*Journal of Computer-Aided Molecular Design* **18:** 511–521, 2004.
DOI 10.1007/s10822-004-5322-1



# Comparison of commercially available genetic algorithms: GAs as variable selection tool

Sabine Schefzick* & Mary Bradley
*Pfizer Global Research and Development, Discovery Technologies, Ann Arbor Laboratories, 2800 Plymouth Road, Ann Arbor, MI 48105, USA*

Received 21 July 2004; accepted in revised form 18 October 2004


*Key words:* genetic algorithms, molecular descriptors, QSAR, Selwood dataset, variable selection

**Summary**

Many commercially available software programs claim similar efficiency and accuracy as variable selection tools. Genetic algorithms are commonly used variable selection methods where most relevant variables can be differentiated from 'less important' variables using evolutionary computing techniques. However, different vendors offer several algorithms, and the puzzling question is: which one is the appropriate method of choice? In this study, several genetic algorithm tools (e.g. GFA from Cerius2, QuaSAR-Evolution from MOE and Partek's genetic algorithm) were compared. Stepwise multiple linear regression models were generated using the most relevant variables identified by the above genetic algorithms. This procedure led to the successful generation of Quantitative Structure–activity Relationship (QSAR) models for (a) proprietary datasets and (b) the Selwood dataset.

## Introduction

Quantitative structure–activity relationships (QSAR) were first introduced by Hansch in 1963 [1, 2]. QSAR models assume that variation of the dependent variable (usually the biological activity) can be explained using structural descriptors that characterize structural differences in the given dataset. Undoubtedly, QSAR methods have permeated the world of computational chemistry, and are extremely useful when no structural information for the receptor exists. Hence computational approaches have been divided into Ligand-based and Structure-based drug design. A prerequisite for a Structure-based drug design approach is an understanding of the molecular recognition process in the protein–ligand complex. If the three-dimensional structure of a given protein is known, this information can be directly exploited for the retrieval and design of new ligands. Ligand-based approaches are applied in the early stage of drug discovery projects, where only limited information about the three-dimensional structure of the protein exists. To emphasize the importance of QSAR, we used SciFinder and searched for the keyword 'quantitative structure–activity relationship'. As expected, an exponential increase in published papers was observed (Figure 1). During the last 30 years more than 170,000 papers were published regarding QSAR modeling.

QSAR has proven to be an effective tool for drug discovery and development, and continues to be a popular research area [3, 4]. Using commercial and readily available academic tools it is relatively simple to calculate thousands of chemical descriptors in a short period of time. These descriptors span the gamut from abstract graph-theoretical descriptors to physical and chemical descriptors such as molecular weight, atom and bond counts, and more sophisticated three-dimensional descriptors [5].

*To whom correspondence should be addressed. Fax: +1-734-622-2782; E-mail: sabine.schefzick@pfizer.com

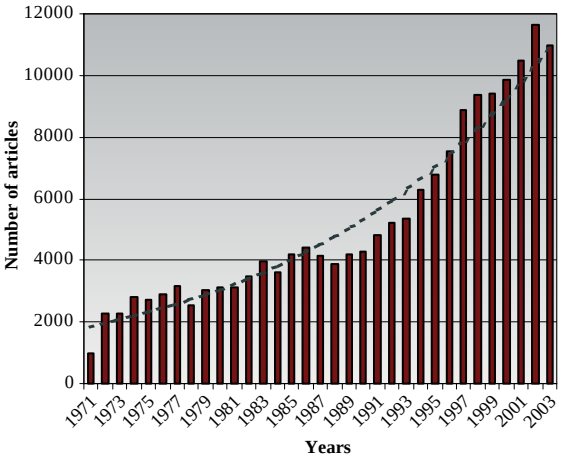


*Figure 1*. Histogram depicting the exponential increase of published papers containing the keyword 'quantitative structure–activity relationship'. The dotted line represents an exponential regression line.

When using QSAR methods, one attempts to find a mathematical relationship between the biological property of interest and these chemical descriptors. Multivariate methods such as multiple linear regression (MLR) are commonly used for this purpose. Typical QSAR datasets contain on the order of hundreds of observations, and as a rule of thumb for applying MLR methods, one should have at least five times as many observations as variables to prevent overfitting the data. Overfitting refers to models that do not obey the principles of parsimony; meaning, more terms or more complicated terms than necessary are used [6]. Hence, overfitting refers to the tendency to give misleadingly good 'goodness of fit' statistics. Overfitting is occurring, when the standard deviation of the resulting QSAR equation is smaller than the error of the dependent variable [7]. Therefore, models created with too many variables perform usually poorly when predicting the activity of compounds outside of the training set. In addition, descriptors can often be correlated to each other and therefore introduce more noise than signal to the model.

