRK2 protein are developed by using machine learning methods. In particular, several regression and classification QSAR models are inferred and their performances contrasted in terms of accuracy and model complexity.

2. Material and Methods

Several QSAR models inferred by feature selection techniques, both for classification and regression models, are described. Figure 1 presents a scheme of the experimental design followed in this work. The database is a compilation of 67 compounds previously synthesized in our research group and tested as inhibitors of LRRK2 enzyme [10]. In this assay, LRRK2 kinase activity was measured as the percentage of enzyme inhibition for

every compound as it is further discussed in the previous reference. This information is available on http://lidecc.cs.uns.edu.ar/pacbb2018.LRRK2/structures_activity.html.

Database analysis and drug-like properties calculation

A crucial step in QSAR studies is to collect a representative set of compounds in order to include a diversity physicochemical space. With the aim of analyzing the dataset, we have performed a characterization of the compounds, both from a physicochemical perspective and a drug-like point of view. Physicochemical and drug-like properties of this dataset were calculated using Qikprop and the most representative descriptors were analyzed to show the diversity of the dataset.

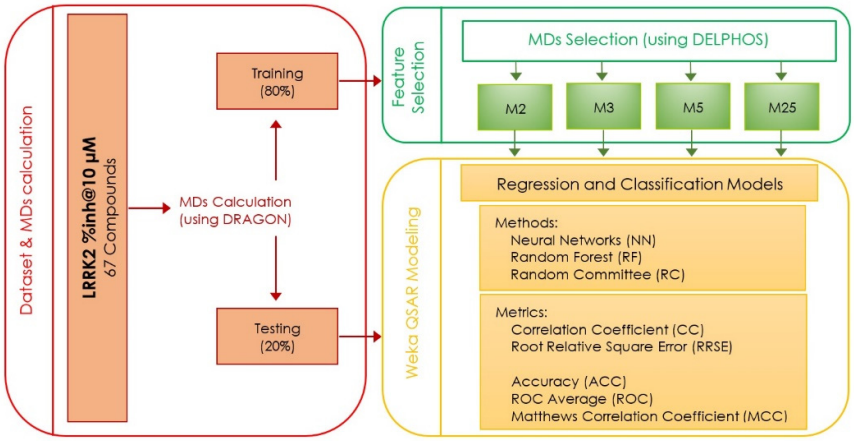

Fig. 1. Scheme of the *in silico* experiments reported for predicting the activity of LRRK2 inhibitors.

Some of these parameters are plotted in Fig. 2, where it can be observed a wide dispersion of 2 different key properties in drug discovery such as logP (x axis) and H-bond acceptors (accptHB y-axis). Compounds are colored taking into account their stars values. This parameter represents the number of properties or descriptors values calculated that fall outside the 95% range of similar values for known drugs. For this reason, a large value of stars suggests that a molecule is less drug-like. In this case, all compounds present a value equal or lower than 2, which means that the complete database is based on drug-like structure. Furthermore, taking into account Lipinski rule of 5 [11], the 67 compounds present 1 or none violations of the rules, which means that the molecules have properties that make them likely orally active drug in humans. Therefore, and after the analysis carried out, we can conclude that the database is diverse in terms of physicochemical properties and all the compounds are drug like.

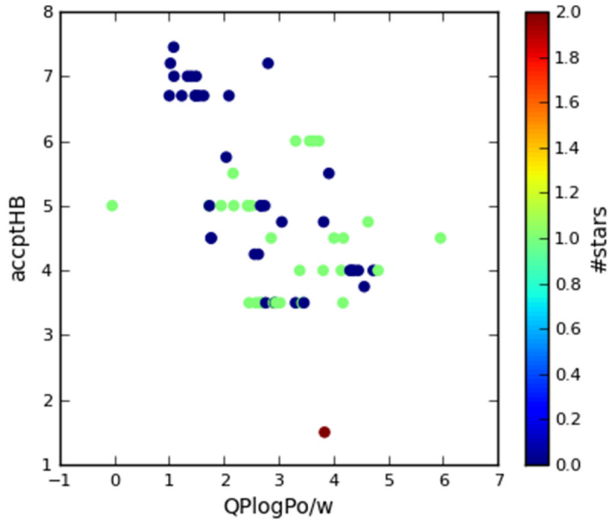

Fig. 2. Dispersion of the database regarding PQ properties, colors are defined by stars.

