ARTÍCULO

An Evolutionary Approach for Feature Selection applied to ADMET Prediction

Axel J. Soto[1,2,3], Rocío L. Cecchini[1,3], Gustavo E. Vazquez[1], Ignacio Ponzoni[1,2]

[1]Laboratorio de Investigación y Desarrollo en Computación Científica (LIDeCC),
Departamento de Ciencias e Ingeniería de la Computación (DCIC)
Universidad Nacional del Sur – Av. Alem 1253 – 8000 – Bahía Blanca (Argentina)
[2]Planta Piloto de Ingeniería Química (PLAPIQUI)
Universidad Nacional del Sur – CONICET
Complejo CRIBABB – Camino La Carrindanga km.7 – CC 717 – Bahía Blanca (Argentina)
[3]These authors contributed equally to this manuscript
{saj,rlc,gev,ip }@cs.uns.edu.ar

Abstract

Feature selection methods look for the selection of a subset of features or variables in a data set, such that these features are the most relevant for predicting a target value. In chemoinformatics context, the determination of the most significant set of descriptors is of great importance due to their contribution for improving ADMET prediction models. In this paper, an evolutionary-based approach for descriptor selection aimed to physicochemical property prediction is presented. In particular, we propose a genetic algorithm with a fitness function based on decision trees, which evaluates the relevance of a set of descriptors. Other fitness functions, based on multivariate regression models, were also tested. The performance of the genetic algorithm as a feature selection technique was assessed for predicting logP (octanol-water partition coefficient), using an ensemble of neural networks for the prediction task. The results showed that the evolutionary approach using decision trees is a promising technique for this bioinformatic application.

Keywords: Feature Selection, Genetic Algorithms, QSAR, hydrophobicity

1. Introduction

Historically, when a new drug had to be developed, a 'serial' process started where drug potency (activity) and selectivity were examined first [1]. Many of the candidate compounds failed at later stages due to ADMET (absorption, distribution, metabolism, excretion and toxicity) behavior in the body. ADMET properties are related to the way that a drug interacts with a large number of macromolecules and they correspond to the principal cause of failure in drug development [1]. In this way, a compound can be promising at first based on its molecular structure, but other factors such as aggregation, limited solubility or limited uptake in the human organism turn it useless as a drug.

In Di Masi [2], data was collected from a survey of 10 pharmaceutical companies and the estimated out-

of-pocket cost per new drug is of about U$S 400 million and compared with earlier studies, there is an increase in the annual rate of the costs (7.3%) above general price inflation. In ref. [3] a description of drug development and research costs are detailed by the therapeutic category.

Currently, the failure rate of a potential drug before reaching the market is still high. The main problem resides in the difficulty to know the rules that govern ADMET behavior in the human body. For these reasons, interest in Quantitative Structure-Activity Relationships (QSAR) and Quantitative Structure-Property Relationships (QSPR) given by the scientific and industrial community has grown considerably in last decades. These both approaches comprise the methods by which chemical structure parameters are quantitatively correlated with a well defined process, such as biological activity or any other experiment. QSAR has evolved over a period of 30 years from simple regression models to different computational intelligence models that are now applied to a wide range of problems [4] [5]. Nevertheless, the accuracy of the ADMET property estimations remain as a challenging problem [6].

In this context, hydrophobicity is one of the most extensively modeled physicochemical property since the difficulty of experimentally determine its value, and also because it is directly related with ADMET properties [4] [7]. This property is traditionally expressed in terms of the logarithm of the octanol-water partition coefficient (logP).

QSAR methods developed by computer means are commonly named as *in silico* methods. These *in silico* methods, clearly cheaper than *in vitro* experiments, allow to examine thousands of molecules in shorter time and without the necessity of intensive laboratory work. Although *in silico* methods are not pretended to replace high-quality experiments at least in the short term, some computer methods have demonstrated to obtain as good accuracy as well-established experimental methods [8]. Moreover, one of the most important features of this approach is that a candidate drug (or a whole library) can be tested before being synthesized. Due to the gains in saved labour time, *in silico* predictions considerably help to reduce the large percentage of leads that fail in later stages of their development, and to avoid the amount of time and money invested in compounds that will not be successful.