Hence it has become important to either identify relevant descriptors via variable selection tools or to apply more advanced mathematical methods such as partial least squares analysis (PLS) to obtain high quality QSAR models [8]. Variable selection algorithms range from simple approaches such as forward selection [9, 10] and backward eliminations [11, 12] to stochastic optimization algorithms such as simulated annealing [13], genetic algorithms [14–17] and evolutionary programming [18]. In this paper, we will examine the performance of three different commercially available genetic algorithms: (genetic function approximation) GFA provided through Cerius2 [19], a genetic algorithm (GA) provided via MOE [20] and a GA embedded in Partek [21]. The most relevant descriptors identified by each GA are used to generate QSAR models for (a) our internal HPLC retention time datasets and (b) the Selwood dataset [9]. Two other datasets [22, 23] have been examined by this method, which yielded qualitatively similar results (data not shown).

## Methods

John Holland introduced the concept of genetic algorithms in his book 'Adaption in Natural and Artificial Systems' in 1975 [24]. Genetic algorithms (GAs) are guided stochastic optimization algorithms, which represent computational analogs as adaptive life forms. GAs loosely model Darwin's evolutionary theory by creating a population of solutions that undergoes selection in the presence of variation-inducing operators such as mutation and recombination (crossover). A fitness function is used to evaluate individuals, and reproductive success varies with fitness. While most stochastic search methods operate on a single solution to the problem, genetic algorithms operate on a population of solutions.

The pseudocode for genetic algorithms is depicted in Figure 2. First a random population of solutions (in this case, multiple QSAR equations) is created. Each chromosome (solution) will be evaluated using a specific fitness function. The chromosome's likelihood of reproduction depends on the fitness of the individual (chromosome). Chromosomes with a poor fitness will not be selected as frequently to create offspring. In a genetic algorithm, children will be produced using mainly crossover procedures, whereas an evolutionary program focuses primarily on mutation [25]. Once a set of children is produced in a genetic algorithm, some mutations can also occur. If the fitness of the children exceeds the fitness of any chromosome in the population, the offspring

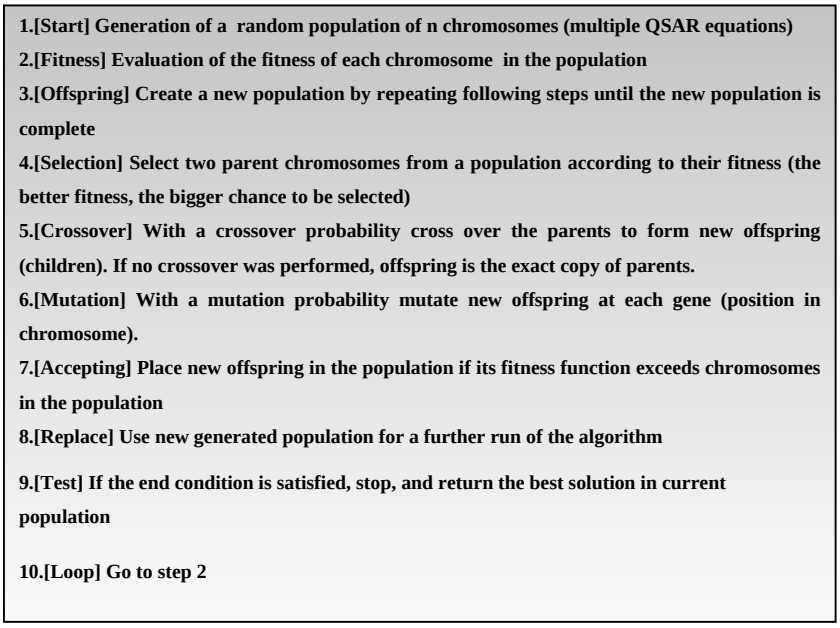


*Figure 2.* Overview of necessary steps for every genetic algorithm.

will replace the chromosome in the population pool, and the algorithm can continue by selecting new parents.

Typically the major difference between GAs is the choice of the fitness function. The fitness of a chromosome can be assessed in different ways including the lack-of-fit (LOF), cross-validated correlation coefficient $q^2$, correlation coefficient $r^2$ and adjusted $r^2$. In this study, we evaluated the performance of three variable selection tools offered by Partek [21], Cerius2 [19], and MOE [20].

Partek is a pattern recognition software, which integrates data mining, statistical analysis, and interactive visualization. The genetic algorithm tool is one option available for variable selection. Partek's GA output includes a rank, a score, the number of selected variables, the column number and name of these variables.

Accelrys commercializes a GA module implementing the GFA as extension of its QSAR capabilities in Cerius2 [15, 26, 27]. This algorithm combines the ideas of Holland's GA with the thoughts of Friedman's multivariate adaptive regression splines (MARS) algorithm [28]. Friedman's MARS algorithm is a statistical technique, which is used to provide an enhanced fitness score for the GA. The fitness, called the lack-of-fit (LOF) score, automatically penalizes models with too many features. The LOF can be calculated as depicted in Equation 1,

$$\mathrm{LOF} = \frac{\mathrm{LSE}}{\left(1 - \frac{c + dp}{M}\right)} \tag{1}$$

where $c$ is the number of basis functions in the model, $d$ is the smoothing parameter, $p$ is the total number of features contained in all basic functions and $M$ is the number of samples in the dataset. In comparison to the commonly used least-squares error (LSE), the LOF has a distinct optimum and cannot be improved by adding more descriptors.

