17th European Symposium on Computer Aided Process Engineering – ESCAPE17
V. Plesu and P.S. Agachi (Editors)

QSAR Analysis of 1,4-Dihydropyridine Calcium Channel Antogonists

Pınar Kahraman and Metin Türkay*

College of Engineering and the Center for Computational Biology and Bioinformatics, Koç University, Rumelifeneri Yolu, Sarıyer, İstanbul, 34450, TURKEY.
pkahraman@ku.edu.tr, mturkay@ku.edu.tr

Abstract

The early prediction of activity related characteristics of drug candidates is an important problem in drug design. The activities of drug candidates are classified as low or high depending on their IC_{50} values. Since experimental determination of IC_{50} values for a vast number of molecules is both time consuming and expensive, computational approaches are employed. In this paper, we present a novel approach to classify the activities of drug molecules. We use hyper-boxes classification method in combination with partial least squares regression to determine the most relevant molecular descriptors of the drug molecules in efficient classification. The effectiveness of the approach is illustrated on DHP derivatives. The results indicate that the proposed approach outperforms the other approaches reported in the literature.

Keywords: Drug design, QSAR analysis, data classification, mixed-integer programming

1. Introduction

The early prediction of activity related characteristics of drug candidates is an important problem in drug design. A large ratio of the capital spent while commercializing a drug is spent on unsuccessful candidate drugs. Therefore, eliminating molecules with undesired properties beforehand has been one of the central research subjects in structure based drug design. Since the number of

possible drug candidates is often in the order of millions, computerized methods are used for prediction of activities. One way is to study chemical structures of the candidate molecules and to predict the activity levels of drug candidates based on them.

One of the data driven methods that is widely used in drug design is QSAR (quantitative structure-activity relationship). QSAR is the effort of understanding correlation between the chemical structure of a molecule and its biological and chemical activities such as biotransformation ability, reaction ability, solubility or target activity. The main assumption in QSAR is that structurally similar molecules tend to have similar activities and that molecules with unknown properties can be compared to structures with known properties. 3D structures of molecules may be used to find many candidate molecules that will fit into the target binding site, which can be constructed using a variety of methods. The problem of early prediction of properties of drug candidates becomes a machine learning problem when there are a number of structurally similar molecules of known activities that fit into the binding site. The activity of the molecules is usually classified into two classes: high or low active based on their toxicity. The reason for this binary classification is that the numerical values for biological activities are not available in most cases.

In this paper, we consider a subgroup of a class of drugs called calcium channel blockers that inhibit the Ca^{+2} flux into the cell. Calcium channel antagonists affect on many excitable cells, like heart muscle cells, vein muscle cells and neuron cells. The special group of antagonists that we concentrate on in this paper are the 1,4-dihydropyridine calcium channel antagonists, also called DHP derivatives. These antagonists are mostly used for the treatment of cardiovascular diseases, such as hypertension and exertional angina.[1] The structural analysis of calcium channel blockers summarizes conformational analysis on a set of 1,4-dihydropyridine derivatives.[2] In this work, a seven-descriptor model is built and the least square support vector machines (LSSVM) method is used in both obtaining the model and in classifying the molecules based on the model. To our knowledge, the most results on the same data set has been released in 2006.[3]

This paper presents a new methodology for early prediction of drug behavior. We use a sequence of methods for characterization of activity levels of drug candidates: COSESSA,[4] for feature generation, PLS for feature selection and a novel mixed-integer programming based classification method[5] for the classification of non-separable data that minimizes misclassifications considerably. We apply this approach to 1,4-dihydropyridine calcium channel blockers for comparison purposes.

2. Strategies, Models and Methods

In a QSAR analysis, the method used in each step has major importance for the success of the study. As well as the classifier, the process used to determine the

numeric molecular attributes (i.e. the descriptors) and the regression method that selects the most relevant descriptors contribute to the efficiency of the analysis. This paper uses the program CODESSA[4] for descriptor calculations, which is reliable and widely used software in QSAR analyses. The molecular structures of drug candidates are constructed and then optimized by energy minimization using HyperChem. The optimized molecular structures are then processed in CODESSA[4] to generate descriptors for each molecule.

