*Journal of Computer-Aided Molecular Design* **18:** 483–493, 2004.
DOI 10.1007/s10822-004-5321-2

# Genetic algorithms and self-organizing maps: a powerful combination for modeling complex QSAR and QSPR problems

Ersin Bayram[a], Peter Santago II[a], Rebecca Harris[b], Yun-De Xiao[b], Aaron J. Clauset[c] & Jeffrey D. Schmitt[b,*]

[a]*Department of Biomedical Engineering, Wake Forest University, Medical Center Blvd., Winston-Salem, NC 27157-1022, USA;* [b]*Molecular Design Group, Targacept, Inc., 200 East First Street, Suite 300, Winston-Salem, NC 27101-4165, USA;* [c]*Computer Science Department, the University of New Mexico, Albuquerque, NM, USA*

Received 18 May 2004; accepted in revised form 28 October 2004

*Key words:* genetic algorithm, neural networks, QSAR, QSPR, supervised self-organizing maps, variable selection

## Summary

Modeling non-linear descriptor-target activity/property relationships with many dependent descriptors has been a long-standing challenge in the design of biologically active molecules. In an effort to address this problem, we couple the *supervised self-organizing map* with the *genetic algorithm*. Although self-organizing maps are non-linear and topology-preserving techniques that hold great potential for modeling and decoding relationships, the large number of descriptors in typical quantitative structure–activity relationship or quantitative structure–property relationship analysis may lead to spurious correlation(s) and/or difficulty in the interpretation of resulting models. To reduce the number of descriptors to a manageable size, we chose the genetic algorithm for descriptor selection because of its flexibility and efficiency in solving complex problems. Feasibility studies were conducted using six different datasets, of moderate-to-large size and moderate-to-great diversity; each with a different biological endpoint. Since favorable training set statistics do not necessarily indicate a highly predictive model, the quality of all models was confirmed by withholding a portion of each dataset for external validation. We also address the variability introduced onto modeling through dataset partitioning and through the stochastic nature of the combined genetic algorithm supervised self-organizing map method using the *z*-score and other tests. Experiments show that the combined method provides comparable accuracy to the supervised self-organizing map alone, but using significantly fewer descriptors in the models generated. We observed consistently better results than partial least squares models. We conclude that the combination of genetic algorithms with the supervised self-organizing map shows great potential as a quantitative structure–activity/property relationship modeling tool.

## Introduction

Because of scientific advances, such as the sequencing of the human genome, the number of potential drug targets is increasing exponentially [1]. Therefore, it is of critical importance to the pharmaceutical industry that fast and accurate means of predicting the biological properties of small molecules be developed, since prototyping is both expensive and time-consuming. QSAR/QSPR methodology seeks to relate calculable numerical descriptors (features) to biological activity/property [2], but investigators have long struggled to find the ideal mapping tool(s). The

*To whom correspondence should be addressed. Tel.: +1-336-480-2124; Fax: +1-336-480-2107; E-mail: jeff.schmitt@targacept.com

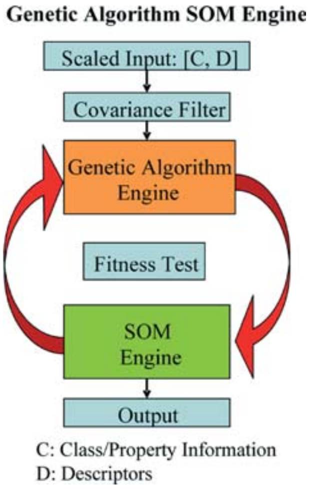

*Figure 1.* Block diagram of the proposed genetic algorithm and supervised self-organizing map based QSAR/QSPR analysis tool.

first, and possibly the most important issue, is that widely utilized linear techniques cannot adequately model the descriptor-target activity/property relationships in biological systems, since the underlying phenomena exist in the non-linear domain. The curse of dimensionality is another major challenge in QSAR/QSPR modeling. The molecular descriptors that comprise QSAR/QSPR models are based on measured or calculated physicochemical descriptors (including linear free energy, graph theoretical, topological, electronic, molecular interaction fields, etc.), and can easily number in the thousands. Indeed, because the predictive ability of standard QSAR/QSPR methodologies usually decreases as a result of having too many descriptors, a descriptor selection strategy is frequently employed. Linear filters utilizing correlation between descriptors, such as multiple linear regression [3] or partial least squares (PLS) [4, 5], offer computationally inexpensive means of reducing descriptor space. Sophisticated, but computationally intensive non-linear selection strategies have also been investigated, as they offer important and exciting advantages. One innovative approach, the genetic functional algorithm (GFA), uses a genetic algorithm to build populations of predictive equations while mutations act on the population to introduce non-linear basis functions [6]. Machine learning techniques such as artificial neural networks have also been widely applied to QSAR data to achieve non-linear mapping [7]. In this paper, we propose a new QSAR/QSPR methodology that couples supervised self-organizing maps (sSOMs) [8] with the genetic algorithm. Figure 1 shows the block diagram of the system. The sSOM clusters compounds according to descriptors and target property, thereby generating a QSAR/QSPR model; while the GA selects descriptors to be included in the model.