Finally, the QuaSAR-Evolution algorithm provided by MOE [29] supports multiple fitness evaluation scores: the LOF, cross-validated correlation coefficient ($q^2$), correlation coefficient $r^2$ and the adjusted $r^2$. Nevertheless, to be consistent when comparing these different GA approaches, only the LOF function was used.

## Datasets and evaluations strategy

Two different datasets are chosen to evaluate the performance of the above GAs. First, we tested all three different GAs on 89 separate in-house datasets (approximately 43 compounds and 673 descriptors), which are used to predict high

performance liquid chromatograph (HPLC) retention times of solutes under various HPLC separation conditions [30]. This Quantitative Structure Retention Relationship (QSRR) study is used to estimate the retention time of structurally diverse solutes under 89 different liquid chromatography mass spectrometry (LC/MS) conditions. Sixty-two compounds were analyzed using 18 commonly used HPLC columns under five different gradient conditions, however, on average only 43 compounds were successfully separated under a given HPLC method. The solute retention time was used as the dependent variable, and more than 1000 molecular descriptors were initially calculated for this compound set to generate QSRR models. After the removal of descriptors with zero variance and correlated variables, 89 datasets with approximately 43 compounds and 673 descriptors were obtained. Being able to accurately predict the retention time of all compounds of a combinatorial library under all conditions, allows us to lay out the combinatorial libraries so that the libraries are characterized and purified in a timely and efficient manner.

The performance of the genetic algorithms was compared with variable importance projection (VIP) [31]. VIP can be compared to multidimensional partial least squares (PLS) analysis where 89 dependent variables and 673 numeric descriptors are used simultaneously to find a structure–activity relationship. However, not all solutes were successfully separated under all the HPLC conditions; hence several data points were missing in the matrix. A multiple imputation procedure was applied to fill in the missing data.

Secondly, the Selwood dataset [9] was chosen. Multiple research groups [9, 25, 27, 32–35] have pursued structure–activity approaches for characterizing antifilarial antimycin analogues using different variables selection tools and/or various mathematical approaches. Here, we compare the results of using genetic algorithms for variable selection with other published methods.

For all HPLC datasets, Dragon descriptors [36] were calculated based on Concord [37] generated three-dimensional structures. For the Selwood dataset, the data were obtained from 'The QSAR and Modelling Society' web page [38]. A brief summary of the descriptors used by Selwood is listed in Table 1. Following the calculation of structural features, every genetic algorithm was applied three times to the datasets. It should be noted that a GA generates the initial population of chromosomes using a random seed. Hence, the results from different GA runs will differ. Independent of the software program used to perform the variable selection task, the parameters for the GA runs were kept as consistent as possible (Figure 3). For example, the same mutation frequency (0.05) was used here as in the original GFA [27]. However, the parameters chosen for the GA runs were geared specifically towards the datasets used for the prediction of HPLC retention times.

*Table 1.* Descriptions of the variables in the Selwood dataset.

| | |
|---|---|
| ATCH1-10 | partial atomic charge for atom 1–10 |
| DIPV_X | dipole vector |
| DIPV_Y | dipole vector |
| DIPV_Z | dipole vector |
| DIPMOM | dipole moment |
| ESDL1-10 | electrophilic superdelocalizability for atom 1–10 |
| NSDL1-10 | nucleophilic superdelocalizability for atom 1–10 |
| VDWVOL | van der Waals volume |
| SURF_A | surface area |
| MOFI_X | principal moments of inertia |
| MOFI_Y | principal moments of inertia |
| MOFI_Z | principal moments of inertia |
| PEAX_X | principal ellipsoid axes |
| PEAX_Y | principal ellipsoid axes |
| PEAX_Z | principal ellipsoid axes |
| MOL_WT | molecular weight |
| S8_IDX | substituent dimensions |
| S8_IDY | substituent dimensions |
| S8_IDZ | substituent dimensions |
| S8_ICX | substituent centers |
| S8_ICY | substituent centers |
| S8_ICZ | substituent centers |
| LOGP | partition coefficient |
| M_PNT | melting point |
| SUM_F | sum of F substituent constant |
| SUM_R | sum of R substituent constant |

The frequency of descriptors selected by three GA runs was monitored to identify the most relevant descriptors. This is a commonly accepted strategy to determine relevant variables. However, it has been pointed out that genetic algorithms search identifies the best combinations of descriptors rather than identifying individual variables [27]. In addition, 20 training and test sets were