In this context, machine learning methods are most preferred given the great amount of existing data and the little understanding of the pharmacokinetic rules of xenobiotics in the human body. Jónsdottir et al [5] detail an extensive review of the many machine learning methods applied to bio- and chemoinformatics.

1.1. Main Goal

One common way of expressing the structural composition of a molecule in a QSAR method is by way of the calculus of whole-molecular descriptors. Each descriptor defines an attribute of the entire molecule, and its value could be obtained by experimental measures or numerical methods. According to the nature of descriptors, they are organized by families. Examples of descriptors are: molecular weight, number of hydrogen atoms, number of alcohols, sum of atomic electronegativities, etc. [9] [10].

One major dilemma when logP is intended to be modeled by QSAR is that, with the exclusion of a few common descriptors, there is no general agreement of which descriptors are relevant or influence the hydrophobic behavior of a compound. This is an important fact, because overfitting and chance correlation could occur as a result of using more descriptors than necessary [11] [12]. Moreover, if less or different than influential descriptors are used, poor models come as a result. In the context of Artificial Intelligence, this topic constitutes a particular case of the *feature selection problem*.

In this way, this work presents a sound approach for inferring most influential descriptors for physicochemical properties. This novel technique is obtained by the application of a genetic algorithm in conjunction with a decision tree for the determination of the fitness function. Additionally, other different approaches for the fitness function are analyzed and compared. This work is organized as follows: next section discusses related issues of feature selection in AI and in chemoinformatics in particular. Section 3 expands the aforementioned idea by the introduction of the genetic algorithm proposed for descriptor selection. In Section 4, the design of the experiments is presented, followed by the results obtained by the applied methods. Finally, in Section 5, main conclusions and future work are discussed.

2. Feature Selection

Feature selection is the common name that is used to comprise all the methods that select from or reduce the set of variables or features used to describe any situation or activity in a dataset. Some authors differentiate variables from features, assuming that variables are the raw entry data, whereas features correspond to processed variables. However both terms will be used here without distinction.

Feature selection have turned into a topic of great importance in many areas, given that, nowadays, applications with datasets of many (even hundreds or thousands) variables have become frequent. Most usual cases where this technique is applied are gene selection from microarray data [13] [14] and text categorization [15] [16]. Confront the *curse of dimensionality* carries some recognized advantages like: reducing the measurement and storage requirements, facilitating visualization and understanding of data, diminishing training and predicting times and also improving prediction performance.

In general, feature selection methods work looking for irrelevant or redundant variables in the data. It can be considered as a necessary pre processing step in machine learning, because it helps to improve the performance of a learning model in several aspects. However, there are cases where developing a predictor is not the final goal. This is the case when the problem consists of finding or ranking all (or most) potentially relevant variables. As it can be elucidated, selecting most relevant variables may be suboptimal for a predictor, especially when relevant variables are redundant. On the other hand, a subset of useful variables may exclude redundant, but relevant, variables [17] [18] [19].

In this context, variable selection methods may be applied in two main ways, in terms of whether variables are individually or globally evaluated. That is, the first of them, works ranking each variable in an isolated way, i.e. these methods rank variables according to their individual predictive power. Pearson correlation [20], the coefficient of determination [13], information theoretic criteria [15] and partial least squares (PLS) [21] are used and correspond to this class.

However, a variable that is useless by itself could be useful in consideration with others variables [19]. In this way, more powerful learning models are obtained, when the feature selection model selects subsets of variables that jointly have good predictive capacity.

A refined division of feature selection methods, especially applied to the latter defined group, is commonly used. They are often divided into filters, wrappers and embedded methods. First of them, filters, select subsets of variables, as a pre-processing step, independently to the selected learning method. Wrappers utilize the learning machine technique of interest as a black box to score subsets of variables in terms of their predictive ability. Finally, embedded methods carry out feature selection in the process of the training of a learning method and are usually tailored to the applied learning method.

