SPECIAL GUEST EDITOR SECTION

Feature Selection Methods in QSAR Studies

MOHAMMAD GOODARZI, BIEKE DEJAEGHER, and YVAN VANDER HEYDEN[1]
Vrije Universiteit Brussel (VUB), Department of Analytical Chemistry and Pharmaceutical Technology, Center for Pharmaceutical Research (CePhaR), Laarbeeklaan 103, B-1090 Brussels, Belgium

A quantitative structure-activity relationship (QSAR) relates quantitative chemical structure attributes (molecular descriptors) to a biological activity. QSAR studies have now become attractive in drug discovery and development because their application can save substantial time and human resources. Several parameters are important in the prediction ability of a QSAR model. On the one hand, different statistical methods may be applied to check the linear or nonlinear behavior of a data set. On the other hand, feature selection techniques are applied to decrease the model complexity, to decrease the overfitting/overtraining risk, and to select the most important descriptors from the often more than 1000 calculated. The selected descriptors are then linked to a biological activity of the corresponding compound by means of a mathematical model. Different modeling techniques can be applied, some of which explicitly require a feature selection. A QSAR model can be useful in the design of new compounds with improved potency in the class under study. Only molecules with a predicted interesting activity will be synthesized. In the feature selection problem, a learning algorithm is faced with the problem of selecting a relevant subset of features upon which to focus attention, while ignoring the rest. Up to now, many feature selection techniques, such as genetic algorithms, forward selection, backward elimination, stepwise regression, and simulated annealing have been used extensively. Swarm intelligence optimizations, such as ant colony optimization and partial swarm optimization, which are feature selection techniques usually simulated based on animal and insect life behavior to find the shortest path between a food source and their nests, recently are also involved in QSAR studies. This review paper provides an overview of different feature selection techniques applied in QSAR modeling.

The quantitative structure-activity relationship (QSAR) approach was first applied in practice around 50 years ago (1, 2). QSAR models describe a relationship between the chemical structure of molecules, described by molecular descriptors (e.g., geometric, steric, and electronic properties) and their corresponding biological activity. QSAR models are used to predict the activity of chemical compounds from their structural properties. Because of the wide use of QSARs for designing drugs, the International Union of Pure and Applied Chemistry defines them as follows: *"Quantitative Structure–Activity Relationships (QSAR) are mathematical relationships linking chemical structure and pharmacological activity in a quantitative manner for a series of compounds. Methods which can be used in QSAR include various regression and pattern recognition techniques"* (3).

Since the introduction of QSAR, many different studies have been made, not only based on so-called two-dimensional (2D; 4) but also on three-dimensional (3D; 5–7) techniques. The major differences between both are the structural parameters that can be used to characterize molecular identities as well as the mathematical procedure used to describe the relationship between descriptors and biological activity (8).

One of the most popular 3D QSAR methods is comparative molecular field analysis (CoMFA), which can be built based on steric and electrostatic field descriptors between the ligand and biological receptor (7). CoMFA and other 3D-QSAR methods have several shortcomings, e.g., in many cases, it is impossible to precisely define a pharmacophore model, and if a nonoptimal alignment of ligands is applied, it may introduce errors in the QSAR model (8). The 3D QSAR may be too computationally expensive to analyze large data sets. For example, alignment of the ligands takes a lot of time, conformational search must be done to find the best conformers, and they affect the final results very much. Sometimes an automated and unambiguous alignment of compounds is not achievable.

The above problems do not occur in 2D QSAR, where we use only zero-dimensional, one-dimensional , 2D, and 3D descriptors. 3D descriptors for 2D QSAR can be calculated only from structures optimized by molecular mechanics and/or quantum chemical calculations. For example, 3D descriptors are: Randic molecular profiles, geometrical descriptors, radial distribution function, 3D-molecule representation of structure based on electron diffraction, weighted holistic invariant molecular and geometry, topology, and atom-weights assembly descriptors. Further, any kind of surfaces (e.g., polar surface area) or volumes (e.g., molecular volume) can be considered as descriptors, as of course, can the quantum chemical descriptors (9).

In spite of the fact that graph theory indexes (i.e., structural formulas of compounds; 10–12) stand for different aspects of molecular structure, their physicochemical meaning is not obvious. Accordingly, 3D descriptors have been developed to address this problem of 2D QSAR techniques. Progressively, with time, the number of descriptors increases; finding reliable descriptors that can be linked to the biological activity in a QSAR model becomes a serious challenge.

Several items can affect the predictive ability of a QSAR model, such as the optimization of the molecular structures

Guest edited as a special report on "Chemometrics in Pharmaceutical Analysis" by Łukasz Komsta.

[1] Corresponding author's e-mail: yvanvdh@vub.ac.be

DOI: 10.5740/jaoacint.SGE_Goodarzi

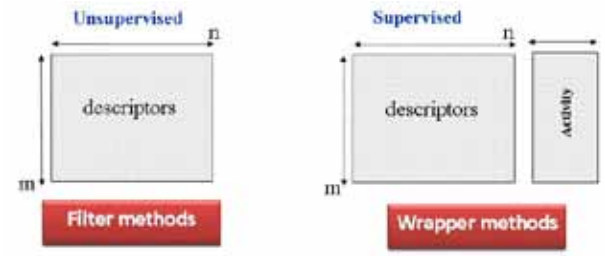

Figure 1. Filter methods: Feature selection based only on the descriptors without contribution of any learning algorithm. Wrapper methods: Selection based on the descriptors and activity using a learning algorithm. m = Numbers of objects (molecules) and *n* = number of variables (descriptors).

and the modeling technique used to make a relation between the descriptors and the biological activity. Several linear and nonlinear techniques have been applied in QSAR studies, such as multiple linear regression (MLR), partial least squares (PLS; 13, 14), nonlinear PLS (15, 16), artificial neural networks (ANN; 17–20), support vector machines (SVM; 21–24), and multivariate adaptive regression splines (MARS; 25, 26). One of the most important tasks, prior to modeling, is the selection of relevant descriptors with maximum information about the compounds and with a minimum collinearity (27).

There are several reasons why feature or variable selection is important. Models with fewer variables are easier to interpret, provide improved performance for new samples, and decrease the risk of overfitting/overtraining. Filter, wrapper, and hybrid methods are the three major categories of feature selection techniques (Figure 1). In fact, a method that reduces the pool of descriptors into a smaller set based on a specified criterion (which is typically based on information content or intervariable correlations) is called a filter feature selection method. Filter methods do not apply any learning machine in the process, and they perform an unsupervised feature selection (28–30).

On the other hand, a linear or nonlinear classifier (or regressor) uses an objective function based on an optimization criterion to select descriptors. These methods are classified into the wrapper techniques (Figure 1; 31–33). Although the wrapper approaches are computationally more expensive than filter methods, their generalization performance is better (33).

Hybrid methods attempt to take advantage of the two approaches by exploiting the different evaluation criteria in different search stages (34–36). Most hybrid approaches are classified as wrapper methods, because there is not much difference between them, and before a wrapper method is applied, a filter method is used to reduce the number of variables. Therefore, this review only discusses the filter and wrapper approaches. The main goal of this review is to provide an overview of recently applied feature selection methods in QSAR studies (Table 1).

Filter Methods

Filter methods are used in the first step after calculating the descriptors to reduce the dimensionality of the dataset. When descriptors are removed from the pool based on mutual correlation, first two descriptors that are highly correlated are indicated from the pool, then one is removed randomly. An alternative approach is to retain the one with the highest correlation to the dependent variable and to remove the other from the data set. The second method seems to be better than the first. Another approach is to remove the descriptors with the lowest variance and the lowest correlation to the dependent variable, and keep those with the highest correlation.

In general, filter methods can be divided into several types, such as distance methods (e.g., using the Euclidean distance measure and Mantaras distance measure; 37, 38); information methods (e.g., entropy, information gain, gain ratio, and normalized gain; 39–41); dependency methods (e.g., correlation coefficient; 42, 43); and consistency methods (e.g., min-features bias; 44). There are still many other approaches that use, for instance, mutual information (45), the Chi-square (χ^2) metric (46), the Kolmogorov-Smirnov statistic (47), the unbalanced correlation score (46), and the Shannon entropy (48), to select features.

Weston et al. (46) performed research on 1909 (training set) organic compounds to evaluate whether they bind to thrombin (a protein involved in blood clotting). Only 42 of the compounds showed a positive result, i.e., interacted with thrombin. Each compound was described by a single response vector comprising a class value ("A" for active, "I" for inactive) and 139 351 features (variables are either 0 or 1), which describe the 3D properties of the compound. The authors referred to examples that bind as having label +1 (and, hence, being called positive examples). Conversely, negative examples (that do not bind) are labeled –1. The test set included 634 compounds, of which 150 were active. The unbalanced correlation score (UCS) method (46) was applied for feature selection. This method ranks the features according to a criterion. The authors set up the criterion to assign a rank to a subset of features rather than just a single value. A feature subset with a high score could thus be chosen for modeling. The authors also compared USC with the Fisher score as a feature selection. The Fisher score is a standard (univariate) correlation score (46). Note that in this score, negative correlations are just as important as positive ones. Complex feature selection methods, such as UCS, may show overfitting, while simple standard methods, such as the Fisher score, do not always perform well. Constructing a criterion that does not overfit but also takes into account the unbalanced nature of the data when selecting features will improve performance.

Liu (49) compared five methods for feature selection, including information gain, mutual information (44), a χ^2-test (45), the odd ratio (49), and the Galavotti-Sebastiani-Simi coefficient (49). Naive Bayesian classifier (50) and SVMs (51) are used to classify the chemical compounds. The data set included the 1909 (training set) and the 634 (test set) organic compounds mentioned above (46).

However, feature selection did not improve the SVM results, and it performed better using all features. Naïve Bayesian classification accuracy results extensively improved after feature selection, especially when the features were selected based on the information gain and χ^2-test methods. Using information gain with a naıve Bayesian classifier, removal of up to 96% of the features yielded an improved classification accuracy measured by sensitivity. The mutual information

approach was found to have a poor performance because of its bias favoring rare features (49, 52).

Whitley et al. (53) proposed an Unsupervised Forward Selection (UFS) approach as feature selection with the aim of eliminating redundancy and reducing multicollinearity. The UFS starts with the two variables that are least correlated and selects additional variables on the basis of their multiple correlation with those already chosen, thus building a subset of variables that is as close to orthogonality as possible. After selecting the descriptors with UFS, continuum regression (54), an algorithm encompassing ordinary least squares regression, principal component regression (PCR) and PLS, was performed. The UFS depends only on the independent variables, and the response variable (dependent variable) is not involved in the selection process. The authors performed UFS on three different datasets, i.e., one with 21 steroid compounds, one with 19 pyrethroid insecticides, and the Selwood dataset with 31 antifilarial antimycin analogs (53). Though performed on three small sets, the authors mentioned that those feature selection methods lead to models with a small number of components (often only one) of a focused set of variables. The obtained models are far easier to interpret than models with several latent variables constructed from a large number of descriptors.

Demel et al. (55) compared several unsupervised feature selection methods, i.e., McCabe coefficient (56), information gain (52), relief (57), correlation-based feature selection (55), and K-nearest neighbor (K-NN; 58), which is a wrapper method. All methods were applied on three data sets. These sets contained ATP-binding-cassette (ABC) transporters, such as ABCB1 (P-glycoprotein), ABCC1 (MRP1), and ABCG2 (P-glycoprotein) substrates. In this study, the authors assigned compounds with correlation coefficients (r) between toxicity and transporter expression lower than –0.3 to be substrates and those with $-0.02 < r < 0.02$ to be nonsubstrates. This way a set of 240 compounds [i.e., 110 (45.8%) substrates and 130 (54.2%) nonsubstrates] was obtained for ABCB1, a set of 227 compounds [i.e., 124 (54.6%) substrates and 103 (45.4%) nonsubstrates] for ABCC, and a set of 198 compounds [i.e., 94 (47.5%) substrates and 104 (52.5%) nonsubstrates] for ABCG2 were selected. The results showed that the wrapper method K-NN outperforms the other feature selection methods. The obtained results indicate that possibly nonlinear behaviors exist in the data sets.