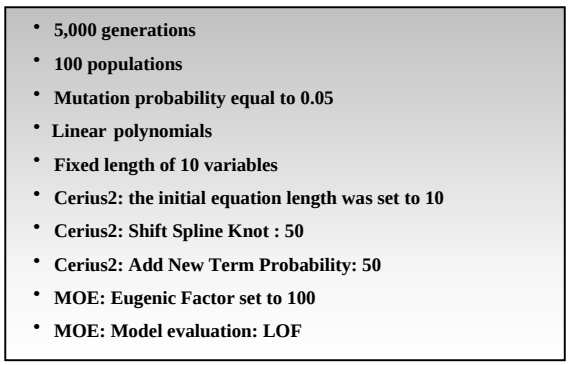


*Figure 3.* Genetic algorithm parameter settings.

created by randomly selecting 20% of the original dataset. Hence, the training sets consist of 80% of the original dataset, whereas the test sets were created from the remaining 20% of the original dataset. The most important descriptors, identified by the frequency count, were selected to generate stepwise linear regression models for each training set. Stepwise multiple linear regression discovers linear relations by automatically selecting only those independent variables, which influence the biological activity most. Hence, a secondary method of variable selection was introduced. Using a two-step elimination process made the comparison between commercial software packages easier. Afterwards, these QSAR models were used to predict the values for the dependent variables in the test sets. The correlation coefficient $r^2$ of the training sets, the cross-validated leave-one-out correlation coefficient $q^2$ of the training sets, and the correlation coefficient $R^2$ of the test sets were used to compare the performance of the three genetic algorithms.

## Results

### *The HPLC datasets*

As mentioned previously, three different commercially available GAs were applied to 89 different datasets using the parameters listed in Figure 3. For each GA, the selected structural features were ranked based on their frequency score. The results for 89 datasets were combined using a sum-ranked fusion method (see Equation 2). Here, $R_i(x)$ denotes the rank position of the descriptor $x$ for a specific HPLC method $i$ and $N$ is the number of different HPLC methods ($N = 89$, considering only analytical conditions). The descriptors with the highest $\mathrm{SUM}_x$ value were considered the most significant descriptors for predicting the solutes retention.

$$\mathrm{SUM}_x = \sum_{i=1}^{N} R_i(x) \tag{2}$$

The $\mathrm{SUM}_x$ values for all possible descriptors were compared using Kendall Tau [39, 40] and the Kappa measures [41]. Basically, we obtained a consensus rank-order for all of the available descriptors for each genetic algorithm. The Kendall Tau measures the correlation between two sets of rankings. The Kendall Tau values span from $-1$ to $+1$, with a positive correlation indicating an agreement between the rankings. Table 2 demonstrates that Kendall Tau and Kappa values for any pairwise combination with VIP are below 0.25, which indicates that descriptors selected by VIP vary significantly from the descriptor selected by any GA. Moreover, the Kendall Tau values in the pairwise comparison of the genetic algorithms of Partek and Cerius2 indicate a good agreement among selected variables. This trend was confirmed by the statistical properties of the best stepwise multiple linear regression models using the different feature sets. Based on the information obtained through variable selection, we concluded that QSAR models generated using descriptors selected by Cerius2's or Partek's GA provide similar results, and variables selected by VIP may generate a statistically different model. Table 3 lists the main statistical parameters for the generated QSAR models. For each dataset, the model with the highest predictability was selected. The results in Table 3 list the statistical averages

*Table 2.* Kendall Tau and Kappa results for sum-ranked frequency score for each GA and VIP for the HPLC data set.

| Variable selection method | By variable | Kendall Tau | Kappa |
|---|---|---|---|
| VIP | Cerius2 GFA | 0.170 | 0.169 |
| VIP | Partek GA | 0.174 | 0.173 |
| VIP | MOE GA | 0.221 | 0.219 |
| MOE GA | Partek GA | 0.369 | 0.360 |
| MOE GA | Cerius2 GFA | 0.467 | 0.454 |
| Partek GA | Cerius2 GFA | 0.719 | 0.719 |

*Table 3*. Best QSAR models identified using different feature selection tools.

| Variable selection method | TRAIN | | | | TEST | | |
|---|---|---|---|---|---|---|---|
| | $r^2$ | $q^2$ | NVars | $F$ value | $R^2$ | $F$ value | Press[a] |
| Partek | 0.990 | 0.874 | 7.7 | 27.8 | 0.813 | 83.0 | 3.06 |
| Cerius2 | 0.993 | 0.866 | 7.5 | 86.2 | 0.807 | 70.7 | 3.46 |
| MOE | 0.962 | 0.780 | 5.8 | 20.5 | 0.762 | 44.2 | 3.83 |
| VIP | 0.952 | 0.695 | 5.4 | 10.9 | 0.666 | 72.4 | 5.46 |

[a]Predictive sum of squares.