QSAR models

A total number of 3224 descriptors were computed using Dragon for the entire database. The experiments were designed following the procedure described in Fig. 1. The dataset was first divided into training (75%) and test set following a stratified sampling. Several subsets of descriptors were selected from the training set using DELPHOS [13]. This tool repeats ten times the random partition of the data in 75/25 to perform the validation of the selected characteristics. In Table 1, a summary of the best molecular descriptors subsets in terms of RAE (Relative Absolute Error) is reported. Using these 4 subsets and different inference methods, a variety of QSAR models were built for regression and classification. The models are computed by WEKA [14] using Neural Networks (NN), Random Forest (RF) and Random Committee (RC) as inference methods, for computing these models default parameters were used in WEKA. Several methods for obtaining the QSAR models were tested due to the fact that recent studies have shown that there does not exist a more advisable strategy for inferring the QSAR from the subsets of descriptors [12]. For classification models, discretization thresholds of target property values were as follows: low activity $\leq$ 50% and high activity >50%. Table 2 shows several metrics computed using WEKA for the best regression and classification QSAR models obtained with each descriptor subset. The performance results for classification models are reported using the accuracy (ACC), namely percentage of cases correctly classified, the average Receiver Operating Characteristic (ROC) and the Matthews Correlation Coefficient (MCC). For regression models, the correlation coefficient (CC) and the root relative square error (RRSE) results are informed.

Table 1. Best molecular descriptors subsets obtained by DELPHOS feature selection tool.

Subset	Cardinality	MDs		Descriptor Type
M2	4	MW		Constitutional indices
		MWC08		Walk and path counts
		BEHp2		Burden eigenvalues
		RDF105p		RDF descriptors
M3	4	MW		Constitutional indices
		JGI2		2D autocorrelation
		HATs6m	R2e	GETAWAY descriptors
M5	5	MW		Constitutional indices
		IC0		Information indices
		ESpm09x		Edge adjacency indices
		JGI3		2D autocorrelation
		L3s		WHIM descriptors
M25	13	MW		Constitutional indices
		HNar	ECC	Topological indices
		GATs7e		2D autocorrelations
		VEZ1	VEp2	2D matrix-based descriptors
		DISPm		Geometrical descriptors
		RDF105p		RDF descriptors
		R8e		GETAWAY descriptors
		B06[N-Br]	B07[C-Cl]	2D atom pairs
		F04[C-C]	F05[O-Cl]	2D atom pairs

Table 2. Performances of the best regression and classification QSAR models for the external validation testing set.

Model	Cardinality	Best Regression QSAR Models			Best Classification QSAR Models			
		Method	CC	RRSE	Method	ACC	ROC	MCC
M2	4	RF	0.55	87.50	RF	68.8	0.69	0.40
M3	4	RC	0.68	74.69	RC	87.5	0.90	0.77
M5	5	RF	0.83	60.82	RC	75.0	0.95	0.52
M25	13	RC	0.44	92.34	RC	75.0	0.73	0.53

The best classification model was inferred from the subset M3 by using Random Committee and achieved 87.5% of correct classification with a ROC value of 0.91, all results are shown in Table 2. In other hand, the best regression model was obtained with the subset M5 by using Random Forest and achieved a correlation coefficient of 0.83 and a RRSE of 60.82. The datasets used to generate the classification and regression models can be found in link http://lidecc.cs.uns.edu.ar/pacbb2018.LRRK2/datasets.html.

The best model found in the regression case contains 5 different descriptors that includes a wide variety of different descriptors classes, from 0D molecular descriptors to 3D. For example, molecular weight is a constitutional index, inside the class of 0D-descriptors, that is obtained from the chemical formula, as they do not consider the tridimensional structure of the ligands. We have also found ESpm09x descriptor, spectral moment 09 from edge adjacency matrix weighted by edge degrees and IC0 information content index (neighborhood symmetry of 0-order) and JGI3: mean topological charge index of order 3 that are 2D descriptors from different families. Finally, the 3D descriptor found in this model is the L3s: 3rd component size directional WHIM index / weighted by I-state. Regarding the best classification model, includes 4 descriptors and we have also found a wide representation of different descriptors families. MW has also been chosen in this model as a 0D descriptor. A very similar descriptor to JGI3 was also found in this model, JGI2, is a 2D descriptor and is related to the charge index of the compounds in the database. Finally, two 3D descriptors were selected by DELPHOS that are R2e: R autocorrelation of lag 2/weighted by Sanderson electronegativity and HATs6m: leverage-weighted autocorrelation of lag 6/weighted by mass. Both are GETAWAY descriptors (GEometry, Topology, and Atom-Weights AssemblY) and are related to electronegativity and mass, key parameters in protein-ligand recognition process for protein inhibition.