2.1 Feature Selection in Bioinformatics

Many several papers successfully applied the feature selection strategy in Bioinformatics related areas, like: drug discovery, QSAR, patient treatment and gene expression patterns analysis. We decided to apply descriptor selection in our work in order to detect which were the ten most useful descriptors for the prediction of logP. We agreed on the use of genetic algorithms (GA), given that they offer a parallel search of solutions, potentially avoiding local minima. Moreover, with a correct design of a fitness function the GA inherently guides the different generations of individuals to a good if not optimal solution.

In this way, and as a product of the review made over the related work in the area, we found some inspiring papers. In ref. [22], several functions are tested in a Genetic Functional Algorithm to determine which express better a structure-property relationship. The work in GA of Leardi *et al.* [23] shown that descriptor selection by stepwise regression is overcome when it is compared to multiple linear regression selected by a GA.

In [24] feature selection is applied to establish which was the subset of the three most influential descriptor that best predict drug activity. They used a genetic algorithm, where a neural network was applied as fitness function. However, this has the drawback of the great amount of time required by the neural network for training. Similar approaches were applied on refs [25] and [26], where supervised self-organizing maps (sSOMs) and principal component analysis (PCA) respectively are coupled with a genetic algorithm.

According to the recent classification, our proposed feature selection method described in next section, belongs to a wrapper method because decision trees are used in the fitness function as the black-box method for assessing its prediction capability.

3. Genetic Algorithm for Descriptor Selection

In order to make the selection of most influential descriptors we implemented a genetic algorithm denominated DS-GA. The main objective of DS-GA is to find a selection of descriptors that results suitable to describe physicochemical behavior.

3.1 GA Design: Chromosome Representation

Binary strings are used to represent the individuals. Each string of length m standing for a feasible descriptor selection, where m is the number of considered descriptors. A nonzero value in *i*th bit position means that the *i*th descriptor is selected (Figure 1).

For this work, we have constrained to a model where only p bits could be set active for each individual at the same time. In other words, the purpose of the DS-GA is to find the p most relevant descriptors for physicochemical prediction.

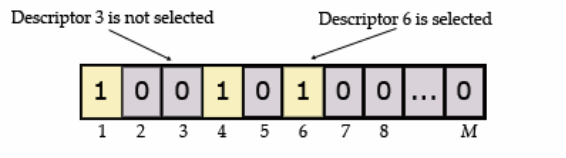

Figure 1. Binary String Representation

3.2 GA Design: Main Characteristics

The initial population is randomly generated imposing the described restriction of exactly p active descriptors on each individual. A one-point crossover is used for the recombination [27]. Non feasible individual could take place after crossover, because the number of nonzero bits may be different than p. This problem is solved by randomly setting or resetting bit locations as needed to be up to p active bits. Since the crossover scheme inherently incorporates bit-flip mutation, we abstained to use an additional scheme of mutation.

In order to decide the best selection method for the DS-GA, we did different experiments with typical selection methods and we concluded than tournament method is appropriate. Furthermore, this method is preferred than others because is particularly easy to implement and its time complexity is *O(n)* [27] [28].

3.3 GA Design: Fitness Function

Taking into account that the DS-GA objective is to determine the most relevant set of p descriptors for predicting a physicochemical property, the fitness function should estimate the accuracy of a prediction method when only the p descriptors are used. In particular, the fitness function employed by DS-GA is presented in equation 1. This formula computes the mean square error of prediction (MSE).

$$\mathrm{F}(\mathcal{P}_{Z_{1,k}}, Z_{2,k}) = \frac{1}{n_2} \sum_{(\vec{x}_i, y_i) \in Z_{2,k}} (y_i - \mathcal{P}_{Z_{1,k}}(\vec{x}_i))^2 \quad (1)$$

Where:

- **Z** is a matrix that represents a compound dataset, where each row and column corresponds to a compound and a descriptor respectively. Last column of **Z** stores the experimental target values for each compound. This column vector is denoted as **y**.
- $\mathcal{P}_Z$ is a predictor method trained with the dataset **Z**.
- $\mathbf{Z_1}$ and $\mathbf{Z_2}$ are compound databases, that are used as learning and validation sets respectively, with corresponding sizes $\mathbf{n_1} \times \mathbf{m}$ and $\mathbf{n_2} \times \mathbf{m}$.
- $\mathbf{Z_{i,k}}$ is a filtered dataset according to the descriptor selection encoded in the *k*th individual. In other words, $\mathbf{Z_{i,k}}$ only contains those variables of $\mathbf{Z_i}$ whose values in the corresponding locations of the *k*th individual's chromosome is 1.
- $\vec{\mathbf{x}}_\mathbf{i}$ is a vector that represents the values of the descriptors for the *i*th compound of a given dataset.
- $\mathbf{y_i}$ is the target value for the *i*th compound of **Z**.

The first argument of the fitness function is the predictor method applied to a given learning set, $\mathcal{P}_{\mathbf{Z}}$. In this work, three different predictor techniques were tested. The first one uses decision trees for evaluating the predictive capacity of the selected descriptors encoded for a given individual. Specifically, the decision trees are used here as regression trees [28] without using any kind of pruning.

A linear and a non-linear regression models were also applied in this paper as first argument of the fitness function. In both cases, a mathematical expression is used to regress the descriptors against the target value. These models incorporate regression coefficients in order to obtain more accurate results. For example, the corresponding linear and non-linear regression model formula to the individual depicted in Figure 1 (only 3 descriptors are selected) is presented in equation 2 and 3 respectively.

$$\beta_1 X_1 + \beta_2 X_4 + \beta_3 X_6 + \beta_0 \quad (2)$$

$$\sum_{j=1}^{4} \beta_{1,j} X_1^{\,j} + \sum_{j=1}^{4} \beta_{4,j} X_4^{\,j} + \sum_{j=1}^{4} \beta_{6,j} X_6^{\,j} + \beta_0 \quad (3)$$

These coefficients (β_i and $\beta_{i,j}$) are adjusted with nonlinear least-squares data fitting by the Gauss-Newton method [29].

4. Experimentation and Results

Our proposal consists of looking for a fixed-size set of descriptors that minimizes the prediction error when they are used as input of a predictor method. Specifically, we propose to find the ten most useful descriptors for logP prediction from the constitutional descriptor family.

In order to measure the performance of the aforementioned objective, we used the output of the DS-GA as input of a neural network ensemble (NNE). This NNE was specially designed for logP prediction. Our goal is to establish whether using the descriptor set obtained by DS-GA involves a significant improvement in the prediction accuracy in relation to other selection criteria. We decided to work with a total set of the first 1200 compounds (CAS-ordered) from the PHYSPROP database [30], 50% of them were used for training, 16% for validation (Set 1) and the remainder (34%) was left for testing divided in two sets (Set 2 and Set 3). The reason of having two different hold-out sets is due the different frequency distributions of logP values in Sets 2 and 3.

For the DS-GA runs we use typical parameter values: population size=45; cross-over probability=0.8; tournament size=3. A phenotypic stopping criterion is used; the DS-GA stops when the highest fitness of the population does not improve during ten generations. With respect to the NNE, each ensemble consists of five NNs, and each one has a three-level architecture with five hidden nodes.

At first, we were interested in testing performance of the DS-GA using decision trees as fitness function. This testing was carried out by comparing the latter fitness function with a random descriptor selection and also with the variants of fitness functions previously described in Section 3. Parameters of fitness functions were established according to 15 predictions done over validation set (Set 1). Once the parameters were established new 15 runs of the DS-GA were carried out. In each selection executed by the DS-GA, 7 runs of the NNE were produced to have enough replicas of the predictions done over validation and test sets. In this way, for each fitness function (and also for the random selection) an average prediction performance could be established, based on 105 (15×7) replicas; and the performance of the best selection of a given fitness function based on the 7 replicas of the NNE.