Previously, unsupervised SOMs have been used for descriptor reduction prior to analysis by other methods such as PLS [9], feed forward neural networks [10, 11], or the fuzzy ARTMAP [12]; to assess similarity or diversity of compounds [13–15]; and to create comparative feature maps [9, 14]. Pintore et al. have used unsupervised SOMs to cluster compounds according to molecular descriptors [16], and have combined this with a separate fuzzy clustering step (because the SOMs were unsupervised) to obtain bioactivity prediction. Our approach, because it utilizes supervised SOMs, provides a unified solution to descriptor selection and non-linear model generation.

The evolutionary algorithm (EA), of which GA is a subtype, simulates the process of natural selection to find good if not optimal solutions to very complex problems. In EA, a set of potential solutions encoded as chromosomes are evaluated and scored *via* a fitness function. Based on the fitness score, better chromosomes are selected to generate the next generation of solutions or ‘offspring’. EA methods can be classified according to the means of offspring generation: the evolutionary programming (EP) method that only allows asexual reproduction (mutations); and GA, that allows sexual reproduction (crossover operation) for offspring generation as well as mutations. EAs have been applied to the development of QSAR models as a descriptor selection strategy. Leardi et al. [17] have shown that multiple linear regression applied to descriptors selected by GA provides better models than those developed for descriptors selected by stepwise regression. The GFA uses a genetic algorithm to build populations of predictive equations by means of non-linear basis functions [6]. Available functions include linear polynomials, quadratic polynomials, splines, and Gaussians. EP has been applied to the descriptor selection problem, with coefficient calculation by least-squares regression [18]. Kubinyi [19] later utilized a combination of systematic and evolutionary methods to generate QSAR models,

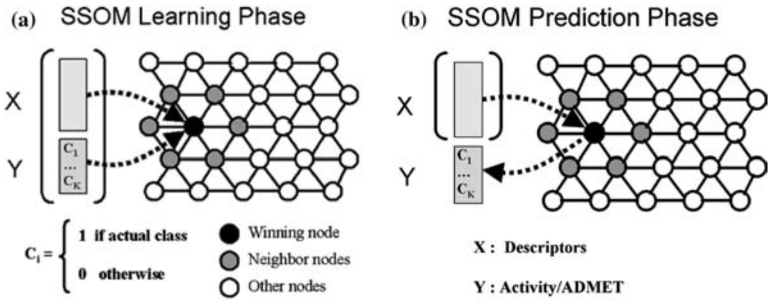

*Figure 2.* sSOM learning and prediction schemes. A hexagonal neighborhood structure is used to define the map's topological connectivity. (a) During the learning phase, actual class information of each training compound is attached to its feature vector in binary format. This combined feature vector is fed into the SOM as input to guide the map organization; (b) during the prediction phase, the map that was created during the learning phase is used to relate the features of the compounds to the unknown class information (activity/property).

where the systematic search was used to limit the number of variables for evolutionary development in an effort to speed up the search process. So and Karplus [20] applied both GFA and EP as descriptor selection tools for neural network based QSAR model generation. Recently, the GA has been applied to improve the non-linear PLS based QSAR modeling [21, 22]. Motivated by the impressive results of GA, we coupled sSOMs with GAs with the aim of developing novel QSAR/ QSPR tools.

## Methods

Matlab 6.5 (Mathworks, Inc., Natick, MA) in combination with the SOM Toolbox [23] was used for all programming [23].

### *Self-organizing maps*

The SOM, also known as a Kohonen network, converts high-dimensional, non-linear statistical relationships into simple geometric relationships in an *n*-dimensional array [8]. This reduced representation seeks to best preserve the input data's original topology and density. Although SOMs are often described as neural networks, it is illuminating to describe them from two different perspectives, neural networks and vector prototyping. A SOM's aim is to find the optimal set of lower dimensional prototype vectors to properly group the high dimensional pattern space (clustering property), while preserving the probability density of the original manifold (topology preservation). The geometric interpretation of this process is that the prototype manifold divides pattern space into cojoint Voronoi cells. The result is that each input vector is mapped to the nearest prototype, known as the best matching unit (BMU), according to a given distance metric; in this paper, Mahalanobis distance metric is used. From a neural network perspective, prototype vectors are the nodes of the network, and BMU is the winning node. Unlike the 'winner takes all' approach found in other neural networks, SOMs update not only the BMU, but also its neighbors according to a weighted neighborhood function. Neighborhood function allows BMU neighbors to be influenced during the training process, resulting in enhanced topology preservation. In our implementation, a hexagonal network structure and Gaussian neighborhood function are utilized.