Salt et al. (59) used both factor analysis (60) and UFS as feature selection prior to predicting the affinity and selectivity of 108 arylpiperazinyl derivatives for both α1 and α2 adrenoceptors (ARs) with a MLR method. For FA, used as a wrapper method in this case because **y** (biological activity) is also involved during the feature selection process, the factors showing a significant ($P < 0.05$) loading for the response variable were selected, and representative descriptors, i.e., those with maximum loading on each factor, were retained. UFS, on the other hand, is based on the correlation matrix between the independent variables, starts by identifying the two least correlated variables, and adds other variables on the basis of their multiple correlation coefficients with those that already have been chosen. The main goal to use UFS in this study was to remove the redundancy from the data matrix (whose initial number of variables greatly exceeded the rank of the matrix).

Principal component analysis (PCA) also can be used for feature selection. PCA defines new latent variables of which the first contains most of the variance of the data. It is a variable reduction technique that allows visualizing the information included in the X matrix (including descriptor values). Usually, based on plotting the PC1 versus PC2 loadings (loading plot), much information from the data set variables can be extracted, because these two PCs contain the highest variances in the data set (61). Some descriptors are clustered in the loading plots, which means that they describe similar information. For the subsequent calculations, the number of descriptors was reduced by choosing a representative one from each cluster of variables and removing those with similar information (62).

Roy et al. (63–69) have been working intensively on feature selection. They compared genetic function approximation (GFA; 70) and PCA as feature selection methods, and MLR as a regression technique on several classes of compounds. For instance, PCA was used for the selection of descriptors for the multiple regression analysis of the acute toxicity of 56 phenylsulfonyl carboxylates to *Vibrio fischeri,* formerly known as *Photobacterium phosphoreum,* which is a Gram-negative rod-shaped bacterium found globally in marine environments. In a second case study, the toxicity of a set of 42 nitroaromatic compounds and in a third case the fish toxicity of 92 substituted benzenes were modeled. Variable selection using GFA was also done prior to modeling the acute toxicity of the 56 phenylsulfonyl carboxylates to *V. fischeri*.

In their studies (63–69), PCA was used for the selection of independent variables. For the selection of variables contributing to the biological activity **y** using (principal component) factor analysis, one has to take the **y** vector also along with the descriptors. Only variables with nonzero loadings in PCs, where biological activity also has nonzero loading, were considered important in explaining the activity.

Bajorath's group (71–73) introduced mutual information differential Shannon entropy (MI-DSE), an improvement of Shannon entropy (SE), as a feature selection approach. The SE method provides a basis for the quantification of the information content of data distributions that can be represented as a histogram. The DSE method was introduced to quantify how much information about a given compound class is contained in the value distribution of a descriptor when compared to another.

Based on DSE, at first a histogram (class specific entropies) is drawn for each class (e.g., classes A and B). Then the two classes A and B are combined into a single histogram. The SE should be calculated for the combined histogram, then the DSE can be obtained as the combined histogram AB minus the average of the individual histograms A and B. The authors found that there was a problem when the two classes are of significantly different size. The combined histogram is too influenced by the larger class, and its distribution is biased. The utility of the MI-DSE approach to identify class-specific information was confirmed by the comparison of the Gaussian distributions of the molecular descriptors and resulting DSE and MI-DSE values and descriptor rankings. The methods are applied on different activities in data sets containing between 30 and 159 compounds, which were selected from the 100 000 ZINC compounds database (a free database of commercially available compounds for virtual screening; 71).

Wrapper Methods

Wrapper methods use the information from both independent

and dependent variables for feature selection. Several methods can be classified in this category. In fact, the number of feature selection methods in this category has been increasing rapidly.

Forward selection and backward elimination, and stepwise regression are three popular and simple feature selection methods. In forward selection, the first selected variable is the one with the highest correlation to the dependent variable. Then the method consecutively adds variables to the model one at a time. This process is terminated when the last variable entering the model has an insignificant regression coefficient or when all variables are included. In contrast, backward elimination starts with all variables in the model and eliminates them one at a time, in which the first eliminated variable is the one with the least significance. This process is terminated when all remaining variables are significant or all but one variable have been deleted. Stepwise regression uses both forward selection and backward elimination. A variable that entered the model in an earlier stage of selection may be deleted at a later stage. Several criteria can be used, but the above approaches are mostly based on F (statistic) value (74).

Xu and Zhang (75) compared several feature selection methods, i.e., forward selection, backward elimination, and stepwise regression, on a data set consisting of 35 nitrobenzenes with corresponding toxic activities. In this study, backward elimination performed better than forward selection and stepwise regression for the selection of molecular descriptors. The authors evaluated them based on the correlation coefficient, and it was shown that backward elimination was better than forward selection and stepwise regression.

Prabhakar (76) introduced a combinatorial protocol interfaced with MLR for feature selection. It was first applied on the Selwood data set (77), and the results were quite acceptable when compared to those obtained by GFA (70) and mutation and selection uncover models (78). If a group of variables is a bundle, then, according to the combination rule, a total of ${}_pC_k$ bundles emerge from p variables with k variables in each bundle (original variable bundle; OVB). A variable may contribute to a model in two different ways: by itself alone, and/or by itself and its functionally transformed term together. To find the influence of a selected function of any variable along with its original form in the model development, the k variables of OVB along with their meaningfully transformed functional variables are adopted for the formation of new bundles. For efficient evaluation of variable bundles, they have to pass four filters. Three of them, interparameter correlation cutoff criteria for variables to stay as a bundle (controls the correlation between independent variables; default value less than or equal to 0.3); a t-value of the variables regression coefficient (evaluates the significance of variables in a bundle); and the square root of adjusted multiple correlation r (to compare the internal explanatory power of bundles with different numbers of variables), are initially adopted in this process. While the first three criteria were used to check the internal consistency, the fourth, i.e., the squared correlation coefficient of cross-validation (leave-one-out is the default option) was applied to address the external consistency.

This method was applied in several cases (79–83) using different classes of compounds. A QSAR study (79) of the HIV-1 reverse transcriptase (RT) inhibitory activity was done for two series of 54 compounds, 2-(2,6-dihalo phenyl)-3-(substituted pyridin-2-yl)-thiazolidin-4-ones and 2-(2,6-dihalophenyl)-3-(substituted phenyl)-thiazolidin-4-ones, belonging to 2,3-diaryl-thiazolidin-4-ones using the combinatorial protocol (CP) interfaced with MLR and PLS analysis. In (80), a series of 5 4-benzyl/benzoyl-pyridin-2-ones along with their anti-HIV activities using CP and genetic algorithms (GA) as feature selection and MLR, PLS, and ANNs as regression techniques were used. A set of 46 benzoylaminobenzoic acid derivatives was studied (81) as β-ketoacyl-acyl-carrier protein (ACP) synthase III inhibitors with CP-MLR applied to build the QSAR model. In (82), the structure-activity models for the myorelaxant activity of 28 cromakalim analogs were developed by the CP-MLR method. The antimycobacterial activity of 31 functionalized alkenol derivatives is also modeled by the CP-MLR method in (83). For all data sets, the squared correlation coefficients were higher than 0.5 (79–83).

The replacement method (RM) is a technique that is proposed for feature selection by Duchowicz et al. (84). It was applied for the first time on a data set consisting of 62 nitroso compounds (85) to model the carcinogenic potency (TD50). The method selected seven descriptors, and an MLR model was then able to explain 84.3% of the experimental variance. In (85) the obtained models were only evaluated by leave-one-out and leave-25%-out cross-validation as internal validation. The RM method is based on replacing a chosen feature of the set by another to minimize the total SD. A set of descriptors is initially chosen randomly (the number of selected variables is less than the number of compounds); linear regression is then applied. One of the features, e.g., xi, is chosen and replaced iteratively by each of the remaining features from the pool, and finally the best set is selected. The variable with the largest relative error in its coefficient is chosen and replaced with all pool features except with the one chosen in the previous iteration. This process is repeated as many times as necessary until the set of features remains unchanged. This method has been used on different classes of compounds (86–90), such as on the antifeedant activities of aurones, chromones, and flavones; the %HIA (human internal absorption) permeability values of 160 organic compounds; 154 non-nucleoside reverse transcriptase inhibitors (NNRTI) of the wild-type HIV-1 virus; the antituberculosis activity of 43 quinoxaline-2-carboxylate 1,4-di-N-oxide derivatives; and the binding affinity constants of flavonoid ligands for the benzodiazepine site of the gamma-aminobutyric acid (GABA)(A) receptor complex. The obtained results indicated the proper ability of RM as feature selection approach. RM gave models with better statistical parameters (correlation coefficient as well as internal and external validation) than the stepwise regression procedure and similar or better ones than the more elaborated GA (86–90).

The replacement method was adapted by Mercader et al. (91) and called the enhanced replacement method (ERM). The ERM follows exactly the same steps as RM but exhibits less propensity for remaining in local minima. First, an initial set of features is chosen randomly, then one of the features is replaced with the remaining one by one, and the set with the smallest value of s is selected. Second, from the resulting set the feature with the largest SD in its coefficient is chosen and substituted by all remaining features from the pool. This procedure is repeated until the set remains unmodified. The ERM and RM are different in the last step. The RM method does not proceed with the optimization if the replacement of the descriptor with the largest error by those in the pool does not decrease the value of s, while ERM chooses the next-smallest s value, which is of

great help for getting out of a local *s* minimum. This algorithm was performed on several data sets with different biological activities. For example, on the inhibition of aldose reductase by 60 flavonoids, on a fluorophilicity data set consisting of 116 organic compounds, on a growth inhibition data set with growth inhibition values to the ciliated protozoan *Tetrahymena pyriformis* by 200 mechanistically diverse phenolic compounds, on a GABA receptor data set containing 78 inhibition data for flavone derivatives, on 100 pED50 antiepileptic activities of enaminones, on 166 aqueous solubilities of drug-like compounds, on 470 pIGC50 aqueous toxicities of heterogeneous aliphatic compounds, on 392 pIGC50 aqueous toxicities of benzene derivatives, on 17 acetylcholinesterase inhibitor activities of substituted indanone and benzylpiperidine analogs, on 35 glass transition temperatures of structurally diverse polymers, on 30 melt transition temperatures of structurally diverse polymers, and on dissociation constants of 88 pharmaceutical compounds. For all models not only internal validation (cross-validation) but also external validation (test set) is taken into account. The models are evaluated based on them, as well as on the squared correlation coefficient and the root mean squared error of prediction (RMSEP). The obtained results were quite acceptable and similar or better than the GA (92–95).

Xue et al. (96) used recursive feature elimination (RFE) as feature selection. RFE is a strategy to select variables based on SVM. At each iteration a linear SVM is trained, followed by removing one or more "bad" features from further consideration. The goodness of the features is determined by the absolute value of the corresponding weights used in the SVM. RFE is used for selecting descriptors to predict P-glycoprotein substrates (P-gp), which facilitates early identification and elimination of drug candidates of low efficacy or high potential resistance from a data set including 116 substrates and 85 nonsubstrates; HIA, including 113 absorbable and 65 nonabsorbable compounds; and TdP agents that cause "torsades de pointes" including 85 TdP and 276 non-TdP agents. When using this feature selection, the dimension of descriptors decreased. For example, the total number of descriptors in the original data set was 159 for all three classes, and was reduced to 22, 27, and 31 for P-gp, HIA, and TdP, respectively. Then SVM based on a Gaussian kernel function is applied as a modeling technique. The RFE reduced the number of descriptors significantly, and the computational speed for the classification increased (96). From another point of view, the prediction accuracies for both P-gp and HIA increased, but for TdP nothing changed when the number of descriptors was decreased by RFE.