	All [a]	Random [b]		Dec. Trees [c]		Linear [d]		Non linear [e]	
	--	Best	Avg	Best	Avg	Best	Avg	Best	Avg
Set 1	1.4430	1.3617	1.5166	1.2118	1.3791	1.3855	1.4066	1.1557	1.2884
Set 2	1.3037	1.2860	1.4318	1.1249	1.3398	1.2851	1.3599	1.2849	1.3213
Set 3	1.0414	1.0787	1.1571	1.0603	1.1227	1.0331	1.109	1.0255	1.1256

Table 1. Mean absolute errors over validation and test sets

Table 1 shows prediction errors obtained for the different selection ways employed for analysis. 'All'[(a)] employs the 47 descriptors from the constitutional family. 'Random' takes 10 descriptors arbitrarily from the same family. Finally, the other selection approaches are: Decision Trees[(c)], Linear[(d)] and Non linear[(e)] regression. The last three also take 10 descriptors, using these methods as fitness functions for the DS-GA. As it is observed, Decision Trees have a satisfactory performance among all different methods, however ANOVA tests and multiple comparisons are needed in order to establish whether the encountered differences are trustful and whether they are not product of variance of the methods.

An ANOVA test is commonly used in order to evidence whether differences in prediction results of each method come up from an intrinsic variation of the methods or from the differences among the methods. So, the one-way ANOVA test breaks up data variation in two sources of variance (S. of v.): *among* variance and *within* variance. A *F* statistic is calculated from the ratio of the mean squares (M.S.) which in turn is calculated from the ratio of the sum of squares (S.S.) and the degrees of freedom (d.f.). Based on the *F* statistic and the d.f.'s a p-value is calculated to determine whether significant differences exist between methods.

4.1 Comparisons of Best Selections of the Different Methods

Two ANOVA tests were carried out, one for each test set (Set 2 and Set 3). The reason of doing these tests separately is that, in other way, normality and homocedasticty of the sets are compromised. Table 2 shows one ANOVA table for each holdout set.

S. of v.	S.S.	d.f.	M.S.	F	p-value
Among	*0,154148444*	*4*	*0,038537111*	*35,9410*	***0,0001***
Within	*0,032166989*	*30*	0,001072233		
Total	*0,186315433*	*34*			

(a)

S. of v.	S.S.	d.f.	M.S.	F	p-value
Among	*0,013058371*	*4*	*0,003264593*	*1,3541*	***0,2732***
Within	*0,072325361*	*30*	0,002410845		
Total	*0,085383732*	*34*			

(b)

Table 2. ANOVA tables for the best selection of each method in: (a) Set 2 (b) Set 3.

As the ANOVA tables show, methods applied over Set 2 behave different whereas predictions of Set 3 are clearly more at the same level. Nevertheless, we decided to run a multiple comparison criteria to identify whether and where differences with respect to the "Dec. Trees" method exist. We decided to apply the Dunnet comparison criterion given that this criterion allows to globally compare all methods against a control method, "Dec. Trees" in this case (Table 3).

"Dec. Trees" outperforms all other methods with a global error confidence of less than 1% in (a), but no significant differences were globally detected in (b) (for 4 comparisons using Dunnet test are (single tailored) 2,25 and 2,97 at 5% and 1% global error, respectively.). Instead, in Set 3 just an individual difference (LSD) could be established between "Non linear" and Random selection at 5%.

Methods	Dec. Trees	Non linear	Linear	Random	All
Statistics	----	9,1373	9,1476	9,1988	10,2126
Results	---	**	**	**	**

(a)

Methods	Non linear	Linear	All	Dec. Trees	Random
Statistics	1.9865	1.5505	1.0759	----	1.0550
Results	Ns	ns	ns	---	ns

(b)

Table 3. Dunnet comparisons of all selection methods against Decision Trees fitness function for (a) Set 2 and (b) Set 3.