### *Supervised self-organizing maps*

Supervised self-organizing maps (sSOMs) incorporate *a priori* knowledge into the neural network training process. In this paper, we adopt the definition of the supervised SOM introduced by Kohonen [8] that is depicted in Figure 2. The self-organizing nature of the system dictates training to be a guided self-learning process. During training, class information of each compound ($\boldsymbol{Y} = [C_1, \ldots, C_K]$) is appended to its descriptor vector ($\boldsymbol{X} = [x_1, \ldots, x_D]$) to form the manifold ($\boldsymbol{Z} = [\boldsymbol{X}^{\mathrm{T}}\ \boldsymbol{Y}^{\mathrm{T}}]$). Here, $\boldsymbol{Y}$ represents a single column binary valued vector containing bioactivity class information, where only the class index to which the compound belongs is set to 1. This data model allows class information to influence the topological ordering of the map during training; and then

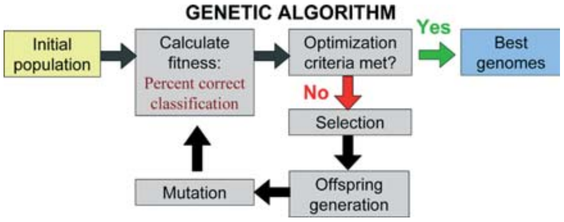

*Figure 3*. The structure of a typical genetic algorithm (GA). GA produces a set of potential solutions rather than a single result and is based on the concept of natural selection. Offspring generation via a crossover operation allows the information exchange between intermediate solutions, while mutations provide an escape mechanism in avoiding local minima. The initial population is selected randomly to avoid bias in the decision mechanism.

the trained map is used for class prediction on unknown compounds. In this paper, Kohonen's batch-training algorithm [8] is used throughout. Training starts with a large neighborhood function to assure proper topological ordering of the SOM, and then the neighborhood kernel is gradually decreased to a specified minimum size during the training.

### *The genetic algorithm*

GA is a stochastic search method mimicking biological evolution that differs substantially from more traditional search and optimization methods. While most stochastic search methods operate on a single solution to a given problem, genetic algorithms operate on a population of solutions (*genomes* or *chromosomes*), applying the 'survival of the fittest' principle to produce ever better solutions. Figure 3 shows the structure of a typical genetic algorithm. With each generation, a new set of solutions is created by selecting the most fit individuals according to a fitness test in the problem domain and breeding them together using mathematical operators inspired by natural selection. This process leads to the evolution of populations of individuals that are better suited to their environment than the individuals that they were created from, just as in natural selection. We decided to use GA as the descriptor selection tool because of the following advantages:

- GA searches a population of solutions in parallel, providing the potential to avoid local minima;
- The objective function and corresponding fitness criteria are sufficient to properly influence the search directionality;
- GA uses probabilistic rather than deterministic transition rules;
- GA can provide a number of potential solutions to a given problem, lending itself well to downstream meta-analysis [6].

### *Genetic algorithm based descriptor selection*

GA is used to reduce the descriptor space to a manageable size for the SOM engine to perform QSAR/QSPR modeling. The fitness function in Figure 1 is further elaborated in Figure 3. Because our goal is to accurately classify compounds based on their bioactivities, a natural choice for a fitness function is the percent correct classification of the training data set by sSOM based on the descriptors selected by the GA. A chromosome of binary strings represents each member of the population of potential solutions. Each descriptor in the manifold is represented by a binary allele value, where 1 indicates that the corresponding descriptor is included in the model and 0 indicates exclusion from the model. The (conservative) number of evolutions chosen in this work (200) is based on Gao's work [24] on GA based binary QSAR analysis. During each evolution, the chromosomes are sorted according to their fitness levels, and the top performers (5% in our implementation) are labeled as 'elite' models and are passed on to the next generation without any modifications. The remaining 95% are bred using crossover operations. The crossover operation starts with the two parent chromosomes and randomly selected allele positions. Each parent is broken into pieces at randomly selected allele positions and the pieces are swapped with a probability of $P_c$ (0.7 in our case). Two new 'offspring' chromosomes containing genetic information from both parents are thus created.

Roulette-wheel selection is used to choose parents for breeding. This stochastic technique is carried out as follows: all chromosomes are mapped to contiguous segments of a unit length line, such that each individual's segment is proportional in size to its fitness. For each offspring to be generated, two random numbers are generated, and the chromosomes whose segments span the random numbers are selected as parents. This method is analogous to a roulette wheel where the size of each slice is proportional to the fitness. Once the offspring are generated, they are subjected to mutation. The mutation operation simply negates the values of randomly selected alleles (2 in our case) with a probability of $P_M$ (0.5 in our case). The evolution process is terminated under one of the following conditions:

1. The specified maximum number of generations is reached (200 in our case); or
2. The algorithm gets stuck in a local minimum, which is characterized by no improvement on best model over a specified number of consecutive iterations (50 in our case).