GA is one of the most popular techniques, which has been used as feature selection approach in QSAR studies as well. The GA is a technique based on natural evolution principles introduced by Holland in 1975 (97) and relies on Darwin's evolution theory. Features play the role of genes, and a set of features is called a chromosome. Each individual object of a population is described by a chromosome of binary values, zeros or ones. The first generation is selected randomly, and the state of each variable is represented by the value 1 (selected) or zero (not selected). The practical application of GAs requires the tuning of some parameters, such as the population size, generation gap, crossover rate, and mutation rate. Crossover is an operation in which a pair of chromosomes is divided, mutually exchanged, and merged. Mutation is a genetic operator (change from a zero to one and vice versa) used to maintain genetic diversity from one generation of a population of algorithm chromosomes to the next.

Several papers have been published that describe GAs to explore and build QSAR models. Hasegwa et al. (98) modified a GA and performed it as feature selection on 35 dihydropyridine (DHP) derivatives with the corresponding inhibitory activities of calcium channel antagonists. The GA is modified in order to protect more informative chromosomes. An extra step, which decides whether a new chromosome replaces the old or not, was introduced after the reproduction step (selection, crossover, and mutation) in the GA. In fact, an informative chromosome is the one with high interval predictivity using a small number of variables. Therefore, the more informative chromosome has a probability to produce a useful QSAR model and should not be eliminated from the population of chromosomes in the GA. Results showed that the prediction of the PLS model was much improved by the GA-PLS analysis. In another study (99), GA-PLS was used as feature selection-modeling approach on a dataset containing 57 benzodiazepines to model their binding affinity to $GABA_{(A)}$ receptor. The PLS model with the selected variables was a significantly better predictor than the one with all descriptors. GA-genetic programming (GP) was used as a feature selection on a dataset of 79 inhibitors of HIV-1 reverse transcriptase, i.e., 1-((2-hydroxyethoxy) methyl)-6-(phenylthio) thymine compounds, by Tang and Li (100). The GP was used in a symbolic regression application to find the appropriate equation prototype and the coefficients of the equation, which are represented by the combination of function nodes and terminal nodes. Nodes are composed of single-variable function nodes (N1), binary function nodes (N2), and leaves (N0). A tree is used to express the individuals. GA was used for optimizing the parameters during the process. It is an algorithm for dealing with all kinds of functional and fractional relationships, which use GP to create new individuals and GA to optimize them simultaneously. In this way, the advantages of GA and GP are fully used. The results showed the applicability of the introduced method in a QSAR study. Cho and Hermsmeier (101) used GA to select subsets of compounds and to group different chemotypes. GA was applied on the molecular electronegativity distance vector of 13 atomic types. Then a QSAR model was built using MLR. The obtained result on a data set containing three molecular systems, i.e., 31 steroids, 58 dipeptides, and 16 COX-2 inhibitors, showed that the prediction ability of GA-MLR is better than that of a PCR method (102). GA coupled to PLS was also performed on 123 organic compounds to model the concentration causing 50% lethality (narcotic activity; 103).

GFA also can be used as a feature selection method (70). This approach involves the combination of the MARS algorithm with a GA to evolve a population of equations that best fits the training set data. It provides an error measure, called the lack-of-fit (LOF) score, that penalizes models with too many features (64). GFA generates a population of equations rather than one single equation to correlate biological activity and physicochemical properties.

GFA starts with the generation of an initial population of equations by a random choice of descriptors. Pairs from the equation population are randomly chosen and "crossovers" are performed. Progeny equations can then be generated, and the fitness of each progeny equation can be assessed by the LOF

Table 1. Different wrapper and filter methods

Feature selection technique	Filter (F) or wrapper (W)	Modeling method	Reference
Unbalanced correlation score (cub)	F		46
Fisher score (fish)	F		46
Information gain	F		49, 55
Mutual information	F		49
χ^2-test	F		49
Odds ratio	F		49
GSS coefficient	F		49
Unsupervised forward selection (UFS)	F		53, 59
McCabe coefficient	F		55
Correlation based feature selection	F		55
K-nearest neighbor	W		55
Relief	F		55
Factor analysis (FA)	F		59
Principal component analysis (PCA)	F		63–69
Mutual information differential Shannon entropy (MI-DSE)	F		71–73
Shannon entropy	F		71–73
Backward elimination	W	MLR	75
Stepwise regression	W	MLR	75, 125
Forward selection	W	MLR	75
Leaps-and-bounds regression	W	MLR	75
Genetic algorithm	W	MLR, PLS, ANN	75, 98, 99, 102, 103, 106–108, 123, 125
Combinatorial protocol (CP)	W	MLR, ANN, PLS	76, 79–83
Replacement method (RM)	W	MLR, ANN, SVM	86–90
Enhanced replacement method (ERM)	W	MLR	91–95
Recursive feature elimination (RFE)	W	SVM	96
Genetic programming (GP)	W	MLR	100, 126, 127
Genetic function approximation (GFA)	W		104, 105, 125
Genetic algorithm Shannon entropy cliques (GASEC)	W	AdaBoost.M1, SVM	109–111
Genetic algorithm-variable subset selection (GA-VSS)	W	ANN, SVM, PLS	120–122
Multi-objective (MO)	W	MLR, ANN, NLR, DT	124
Generalized simulated annealing (GSA)	W	MLR	125, 130–133
k-nearest neighbor (K-NN)	W	NN	133
Successive projection algorithm (SPA)	W	MLR, ANN	134–136
Counter-propagation artificial neural network (CPANN)	W	ANN	138
Ensemble of neural networks (NN)	W	ANN	139
Heuristic multilinear regression	W	MLR, ANN, SVM	96, 147–152
Variable selection and modeling based on the prediction (VSMP)	W	MLR	153
Kolmogorov-Smirnov statistics (KS)	F		155
Factor analysis (FA)	W	PLS	159–162
Uninformative variable elimination (UVE)	W	PLS	163
Ordered predictors selection	W	PLS	164
Bayesian regularized neural network (BRNN)	W	ANN	165, 166
Reverse elimination method-tabu search (REM-TS)	W	PCR, PLS	167, 168

Table 1. (*continued*)

Feature selection technique	Filter (F) or wrapper (W)	Modeling method	Reference
High ranking	F		45
High ranking set cover	F		45
Signal method	W	OLS	45
Ant colony optimization (ACO)	W	MLR, PLS, SVM	171–175
Particle swarms optimization (PSO)	W	MLR, PLS	176–179
Regression coefficients	W	PLS	180
Variable importance in the projection (VIP)	W	PLS	181, 182

measure. When the fitness of the new progeny equation is better, it is preserved (104, 105).

An improvement of GA as feature selection is done in (106) by adding an extra step after selection, cross-over, and mutation. A chromosome with k variables is defined as the most informative when it gives the best prediction among all the chromosomes with the most k variables. Because the more informative chromosome has probability to produce a useful QSAR model (predictive and easily interpretable model), such a chromosome should not be eliminated from the population of chromosomes in the GA. Therefore, this chromosome should be protected, and the authors added an extra step for protecting them. In this extra step, it is decided whether or not a new chromosome can replace an old one. If a new chromosome is protected, the method is replacing the least-fitting nonprotected chromosome (one that can be eliminated from the population). If a new chromosome is nonprotected, it replaces the old one only if its predicative ability is higher than that of the least-fitting nonprotected chromosome; otherwise the new chromosome is rejected. This method was applied on 35 dihydropyridine derivatives to model the molar concentration necessary to inhibit 50% of the contraction of guinea pig ileum induced by methylfurmethide. The results indicated the ability of this method for feature selection. The predictivity of the PLS model is improved by variable selection, and the squared correlation coefficient of prediction is considered to be a good measure of fitness in this GA-PLS computation (106).

Hybrid GA as feature selection, coupled to linear methods, such as MLR and correlation-based feature selection, and nonlinear methods, such as nonlinear decision tree and ANN, were used as feature selection approaches on 170 (cycloalkylpyranone analogs) HIV protease enzyme inhibitors (107). The selected descriptors were quite different, but all methods showed good prediction performance, although the ANN models were better than the MLR and decision tree models.

GA was used for a binary classification of three different data sets, i.e., one with 463 estrogen receptor (ER) ligands and their corresponding relative binding affinity, one with 337 carbonic anhydrase II (CAII) inhibitors, and one with 1608 monoamine oxidase (MAO) inhibitors (108). The GA-based feature selection could improve the performance of the binary classification QSAR models. Binary classification is an approach for the analysis of high throughput screening data by correlating structural properties of compounds with a "binary" expression of biological activity (1 = active and 0 = inactive) and calculating a probability distribution for active and inactive compounds in a training set. The predictive accuracy obtained for the binary classification of CAII was much better than for both ER ligands and MAO inhibitors. The results with or without GA were compared and showed that GA can remove irrelative variables from the dataset. For instance, for ER ligands, CAII inhibitors, and MAO inhibitors obtained results were 91, 90, and 95 with GA and 96, 85, and 87 without GA, respectively, which showed that when irrelevant variables were removed the result did not change a lot.

Wegner et al. (109–111) introduced GA coupled to Shannon entropy cliques (SEC) as feature selection. SEC was used to measure the information content of the descriptors. Clique detection is used to find initial feature sets that are uncorrelated and have a high information content. The sets selected in the clique detection phase then form a population that is optimized using a GA. The authors also reported a number of hybrid approaches to feature selection that combines filter and wrapper methods. In the third paper of their series (111), they analyzed a human intestinal absorption data set including 194 compounds, represented by 2934 descriptors.

GA was performed on several data sets as feature selection, e.g., on the data set of 72 4-(1-methyl-5-nitro-2-imidazolyl) dihydropyridine derivatives with their calcium channel antagonist activities in quinea-pig (112), and the data set of 41 DHP derivatives with their channel antagonist activity in guinea pig ileal (113). GA was also performed as feature selection on 18 sulfa drugs with carcinogenesis activity, for which they found a significant effect of the highest occupied molecular orbital energy on the carcinogenesis activity in the context of the shape of this orbital (114). GA-based feature selection also was performed on 46 nonpeptide HIV-1 protease inhibitors (115), 6-naphthylthio 1-[(2-hydroxyethoxy) methyl)-6-(phenylthio)] thymine derivatives in the prediction of anti-HIV-1 activity (116), 26 diaryl-substituted pyrazoles CCR2 inhibitors (117), 53 structurally diverse compounds that were known or suspected to interact with CYP 3A4 (118), and 70 ligands with their dopamine transporter inhibitors (119). All obtained models based on either PLS or MLR modeling indicated the ability of GA as a feature selection approach, and the obtained models clearly demonstrated good correlations between the structure (descriptors) and the inhibitory activity of the studied compounds.

GA-variable subset selection (VSS) is a method to search for the best ranking within a wide set of predictor variables (120). It was performed on several data sets, e.g., for predicting the acute

toxicity to the fathead minnow (*Pimephales promelas*) of a set of 408 heterogeneous chemicals (121). Several quantitative structure–toxicity (lethal oral dose for mouse) relationship models for 54 benzodiazepine derivatives, using GA-VSS as feature selection and ANNs, SVMs, or PLS as regression techniques, have been developed (122). In that study, the nonlinear models showed better results than the linear (122).

A combination of GA and neural network was used to select a subset of relevant descriptors in (123). The introduced method is different in two ways to what usually is done. The GA was not constrained to a defined number of descriptors. Second, optimization of the neural network architecture was done simultaneously with the variable selection by dynamically modifying the size of the hidden layer (123). Six different data sets were used, three of them were simulated and the rest were 55 benzodiazepine compounds with their biological activities, 50 2-phenyl-4-quinolone and 2-phenyl-1-, 8-naphtyridin-4-one derivatives with their tubulin polymerization inhibitory activities, and a set of 268 organic compounds with known central nervous system activities.

Soto et al. (124) proposed a novel method for feature selection containing two steps. The first step is a multi-objective (MO) wrapper. It provides a framework for solving decision-making problems involving multiple objectives that aim both to maximize predictive capacity and to reduce the number of selected descriptors. The output of the first step is used by the second, also called validation phase, in order to determine which subsets of descriptors are the most relevant for prediction. The mentioned method is applied on three different data sets consisting of 289 compounds with their blood-brain barrier penetration, 127 compounds with their human intestinal absorption, and 442 organic compounds with their hydrophobicity, respectively.