over the 89 HPLC datasets. As previously remarked, QSAR models using Partek's or Cerius's GA indicate high internal consistency with a good external predictability, as indicated by the high $R^2$ values for the test sets. In addition, we observed that QSAR models using variables selected by the VIP method are less reliable. Furthermore, statistically relevant QSAR models are derived using features identified through MOE's GA. Interestingly, the number of variables used for these models is comparable with the VIP algorithm. Hence, fewer descriptors are used than in models obtained from Cerius2's and Partek's GA. The variability between the selected variable sets can occur because of the differences in the GA methods or could also arise from an under-population of certain descriptors. In the case of the HPLC dataset, 673 descriptors were available to be selected for 1000 terms the initial equations ($10 \times 100$). The redundancy in this case is minimal and it might be possible that some of the original descriptors are missing in the initial population.

*Selwood dataset*

In this case, 20 different GA runs were performed using each software package. The variables which were selected at least 15% of the time, were considered as relevant variables and are listed in Table 5. Once again, we used the Kendall Tau and Kappa values to compare the different variable selection tools by examining the pairwise agreement. In addition, we calculated the pairwise Kendal Tau and Kappa values for already published results [9, 25, 27, 32–35]. Figure 4 shows a distribution plot of Kendall Tau values for all 45 possible pairwise combinations. The variables selection results for the seven published methods (listed in Table 4) are found to be similar according to the Kendall Tau and Kappa statistics. Interestingly, in contrast to the results of the HPLC datasets, the relevant variables identified by the MOE are in good agreement with the features selected by Partek. Furthermore, it is important to notice that the majority of the pairwise combination results reside in a Kendall Tau value range of 0–0.4, indicating a poor agreement between the selected variables. We therefore conclude that the different variable selection methods used in this study find different statistical relevant descriptor combinations, and it does not appear to be a straightforward correlation between the resulting combinations.

The identified variables, listed in Table 5, were used to generate the stepwise multiple regression

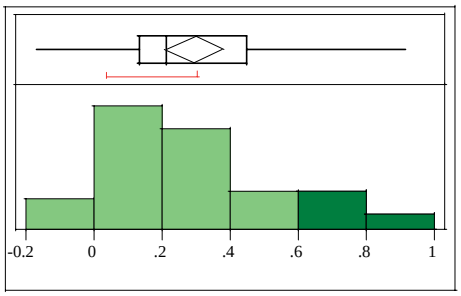


*Figure 4*. Pairwise Kendall Tau distribution of variable selection results (including all published results and results from Partek's, MOE's and Cerius's GA).

*Table 4*. List of the methods with a pairwise Kendall Tau greater than 0.6.

| Variable selection method | By variable | Kendall Tau | Kappa |
|---|---|---|---|
| GAS | Kubinyi | 0.911 | 0.907 |
| McFarland | Wikel | 0.884 | 0.877 |
| Partek GA | MOE GA | 0.792 | 0.790 |
| GAS | Waller | 0.753 | 0.752 |
| Kubinyi | Rogers | 0.733 | 0.699 |
| GAS | Rogers | 0.696 | 0.679 |
| Waller | Kubinyi | 0.676 | 0.669 |

*Table 5.* Comparison of variables identified by different variable selection tools.

| | Selwood | Wikel | McFarland | Rogers | Kubinyi | Waller | GAS | Cerius GFA | MOE GA | Partek GA |
|---|---|---|---|---|---|---|---|---|---|---|
| LOGP | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| M_PNT | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| ATCH2 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ESDL10 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| DIPV_Y | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| DIPV_Z | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ESDL5 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| NSDL2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| S8_1CZ | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| SUM_R | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ATCH4 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| MOFI_Y | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 |
| VDWVOL | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 |
| DIPV_X | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| MOFI_X | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| PEAX_Y | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ATCH5 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| S8_1DX | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ESDL3 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| SUM_F | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| PEAX_X | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 |
| ATCH1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 |
| SURF_A | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 |
| ATCH6 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 |
| MOFI_Z | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| ATCH3 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 |
| ATCH7 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 |
| ESDL8 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 |
| NSDL3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| NSDL6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| ESDL1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| ESDL4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| NSDL4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| ESDL9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

models using 20 training and test sets. The test and training set combinations were kept constant for all variable selection methods. Figure 5 depicts the cross-validated $q^2$ and the correlation coefficient $R^2$ mean with a 95% confidence interval applied to various features selection methods. The number of variables used in the QSAR models is shown in the same illustrative fashion. Every diamond in the plots illustrates a sample mean, while the vertical span of the diamond indicates a 95% confidence interval. All variable combinations indicate structure–activity models with statistically relevant internal consistency, except for the Selwood and Waller variable sets. The variable combinations from MOE generated less predictive and stable models, while the descriptors identified by Partek generated models with the highest $q^2$. The most predictive models were generated using the descriptors identified by the GFA from Cerius2. However, using the GFA in its default setting seems to generate QSAR models with the highest number of selected variables. The Selwood data set required the least number of variables to obtain a model. Across all variables selection tools used in this investigation, a minimum of five variables is necessary to generate statistically relevant models with good internal stability and good predictive power. It is also important to note that the

