Chemometrics and Intelligent Laboratory Systems 176 (2018) 53–65

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics

A new data representation based on relative measurements and fingerprint patterns for the development of QSAR regression models

Irene Luque Ruiz *, Miguel Ángel Gómez Nieto

University of Córdoba, Department of Computing and Numerical Analysis, Campus de Rabanales, Albert Einstein Building, E-14071, Córdoba, Spain

ARTICLE INFO

Keywords:
QSAR
Relative distance matrix
Support vector machine
Regression models
Factor VIIa inhibitors
Interleukin-4 inhibitors

ABSTRACT

Relative distance matrixes represent measurements of the structural characteristics of the molecules, having into account a reference pattern common to the whole data set considered in the development of QSAR regression models. These matrixes store relationships between the data set molecules, measuring the transformation cost between pairs of molecules and a pattern from the common fragments to the entire data set. These measurements are quite related with the activity value changes and, therefore, its use allows the building of robust QSAR regression models. In this paper, we describe the building of relative distance matrixes for the representation of two data sets with clearly different characteristics and previously used as benchmark. Applying Support Vector machine algorithms, several training models and external validation were carried out randomly selecting both sets. The results obtained with correlation coefficient greater than 0.9, low values of error and values of slope and bias close to the ideality have shown the goodness of the presented proposal, clearly improving the results obtained in the literature.

1. Introduction

Many of the QSAR proposals conducted for the development of prediction models are based on the use of fingerprints or descriptors matrixes as input data to the corresponding algorithm. Both fingerprint and descriptor matrixes are non symmetrical and non squared matrixes where each row represents a molecule and each column represents a characteristic extracted from the structural, topological, sterical and so on molecule information [1–4].

Thus, each column of the fingerprint matrix informs about whether the corresponding molecule (row of the matrix) contains a structural fragment having, in this case, the value of 1, and each column of the descriptor matrix informs about the value of an invariant or descriptor obtained from the molecular structure of the molecule corresponding to the row of the matrix.

As observed, in both cases these matrixes used for the representation of a set of molecules store measures extracted from each molecule information independently of the presence or absence of other molecules in the set.

On the other hand, based on the principle that similar molecules show similar properties, similarity matrixes have also been commonly used as input data for the predictions of molecules' activity in QSAR proposals. These symmetrical and squared matrixes can be derived from fingerprint and descriptor matrixes or any other kind of matrixes such as matrixes of molecular fragments, non-isomorphic fragments, etc. [5–9].

In any case, these types of representations of the molecules of a data set contain absolute measurements of the characteristics of the molecules used in the QSAR process and they do not have into account the data set composition.

The use of matrixes containing relative measurements of the characteristics of the molecules of a data set has been also widely considered [10–12]. In this case, each column of the input data matrix contains a molecular characteristic extracted from the molecule represented in the row of the matrix by means of an internal or external structural reference. This structural reference, frame, pattern or core is used to align or fix all the molecules of the data set, carrying out the calculation of 2D/3D invariants having into account this match of the molecule to the pattern [13]. Recently [14], we have demonstrated that the use of relative measurements of the molecules of a data set improves the results of the classification process, generating robust QSAR classification models with a wide applicability domain.

In this paper, we show the improvement that the use of relative distances representing the relationship between the molecules of the data set in the development of regression QSAR models.

* Corresponding author.
E-mail addresses: iluque@uco.es (I. Luque Ruiz), mangel@uco.es (M.Á. Gómez Nieto).

https://doi.org/10.1016/j.chemolab.2018.03.007
Received 10 December 2017; Received in revised form 3 March 2018; Accepted 13 March 2018
Available online 15 March 2018
0169-7439/

The reference frame or pattern used to obtain these relative distances is retrieved from the whole data set characteristics. Thus, data sets are initially represented by a fingerprint matrix and the reference pattern is calculated from the maximum common bits set to 1 of the fingerprint matrix. Weighing the pattern fingerprint and each molecules' fingerprint of the data set, a matrix can be built containing relative distances between each pair of molecules of the data set and this matrix can be efficiently used for the development of QSAR regression models.

This paper has been organized as follows: the introduction section states the background and aims of the research. In section 2 the selected data sets for validating our proposal and the pillars of itself are analyzed, describing the building of the different types of representations of the input data to the QSAR algorithm based on fingerprint, similarity and relative distance matrixes.

In section 3, the experimental results are presented, describing the characteristics and behavior of the different input matrixes considered. In addition, the results obtained for the regression models built are analyzed, carrying out the study and interpretation of outliers. In addition, in section 3, the results obtained for the external validation of the generated regression models for the studied data sets are described, comparing our best results with those using descriptors matrix as input data to the regression algorithm. Finally, in the last section our proposal and the results are discussed.

2. Materials and method

2.1. Experimental data sets

Two data sets with different characteristics have been selected for the analysis and validation of the proposal presented in this paper, aiming to prove its applicability to different chemical spaces: Factor VIIa inhibitors (*F7*) and Interleukin-4 (*IL4*).

The formation of the serine protease FVIIa and its cell associated cofactor tissue factor (TF) complex is the critical step that initiates the sequentially amplified cascade of proteolytic activation steps, which ultimately leads to the generation of the protease thrombin and thrombus formation. Thus, an antithrombotic agent based on the inhibition of FVIIa/TF may have advantages over the inhibition of downstream coagulation proteases such as Factor Xa [15].