4.2 Comparisons of Average Selections Obtained from the Different Methods

Each run of the DS-GA with the same configuration of parameters and same fitness function, may give place to very different selections of descriptors. This implies that, in statistical terms, we have a non controlled source of variance given by each run. For this reason, we carried out a nested ANOVA test, where the method was the principal factor (fixed), and each selection of the method was the nested factor. Table 4 shows this latter test for both test sets. From this table, we can conclude that the nested factor incorporates a variance component ($p < 0.001$). It can be computed that nested factor represents 77% of variance in (a), and 63% in (b). Given latter table, we can calculate the simple ANOVA, averaging all nested replicas (Table 5).

S. of v.	S.S.	d.f.	M.S.	F	p-value
Principal F.	*0,737076291*	*3*	*0,245692097*	*5,8253*	***0,0015***
Nested F.	*2,361903692*	*56*	0,042176851	*30,65412*	***0,0001***
Within	*0,495322256*	*360*	0,001375895		
Total	*3,594302239*	*419*			

(a)

S. of v.	S.S.	d.f.	M.S.	F	p-value
Principal F.	*0,16849875*	*3*	*0,05616625*	*2,3485*	***0,0823***
Nested F.	*1,33926114*	*56*	0,023915377	*18,38*	***0,0001***
Within	*0,468360687*	*360*	0,001301001		
Total	*1,976120577*	*419*			

(b)

Table 4. Nested ANOVA test for (a) Set 2, and (b) Set 3

These analysis show that strong confidence exists, to determine significant differences among methods ($p = 0.0015$ and $p = 0.0823$). So, we calculate the multiple comparisons to determine where and how significant are the differences (Table 6). For Set 2 no globally significant differences were encountered among DS-GA methods, but globally differences were encountered among "Random" and "Dec. Trees"

(critical values for 4 comparisons using Dunnet test (single tailored) are 2,1 and 2,75 at 5% and 1% global error respectively), and also between Random and any DS-GA (using Bonferroni test, not showed). Again, Set 3 did not allow that global differences appeared among methods, but LSD individual differences are encountered when "Linear" and "Dec. Trees" are compared against "Random".

S. of v.	S.S.	d.f.	M.S.	F	p-value
Among	*0,105296613*	*3*	*0,035098871*	*5,8253*	***0,0015***
Within	*0,337414806*	*56*	0,006025264		
Total	*0,442711419*	*59*			

(a)

S. of v.	S.S.	d.f.	M.S.	F	p-value
Among	*0,024071245*	*3*	*0,008023748*	*2,3485*	***0,0823***
Within	*0,19132302*	*56*	0,003416482		
Total	*0,215394264*	*59*			

(b)

Table 5. Simple ANOVA tables from the average of all selections of each method for: (a) Set 2; (b) Set 3.

Methods	Non linear	Dec. Trees	Linear	Random
Statistics	0,6530	----	0,7074	3,2455
Results	ns	----	ns	**

(a)

Methods	Linear	Dec. Trees	Non linear	Random
Statistics	0,7688	----	0,1028	1,2127
Results	ns	----	ns	Ns

(b)

Table 6. Dunnet comparisons of all selection methods against Decision Trees fitness function for (a) Set 2 and (b) Set 3.

4.3 Discussion of results

The preceding results support our hypothesis of to acquire an improving of the prediction results by training only those features that the DS-GA considers as important. The comparisons were divided in two: average results of the 15 different selections of the methods (Section 4.1) and the best selection of each different method (Section 4.2). Also we incorporated the results of the predictions when random selection of descriptors were applied and when all descriptors were used for prediction. It is worth mentioning that when all descriptors were applied, only a single selection exists and thus, no average result could be obtained.