The objective of the next step is to determine a model that will describe the activity in terms of the descriptors. In this paper, we used PLS, which is basically an MLR method closely related to principal component regression. PLS is especially efficient when number of instances is much smaller than the number of descriptors. We used MINITAB[21] for PLS runs, each providing a linear model of the dependent variable. The variables that have coefficients of zero are concluded to have no relationship with the independent variable. Standardized coefficients are considered for indicating relevant variables. Once the model is built and the most relevant descriptors are identified, the next step is the classification of the drugs based on the values of the descriptors.

Classification of drugs was carried out based on the selected descriptors and their values using the hyper-boxes method, which is a mixed-integer programming based model.[5] The hyper-boxes model encloses inputs in hyper-boxes by solving an MILP problem. This approach is used to classify the 45 DHP derivatives as low active and high active, first using the initial selection of descriptors, and then utilizing the descriptors chosen after the significance analysis. Comparison of the hyper-boxes model as a classifier is done with 51 different classification methods available in WEKA[6]. The problem with possible overestimation of the importance of descriptors mentioned above is addressed by making significance tests to the selected descriptor values after the preliminary classification. The significance test examines whether the hypothesis that the variance of the whole set of drugs is equal to the variance of the subset of drugs separated by the classification process can be significantly adopted. We expect the variance of the wholes set to be larger than the variances of the subsets, which become the alternative hypothesis. Analytically, the null hypothesis is $S_{ij}^2 = S_{ik}^2$ and we test for $H_a = S_{ij}^2 > S_{ik}^2$, where j represents the whole data set, and k is one of the classes.

3. Application on DHP Derivatives

The data classification approach presented in the previous section is applied on 45 variants of 1,4-Dihydropyridine Calcium Channel Antagonists (DHP). We present the illustration of the approach and the results in this section.

3.1. The Data Set

The quantitative structure-activity relationship study in this paper is applied to 45 drug molecules that are constructed based on a template molecule. The 45 antagonists are constructed by attaching various fragments to the upper ring X of the template as illustrated in Figure 1. In addition to the fragments, experimental values for $\log(1/IC_{50})$ values are also given in the table. IC_{50} corresponds to the concentration of an inhibitor necessary for 50% inhibition of the targets in vitro. This quantity is used as a measure of drug effectiveness. The lower the effectiveness of the drug is, the smaller the $\log(1/IC_{50})$ value gets.[7] Drugs having $\log(1/IC_{50})$ values lower than 6.72 are classified as low active, indicated by asterisks in the table, and the others as high active.[2] The values of 172 molecular descriptors for each 45 drugs are obtained in CODESA.

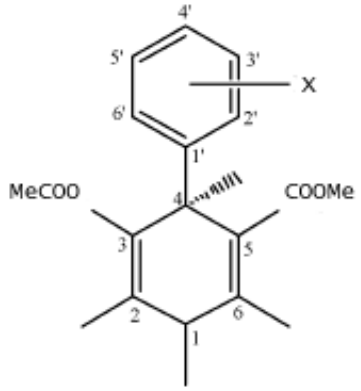

Fig. 3. Template of 1,4-Dihydropyridine Calcium Channel Antagonists.

3.2. Results

Three models were formed in partial least squares: 7, 10 and 15 attribute models. The reason for constructing several models was to increase the accuracy by having different descriptors and models allowing us to replace insignificant descriptors of the 7 and 10-attribute models with the significant ones selected from the 15-attribute model. The relevant variables are chosen based on the absolute values of the standardized coefficients calculated through MINITAB. The most relevant descriptors for the 7, 10 and 15 attribute models are listed with their absolute standardized coefficients and incremental contributions to the R^2 value.

After selecting the descriptors, the hyper-box data classification method was solved for binary classification of drugs, as high or low active. In this step, 66% of the data, i.e. 29 instances, was included in the training set, and the

remaining 16 in the test set. The 10-attribute model achieved 100% accuracy with the hyper-boxes method. However, the result of the first classification run with 7 descriptors is relatively low, indicating a possible existence of some descriptors with low significance in terms of classifying the drugs as low active and high active. In the corresponding classification, the hyper boxes method placed one molecule in both classes, which is the reason of the "half placements". It can be deduced from the results that as the number of descriptors used increases, the accuracy of the classification process increases, since more of the dependent variable is explained. After these preliminary classification runs, a significance test is conducted on class variances.

Table 1. Significance test results for the initial classification run for 7 descriptors.