So et al. (125) compared forward regression, GFA, GSA, and GA-neural network (GA-NN) on a data set containing 56 progestagenic steroid compounds with their relative binding affinities. It was discovered that using FR and GFA was good for an initial screening of the data set, but the result was not good enough because of the nonlinear behavior of the data set. Although excellent results were obtained by GSA, the best were found using GA-NN.

GP (126) is in fact, the same as GA, but the main difference is the representation of a potential solution. In GP, an individual selection is presented as a tree, while in GA it is represented as binary strings of 0 and 1. This makes GP more complicated than GA. However, it was used in QSAR studies (127) to select descriptors from a large pool. The advantage in GP is that the number of required terms does not have to be specified, and the drawback is that the penalty function, which controls the model complexity, has to be calibrated for each data set. However, Nicolotti et al. (127) designed it to derive a single linear model that represents an appropriate balance between the variance and the number of descriptors selected for the model. The authors represented a further drawback, i.e., a single solution is found, which represents one particular compromise solution, while typically a family of different compromise solutions exists. However, the authors presented MOGP, which exploits the population nature of GP to optimize a family of solutions in parallel. In the MOGP method a family of equivalent models is found, where each model represents one particular compromise between accuracy and complexity. The authors applied it on several data sets; in each case, a variety of different models was found. In the case of the Selwood data set (77), these models include "best" models previously reported in the literature (127).

Simulated annealing is based on the Metropolis Monte Carlo algorithm (128), which has been extensively used. In the Metropolis algorithm each iteration is composed of a random perturbation of the actual configuration and the computation of the corresponding energy variation (ΔE; 129). It starts from an initial state and introduces perturbations or random moves by adding or removing a single variable, which creates a new state. Movements with a value lower than the cost function are always accepted, and those with a higher value with less probability might be accepted in some cases. The acceptance probability is based on a parameter so-called temperature (T). The higher the value of T, the more likely that a movement with value higher than the cost function is accepted.

This method has been performed on several data sets to model different biological activities such as the antitubercular activities of quinoxaline compounds (130), the antituberculosis activities of a series of nitrofuranyl amides (131), the PDE-5 inhibition of a series of substituted pyrido[3,2-b]pyrazinones (132), and 58 estrogen receptor ligands (133).

The K-NN QSAR approach (133) explores formally the active analog approach, which implies that similar compounds display similar profiles of pharmacological activities. In K-NN an unknown pattern is classified according to the majority of the class memberships of its *K*-nearest neighbors in the training set (where nearest is based on a distance metric). The procedure starts by calculating the distances between an unknown object and all the objects in the training set. Then some objects (K) from the training set, which are the most similar to the unknown object, are selected according to the calculated distances. Finally, the unknown object is classified with the group to which a majority of the selected objects belongs. The optimal subset size k is selected based on the classification of a test set or by the leave-one-out cross-validation (133).

Akhlaghi and Kompany-Zareh (134) applied the successive projection algorithm (SPA; 135) as a feature selection on a series of 107 1-(2-hydroxyethoxy-methyl)-6-(phenylthio) thymine derivatives as NNRTIs. They coupled this SPA with backward elimination and radial basis function neural network to model anti-HIV activity. From 160 molecular descriptors that have been used as initial data to derive a new QSAR model, only a subset of 11 descriptors was selected using SPA with backward elimination. It was concluded that hydrophobic, electronic, and geometric descriptors were important for the anti-HIV activity of the mentioned class of compounds. Based on the obtained results, the reliability of SPA to select variables without losing useful information was concluded. In another work, the authors proposed the correlation weighted successive projections algorithm (136) as a modified version of SPA that was applied on same data set as in the previous study. In the proposed procedure the correlation coefficient of each descriptor with the activities was an additional criterion for selection of descriptors. Compared to SPA, results with a lower root mean squared error of prediction are obtained using a lower number of selected variables (six variables).

The counter-propagation artificial neural network (CPANN) technique was used by Jezierska et al. (137) for feature selection on a set of 95 aromatic and heteroaromatic amines with a mutagenic activity. They removed the variables that were

constant for more than 80% of the molecules prior to CPANN. The number of the independent variables was reduced from 275 to 240, and then different dimensions of networks (ranging from 5*5 to 7*7) and of the number of learning epochs (ranging from 100 to 1300) were tested. Finally, the 6*6 network dimension with 1000 epochs was chosen. The first descriptor from each neuron was selected. First descriptor means the one closest to the neuron weight. Each molecule is represented as a vector of *m* elements, where *m* is equal to the number of descriptors. In the transposed matrix, the descriptors are stored in rows, which means that each row represents an *N*mol-dimensional vector of one of the *m* descriptors. For example, the first row contains *N*mol values of the first descriptor with the vector components corresponding to the values of this first descriptor in each of the *N*mol molecules. This transposed matrix is normalized by columns and then introduced to the Kohonen Network that is trained until a limiting error is reached (138), and 36 descriptors were selected for further study and modeling by Kohonen Network using the transposed matrix.

Tetko et al. (139) used an ensemble of NNs to find the best set of variables while avoiding chance correlation. Model selection, as well as feature selection, was done in different ways, e.g., based on a sensitivity determination followed by pruning. The sensitivities of all variables are calculated (139), and the less sensitive are detected and pruned. Five different algorithms are designed to estimate the importance of the features. They can be divided into two different categories, such as sensitivity (140–142) and penalty term (143, 144) methods. Sensitivity methods measure the importance of weights, and the elements with the smallest sensitivities are deleted. The second group modifies the error function by introducing penalty terms (these terms drive some weights to zero during training). Pruning is stopped depending on the minimal RMSEP value, and this value is always determined with some final precision. Cascade-correlation was also used (145) for optimizing the NN architecture, in which one starts with a small network and dynamically adds new neurons until the analyzed problem is solved. However, the optimal set of descriptors could be determined using an S/N method (146).

Heuristic MLR (147) was performed on several data sets (96, 148–152) to find the best set of selectors and to model. Collinearity is checked and avoided. For example, two descriptors that are intercorrelated above 0.8 are never involved in the same model. The subset size was also considered (149–151). When adding descriptors to the model did not improve the squared correlation coefficient result anymore, the optimum set was obtained. The regression was done on different data sets, e.g., for the prediction of the affinity of a diverse set of 94 drugs binding to human serum albumin (96), of the neuraminidase inhibition of 46 influenza viruses (148), of the anticancer activity of 35 2,5-disubstituted 9-aza-anthrapyrazoles (9-aza-APs) (149), of the percent inhibitions toward HT-29 of triaminotriazine drugs (150), and of the Gp120-co-receptor (CCR5) binding affinity of 79 substituted 1-(3,3-diphenylpropyl)-piperidinyl amides and ureas (151). The performance of two heuristic methods, the best multilinear regression approach and the heuristic back-propagation neural network, was evaluated in developing QSAR models (152), where the toxicity of a diverse data set of 1371 organic chemicals is modeled. The descriptor selection algorithm started by evaluating ANN models with one descriptor as input. The best models were then selected in the next step, where a new descriptor was added to the input layer, and the number of hidden units was increased by one. Again, the best models were selected, and this stepwise procedure was repeated until the addition of new input parameters did not improve the model significantly. Since ANN models are quite likely to converge to some local minima, each model was retrained 30 times, and the model with the lowest error was selected.

Variable selection and modeling based on the prediction (VSMP) is another feature selection method similar to forward selection that was introduced by Liu et al. (153). The various optimal subsets are searched based on the squared correlation coefficient or the RMSEP. Two main steps are needed to search the best subset from the descriptor pool. The first is the selection of various subsets based on either correlation coefficient or RMSEP. Then the best subset from all subsets is determined. This is based on high correlation statistics for both internal and external validation.

SVM (154) is recently a popular machine learning tool that has been applied in many fields. SVM can be used for regression and for binary or multiclass classification problems. RFE (155) is a method that has been used as feature selection. In RFE the variables will be removed if they do not substantially change the objective function (155, 156). Although RFE is a suitable and fast algorithm for selecting variables, it uses a greedy strategy to perform backward elimination that can lead to suboptimal solutions (156).

RFE-SVM was improved in a method called incremental regularized risk minimization (157). This approach will put the removed descriptors from the pool in one set and the selected descriptors in another. Then it is evaluated whether coupling the first set of variables with the second can improve the accuracy of the SVM (157). SVM can also be performed on all features, and then the least important are deleted using Kolmogorov-Smirnov (KS) statistics (158). KS statistics can be used as a feature selection method (157) based on the cumulative fraction function, which represents the dependency of the percentage of samples whose feature values are below a certain threshold on the position of the threshold value in the sorted list of feature values.

The combination of FA and PLS is also possible. FA is used for the initial selection of descriptors, after which PLS is performed. FA is a tool to evaluate the relationship among variables. It reduces variables into a few latent factors from which important variables are then selected for PLS regression. Most of the time, a leave-one-out method is used as a tool to select the optimum number of components for PLS. FA-PLS was performed on several data sets, e.g., for the prediction of the reverse transcriptase inhibition for 70 tetrahydroimidazo (4,5,1-*jk*)(1,4)benzodiazepine derivatives, of the anti-HIV-1 integrase inhibition for 36 styrylquinoline derivatives, of the inhibition by 41 substituted phenols of the germination rate of *Cucumis sativus*, and of the CCR5 binding affinity for 79 1-(3,3-diphenylpropyl)-piperidinyl amides and ureas (159–162).

Uninformative variable elimination (UVE) is a method to exclude uninformative variables from the pool that has high variance but small covariance with the dependent variable (biological activity). This method is most frequently coupled with the PLS method to reduce the complexity and/or to improve the predictive ability of the model. UVE uses a cutoff value for the PLS regression coefficients, which is determined by adding irrelevant descriptors to the original data and evaluating their

corresponding PLS coefficients. This method has the ability to improve the model by removing the uninformative explanatory descriptors from the pool (163). The dataset of (163) includes 202 non-nucleotide NNRTIs that belong to two chemical classes of compounds, known as diaryltriazine- (DATA) and diarylpyrimidine (DAPY)-like inhibitors. The groups are represented by 78 and 124 NNRTIs, respectively. The selected NNRTIs have high inhibitory activities against the wild-type HIV and four mutant strains (181C, 103N, 100I, and 188L; 163).

Sorting variables by using informative vectors also can be used as a feature selection method (164). The main idea of this strategy is to obtain an informative vector, e.g., regression vector, correlation vector, residual vector, variable influence on projection (VIP), net analyte signal, covariance procedures vector, and S/N, that contains information about the location of the best variables for prediction. The approach was applied on 14 molecular descriptors for 48 HIV-1 protease inhibitors (164).

A Bayesian regularized neural network (BRNN) with a sparse laplacian prior (165) as an efficient method for supervised feature selection is used to model the blood-brain barrier partition for 106 organic compounds. It concerns in vivo (in rats) measurements of the partition coefficient of the compound between the brain and the blood (165). BRNN was achieved in an analogous way as the expectation maximization algorithm (165, 166) by progressively setting low relevance weights representing zero. When all weights connecting a given input node are zero, the node and its descriptor are effectively pruned out of the model.

The reverse elimination method-tabu search (REM-TS) is used to select reliable descriptors from a pool (167). This method adds or eliminates one variable/iteration, as forward selection or backward elimination. In each iteration, the complete neighborhood of the current trail solution is searched. This neighborhood is generated by systematically changing the status of every variable (in-out or out-in), one after another. The neighborhood solution that yields the largest improvement in objective function is accepted as the trail solution for the next iteration. If no improving neighborhood solution exists, the one that results in the mildest detrimental move is chosen where in tabu search a move is a transition from one trail subset to another (167). The tabu list keeps track of previously explored solutions and prevents the search from returning to a previously visited solution. This approach was performed in (167) on a set of organic compounds to model their partition coefficient.