### *Datasets*

To demonstrate the effectiveness of the sSOM method, six real data sets covering different biological domains were used:

1. $\alpha 4\beta 2$ neuronal nicotinic receptor affinity ($K_i$) for 148 compounds [25];
2. D2 dopamine receptor affinity for 207 compounds [26];
3. D3 dopamine receptor affinity for 207 compounds [26];
4. Dihydrofolate reductase (DHFR) affinity for 256 compounds [27, 28];
5. Topliss dataset for 272 compounds [29];
6. Growth inhibition ($GI_{50}$) data from the most diverse 2400 compounds of the National Institute of Cancer's (NCI) Anticancer Screening Database [30].

Because pharmacologically meaningful bin populations in the congeneric DHFR dataset were heavily skewed, molecules from the more populated bins were randomly eliminated to achieve an even distribution, resulting in a final set of 135 molecules. Figure 4 displays the histograms of the six datasets. The published target property of Topliss is categorical (4-way binning = {1.5, 2.5, 3.5}) and was used without modification. Datasets 1–4 and 6 were partitioned into categorical bins in a way that provided the most even distribution of samples among the bins. The following binning and partitions were used: $\alpha 4\beta 2$, 3-way = {200, 1000}; D2, 3-way = {100, 1000}; D3, 3-way =

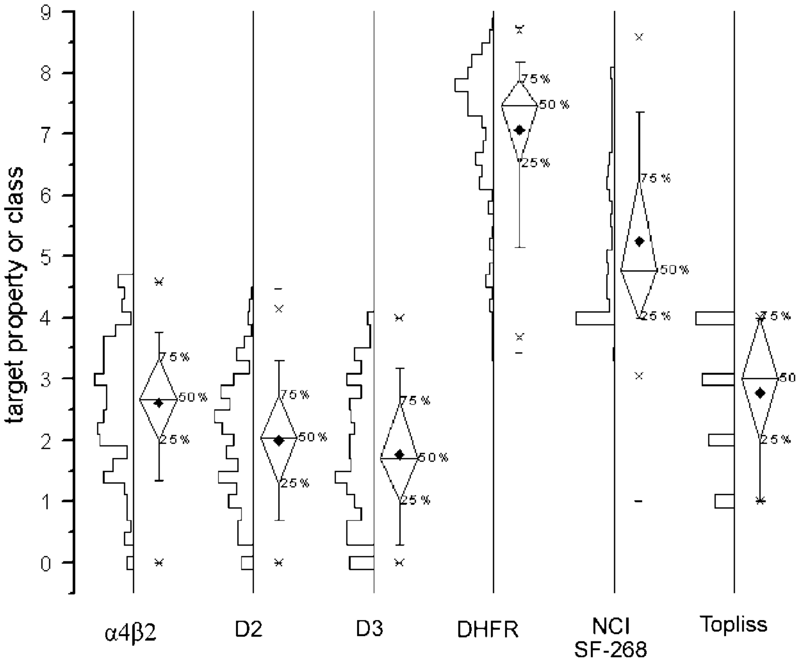

*Figure 4.* Dataset target value distribution. The *y*-axis is log($K_i$) for $\alpha 4\beta 2$, D2, and D3, log($IC_{50}$) for DHFR, log($GI_{50}$) for NCI SF-268, and the actual class distribution for Topliss. For each dataset a box chart and histogram show distribution of target property; 25th, 50th and 75th percentile values label the boxes and whiskers are placed at +/− SD (standard deviation); (◆) data mean; (x) outliers; (–) minimum and maximum values.

{100, 1000}; DHFR, 3-way = {6.75, 7.75}; NCI SF-268, 3-way = {4.2, 6}.

The neuronal nicotinic and dopamine datasets are typical of datasets encountered in drug design and represent a moderate degree of molecular diversity and have been shown to evade attempts at linear modeling with generic 1D- and 2D-descriptors. The D2 and D3 datasets are chosen because the differentiation between them is typical of that encountered during the optimization of drug-like molecules, where specificity towards one of many receptor/enzyme isoforms is critical. The Topliss dataset represents an important ADMET dataset with a high degree of structure diversity and pharmacological activity. In contrast, the well-studied DHFR dataset comprises a congeneric series of molecules and is included as a means to benchmark against published non-linear methods.

The NCI SF-268 dataset is both sparse and non-linear and represents a large, highly diverse set of compounds. The NCI database contains approximately 48,000 compounds and data describing their ability to inhibit growth in 60 different human tumor cell lines [30]. Growth inhibition is expressed as $GI_{50}$, or the concentration that causes a 50% reduction in net protein production. The SF-268 human CNS cell line was selected for this study and approximately 30,000 three-dimensional structures for which data exists were downloaded. Of these, all molecules with heavy and/or metallic atoms (ones that were not handled by the commercial software used for descriptor generation) were eliminated, as were mixtures and those molecules possessing a molecular weight greater than 800. The resulting 27,425 compounds were divided into three groups according to their $\log(GI_{50})$ activity: actives compounds (those with a value of at least 6), moderates (those with a value between 4.2 and 6), and inactives (those with a value less than 4.2). If the database contained more than one $GI_{50}$ value entry for a given compound, then the value with the lowest standard deviation was selected. To keep bin sizes as even as possible and avoid model bias, 800 compounds were selected to form each group. This selection was done using descriptor-space diversity analysis based on Mahalanobis distance.