ATTRIBUTE	CLASS	SAMPLE VAR	P-VALUE
moment of inertia c	all together	1.86E-12	
	high class	2.45E-12	0.69553
	low class	2.14E-13	0.02475
zx shadow / zx rectangle	all together	4.78E-09	
	high class	5.43E-09	0.59965
	low class	4.98E-09	0.58132
yzshadow	all together	3.32E-05	
	high class	1.71E-05	0.13479
	low class	7.55E-05	0.89063
moment of inertia b	all together	3.07E-12	
	high class	3.66E-12	0.63273
	low class	1.49E-12	0.25311
rel. no. of double bonds	all together	7.14E-10	
	high class	6.57E-10	0.45377
	low class	1.43E-09	0.85419
minimum partial charge	all together	3.46E-12	
	high class	1.73E-13	0.00001
	low class	1.04E-11	0.94724
xy shadow / xy rectangle	all together	3.80E-09	
	high class	4.22E-09	0.58499
	low class	1.45E-09	0.18189

Corresponding *p*-values of the 7-descriptor model are provided in Table 1, where a *p*-value below a certain α value means that the null hypothesis that the variance of the values of the corresponding descriptor for all 45 drugs is equal to the variance of the values for the indicated class, can be rejected with 1-α confidence against the alternative hypothesis that the variance of the whole set is larger than the variance of the classified set. It can be seen that different descriptors are significant for different classes. "Moment of inertia c" and "xy shadow / xy rectangle" have *p*-values smaller than 0.2 for the low class, which means that the molecules in this class have very similar values for these descriptors. For "minimum partial charge", the *p*-value for the high class is significantly low, which indicates that this descriptor is significant for the high class. However, *p*-values for "relative number of double bonds" and "zx shadow / zx rectangle" are considerably large for both classes. As the number

of descriptors used in classification increases, either the significance for the high class, the significance for the low class, or both improve, i.e. the corresponding *p*-values decrease, for the descriptors still surviving in the larger models, increasing the accuracy of classification analysis. After the significance analysis, the set of molecular descriptors are identified for the 7-descriptor model.
The hyper-boxes model is once more implemented to the 45 molecules. A classification accuracy of 100% is obtained using the proposed approach. From these values, it can be deduced that such a significance analysis and an adjustment in descriptor selection pays off with higher accuracy levels.

4. Conclusions

In this paper, a novel approach for the early prediction of the behavior of drug molecules is presented. The 45 calcium channel antagonists are chosen from literature, which have been studied widely. The steps that constitute the method are compared with those that are available in literature using the same data set. It is seen that with the presented approach, a 7-attribute model is enough to reach a 100% accuracy in classifying the data set into high active and low active, and the proposed sequence of methods seems to be providing the best results among the studies that have been published so far. Moreover, a novel classifier, the MILP based hyper-boxes method, is proved to be highly accurate and superior to all of the classifiers available in WEKA and the results reported in literature.

References

1. Takahata, Y.; Costa, M. C. A.; Gaudio, A. C. (2003). Comparison between neural networks (NN) and principle component analysis (PCA): structure activity relationships of 1,4-dihydropyridine calcium channel antagonists (nifedipine analogues). *J. Chem. Inf. Comput. Sci,* 43, 540-544.
2. Yao, X.; Liu, H.; Zhang, R.; Liu, M.; Hu, Z.; Panaye, A.; Doucet, J. P.; Fan, B. (2005). QSAR and classification study of 1,4-dihydropyridine calcium channel antagonists based on least squares support vector machines. *Mol. Pharm.,* 2(5), 348-356.
3. Si, H. Z.; Wang, T.; Zhang, K. J.; Hu, Z. D.; Fan, B. (2006). QSAR study of 1,4-dihydropyridine calcium channel antagonists based on gene expression programming. *Bioorg. Med. Chem.,* 14, 4834-4841.
4. Katritzky, A. R.; Lobanov, V. S.; Karelson, M. (1997). *Comprehensive Descriptors for Structural and Statistical Analysis, Reference Manual,* version 2.0 and 2.13; University of Florida: Gainsville, FL.
5. Üney, F.; Türkay, M. (2006). A mixed-integer programming approach to multi-class data classification problem. *Eur. Jour. Oper. Res.,* 173(3), 910-920.
6. WEKA 3: Data Mining Software in Java. The University of Waikato, 2005.
7. Patankar, S. J.; Jurs, P. C. (2000). Prediction of IC50 values for ACAT inhibitors from molecular structure. *J. Chem. Inf. Comput. Sci,* 40, 706-723.