Results obtained for Set 2 and Set 3 are quite different in the sense that in Set 3 no significant differences were obtained with the use of DS-GA whereas in Set 2, DS-GA always outperforms "Random" and "All" (when applicable) selection criteria. Moreover, decision trees fitness function is not overcome, at least in our experiments, by no other applied selection method.

Set 3 prediction errors were deliberately kept in this paper, in spite of the obtained results, since it is a possible scenario in logP prediction. Our opinion about the lack of improving with the descriptor selection algorithm is that the constitutional descriptor family is not sufficient to describe the logP behavior. Notwithstanding, it is important to note that DS-GA (with or without decision trees) do not harm the prediction results in the case of missing influential descriptors.

5. Conclusions and Future Work

The present work proposes a novel variant of a methodology for improving the prediction errors in structure-property relationships. This approach allows to detect which descriptors are the most influential to the prediction of the molecule hydrophobicity. We also establish a statistical comparative analysis, among other ways of selecting descriptor with a genetic algorithm.

It is clear that our proposal is not restricted to logP, because this method could also be applied to any physicochemical property. In addition, to the reduction of the prediction errors, this strategy allows to know which are the descriptors that best encode a specific property.

The combination of several machine learning methods often outperforms the capacity of single individually-applied classifiers [31]. The key contributions of our work are the proposal and the comparison of the use of decision trees in the fitness function. This feature allows a fast evaluation of whether the descriptors possessed by an individual are able to obtain good prediction capacity.

As future work, it would be interesting to experiment this proposal with the aggregation of other descriptor families. In this context, we are evaluating other descriptors that express interactions between functional groups in molecules. Moreover, DS-GA could also be developed to detect the most adequate number of descriptors to be taken into account for a predictor method, instead of fixing to a specific number. At this moment, we are also considering to use other AI methods as feature selection techniques.

Acknowledgements

Authors acknowledge the "Agencia Nacional de Promoción Científica y Tecnológica" from Argentina, for Grant PICTO-UNS Nº 917. It was awarded to research projects as part of the "Programa de Modernización Tecnológica, Contrato de Préstamo BID 1728/OC-AR". They would also like to aknowledge SeCyT (UNS) for Grant PGI 24/N019.

References

[1] H.E. Selick, A.P. Beresford, M.H. Tarbit. The Emerging Importance of Predictive ADME Simulation in Drug Discovery. *Drug Discovery Today* 7, 2:109-116, 2002.

[2] J.A DiMasi, R.W. Hansen, H.G. Grabowski. The Price of Innovation: New Estimates of Drug Development Costs. *Journal of Health Economics*, 22:151-185, 2003.

[3] J.A. DiMasi, H.G. Grabowski, J. Vernon. R&D Costs and Returns by Therapeutic Category. *Drug Information Journal*, 38, 3: 211-223, 2004.

[4] J. Taskinen, J. Yliruusi. Prediction of Physicochemical Properties Based on Neural Network Modeling. *Advanced Drug Delivery Reviews*, 55, 9: 1163-1183, 2003.

[5] S.Ó. Jónsdottir, F.S. Jørgensen, S. Brunak. Prediction Methods and Databases Within Chemoinformatics: Emphasis on Drugs and Drug Candidates. *Bioinformatics*. 21:2145-2160, 2005.

[6] I.V. Tetko, P. Bruneau, H.W. Mewes, D.C. Rohrer, G.I Poda. Can we estimate the accuracy of ADME-Tox predictions? *Drug Discovery Today*, 11:700-707, 2006.

[7] J.J. Huuskonnen, D.J. Livingstone, I.V. Tetko.: Neural Network Modeling for Estimation of Partition Coefficient Based on Atom-Type Electrotopological State Indices. *Journal of Chemical Information and Computer Science*, 40:947-995, 2000.

[8] S. Agatonovic-Kustrin, R.J. Beresford. Basic Concepts of Artificial Neural Network (ANN) Modeling and its Application in Pharmaceutical Research. *Journal of Pharmaceutical and Biomedical Analysis*, 22, 5: 717-727, 2000.