Datasets contained 201 (α4β2), 151 (D2), 153 (D3), 171 (DHFR), 450 (Topliss) and 562 (NCI) calculated 1D and 2D descriptors derived from QSARIS (version 1.1, MDL Information Systems), Cerius2 (version 4.8, Accelrys), Volsurf (version 3.0.11, Molecular Discovery Ltd.) and Dragon (version 4.0, Milano Chemometrics and QSAR Research Group). By forming a diverse and challenging database for testing our QSAR/ADMET models, we ensure that comparison studies will provide a reliable measure of model performance.

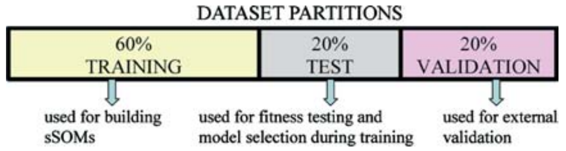
DATASET PARTITIONS
60% TRAINING
20% TEST
20% VALIDATION
used for building sSOMs
used for fitness testing and model selection during training
used for external validation

*Figure 5.* Every dataset is partitioned into three sets: training (60%), test (20%), and validation (20%). Training data are used for building sSOM during training phase; test data are not included in the supervision process, but rather used in the process of selecting models according to their fitness during training. The models were blinded to the validation data, which was used as the ultimate measure of model predictivity (and generalizability).

## Results

To account for sampling error, models were generated from 20 different training/test set partitions. Furthermore, because GAs provide multiple solutions, we have formed each model by merging the results of the top 10 sub-models from each GASOM training procedure. Classification results from the top 10 sub-models are used to vote for the most likely class membership of each observation. The degree to which a given sub-model influences the voting procedure is determined by its fitness. Although QSAR model quality is almost always reported in terms of a training and test set statistics (such as classification rate, $r^2$ and $q^2$), these measures do not necessarily indicate a highly predictive model [31]. For this reason, we used an external validation set in order to provide a rigorous check on model quality. Figure 5 demonstrates the dataset partitioning process and function of each partition. Partitioning was performed in such a way that descriptor and activity diversity were retained within each partition while equal distribution was also preserved among the target bins.

Table 1 summarizes the descriptor reduction behavior of the GASOM. Results indicate that GASOM models only use 15% to 25% of the

*Table 1.* Twenty different models were created with GASOM using different training/test sets, and the number of descriptors after descriptor reduction is reported as mean ± SD of the 20 different models.

| Dataset | Number of descriptors | |
|---|---|---|
| | Initial | After GASOM's descriptor reduction |
| α4β2 | 201 | 41.8 ± 6.5 |
| D2 | 151 | 31.8 ± 4.2 |
| D3 | 153 | 34.8 ± 5.5 |
| DHFR | 171 | 34.5 ± 6.7 |
| Topliss | 450 | 50.6 ± 9.0 |
| NCI SF-268 | 562 | 77.1 ± 9.9 |

available descriptors, enabling a significant reduction in model complexity. This degree of descriptor reduction is particularly useful in model interpretability. As an example, the top five most frequently used descriptors taken from the 20 DHFR models are as follows (percentage of use among the models is given in parentheses): radius of gyration (49%), charge of most positive atom in molecule (51%), sum of all {=C<} E-State values in molecule (52%), calculated LogP (58%), and net molecular dipole (64%).

Table 2 shows the percent correct classification of sSOMs (without descriptor selection), PLS, the GASOMs, and a bootstrapped GASOMs. Bootstrapping is achieved by using our GA algorithm to create new models using 20 different training/test set partitions of the original data. In this procedure we only keep the best model (based on %TS) resulting in a pool of 20 models, each trained on partially different training/test data. These models vote for the most likely class membership of each observation in the external validation data; the degree to which a given model influences the voting procedure is determined by its fitness on the test data. Unfortunately, bootstrapping did not provide a significant improvement on external validation accuracies (see Table 2).

Our results demonstrate that, except for α4β2, sSOMs and GASOMs perform better than PLS, especially in the case of D2, D3, and DHFR. Results also highlight that the accuracies of sSOMs without descriptor reduction and GASOMs on the blinded validation data are within a few percent of each other. Not only are native sSOMs immune to underdetermined datasets, when coupled with GA they have similar predictive power with a significantly reduced number of descriptors (see Table 2). A reduced descriptor space also translates into valuable time savings, as one needs to measure or calculate only 15–25% of the descriptors to predict the target values. The computational cost of descriptor calculation can become significant with large virtual screening exercises. GASOMs without parameter optimization appear to be a flexible and powerful tool, so there is every reason to believe that the results described herein will easily be improved.