On the other hand, Interleukin-4 is a cytokine that regulates multiple biological functions. It can regulate proliferation, differentiation, and apoptosis in several cell types of haematopoietic and non-haematopoietic origin. It has a critical role in the regulation of Th0 cell differentiation during a normal immune response, and IL-4-driven Th2 cells direct host responses against parasitic infections and immune diseases including allergy, autoimmunity and cancer [16,17].

Source data of Factor VIIa inhibitors (*F7*) and Interleukin-4 (*IL4*) inhibitors data sets were extracted from a previous work [18]. In this paper, authors recovered both data sets from GOSTAR databases and used Free-Wilson methodology applying non linear Support Vector Machine algorithms to some data models of the data sets based on Pipeline Pilot extended-connectivity fingerprint (ECFP) [19], signatures of R-Groups and physicochemical descriptors [20].

Authors obtained regression models for the whole data sets reaching values of $r^2 = 0.73$ for *F7* and $r^2 = 0.74$ for *IL4* data sets. However, when the data sets are split having into account the similar signature/fragments the correlation coefficient for *F7* data set increases until values greater than 0.8.

Both data sets have also been considered by Norinder et al. [21] to propose QSAR regression models using conformal prediction as mathematical support to establish a high applicability domain. In this study, authors applied random forest algorithm (RF) obtaining values of $r^2 = 0.66$ for *F7* and $r^2 = 0.53$ for *IL4* data sets in external validations with percentages of accuracy between 80% and 90% at confidence level between 0.8 and 0.9 respectively.

Later, Cortés-Ciriano et al. [22] uses *F7* and *IL4* data sets in a study of the influence of the experimental error of the bioactivity values regarding the QSAR algorithm used in the building of the prediction models. In this experimental, twelve data sets and twelve algorithms were considered. Each data set was split in a training subset (~85%) and a test subset (~15%), and the experimental values of the response (dependent) variable of the training subset were modified by means of a noise function. The values of the noise added depend of the values of the response variable of the molecules considered in the training subset. The results have shown that some algorithms generate better models than others despite the experimental error of the molecular bioactivity existing in the data sets.

F7 and *IL4* data sets have been once again considered by Cortés-Ciriano et al. [23] for the study of the improvement of the statistics results of the prediction models thanks to the data augmentation. In this work, once again the original data sets were split in a subset for training and other subset for external validation. Using a technique similar to y-randomizing [24], authors modify the values of the predictors or/and response variables using a noise function and, in addition, from none to five replications are performed to the training subset, thus augmenting the number of molecules used for the model building. Each replication implies the duplication of the training subset with the same number of molecules but with values of the response variable modified by the noise function. Thanks to this technique an improvement close to 10% is obtained in the RMSE values. In the case of *F7* data set, the lower RMSE obtained was 0.41 using a noise of 0.1 for descriptors, 0.2 for bioactivities and using five replications of the training subset. In the case of *IL4* data set, the lower RMSE value was 0.29 using a noise equal to zero for predictor variables, 0.05 for bioactivities and three replications.

With the aim of improving the prediction QSAR models, *F7* and *IL4* data sets have been considered with others 27 data sets in a study using ligand efficiency indices (LE) instead of pIC_{50} values as response variable [25]. Prediction models using eleven different LE and pIC_{50} as dependent variable were generated using Partial Least Squares Regression (PLS), Support Vector Machine (SVM), Gradient Boosting Machines (GBM) and Random Forest (RF). Data sets were split in six partitions, using five partitions for the training subset and separating a partition to be used later as test subset. The use of LEs as response variables clearly increased the predictive power of the regression models in comparison to the use of pIC_{50} as dependent variable. Thus, authors obtain an increase of r^2_{test} of ~0.3 and a decrease of the normalized RMSE >0.1 units in some cases. However, these results were later studied by Sheridan [26] using the same 29 data sets and others in-house data sets, demonstrating that although the use of ligand efficiency indexes by means of the transformation of pIC_{50} values improves the prediction capability of the QSAR models, this improvement is due to a statistical artifact occurring when: i) LE strongly correlates (positively or negatively) with the computable physical property included in the LE definition, ii) the property in the LE definition is easier to predict than pIC_{50}.

As observed, *F7* and *IL4* data sets can be considered as benchmark data sets having been used as reference in important proposals of the development of QSAR prediction models. In all the works above described in which *F7* and *IL4* data sets have been considered, the number of molecules used of *F7* and *IL4* data sets has been the same as the one used in this work. Thus, in these experiments, the QSAR models have been developed splitting the original data sets in two independent subsets for training and test and no other external subsets have been considered.

The different characteristics of these data sets can be observed in Fig. 1. *F7* data set is composed of 365 molecules with pIC_{50} activity values in the interval from 4.04 to 8.18, distributed as the histogram of Fig. 1 (left) shows. As it can be observed, this data set shows a moderate distribution of the molecules' data set along the whole interval except for the lower (<4.25) and higher activities (>7.7) values, in which a few number of molecules are found.

IL4 data set is composed of 665 molecules with pIC_{50} activity values