In a modified tabu search (168), if the neighbor solution is not in the tabu list, it is selected to be the new current solution. However, this solution is often worse than the current best solution; thus, usually a local minimum is reached. To improve the performance, the information-sharing mechanism (168), among the best previous solutions of all iterations and the current solution, is introduced in the step of generating neighbors of a given solution. The neighbors are generated by moving the given solution toward the best solution of all iterations; the move function directs the moving of the solution (167, 168). The modified tabu search was performed in the modeling of the observed toxicity to *Chlorella vulgaris* in a novel short-term assay for 65 aromatic chemicals (168).

Lanctot et al. (45) introduced a wrapper method, the so-called signal method, that can collect an ensemble of meaningful descriptors from a large pool. In wrapper methods, the first step is a filter method that reduces the initial number of descriptors and helps to decrease the computational cost of the wrapper method. However, the signal method consists of two steps. The first evaluates for each descriptor the correlation with the activity based on either mutual information or an χ^2 metric. The second creates an ensemble model using only the high ranked descriptors for modeling. Two different ensemble methods were applied in (45), i.e., a high ranking and high ranking set cover.

The high ranking method will select features to create the ensemble, only based on rank. The high ranking set cover method selects features to create the ensemble based on rank, but a feature is only added if it is contained in an active molecule that has not yet been covered by the growing ensemble.

This signal method was applied on a thrombin data set consisting of 6509 compounds, and including 41 active and 6468 inactive compounds. Two different binary descriptor spaces can be considered: one is a pharmacophore fingerprint that represents all possible combinations of up to four pharmacophoric descriptors and the distances between them, and the second is a shape-feature fingerprint. The authors found that the combination of the high ranking set cover ensemble method with the Chi-square ranking metric gave the best result.

K-means clustering based on Fisher discriminant ratio (used as a class separability criterion and implemented in a *k*-means clustering algorithm; 169) was used simultaneously as feature selection and modeling technique on a set of 221 HIV-1 protease inhibitors (170). The total number of molecular descriptors computed for each inhibitor was 43. SE (43) was also applied on this dataset. It was concluded that the *k*-means clustering scheme performs well in combination with the Fisher discriminant ratio (169).

Ant colony optimization (ACO) is a class of optimization algorithms based on the actions of ants. It is an area of study within what is called swarm intelligence. The basic idea in the ACO algorithm is the simulation of the behavior of real ant colonies. Ants are capable of finding the shortest route between a food source and their nest without using visual information. Hence, no global world model is obtained. Ants deposit pheromones along their trail to a food source. At a decision point, they make a probabilistic choice based on the amount of pheromone along each search branch. Over time, the shortest route will have the highest rate of ant traversal. In the variable selection problem, *m* ants select one variable, then every ant moves to another variable according to the probability defined (based on a heuristic approach and the amount of pheromone). After each iteration, the amount of pheromone is updated based on the best subsets found. This process is terminated after a fixed number of iterations, or when the system has converged. Different papers have been published using ACO as a feature selection approach (171–175), e.g., to model the HIV-1 activities for a series of 43 3-(3,5-dimethylbenzyl)uracil derivatives, the inhibiting action on the epidermal growth factor receptor tyrosine kinase of 61 analogs of 4-(3-bromoanilino)-6,7-dimethoxyquinazoline, and the inhibitory action of a series of 111 thiocarbamates, i.e., non-nucleoside HIV-1 reverse transcriptase inhibitors. ASO was also applied on the Selwood data set (177) and to the modeling of the rate constants of *o*-methylation of 36 phenal derivatives

Another swarm intelligence algorithm is the so-called particle swarms optimization (PSO), which is based on simulating the social behavior of bird flocking. PSO is used as feature selection in QSAR studies. Particles are generated

numbers (in the range between zero and one) having random positions in the variables space. Particle swarms explore the search space through a population of individuals (particles) that adapt by returning stochastically toward previously successful regions. They are influenced by the success of their neighbors, i.e., by "flying" through a multidimensional search space. Each particle keeps track of its coordinates in the problem space, which are associated with the best solution (fitness) it has achieved so far. Their movement is stochastic and is influenced by the individuals' own memories as well as the memories of their peers. PSO starts with a set of particles with random locations and velocity vectors. These particles "move" through the search space and record the best solutions encountered. A number of particles is defined having random positions and velocities, which change in time within a multidimensional space according to definite rules. The particles are stochastically drawn toward positions that are a trade-off between their own previous best performance and the best previous performance of their neighbors. The coordinates of the position of each particle indicate the relative weight of a given variable in building the regression model. In this approach, the elements of the location vectors can only take the values 1 and 0, indicating whether the feature is selected or not in the *i*th particle (subset) (176–179). PSO has been applied on several data sets, e.g., to model the carcinogenic potencies of a set of 41 aromatic amines (176), the angiotensin II antagonism of a set of 38 4*H*-1,2,4-trialzoles (176), the antifilarial activity of antimycin analogs (177), the binding affinities of ligands to benzodiazepine/GABA(A) receptors (177), the inhibition of dihydrofolate reductase by pyrimidines (175), the affinity to a benzodiazepine receptor of a series of 58 2-aryl(heteroaryl)-2,5-dihydropyrazolo(4,3-*c*) quinolin-3-(3*H*)-ones (178), the inhibition of dihydrofolate reductase by 111 2,4-diamino-5-(3,4-dichlorophenyl)-6-substituted pyrimidines (179), the inhibition of epidermal growth factor receptor tyrosine kinase 4-(X-phenylamino-)-Y-quinazoline by 61 compounds (179), and the nonpeptide angiotensin II antagonism of a set of 85 1,2,4-triazoles.

Feature selection can also be made based on regression coefficients. For example, the features that have the least standardized values of the regression coefficients can be deleted, and a new model developed with a reduced set of features. This can be done for different regression methods, such as PLS, MLR, and SVM (180). VIP scores, which estimate the importance of each variable in the projection used in a PLS model, also can be used as a feature selection approach. The value indicating the contribution of each predictor variable to a model is evaluated to decide on the (un)importance of the variables. A feature with an average value of the squared VIP scores close to or above 1 can be considered important in a model (181, 182).

Conclusions

The main purpose of feature selection is to reduce the number of features in a statistical model, while the accuracy of the model has to be kept high. In fact, feature selection is quite challenging. For example, redundancy exhibited by the multitude of features tends to exert an undue influence in the analysis, giving rise to misleading associations between the features, e.g., the existence of an inherent nonlinearity between most features and the biological activity. On the other hand, many feature-selection processes are computationally intensive, when the number of variables is high, because of, for example, many possible iterations to test.

In fact, for any feature to be selected, certain heuristic criteria need to be satisfied. For instance, a feature should be informative about the output but should not be strongly correlated to other selected features. A feature should carry as much information as possible about the molecular structure, but have little multicollinearity to other features.

However, there is very little consensus on which method should be preferred. One possible reason is the different behavior of data sets, which are sometimes linear and sometimes nonlinear. Regression methods also are linear or nonlinear. However, often a given feature selection method is applied both with linear and nonlinear regression techniques. Therefore, it might be interesting to evaluate whether given feature selection techniques are considered better with either linear or nonlinear methods.

It is difficult to evaluate and compare the ability of the many feature selections. Thus, some research using different feature selection techniques on several data sets is recommended to find a basic idea about their mutual abilities. Selection based on a nonlinear method or meant for nonlinear modeling should also be taken into account.

Acknowledgments

Bieke Dejaegher is a postdoctoral fellow of the Fund for Scientific Research, Vlaanderen, Belgium.

References

(1) Hansch, C.L.A. (1979) in *Substituent Constants for Correlation Analysis in Chemistry and Biology*, John Wiley & Sons, New York, NY

(2) Abraham, D.J. (2003) in *Burger's Medicinal Chemistry and Drug Discovery,* Vol. 1, 6th Ed., C.D. Selassie (Ed.), John Wiley & Sons, New York, NY

(3) IUPAC Recommendations: Glossary of Terms Used in Computational Drug Design (1997) *Pure Appl. Chem.* **69**, 1137–1152 (IUPAC Recommendations 1997), http://goldbook.iupac.org/QT06977.html (accessed on July 6, 2011)

(4) Tropsha, A. (2010) *Mol. Inf.* **29**, 476–488. http://dx.doi.org/10.1002/minf.201000061

(5) Zhang, L., Tsai, K.C., Du, L., Fang, H., Li, M., & Xu, W. (2011) *Curr. Med. Chem.* **18**, 923–930. http://dx.doi.org/10.2174/092986711794927702

(6) Clark, R.D., & Norinder, U. (2011) *Comput. Mol. Sci*. **2**, 108–113

(7) Cramer, R.D., Patterson, D.E., & Bunce, J.D. (1988) *J. Am. Chem. Soc*. **110**, 5959–5967. http://dx.doi.org/10.1021/ja00226a005

(8) Golbraikh, A., & Tropsha, A. (2003) *J. Chem. Inf. Comput. Sci.* **43**, 144–154. http://dx.doi.org/10.1021/ci025516b

(9) Todeschini, R., & Consonni, V. (2000) *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim, Germany. http://dx.doi.org/10.1002/9783527613106

(10) Randic, M. (1975) *J. Am. Chem. Soc*. **97**, 6609–6615. http://dx.doi.org/10.1021/ja00856a001

(11) Kier, L.B., & Hall, L.H. (1976) *Molecular Connectivity in Chemistry and Drug Research,* Academic Press, New York, NY

(12) Kier, L.B., & Hall, L.H. (1986) *Molecular Connectivity in Structure-Activity Analysis,* Research Studies Press, Chichester, UK

(13) Luco, J.M., & Ferretti, F.H. (1997) *J. Chem. Inf. Comput. Sci.* **37**, 392–401. http://dx.doi.org/10.1021/ci960487o

(14) Goodarzi, M., & Freitas, M.P. (2009) *J. Chemometr. Intell. Lab. Syst*. **96**, 59–62. http://dx.doi.org/10.1016/j.chemolab.2008.11.007

(15) Li, T., Mei, H., & Cong, P. (1999) *J. Chemometr. Intell. Lab. Syst*. **45**, 177–184. http://dx.doi.org/10.1016/S0169-7439(98)00102-6

(16) Zhou, Y.P., Jiang, J.H., Lin, W.Q., Xu, L., Wu, H.L., Shen, G.L., & Yu, R.Q. (2007) *Talanta* **71**, 848–853. http://dx.doi.org/10.1016/j.talanta.2006.05.058

(17) Fernández, M., Caballero, J., & Tundidor-Camba, A. (2006) *Bioorg. Med. Chem*. **14**, 4137–4150. http://dx.doi.org/10.1016/j.bmc.2006.01.072

(18) Hu, R., Doucet, J.P., Delamar, M., & Zhang, R. (2009) *Eur. J. Med. Chem*. **44**, 2158–2171. http://dx.doi.org/10.1016/j.ejmech.2008.10.021

(19) Shahlaei, M., Fassihi, A., & Saghaie, L. (2010) *Eur. J. Med. Chem*. **45**, 1572–1582. http://dx.doi.org/10.1016/j.ejmech.2009.12.066

(20) Goodarzi, M., Deshpande, S., Murugesan, V., Katti, S.B., & Prabhaka, Y.S. (2009) *QSAR Comb. Sci*. **28**, 1487–1499. http://dx.doi.org/10.1002/qsar.200960074

(21) Fatemi, M.H., & Gharaghani, S. (2007) *Bioorg. Med. Chem*. **15**, 7746–7754. http://dx.doi.org/10.1016/j.bmc.2007.08.057

(22) Hernández, N., Kiralj, R., Ferreira, M.M.C., & Talavera, I. (2009) *J. Chemometr. Intell. Lab. Syst*. **98**, 65–77. http://dx.doi.org/10.1016/j.chemolab.2009.04.012

(23) Zhao, C.Y., Zhang, H.X., Zhang, X.Y., Liu, M.C., Hu, Z.D., & Fan, B.T. (2006) *Toxicology* **217**, 105–119. http://dx.doi.org/10.1016/j.tox.2005.08.019