Although GASOM's validation set accuracy is comparable to sSOM without descriptor reduction, it is consistently and significantly better on the training and test set results. This observation indicates GASOM's tendency towards overfitting. Because we coupled two very powerful non-linear methods and implemented supervised learning, GASOM managed to perform well on training and test data even though the training period was relatively short (a very conservative number of generations).

We have also analyzed the descriptors used by the GASOM models to assess the extent to which non-linear data contributed to their selection. Figure 6 shows the most frequently used 10 descriptors in generated models. The vertical axis of each graph shows the percent usage, while the horizontal axis represents the linear correlation ($r^2$) of descriptors to the target bioactivity/property. Plots reveal that the GASOM does make use of descriptors with low linear correlation to target bioactivity/property and selects descriptors from a wide spectrum of correlation values. Therefore, we conclude that linear correlation is not a sufficient criterion for descriptor selection. Contextual information – how much a descriptor contributes to the model when used in combination with other descriptors – should be considered in the descriptor selection process.

There are three main sources of GASOM variability: (1) training/test/validation set sampling; (2) the nondeterministic nature of the GA; and (3) the random initialization procedure used to position the prototype vectors in sSOM training. The following experiment describes our effort to assess the relative contribution of these sources. To measure variability arising due to the results of data portioning and the GA, 20 different models were created, each from a different training/test/validation set partition. To assess variation due to

*Table 2.* Classification results for PLS, sSOMs, GASOMs and bootstrapped GASOM (BS GASOM). %Correct: average percent correct classification of 20 different runs with different training and test partitions.

| | Classification power comparison: PLS, sSOM without descriptor reduction, GASOM and bootstrapped GASOM | | | | | |
|---|---|---|---|---|---|---|
| | Training | | Test | | Validation | |
| | %Correct | Best | %Correct | Best | %Correct | Best |
| α4β2 3-way | | | | | | |
| PLS | 49 ± 7 | 56 | 43 ± 7 | 57 | 53 ± 6 | 66 |
| sSOM | 84 ± 3 | 90 | 39 ± 11 | 61 | 46 ± 7 | 59 |
| GASOM | 98 ± 1 | 100 | 57 ± 8 | 74 | 48 ± 10 | 69 |
| BS GASOM | 94 ± 3 | 97 | 60 ± 6 | 70 | 52 | |
| D2 3-way | | | | | | |
| PLS | 35 ± 2 | 41 | 31 ± 6 | 42 | 33 ± 4 | 44 |
| sSOM | 80 ± 2 | 84 | 54 ± 7 | 67 | 51 ± 5 | 61 |
| GASOM | 87 ± 2 | 92 | 65 ± 7 | 76 | 50 ± 6 | 59 |
| BS GASOM | 84 ± 3 | 91 | 65 ± 4 | 73 | 51 | |
| D3 3-way | | | | | | |
| PLS | 24 ± 6 | 37 | 20 ± 6 | 33 | 26 ± 3 | 37 |
| sSOM | 86 ± 2 | 89 | 65 ± 7 | 82 | 60 ± 6 | 73 |
| GASOM | 91 ± 2 | 95 | 77 ± 7 | 91 | 60 ± 6 | 71 |
| Bs GASOM | 89 ± 3 | 94 | 78 ± 6 | 91 | 63 | |
| DHFR 3-way | | | | | | |
| PLS | 58 ± 4 | 68 | 49 ± 6 | 62 | 50 ± 4 | 56 |
| sSOM | 91 ± 2 | 95 | 61 ± 8 | 76 | 65 ± 5 | 70 |
| GASOM | 97 ± 2 | 100 | 68 ± 9 | 86 | 62 ± 4 | 70 |
| BS GASOM | 93 ± 3 | 99 | 74 ± 7 | 95 | 60 | |
| Topliss 4-way | | | | | | |
| PLS | 36 ± 4 | 41 | 33 ± 5 | 42 | 31 ± 3 | 35 |
| sSOM | 74 ± 2 | 78 | 40 ± 6 | 51 | 39 ± 5 | 48 |
| GASOM | 96 ± 2 | 99 | 53 ± 6 | 63 | 41 ± 4 | 48 |
| BS GASOM | 92 ± 2 | 97 | 54 ± 5 | 63 | 44 | |
| NCI SF-268 3-way | | | | | | |
| PLS | 55 ± 1 | 57 | 53 ± 3 | 57 | 53 ± 1 | 54 |
| sSOM | 72 ± 1 | 73 | 55 ± 2 | 60 | 57 ± 2 | 60 |
| GASOM | 86 ± 1 | 88 | 59 ± 2 | 63 | 58 ± 2 | 61 |
| BS GASOM | 82 ± 2 | 85 | 59 ± 2 | 64 | 62 | |

Best: the most accurate classification percentage. %Correct values are given in mean ±SD and rounded to the closest integer. Bootstrapped GASOM results are based on 20 different models generated by 20 randomly selected different partitions of training and test data.

the sSOM initialization, we picked two representatives among the 20 generated models: the best model and the model that is closest in accuracy to the 20 models' average and re-generated 19 more models for both cases using the identical conditions as the initial actual runs. In the end, three different groups of runs had been completed, each possessing 20 models: the 'average repeats', the 'best repeats', and the 'actual runs'.