Furthermore, we analyzed the relationship among the descriptors in statistical terms by using VIDEAN, which is a visual analytics tool for the study of molecular descriptor subsets. It shows the relationships and interactions between the descriptors and the target property in statistical terms. The analysis of the pair correlation between the five descriptors that conform the regression model and the four descriptors that are part of the best classification model is presented in Fig. 3 using Kendall tau metric.

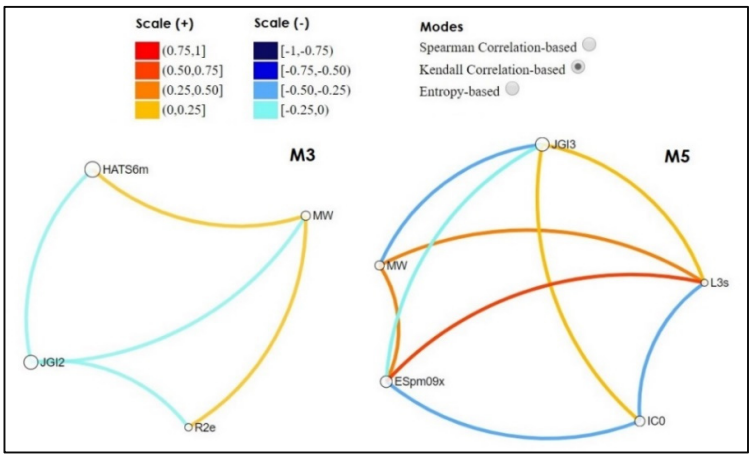

Fig. 3. Kendall correlation analysis among the descriptors that conforms the subsets M3 y M5.

In this type of analysis, the main goal is to identify models with low correlation among descriptors that means low redundant information. In this case, we observe a clear majority of light tones of links, both blue and orange, that make connection between the descriptors (nodes) present in the models. This fact, demonstrate the low data redundancy in both models that have been selected as the best ones.

All the experimental section and protocols regarding database analysis and property calculation as well as QSAR models both building and analysis can be found at: http://lidecc.cs.uns.edu.ar/pacbb2018.LRRK2/supplementary_material.html.

3. Conclusions

Parkinson's disease constitutes one of the neurodegenerative disorders with higher impact in elder population around the world. An auspicious target for its pharmacological treatment is the leucine-rich repeat kinase 2 (LRRK2). Several studies proposed that LRRK2 inhibitors can be a beneficial strategy for preventing neurodegeneration. For this reason, in this paper, QSAR models have been developed with the aim to use them as useful filters for virtual screening to identify potential inhibitors of LRRK2 protein. These models were obtained by machine learning methods over data from an in house chemical library.

The computational approach used in this work follows two main steps: first, alternative subsets of molecular descriptors relevant for structural characterization of LRRK2 inhibitors are identified by a feature selection method; secondly, QSAR models are learned from these subsets by applying several supervised learning algorithms. The performance of these QSAR models was contrasted by traditional metrics.

The molecular descriptor subsets associated with the regression and classification models that reached the best performances were analyzed in statistical and physicochemical terms. From the analysis, it is possible to observe that the selected subset has low cardinality but cover a wide spectrum of the molecular descriptor classes, contributing in this way with meaningful and diverse structural information to the models. Besides, the visual analytics study reveals that the selected molecular descriptors provides non-redundant information in statistical terms.

Nevertheless, even when these QSAR models achieve high accuracies, it is important to mention that these models have been learned from datasets integrated by a reduced number of chemical compounds, which can limit the generalization properties of these predictive models. For this reason, our advice for potential practitioners of these models is to employ applicability domain methods over their testing compounds before apply these models. As future work, we hope to extend our in house chemical library for LRKK2 in order to improve the generalizability of these achievements.