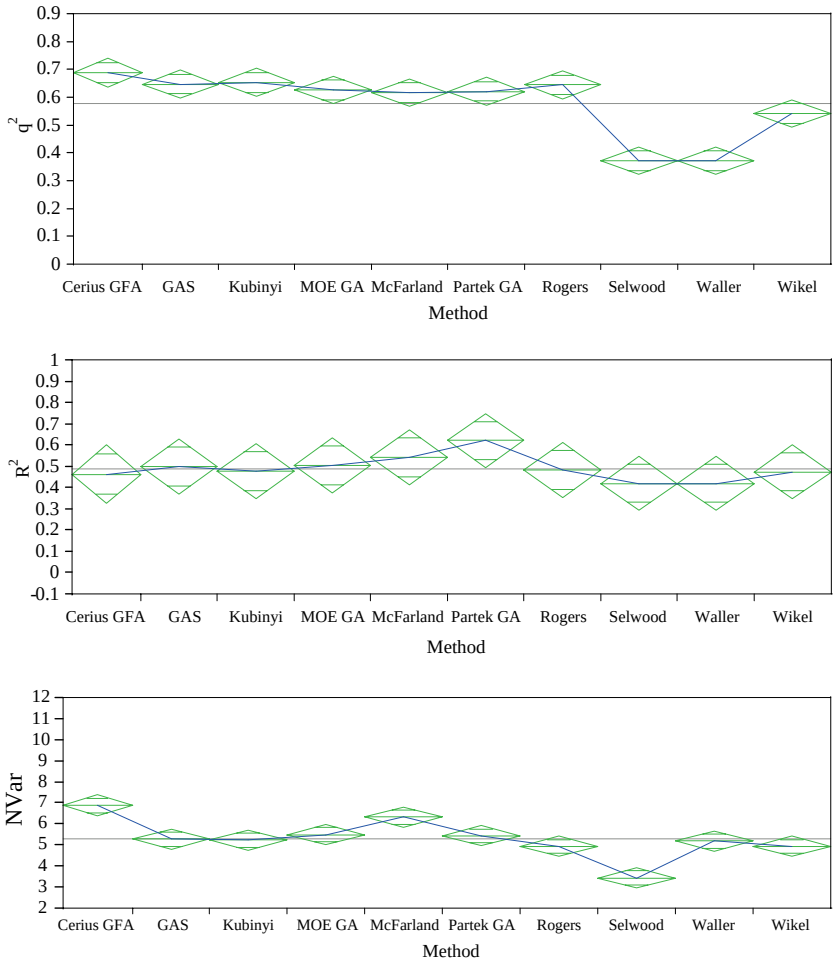


*Figure 5*. Pictorial presentation of the cross-validated $q^2$, correlation coefficient $R^2$ and 'number of variables' mean and standard deviation applied to various methods.

variable selection algorithm published by Rogers is identical to the genetic function algorithm imbedded in Cerius2. Nevertheless, using the default settings of Cerius2, the obtained variable selection sets as well as the generated QSAR models differ from Roger's results, probably because we used the default $d = 1$, whereas Rogers and Hopfinger used $d = 2$ [27]. Overall, we found that the variables selected by GAS, Kubinyi, MOE, McFarland, Partek and Rogers delivered the most reliable QSAR results, indicated by high $q^2$ and $R^2$ values and a low number of variables necessary for the models.

Every genetic algorithm creates an initial set of chromosomes on a random basis. Hence, the GA solution will vary from one run to another. To investigate the reproducibility of the GA the genetic algorithm of Cerius2 was applied 100 times to the Selwood dataset. Only the 'best' solution from each run was collected and transformed into a binary descriptor fingerprint (descriptors used for the QSAR models are assigned values of one, the remaining descriptors values of zero). After 100 repetitive GA runs, 12 variables were never selected: ATCH1, DIPV_Y, IPV_Z, NSDL5, NSDL7, NSDL9, NSDL10, PEAX_Y, S8_1DX, S8_1CX, S8_1CY, SUM_R (for descriptions see Table 1). The partition coefficient Log*P* was selected in every GA run. Table 6 lists the 10 most frequently selected descriptors and their total frequency count after 100 GA runs. The Kendall Tau was used to identify the degree of similarity in all the combinations. Out of 4950 pairwise combinations, 1421 GA combinations were identical,

*Table 6*. Ten most frequently selected descriptors and their total frequency count after 100 genetic algorithm runs.

| Descriptors | Frequency count |
|---|---|
| LOGP | 100 |
| ATCH4 | 88 |
| MOFI_Z | 74 |
| NSDL6 | 67 |
| ESDL8 | 61 |
| NSDL3 | 60 |
| ATCH5 | 56 |
| ATCH6 | 53 |
| ESDL10 | 52 |
| DIPV_X | 30 |

indicating that approximately every third GA run produces an equivalent solution.