[9] R. Todeschini, V. Consonni. Handbook of Molecular Descriptors. *Wiley-VCH, Weinheim Germany*, 2000.

[10] D.J. Livingston. The Characterization of Chemical Structures Using Molecular Properties. A Survey. *Journal of Chemical Information and Computer Science*, 40:195-209, 2000.

[11] I.V. Tetko, D.J. Livingstone, A.I. Luik. Neural Networks Studies. 1. Comparison of Overfitting and Overtraining. *Journal of Chemical Information and Computer Science*, 35: 826-833, 1995.

[12] J.G. Topliss, R.P. Edwards. Chance Factors in Studies of Quantitative Structure-Activity Relationships. *Journal of Medicinal Chemistry* 22, 10:1238-1244, 1979.

[13] T.S. Furey, N. Cristianini, N. Duffy, D.W. Bednarski, M. Schummer, D. Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16:906-914, 2000.

[14] I. Guyon, J. Weston, S. Barnhill, V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389-422, 2002.

[15] R. Bekkerman, R. El-Yaniv, N. Tishby, Y. Winter. Distributional word clusters vs. words for text categorization. *Journal of Machine Learning Research*, 3: 1183-1208, 2003.

[16] G. Forman. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 3:1289-1306, 2003.

[17] R. Kohavi, G. John. Wrappers for feature selection. *Artificial Intelligence*, 97:273-324, 1997.

[18] A. Blum, P. Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97:245-271, 1997.

[19] I. Guyon, A. Elisseeff. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*. 3:1157-1182, 2003.

[20] J. Weston, A. Elisseeff, B. Schoelkopf, M. Tipping. Use of the zero norm with linear models and kernel methods. *Journal of Machine Learning Research*, 3:1439-1461, 2003.

[21] W. Lindberg, J. Persson, S. Wold. Partial Least-Squares Method for Spectrofluorimetric Analysis of Mixtures of Humic Acid and Ligninsulfonate. *Analitical Chemistry*, 55:643-648, 1983.

[22] D.R. Rogers, A.J. Hopfinger. Application of Genetic Function Approximation to Quantitative Structure-Activity Relationships and Quantitative Structure-Property Relationships. *Journal of Chemical Information and Computer Science*, 34:854-866, 1994.

[23] R. Leardi, R. Boggia, M. Terrile. Genetic Algorithms as a Strategy for Feature selection. *Journal of Chemometrics*, 6:267-281, 1992.

[24] S-S. So, M. Karplus M. Evolutionary Optimization in Quantitative Structure-Activity Relationship: An Application of Genetic Neural Networks. *Journal of Medicinal Chemistry*, 39:1521-1530, 1996.

[25] E. Bayram, P. Santago, R. Harrisb, Y. Xiaob, A.J. Clausetc, J.D. Schmittb. Genetic algorithms and self-organizing maps: a powerful combination for modeling complex QSAR and QSPR problems. *Journal of Computer Aided Molecular Design*, 18:483-493, 2004.

[26] B.K. Lavine, C.E. Davidson, A.J. Moores. Innovative genetic algorithms for chemoinformatics. *Chemometrics and Intelligent Laboratory Systems*, 60:161– 171, 2002.

[27] D.E. Goldberg, K. Deb. A comparative analysis of selection schemes used in genetic algorithms, in Foundations of Genetic Algorithms. *San Mateo, CA: Morgan Kaufmann*, 69-93, 1991.

[28] L. Breiman. Classification and Regression Trees, *Chapman & Hall, Boca Raton*, 1993.

[29] K. Madsen, H.B. Nielsen, O. Tingleff: Methods for Non-Linear Least Squares Problems. *Technical University of Denmark. Informatics and Mathematical Modeling. 2nd Edition*, 2004.

[30] The Physical Properties Database (PHYSPROP) is marketed by Syracuse Research Corporation (SRC), North Syracuse, USA at URL http://www.syrres.com/esc/

[31] D. Wolpert. Stacked Generalization. *Neural Networks*. 5:241-260, 1992.