For each dataset, average absolute *z*-scores for test and validation sets are calculated by averaging the absolute *z*-scores of 20 models in each group (results are given in Table 3). Average absolute *z*-score tests of the actual runs reflect the variation due to data partitioning plus stochastic nature of the GA itself. *Z*-score results on average repeats and best repeats represent the variation only due to the sSOM. Unfortunately, the stochastic nature of the GASOM makes it impossible to relate the *z*-scores on actual runs to those of repeats in a quantitative manner that estimates the randomness due to data portioning alone. That said, the similar

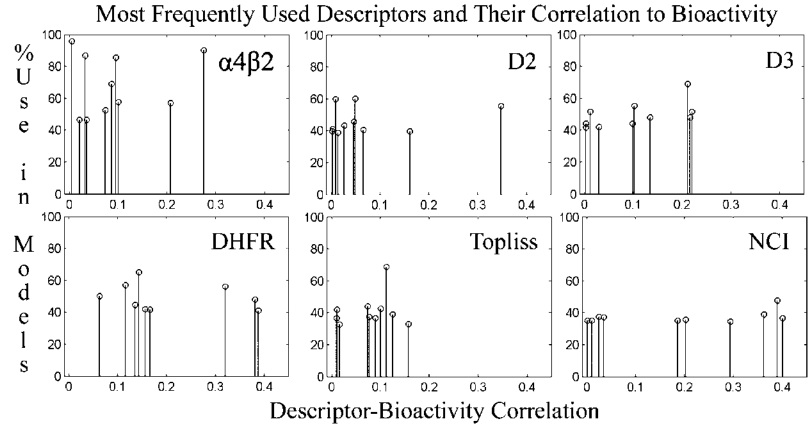

*Figure 6.* Plots of the 10 most frequently GASOM selected descriptors for each dataset *versus* their correlation to the target bioactivity/property ($r^2$). In all datasets, GASOM consistently utilized very low correlation descriptors. For instance, among the top 10 descriptors of the α4β2 dataset, the most frequently used descriptor (almost all of the models used this descriptor) has the lowest correlation to the bioactivity.

*Table 3.* Average absolute *z*-score tests of actual runs reflect the variation due to data partitioning as well as stochastic nature of the method itself.

| | GASOM average absolute *z*-score results on test and validation datasets | | | | | |
|---|---|---|---|---|---|---|
| | Test data runs | | | Validation data runs | | |
| | Actual | Average repeats | Best repeats | Actual | Average repeats | Best repeats |
| α4β2 | 0.782 | 0.730 | 0.812 | 0.805 | 0.790 | 0.733 |
| D2 | 0.799 | 0.808 | 0.811 | 0.782 | 0.789 | 0.702 |
| D3 | 0.786 | 0.858 | 0.814 | 0.719 | 0.726 | 0.753 |
| DHFR | 0.809 | 0.736 | 0.731 | 0.750 | 0.816 | 0.794 |
| Topliss | 0.741 | 0.736 | 0.826 | 0.847 | 0.771 | 0.773 |
| NCI SF-268 | 0.850 | 0.794 | 0.833 | 0.856 | 0.814 | 0.766 |

*z*-Score results on 'average repeats' and 'best repeats' represent the variation due to the stochastic nature of GASOM.

range of values suggests that variation due to data portioning is not significant in these experiments. The degree of observed immunity to sampling error is desirable in training based methods, especially in QSAR because of the limited size training sets. Wilcoxon ranksum was also performed on the pairwise permutations of three groups to test the null hypothesis that distribution medians are equal at the 5% significance level. The results in Table 4, while largely in agreement with the *z*-score results, do show that variation due to the sSOM initialization process is significant in some cases. *In any stochastic approach where method dependent variations might affect results, a cross-validation approach or multiple runs are needed to derive reasonable and robust conclusions.*

In an effort to better understand the relative merits of the sSOM technique, we investigated sSOM's sensitivity to noise. The 40 most frequently used descriptors were selected from the 20 GASOM training runs on the DHFR dataset. The performance of the sSOM was evaluated as the descriptor manifold was gradually diluted with scrambled descriptors (row vectors were scrambled). Figure 7 shows the external validation accuracy of sSOM with increasing noise levels. Results demonstrate that up to 60% noise sSOM model predictivity is robust. We are currently investigating this promising observation with other datasets to see if sSOMs, in general, provide robust classification with noisy QSAR data.