An important factor to consider when dealing with genetic algorithms is the number of iterations performed during one GA run. A sufficient number of generations must be created during the GA run to reach a constant fitness function value. If the fitness function does not converge, the genetic algorithm will not reach a global minimum and the solutions obtained from each run may differ dramatically. As observed from the plot in Figure 6, only 3000 iterations are necessary to obtain stable results in the case of the Selwood dataset. It is therefore recommended by the authors to always perform one longer GA run before performing multiple GA runs in an automated fashion.

## Conclusions

Because of the continually increasing number of molecular descriptors available, variable selection tools and genetic algorithms (GAs) in particular have come to the spotlight. The overwhelming quantities of descriptors available for computational analyses have brought new challenges. Using available variable selection tools allows one to first identify the relevant descriptors before constructing a structure–activity relationship. In this study three commercially available GAs were compared in their performance using four different datasets. Quantitative structure–activity relationship models were derived using all three variable selection tools. We found that the set of variables identified as 'relevant' by the various different GA approaches are significantly different. It has to be noted that the possibility to identify multiple solutions (different variable combinations) is an advantage of GAs. The physical meaning of two different variable sets can still be the same, if the set

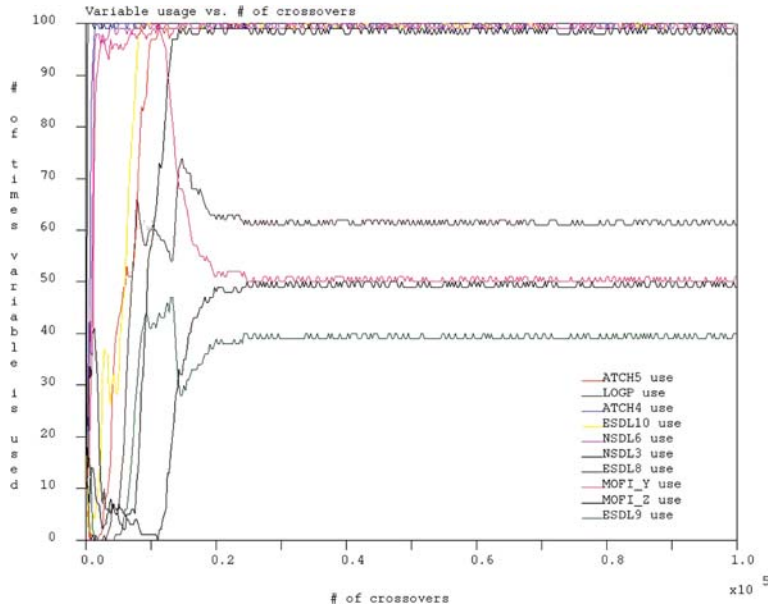


*Figure 6*. Plot illustrating how many numbers of crossover or iterations are needed to obtain stable results because the fitness function converged to a constant number.

of descriptors correlate to each other [42]. The pairwise comparisons between descriptor fingerprints consisting of the frequency counts of all possible descriptors confirmed our assumptions. Stepwise variable reduction was performed by first applying a genetic algorithm to obtain a reduced set of variables, and then by generating stepwise multiple linear regression to generate QSAR models. Since it was possible to generate statistically relevant structure–activity models for all of the datasets, it appears that the performance of GAs is dataset independent. However, the results of a genetic algorithm definitely depend on the number of generations allowed to evolve. GAs that are terminated before the fitness function converges, may not result in a optimal combination of descriptors. Nevertheless, one should not expect to obtain the same optimal solution for every GA run. As we showed in the case of the Selwood dataset, every third GA run will produce the same results. These findings appear to be dataset dependent; further, the random generation of the first generation in genetic algorithms ensures that the results will differ between runs even when using the same method.

The ease and user-friendliness of the genetic algorithm tools is, however, significantly different. For example, it is not possible to set up scripts to automatically perform multiple GA runs in MOE or Partek, but it is for Cerius2. In addition, descriptors have to be pre-computed and imported into Partek. Moreover, speed is an important factor. Even though, the authors did not perform any quantitative speed comparison runs; it became obvious that Cerius2 performed the most crossover iterations per minute. Furthermore, it should be noted that this was not dataset dependent. For both data sets, the Cerius2 genetic function approximation finished significantly faster than the MOE GA and Partek's GA (Cerius2 < MOE < Partek).

To recap, our findings indicate that independently of which variable selection combinations and methods are applied to a dataset, it will be advantageous in creating robust QSAR models.

## Acknowledgement

The authors would like to thank Kjell Johnson for his fruitful discussions around nonparametric analysis methods.