(24) Zhou, Y.P., Jiang, J.H., Lin, W.Q., Zou, H.Y., Wu, H.L., Shen, G.L., & Yu, R.Q. (2006) *Eur. J. Pharm. Sci*. **28**, 344–353. http://dx.doi.org/10.1016/j.ejps.2006.04.002

(25) Deconinck, E., Xu, Q.S., Put, R., Coomans, D., Massart, D.L., & Vander Heyden, Y. (2005) *J. Pharm. Biomed. Anal*. **39**, 1021–1030. http://dx.doi.org/10.1016/j.jpba.2005.05.034

(26) Nguyen-Cong, V., & Dang, G.V. (1996) *Eur. J. Med. Chem*. **31**, 797–803. http://dx.doi.org/10.1016/0223-5234(96)83973-0

(27) Venkatraman, V., Dalby, A.R., & Yang, Z.R. (2004) *J. Chem. Inf. Comput. Sci.* **44**, 1686–1692. http://dx.doi.org/10.1021/ci049933v

(28) Liu, H., & Tu, L. (2005) *IEEE Trans. Knowl. Data Eng.* **17**, 491–502. http://dx.doi.org/10.1109/TKDE.2005.135

(29) Das, M., & Liu, H. (1997) *Intell. Data Anal.* **1**, 131–156. http://dx.doi.org/10.1016/S1088-467X(97)00008-5

(30) Hall, M.A. (2000) in *Proceedings of the 17th International Conference on Machine Learning*, Morgan Kaufman Publishers, San Francisco, CA, June 29 to July 2, 2000, pp 359–366

(31) Guan, S., Liu, J., & Qi, Y. (2004) *J. Intell. Syst*. **13**, 15–44. http://dx.doi.org/10.1515/JISYS.2004.13.1.15

(32) Sivagaminathan, R.K., & Ramakrishnan, S. (2007) *Expert Syst. Appl*. **33**, 49–60. http://dx.doi.org/10.1016/j.eswa.2006.04.010

(33) Hsu, C., Huang, H., & Schuschel, D. (2002) *IEEE Trans. Syst. Man Cybern. B* **32**, 207–212. http://dx.doi.org/10.1109/3477.990877

(34) Das, S. (2001) in *Proceedings of the 18th International Conference on Machine Learning*, Morgan Kaufmann Publishers, San Francisco, CA, pp 74–81

(35) Ng, A.Y. (1998) in *Proceedings of the 15th International Conference on Machine Learning*, Morgan Kaufmann Publishers, San Francisco, CA, pp 404–412

(36) Dash, M., & Liu H. (1997) *Intell. Data. Anal.* **1**, 131–156. http://dx.doi.org/10.1016/S1088-467X(97)00008-5

(37) Breiman, L., Friedman, J.H., Olshen, R.A., & Stone, C.J. (1984) *Classification and Regression Trees*, The Wadsworth Statistics/Probability Series, Wadsworth, Belmont, CA

(38) De Mantaras, R.L. (1991) *Mach. Learn*. **6**, 81–92. http://dx.doi.org/10.1023/A:1022694001379

(39) Jun, B.H., Kim, C.S., Song, H.Y., & Kim, J. (1997) *IEEE T. Pattern Anal.* **19**, 1371–1375. http://dx.doi.org/10.1109/34.643896

(40) Quinlan, J. (1986) *Mach. Learn.* **1**, 81–106

(41) Shannon, C.E. (1948) *AT&T Tech. J.* **27**, 379–423

(42) Hart, A. (1984) in *Research and Development in Expert Systems*, M. Bramer (Ed.), Cambridge University Press, Cambridge, MA

(43) Mingers, J. (1987) *J. Oper. Res. Soc*. **38**, 39–47

(44) Zhou, X., & Dillon, T.S (1991) *IEEE T. Pattern Anal.* **13**, 834–841. http://dx.doi.org/10.1109/34.85676

(45) Lanctot, J.K., Putta, S., Lemmen, C., & Greene, J. (2003) *J. Chem. Inf. Comput. Sci.* **43**, 2163–2169. http://dx.doi.org/10.1021/ci034129e

(46) Weston, J., Perez-Cruz, F., Bousquet, O., Chapelle, O., Elisseeff, A., & Scholkopf, B. (2003) *Bioinformatics* **19**, 764–771. http://dx.doi.org/10.1093/bioinformatics/btg054

(47) Rassokhin, D.N., & Agrafiotis, D.K. (2000*) J. Mol. Graphics Modell*. **18**, 368–382. http://dx.doi.org/10.1016/S1093-3263(00)00063-2

(48) Bender, A., Mussa, H.Y., & Glen, R.C. (2004) *J. Chem. Inf. Comput. Sci.* **44**, 170–178. http://dx.doi.org/10.1021/ci034207y

(49) Liu, Y. (2004) *J. Chem. Inf. Comput. Sci.* **44**, 1823–1828. http://dx.doi.org/10.1021/ci049875d

(50) Bender, A., Mussa, H.Y., Glen, R.C., & Reiling, S. (2004) *J. Chem. Inf. Comput. Sci.* **44**, 170–178. http://dx.doi.org/10.1021/ci034207y

(51) Li, H., Liang, Y., & Xu, Q. (2009) *J. Chemometr. Intell. Lab. Syst*. **94**, 188–198. http://dx.doi.org/10.1016/j.chemolab.2008.10.007

(52) Yang, Y., & Pederson, J.O. (1997) *International Conference on Machine Leaning (ICML'97),* Morgan Kaufmann Publishers, San Francisco, CA, July 8–12, 1997

(53) Whitley, D.C., Ford, M.G., & Livingstone, D.J. (2000) *J. Chem. Inf. Comput. Sci.* **40**, 1160–1168. http://dx.doi.org/10.1021/ci000384c

(54) Qannari, E., & Hanafi, M. (2005) *J. Chemometr*. **19**, 387–392. http://dx.doi.org/10.1002/cem.942

(55) Demel, M.A., Janecek, A.G.K., Gansterer, W.N., & Ecker, G.F. (2009) *QSAR Comb. Sci*. **28**, 1087–1091. http://dx.doi.org/10.1002/qsar.200860191

(56) McCabe, G.P. (1984) *Technometrics* **26**, 137–144. http://dx.doi.org/10.2307/1268108

(57) Zheng, H., & Zhang, Y. (2008) *Adv. Space Res*. **41**, 1960–1964. http://dx.doi.org/10.1016/j.asr.2007.08.033

(58) Kohavi, R., & John, G.H. (1997) *Artif. Intell*. **97**, 273–324. http://dx.doi.org/10.1016/S0004-3702(97)00043-X

(59) Salt, D.W., Maccari, L., Botta, M., & Ford, M.G. (2004) *J. Comput. Aided Mol. Design* **18**, 495–509. http://dx.doi.org/10.1007/s10822-004-5203-7

(60) Malinowski, E.R. (2002) *Factor Analysis in Chemistry*, John Wiley & Sons, New York, NY

(61) Hotelling, H. (1933) *J. Educ. Psychol*. **24**, 417–441. http://dx.doi.org/10.1037/h0071325

(62) Djakovic–Sekulic, T., Lozanov-Crvenkovic, Z., & Perisic-Janjic, N. (2008) *Novi Sad. J. Math*. **38**, 39–46

(63) Bhattacharya, P., & Roy, K. (2005) *Bioorg. Med. Chem. Lett*. **15,** 3737–3743. http://dx.doi.org/10.1016/j.bmcl.2005.05.051

(64) Roy, K., & Popelier, P.L.A. (2008) *Bioorg. Med. Chem. Lett*. **18**, 2604–2609. http://dx.doi.org/10.1016/j.bmcl.2008.03.035

(65) Roy, K., & Ghosh, G. (2004) *QSAR Comb. Sci*. **23**, 526–535. http://dx.doi.org/10.1002/qsar.200430891

(66) Roy, K., & Ghosh, G. (2004) *QSAR Comb. Sci*. **23**, 99–108. http://dx.doi.org/10.1002/qsar.200330864