Acknowledgments

This work is kindly supported by CONICET, grant PIP 112-2012-0100471 and UNS, grants PGI 24/N042 and PGI 24/ZM17. We also acknowledge MECD, VSP grant FPU15/01465 and Banco Santander for VSP fellowship AY21/17-D-27 in the "Becas Iberoamerica-Santander Investigación" program.

References

1. Zimprich, A., Biskup, S., Leitner, P., Lichtner, P., Farrer, M., Lincoln, S., Kachergus, J., Hulihan, M., Uitti, R.J., Calne, D.B., Stoessl, A.J., Pfeiffer, R.F., Patenge, N., Carbajal, I.C., Vieregge, P., Asmus, F., Muller-Myhsok, B., Dickson, D.W., Meitinger, T., Strom, T.M., Wszolek, Z.K., Gasser, T.: Mutations in LRRK2 cause autosomal-dominant parkinsonism with pleomorphic pathology. Neuron 44, 601-607 (2004)
2. Gilligan, P.J.: Inhibitors of leucine-rich repeat kinase 2 (LRRK2): progress and promise for the treatment of Parkinson's disease. Curr Top Med Chem 15, 927-938 (2015)
3. Estrada, A.A., Sweeney, Z.K.: Chemical Biology of Leucine-Rich Repeat Kinase 2 (LRRK2) Inhibitors. J Med Chem 58, 6733-6746 (2015)
4. Cookson, M.R.: LRRK2 Pathways Leading to Neurodegeneration. Curr Neurol Neurosci Rep 15, 42 (2015)
5. Smith, W.W., Pei, Z., Jiang, H., Moore, D.J., Liang, Y., West, A.B., Dawson, V.L., Dawson, T.M., Ross, C.A.: Leucine-rich repeat kinase 2 (LRRK2) interacts with parkin, and mutant LRRK2 induces neuronal degeneration. Proc Natl Acad Sci U S A 102, 18676-18681 (2005)
6. Volpicelli-Daley, L.A., Abdelmotilib, H., Liu, Z., Stoyka, L., Daher, J.P., Milnerwood, A.J., Unni, V.K., Hirst, W.D., Yue, Z., Zhao, H.T., Fraser, K., Kennedy, R.E., West, A.B.: G2019S-LRRK2 Expression Augments alpha-Synuclein Sequestration into Inclusions in Neurons. J Neurosci 36, 7415-7427 (2016)
7. Lima, A., Philot, E., Trossini, G., Scott, L., Maltarollo, V., Honorio, K.: Use of machine learning approaches for novel drug discovery. Expert Opin Drug Discov 11, 225-239 (2016)
8. Kahn, I., Lomaka, A., Karelson, M.: Topological Fingerprints as an Aid in Finding Structural Patterns for LRRK2 Inhibition. Mol Inform 33, 269-275 (2014)
9. Pourbasheer, E., Aalizadeh, R.: 3D-QSAR and molecular docking study of LRRK2 kinase inhibitors by CoMFA and CoMSIA methods. SAR QSAR Environ Res 27, 385-407 (2016)
10. Salado, I.G., Zaldivar-Diez, J., Sebastian-Perez, V., Li, L., Geiger, L., Gonzalez, S., Campillo, N.E., Gil, C., Morales, A.V., Perez, D.I., Martinez, A.: Leucine rich repeat kinase 2 (LRRK2) inhibitors based on indolinone scaffold: Potential pro-neurogenic agents. Eur J Med Chem 138, 328-342 (2017)
11. Lipinski, C.A., Lombardo, F., Dominy, B.W., Feeney, P.J.: Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev 46, 3-26 (2001)
12. Eklund, M., Norinder, U., Boyer, S., Carlsson, L. Choosing feature selection and learning algorithms in QSAR. J Chem Inf Model 54, 837-843 (2014)
13. Soto, A.J., Martínez, M. J., Cecchini, R. L., Vazquez, G. E. & Ponzoni, I.: DELPHOS: Computational Tool for Selection of Relevant Descriptor Subsets in ADMET Prediction. 1st International Meeting of Pharmaceutical Sciences (2010)
14. Eibe Frank, M.A.H., and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.