## References

1. Hansch, C. and Steward, A.R., J. Med. Chem., 44 (1964) 691.
2. Hansch, C., Muir, R.M., Fujita, T., Maloney, P.P., Geiger, F. and Streich, M., J. Am. Chem. Soc., 85 (1963) 1817.
3. Keseru, G.M., Mol. Div., 7 (2003) 1.
4. Norinder, U. and Hogberg, T., Textbook of Drug Design and Discovery, Taylor & Francis, New York, NY, 2002, p. 117.
5. Todeschini, R. and Consonni, V. (Eds.), Handbook of Molecular Descriptors, Wiley-VCH, Weinheim, Germany, 2000.
6. Hawkins, D.M., J. Chem. Inf. Comput. Sci., 44 (2004) 1.
7. Wold, S., Quant. Struct.-Act. Relat., 10 (1991) 191.
8. Cramer III, R.D., Perspect. Drug Discov. Design, 1 (1993) 269.
9. Selwood, D.L., Livingstone, D.J., Comley, J.C., O'Dowd, A.B., Hudson, A.T., Jackson, P., Jandu, K.S., Rose, V.S. and Stables, J.N., J. Med. Chem., 33 (1990) 136.
10. Whitley, D.C., Ford, M.G. and Livingstone, D.J., J. Chem. Inf. Comput. Sci., 40 (2000) 1160.
11. Livingstone, D.J. and Rahr, E., Quant. Struct.-Act. Relat., 8 (1989) 103.
12. Kikuch, O., Quant. Struct.-Act. Relat., 6 (1987) 179.
13. Sutter, J.M., Dixon, S.L. and Jurs, P.C., J. Chem. Inf. Comput. Sci., 35 (1995) 77.
14. Sumida, B.H., Houston, A.I., McNamara, J.M. and Hamilton, W.D., J. Theor. Biol., 147 (1990) 59.
15. Rogers, D., In Proceedings of the Fourth International Conference on Genetic Algorithms, San Diego, CA, 1991.
16. Leardi, R., Boggia, R. and Terrile, M., J. Chemometr., 6 (1992) 267.
17. Mitchell, M. (Ed.), An Introduction to Genetic Algorithm, The MIT Press, London, UK, 1999.
18. Luke, B.T., J. Chem. Inf. Comput. Sci., 34 (1994) 1279.
19. Accelrys, Cerius2, San Diego, CA, 2003.
20. ChemicalComputingGroup, MOE, Quebec, Canada, 2003.
21. Partek, Partek Pro, St. Charles, MO, 2003.
22. Chavatte, P., Yous, S., Marot, C., Baurin, N. and Lesieur, D., J. Med. Chem., 44 (2001) 3223.
23. Cavalli, A., Poluzzi, E., De Ponti, F. and Recanatini, M., J. Med. Chem., 45 (2002) 3844.
24. Holland, J.H. (Ed.), Adaptation in Natural and Artificial Systems, The MIT Press, Cambridge, MA, 1975.
25. Kubinyi, H., Quant. Struct.-Act. Relat., 13 (1994) 285.
26. Rogers, D., Adv. Neural Inform. Process. Syst., 4 (1992) 1088.
27. Rogers, D. and Hopfinger, A.J., J. Chem. Inf. Comput. Sci., 34 (1994) 854.
28. Friedman, J.H., Ann. Statist., 19 (1991) 1.
29. ChemicalComputingGroup, Svl Exchange Webpage. http://svl.chemcomp.com/, 2003.
30. Schefzick, S., Kibbey, C. and Bradley, M.P., J. Comb. Chem., 6 (2004) 916.
31. Wold, S., In v.d. Waterbeemd H. (Ed.), QSAR: Chemometric Methods in Molecular Design, Verlag-Chemie, Weinheim, Germany, 1994.
32. Cho, S.J. and Hermsmeier, M.A., J. Chem. Inf. Comput. Sci., 42 (2002) 927.
33. Kubinyi, H., Quant. Struct.-Act. Relat., 13 (1994) 393.

34. Waller, C.L. and Bradley, M.P., J. Chem. Inf. Comput. Sci., 39 (1999) 345.
35. McFarland, J.W. and Gans, D.J., Quant. Struct.-Act. Relat., 13 (1994) 11.
36. Todeschini, R., Dragon, Milan, Italy, 2003.
37. Pearlman, R.S., Tripos Inc., St. Louis, MO.
38. The QSAR and Modelling Society. http://www.ndsu.nodak.edu/qsar_soc/, 2003.
39. Kendall, M. (Ed.), Rank Correlation Methods, Charles Griffin and Co., London, UK, 1955.
40. Brown, M.B. and Benedetti, J.K., J. Am. Stat. Assoc., 72 (1977) 309.
41. Agresti, A. (Ed.), Categorical Data Analysis, John Wiley & Sons, Inc., New York, 1990.
42. Todeschini, R., Consonni, V. and Pavan, M., Chemometr. Intell. Lab. Syst., 70 (2004) 55.