(67) Roy, K., & Ghosh, G. (2004) *J. Chem. Inf. Comput. Sci*. **44**, 559–567. http://dx.doi.org/10.1021/ci0342066
(68) Roy, K., Au, D., & Sengupta, C. (2002) *Drug Design Discov*. **18**, 23–31. http://dx.doi.org/10.1080/10559610213503
(69) Roy, K., & Ghosh, G. (2005) *Bioorg. Med. Chem.* **13**, 1185–1194. http://dx.doi.org/10.1016/j.bmc.2004.11.014
(70) Rogers, D., & Hopfinger, A.J. (1994) *J. Chem. Inf. Comput. Sci*. **34**, 854–866. http://dx.doi.org/10.1021/ci00020a020
(71) Wassermann, A.M., Nisius, B., Vogt, M., & Bajorath, J. (2010) *J. Chem. Inf. Model.* **50**, 1935–1940. http://dx.doi.org/10.1021/ci100319n
(72) Godden, J.W., & Bajorath, J. (2001) *J. Chem. Inf. Comput. Sci*. **41**, 1060–1066. http://dx.doi.org/10.1021/ci0102867
(73) Godden, J.W., & Bajorath, J. (2003) *QSAR Comb. Sci*. **22**, 487–497. http://dx.doi.org/10.1002/qsar.200310001
(74) Grechanovsky, E., & Pinsker, I. (1995) *Comput. Stat. Data. Anal*. **20**, 239–263
(75) Xu, L., & Zhang, W.J. (2001) *Anal. Chim. Acta* **446**, 475–481. http://dx.doi.org/10.1016/S0003-2670(01)01271-5
(76) Prabhakar, Y.S. (2003) *QSAR Comb. Sci*. **22**, 583–595. http://dx.doi.org/10.1002/qsar.200330814
(77) Selwood, D.L., Livingstone, D.J., Comley, J.C.W., O'Dowd, A.B.T., Hudson, A., Jackson, P., Jandu, K.S., Rose, V.S., & Stables, J.N. (1990) *J. Med. Chem*. **33**, 136–142. http://dx.doi.org/10.1021/jm00163a023
(78) Kubinyi, H. (1994) *Quant. Struct. Act. Relat*. **13**, 285–294
(79) Prabhakar, Y.S., Solomon, V.R., Rawal, R.K., Gupta, M.K., & Katti, S.B. (2004) *QSAR Comb. Sci*. **23**, 234–244. http://dx.doi.org/10.1002/qsar.200330854
(80) Deshpande, S., Singh, R., Goodarzi, M., Katti, S.B., & Prabhakar, Y.S. (2011) *J. Enzyme Inhib. Med. Chem.* **26**, 696–705. DOI:10.3109/14756366.2010.548328
(81) Singh, S., Soni, L.K., Gupta, M.K., Prabhakar, Y.S., & Kaskhedikar, S.G. (2008) *Eur. J. Med. Chem*. **43**, 1071–1080. http://dx.doi.org/10.1016/j.ejmech.2007.06.018
(82) Sharma, S., Prabhakar, Y.S., Singh, P., & Sharma, B.K. (2008) *Eur. J. Med. Chem*. **43**, 2354–2360. http://dx.doi.org/10.1016/j.ejmech.2008.01.020
(83) Gupta, M.K., Sagar, R., Shaw, A.K., & Prabhakar, Y.S. (2005) *Bioorg. Med. Chem*. **13**, 343–351. http://dx.doi.org/10.1016/j.bmc.2004.10.025
(84) Duchowicz, P.R., Castro, E.A., & Fernández, F.M. (2006) *MATCH Commun. Math. Comput. Chem*. **55**, 179–192
(85) Morales, A.H., Duchowicz, P.R., Pérez, M.Á.C., Castro, E.A., Cordeiro, M.N.D.S., & González, M.P. (2006) *J. Chemometr. Intell. Lab. Syst*. **81**, 180–187. http://dx.doi.org/10.1016/j.chemolab.2005.12.002
(86) Duchowicz, P.R., Goodarzi, M., Ocsachoque, M.A., Romanelli, G.P., Ortiz, E.D.V., Autino, J.C., Bennardi, D.O., Ruiz, D.M., & Castro, E.A. (2009) *Sci. Total. Environ*. **408**, 277–285. http://dx.doi.org/10.1016/j.scitotenv.2009.09.041
(87) Talevi, A., Goodarzi, M., Ortiz, E.V., Duchowicz, P.R., Bellera, C.L., Pesce, G., Castro, E.A., & Bruno-Blanch, L.E. (2011) *Eur. J. Med. Chem*. **46**, 218–228. http://dx.doi.org/10.1016/j.ejmech.2010.11.005
(88) Duchowicz, P.R., Fernández, M., Caballero, J., Castro, E.A., & Fernández, F.M. (2006) *Bioorg. Med. Chem*. **14**, 5876–5889. http://dx.doi.org/10.1016/j.bmc.2006.05.027
(89) Vicente, E., Duchowicz, P.R., Benítez, D., Castro, E.A., Cerecetto, H., González, M., & Monge, A. (2010) *Bioorg. Med. Chem. Lett*. **20**, 4831–4835. http://dx.doi.org/10.1016/j.bmcl.2010.06.101
(90) Goodarzi, M., Duchowicz, P.R., Wu, C.H., Fernandez, F.M., & Castro, E.A. (2009) *J. Chem. Inf. Model.* **49**, 1475–1485. http://dx.doi.org/10.1021/ci900075f
(91) Mercader, A.G., Duchowicz, P.R., Fernández, F.M., & Castro, E.A. (2008) *J. Chemometr. Intell. Lab. Syst*. **92**, 138–144. http://dx.doi.org/10.1016/j.chemolab.2008.02.005
(92) |Mercader, A.G., & Pomilio, A.B. (2010) *Eur. J. Med. Chem*. **45**, 1724–1730. http://dx.doi.org/10.1016/j.ejmech.2010.01.005
(93) Mercader, A.G., Duchowicz, P.R., Fernández, F.M., Castro, E.A., Bennardi, D.O., Autino, J.C., & Romanelli, G.P. (2008) *Bioorg. Med. Chem*. **16**, 7470–7476. http://dx.doi.org/10.1016/j.bmc.2008.06.004
(94) Mercader, A.G., Duchowicz, P.R., Fernandez, F.M., & Castro, E.A. (2010) *J. Chem. Inf. Model.* **50**, 1542–1548. http://dx.doi.org/10.1021/ci100103r
(95) Mercader, A.G., Goodarzi, M., Duchowicz, P.R., Fernandez, F.M., & Castro, E.A. (2010) *Chem. Biol. Drug. Design* **76**, 433–440. http://dx.doi.org/10.1111/j.1747-0285.2010.01033.x
(96) Xue, Y., Li, Z.R., Yap, C.W., Sun, L.Z., Chen, X., & Chen, Y.Z. (2004) *J. Chem. Inf. Comput. Sci.* **44**, 1630–1638. http://dx.doi.org/10.1021/ci049869h
(97) Holland, J. (1975) *Adaptation in Natural and Artificial Systems*, The Michigan University Press, Ann Arbor, MI
(98) Hasegawa, K., Miyashita, Y., & Funatsu, K. (1997) *J. Chem. Inf. Comput. Sci.* **37**, 306–310. http://dx.doi.org/10.1021/ci960047x
(99) Hasegawa, K., & Funatsu, K. (1998) *J. Mol. Struc.-THEOCHEM*. **425**, 255–262. http://dx.doi.org/10.1016/S0166-1280(97)00205-4
(100) Tang, K., & Li, T. (2002) *J. Chemometr. Intell. Lab. Syst*. **64**, 55–64. http://dx.doi.org/10.1016/S0169-7439(02)00050-3
(101) Cho, S.J., & Hermsmeier, M.A. (2002) *J. Chem. Inf. Comput. Sci.* **42**, 927–936. http://dx.doi.org/10.1021/ci010247v
(102) Liu, S.S., Yin, C.S., & Wang, L.S. (2002) *J. Chem. Inf. Comput. Sci*. **42**, 749–756. http://dx.doi.org/10.1021/ci010247v
(103) Sagrado, S., & Cronin, M.T.D. (2008) *Anal. Chim. Acta* **609**, 169–174. http://dx.doi.org/10.1016/j.aca.2008.01.013
(104) Maccari, L., Magnani, M., Strappaghetti, G., Corelli, F., Botta, M., & Manetti, F. (2006) *J. Chem. Inf. Model.* **46**, 1466–1478. http://dx.doi.org/10.1021/ci060031z
(105) Roy, K., & Leonard, J.T. (2005) *J. Chem. Inf. Model.* **45**, 1352–1368. http://dx.doi.org/10.1021/ci050205x
(106) Hasegawa, K., Miyashita, Y., & Funatsu, K. (1997) *J. Chem. Inf. Comput. Sci.* **37**, 306–310. http://dx.doi.org/10.1021/ci960047x
(107) Reddy, A.S., Kumar, S., & Garg, R. (2010) *J. Mol. Graphics Modell*. **28**, 852–862. http://dx.doi.org/10.1016/j.jmgm.2010.03.005
(108) Gao, H., Lajiness, M.S., & Drie, J.V. (2002) *J. Mol. Graphics Modell*. **20**, 259–268. http://dx.doi.org/10.1016/S1093-3263(01)00122-X
(109) Wegner, J.K., Fröhlich, H., & Zell, A. (2004) *J. Chem. Inf. Comput. Sci*. **44**, 931–939. http://dx.doi.org/10.1021/ci034233w
(110) Wegner, J.K., & Zell, A. (2003) *J. Chem. Inf. Comput. Sci*. **43**, 1077–1084. http://dx.doi.org/10.1021/ci034006u
(111) Wegner, J.K., Fröhlich, H., & Zell, A. (2004) *J. Chem. Inf. Comput. Sci*. **44**, 921–930. http://dx.doi.org/10.1021/ci0342324
(112) Hemmateenejad, B., Miri, R., Akhond, M., & Shamsipur, M. (2002) *J. Chemometr. Intell. Lab. Syst*. **64**, 91–99. http://dx.doi.org/10.1016/S0169-7439(02)00068-0
(113) Mohajeri, A., Hemmateenejad, B., Mehdipour, A., & Miri, R. (2008) *J. Mol. Graph. Modell*. **26**, 1057–1065. http://dx.doi.org/10.1016/j.jmgm.2007.09.002
(114) Deeb, O., Hemmateenejad, B., Jaber, A., Garduno-Juarez, R., & Miri, R. (2007) *Chemosphere* **67**, 2122–2130. http://dx.doi.org/10.1016/j.chemosphere.2006.12.098
(115) Deeb, O., & Goodarzi, M. (2010) *Chem. Biol. Drug Design* **75**, 506–514. http://dx.doi.org/10.1111/j.1747-0285.2010.00953.x
(116) Riahi, S., Pourbasheer, E., Dinarvand, R., Ganjali, M.R., &

Norouzi, P. (2009) *Chem. Biol. Drug Design* **74**, 165–172. http://dx.doi.org/10.1111/j.1747-0285.2009.00843.x
(117) Saghaie, L., Shahlaei, M., Fassihi, A., Madadkar-Sobhani, A., Gholivand, M.B., & Pourhossein, A. (2011) *Chem. Biol. Drug Design* **77**, 75–85. http://dx.doi.org/10.1111/j.1747-0285.2010.01053.x
(118) Wanchana, S., Yamashita, F., & Hashid, M. (2003) *Pharm. Res*. **20**, 1401–1408. http://dx.doi.org/10.1023/A:1025702009611
(119) Hoffman, B.T., Kopajtic, T., Katz, J.L., & Newman, A.H. (2000) *J. Med. Chem*. **43**, 4151–4159. http://dx.doi.org/10.1021/jm990472s
(120) Pavan, M., Consonni, V., Gramatica, P., & Todeschini, R. (2006) in *Partial Order in Chemistry and Environmental Sciences*, R. Brüggemann & L. Carlsen (Eds), Chapter 3, Springer-Verlag, Berlin, Germany, pp 181–217
(121) Pavan, M., Netzeva, T.I., & Worth, A.P. (2006) *SAR QSAR Environ. Res*. **17**, 147–171. http://dx.doi.org/10.1080/10659360600636253
(122) Funar-Timofe, S., Ionescu, D., & Suzuki, T. (2010) *Toxicol. in Vitro* **24**, 184–200. http://dx.doi.org/10.1016/j.tiv.2009.09.009
(123) Yasri, A., & Hartsough, D. (2001) *J. Chem. Inf. Comput. Sci.* **41**, 1218–1227. http://dx.doi.org/10.1021/ci010291a
(124) Soto, A.J., Cecchini, R.L., Vazquez, G.E., & Ponzoni, I. (2009) *QSAR Comb. Sci*. **28**, 1509–1523. http://dx.doi.org/10.1002/qsar.200960053
(125) So, S.S., Helden, S.P.V., Geerestein, V.J.V., & Karplus, M. (2000) *J. Chem. Inf. Comput. Sci.* **40**, 762–772. http://dx.doi.org/10.1021/ci990130v
(126) Koza, J.R. (1993) *Genetic Programming*, The MIT Press, Cambridge, MA
(127) Nicolotti, O., Gillet, V.J., Fleming, P.J., & Green, D.V.S. (2002) *J. Med. Chem.* **45**, 5069–5080. http://dx.doi.org/10.1021/jm020919o
(128) Sutter, J.M., & Kalivas, J.H. (1993) *Microchem. J*. **47**, 60–66. http://dx.doi.org/10.1006/mchj.1993.1012
(129) Laarhoven, P.J.V., & Aarts, E.H. (1987) *Simulated Annealing: Theory and Applications*, Kluwer Academic Publishers, Boston, MA
(130) Ghosh, P., & Bagchi, M.C. (2009) *Curr. Med. Chem*. **16**, 4032–4048. http://dx.doi.org/10.2174/092986709789352303
(131) Ghosh, P., & Bagchi, M.C. (2009) *Mol. Simulat*. **35**, 1185–1200. http://dx.doi.org/10.1080/08927020903033141
(132) Bhaisare, M., Karthikeyan, C., Tanwar, O., Waghulde, S., & Laddha, S. (2010) in *14th Int. Electron. Conf. Synth. Org. Chem.* **b007**, Sciforum Electronic Conferences Series, November 1–30, pp 1–25
(133) Zheng, W., & Tropsha, A. (2000) *J. Chem. Inf. Comput. Sci.* **40**, 185–194. http://dx.doi.org/10.1021/ci980033m
(134) Akhlaghi, Y., & Kompany-Zareh, M. (2006) *J. Chemometr*. **20**, 1–12. http://dx.doi.org/10.1002/cem.971
(135) Galvão, R.K.H., Araújo, M.C.U., Jose, G.E., Pontes, M.J.C., Silva, E.C., & Saldanha, T.C.B. (2005) *Talanta* **67**, 736–740. http://dx.doi.org/10.1016/j.talanta.2005.03.025
(136) Kompany-Zareh, M., & Akhlaghi, Y. (2007) *J. Chemometr.* **21**, 239–250. http://dx.doi.org/10.1002/cem.1073
(137) Jezierska, A., Vracko, M., & Basak, S.C. (2004) *Mol. Divers*. **8**, 371–377. http://dx.doi.org/10.1023/B:MODI.0000047502.66802.3d
(138) Roncaglioni, A., Novic, M., Vracko, M., & Benfenati, E. (2004) *J. Chem. Inf. Comput. Sci*. **44**, 352–358. http://dx.doi.org/10.1021/ci030421a
(139) Tetko, I.V., Villa, A.E.P., & Livingstone, D.J. (1996) *J. Chem. Inf. Comput. Sci.* **36**, 794–803. http://dx.doi.org/10.1021/ci950204c
(140) LeChun, Y., Denker, J.S., & Solla, S.A. (1990) in *Advances in Neural Processing Systems 2 (NIPS*2)*, D.S. Touretzky (Ed.), Morgan-Kaufmann Publishers, San Mateo, CA, pp 598–605
(141) Hassibi, B., & Stork, D. (1993) in *Advances in Neural Processing Systems 5 (NIPS*5)*, S.J. Hanson, J. Cowan, & L. Giles (Eds), Morgan-Kaufmann Publishers, San Francisco, CA, pp 164–171
(142) Hansen, L.K., & Rasmussen, C.E. (1994) *Neural. Comp*. **6**, 1223–1233. http://dx.doi.org/10.1162/neco.1994.6.6.1223
(143) Chauvin, Y.A. (1989) in *Advances in Neural Processing Systems 1 (NIPS*1)*, D.S. Touretzky (Ed.), Morgan-Kaufmann Publishers, San Francisco, CA, pp 519–526
(144) Weigen, S.A., Rumelhart, D.E., & Huberman, B.A. (1991) in *Advances in Neural Processing Systems 3 (NIPS*3)*, R. Lippann, J. Moody, & D.S. Touretzek (Eds), Morgan-Kaufmann Publishers, San Francisco, CA, pp 875–882
(145) Kovalishyn, V.V., Tetko, I.V., Luik, A.I., Kholodovych, V.V., Villa, A.E.P., & Livingstone, D.J. (1998) *J. Chem. Inf. Comput. Sci.* **38**, 651–659. http://dx.doi.org/10.1021/ci980325n
(146) Turner, J.V., Cutler, D.J., Spence, I., & Maddalena, D.J. (2003) *J. Comput. Chem*. **24**, 891–897. http://dx.doi.org/10.1002/jcc.10148
(147) Katritzky, A.R., Oliferenko, A., Lomaka, A., & Karelson, M. (2002) *Bioorg. Med. Chem. Lett*. **12**, 3453–3457. http://dx.doi.org/10.1016/S0960-894X(02)00741-2
(148) Lü, W.J., Chen, Y.L., Ma, W.P., Zhang, X.Y., Luan, F., Liu, M.C., Chen, X.G., & Hu, Z.D. (2008) *Eur. J. Med. Chem*. **43**, 569–576. http://dx.doi.org/10.1016/j.ejmech.2007.04.011
(149) Slavov, S., Atanassova, M., & Galabov, B. (2007) *QSAR Comb. Sci*. **26**, 173–181. http://dx.doi.org/10.1002/qsar.200530216
(150) Liu, K., Xia, B., Ma, W., Zheng, B., Zhang, X., & Fan, B. (2008) *QSAR Comb. Sci*. **27**, 425–431. http://dx.doi.org/10.1002/qsar.200730045
(151) Yuan, Y., Zhang, R., Hu, R., & Ruan, X. (2009) *Eur. J. Med. Chem*. **44**, 25–34. http://dx.doi.org/10.1016/j.ejmech.2008.03.004
(152) Kahn, I., Sild, S., & Maran, U. (2007) *J. Chem. Inf. Model.* **47**, 2271–2279. http://dx.doi.org/10.1021/ci700231c
(153) Liu, S.S., Liu, H.L., Yin, C.S., & Wang, L.S. (2003) *J. Chem. Inf. Comput. Sci.* **43**, 964–969
(154) Vapnik, V. (1998) *Statistical Learning Theory*, John Wiley & Sons, New York, NY
(155) Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002) *Machine Learning* **46**, 389–422. http://dx.doi.org/10.1023/A:1012487302797
(156) Cao, D.S., Xu, Q.S., Liang, Y.Z., Chen, X., & Li, H.D. (2010) *J. Chemometr. Intell. Lab. Syst*. **103**, 129–136. http://dx.doi.org/10.1016/j.chemolab.2010.06.008
(157) Frohlich, H., Wegner, J.K., & Zell, A. (2004) *QSAR Comb. Sci*. **23**, 311–318. http://dx.doi.org/10.1002/qsar.200410011
(158) Byvatov, E., & Schneider, G. (2004) *J. Chem. Inf. Comput. Sci.* **44**, 993–999. http://dx.doi.org/10.1021/ci0342876
(159) Mandal, A.S., & Roy, K. (2009) *Eur. J. Med. Chem*. **44**, 1509–1524. http://dx.doi.org/10.1016/j.ejmech.2008.07.020
(160) Leonard, J.T., & Roy, K. (2008) *Eur. J. Med. Chem*. **43**, 81–92. http://dx.doi.org/10.1016/j.ejmech.2007.02.021
(161) Roy, K., & Ghosh, G. (2006) *QSAR Comb. Sci*. **25**, 846–859. http://dx.doi.org/10.1002/qsar.200510211
(162) Leonard, J.T., & Roy K. (2006) *Bioorg. Med. Chem. Lett*. **16**, 4467–4474. http://dx.doi.org/10.1016/j.bmcl.2006.06.031
(163) Daszykowski, M., Stanimirova, I., Walczak, B., Daeyaert, F., de Jonge, M.R., Heeres, J., Koymans, L.M.H., Lewi, P.J., Vinkers, H.M., Janssen, P.A., & Massart, D.L. (2005) *Talanta* **68**, 54–60. http://dx.doi.org/10.1016/j.talanta.2005.04.071
(164) Teófilo, R.F., Martins, J.P.A., & Ferreira, M.M.C. (2009) *J. Chemometr*. **23**, 32–48. http://dx.doi.org/10.1002/cem.1192
(165) Burdena, F.R., & Winkler, D.A. (2009) *QSAR Comb. Sci*. **28**, 1092–1097. http://dx.doi.org/10.1002/qsar.200810202

(166) Burdena, F.R., & Winkler, D.A. (2009) *QSAR Comb. Sci*. **28**, 645–653. http://dx.doi.org/10.1002/qsar.200810173

(167) Baumann, K., Albert, H., & Korff, M.V. (2002) *J. Chemometr*. **16**, 339–350. http://dx.doi.org/10.1002/cem.730

(168) Shen, Q., Shi, W.M., & Kong, W. (2010) *Artif. Intell. Med*. **49**, 61–66. http://dx.doi.org/10.1016/j.artmed.2010.01.004

(169) Fisher, R.A. (1936) *Ann. Eugenics* **7**, 179–188. http://dx.doi.org/10.1111/j.1469-1809.1936.tb02137.x

(170) Lin, T.H., Li, H.T., & Tsai, K.C. (2004) *J. Chem. Inf. Comput. Sci.* **44**, 76–87. http://dx.doi.org/10.1021/ci030295a

(171) Goodarzi, M., Freitas, M.P., & Jensen, R. (2009) *J. Chemometr. Intell. Lab. Syst*. **98**, 123–129. http://dx.doi.org/10.1016/j.chemolab.2009.05.005

(172) Shi, W.M., Shen, Q., Kong, W., & Ye, B.X. (2007) *Eur. J. Med. Chem*. **42**, 81–86. http://dx.doi.org/10.1016/j.ejmech.2006.08.001

(173) Goodarzi, M., Freitas, M.P., & Vander Heyden, Y. (2011) *Anal. Chim. Acta* **705**, 166–173. doi:10.1016/j.aca.2011.04.046

(174) Shamsipur, M., Zare-Shahabadi, V., Hemmateenejad, B., & Akhond, M. (2009) *Anal. Chim. Acta* **646**, 39–46. http://dx.doi.org/10.1016/j.aca.2009.05.005

(175) Goodarzi, M., Freitas, M.P., & Jensen, R. (2009) *J. Chem. Inf. Model*. **49**, 824–832. http://dx.doi.org/10.1021/ci9000103

(176) Lin, W.Q., Jiang, J.H., Shen, Q., Shen, G.L., & Yu, R.Q. (2005) *J. Chem. Inf. Model.* **45**, 486–493. http://dx.doi.org/10.1021/ci049890i

(177) Agrafiotis, D.K., & Cedeño, W. (2002**)** *J. Med. Chem.* **45**, 1098–1107. http://dx.doi.org/10.1021/jm0104668

(178) Lin, L., Lin, W.Q., Jiang, J.H., Zhou, Y.P., Shen, G.L., & Yu, R.Q. (2005) *Anal. Chim. Acta* **552**, 42–49. http://dx.doi.org/10.1016/j.aca.2005.07.033

(179) Lin, W.Q., Jiang, J.H., Shen, Q., Wu, H.L., Shen, G.L., & Yu, R.Q. (2005) *J. Chem. Inf. Model.* **45**, 535–541. http://dx.doi.org/10.1021/ci049642m

(180) Roy, P.P., & Roy, K. (2008) *QSAR Comb. Sci*. **27**, 302–313. http://dx.doi.org/10.1002/qsar.200710043

(181) Schefzick, S., & Bradley, M. (2004) *J. Comput. Aided Mol. Design* **18**, 511–521. http://dx.doi.org/10.1007/s10822-004-5322-1

(182) Ghafourian, T., & Cronin, M.T.D. (2006) *QSAR Comb. Sci*. **25**, 824–835. http://dx.doi.org/10.1002/qsar.200510153

Abbreviations

0D	Zero-dimensional
1D	One-dimensional
2D	Two-dimensional
3D	Three-dimensional
3DMorSE	3D-Molecule representation of structure based on electron diffraction
ABC	ATP-binding-cassette
ACO	Ant colony optimization
AMA	Antimycin analogues
ANNs	Artificial neural networks
BBB	Blood-brain barrier
BMLR	Best multilinear regression
BRNN	Bayesian regularized neural network
BzR	Benzodiazepine receptor
CA II	Carbonic anhydrase II
CCR5	Gp120-co-receptor
CNS	Central nervous system
CoMFA	Comparative Molecular Field Analysis
CovProc	Covariance procedures vector
CPANN	Counter-propagation artificial neural network
CP-MLR	Combinatorial protocol multiple linear regression
CWSPA	Correlation weighted successive projections algorithm
DHP	Dihydropyridine
EGFR	Epidermal growth factor receptor
ERM	Enhanced Replacement Method
FA	Factor Analysis
FR	Forward regression
FSR	Forward Stepwise Regression
GA	Genetic Algorithms
GA-GP	Genetic algorithm-genetic programming
GA-NN	Genetic algorithm neural network
GA-VSS	Genetic algorithm-variable subset selection
Getaway	GEometry, Topology, and Atom-Weights AssemblY
GFA	Genetic function approximation
GP	Genetic programming
GSA	Generalized simulated annealing
GSS	Galavotti-Sebastiani-Simi
hBNN	Heuristic back-propagation neural network
HIV-1	*Human immunodeficiency virus* type-1
HOMO	Highest occupied molecular orbital
HTS	High throughput screening
IRRM	Incremental Regularized Risk Minimization
K-NN	K-nearest neighbor
KS	Kolmogorov-Smirnov
LOF	Lack-of-fit
MARS	Multivariate adaptive regression splines
MI-DSE	Mutual Information Differential Shannon Entropy

Abbreviations (*continued*)

MLR	Multiple linear regression
MO	Multiobjective
MOGP	Multiobjective Genetic programming
MUSEUM	Mutation and Selection Uncover Models
NAS	Net analyte signal
NNRTI	Non-nucleoside reverse transcriptase inhibitors
N-PLS	Nonlinear partial least squares
PCA	Principal component analysis
PCR	Principal component regression
PCs	Principal components
P-gp	P-glycoprotein substrates
PLS	Partial least squares
PSO	Particle Swarms Optimization
QSAR	Quantitative Structure-Activity Relationship
RDF	Radial Distribution Function
REMTS	Reverse Elimination Method-Tabu Search
RFE	Recursive feature elimination
RM	Replacement method
RMSEP	Root mean squared error of prediction
RT	Reverse Transcriptase
SE	Shannon Entropy
SEC	Shannon entropy cliques
SPA	Successive projection algorithm
SN	Signal-to-noise ratios vector
SVMs	Support vector machines
TIBO	Tetrahydroimidazo benzodiazepine
UCS	Unbalanced correlation score
UFS	Unsupervised forward selection
UVE	Uninformative variable elimination
VIP	Variable influence on projection
VSMP	Variable selection and modeling based on the prediction
WHIM	Weighted Holistic Invariant Molecular