*Table 4*. The null hypothesis ‘medians are equal’ is tested at the 0.05 confidence level.

| | GASOM Wilcoxon ranksum test results on test and validation datasets | | | | | |
|---|---|---|---|---|---|---|
| | Test data runs | | | Validation data runs | | |
| | Actual vs. Average Repeats | Actual vs. Best Repeats | Average vs. Best Repeats | Actual vs. Average Repeats | Actual vs. Best Repeats | Average vs. Best Repeats |
| α4β2 | 0.264 | 0.781 | 0.451 | 0.171 | 0.158 | 0.002 |
| D2 | 0.039 | 0.059 | 0.000 | 0.058 | 0.003 | 0.000 |
| D3 | 0.123 | 0.000 | 0.000 | 0.457 | 0.629 | 0.912 |
| DHFR | 0.005 | 0.003 | 0.653 | 0.006 | 0.684 | 0.103 |
| Topliss | 0.291 | 0.325 | 0.945 | 0.902 | 0.184 | 0.193 |
| NCI SF-268 | 0.006 | 0.000 | 0.464 | 0.626 | 1.000 | 0.456 |

The numbers represent the confidence levels of the corresponding ranksum tests and gray shaded boxes point to the tests for which the null hypothesis can be rejected.

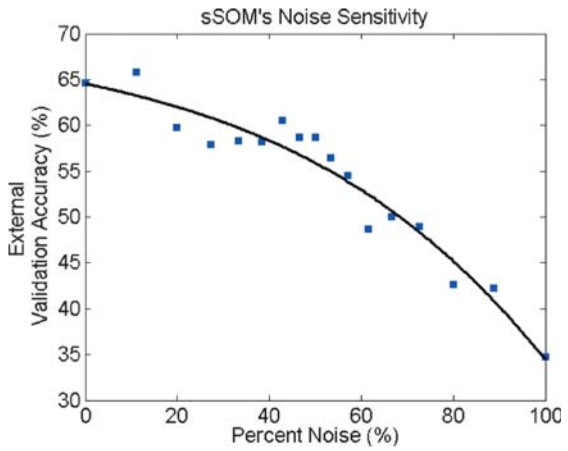

*Figure 7*. Noise sensitivity of sSOM on the DHFR dataset. The solid line represents best-fit using a cubic polynomial to the noise (■) results. Accuracy on external validation data gradually decreases with increasing noise level.

## Discussion and conclusions

A novel QSAR analysis tool has been proposed that couples GA based descriptor selection with sSOM QSAR modeling. This method has been compared to commonly used linear PLS methodology using six different real datasets, of moderate-to-large size and moderate-to-great diversity, and with varying biological endpoints. The results conclusively demonstrate the promising predictive power of the supervised GASOM methodology over PLS.

The results reported herein demonstrate that the GA is an effective tool for descriptor selection. The fitness function that drives the GASOM engine could be modified to favor models with even fewer descriptors in cases where simpler and more explanatory models may be required. In our experiments, GASOM training was driven solely by accuracy. Thus, more complicated models were picked even if they provided only modest accuracy increase. After adding a complexity penalty term to the model selection criteria, typical model sizes were reduced from around 30 descriptors to 8–15 descriptors (data not shown). The penalty term employed was very simple: the ratio of the number of descriptors in the model to the total number of descriptors. We are currently investigating the use of alternative penalty functions.

Our results also reveal that the GASOM is overfitting to the training data – many of our models did not generalize well to the external validation sets – even when bootstrapping is employed. In order to conduct the best possible validation of models, we tested GASOM on real QSAR datasets. But, with the exception of the NCI dataset, the training data were overdetermined enough (few observations, many descriptors) to set the stage for overfitting. Insufficient data is a problem typical of QSAR modeling techniques, often exacerbated by the complexity and non-linear nature of the QSAR data. If we had used synthetic data and had tested the performance on a simpler problem, we anticipate that the performance decrement would have been much smaller. Interestingly, there appears to be no strong relationship between number of observations in a given dataset and degree of overfitting (although the NCI dataset showed the lowest value for {training% correct − validation% correct}).

More importantly, the sSOM itself may drive overfitting. The training mechanism is a guided self-learning process, in that class assignments are attached to descriptor vectors during training. This process creates competition between preservation of topology and the creation of a lookup table. Because we have done little to optimize the sSOM parameters, it is likely that the latter is winning out, leading to poor generalization. This has proven to be a challenging but resolvable problem for the GASOM. Besides implementing a simple penalty term on model complexity, we have conducted a preliminary analysis of increasing the number of generations with encouraging results. Even with overfitting, external validation indicates that GASOM has overall higher accuracy than the PLS approach. It is important to restate that very little has been done to systematically optimize the parameters associated with the GASOM.

The results reported here highlight the need for rigorous testing through external validation in QSAR/QSPR modeling; as we and others [32] have previously reported, internal test sets tend toward overestimating model quality. To *get it right*, experimental design should include not only external validation procedure(s) but also assess impact of stochastic components in the modeling method, if they exist.

## Acknowledgements

This work is supported by the National Science Foundation under Grant No. 0139787. Authors are thankful to reviewers for their time and conscientious effort in providing useful suggestions.

## References

1. Barnett, S., Silicon Rally: The race to e-R&D, Pharma 2005, PriceWaterhouseCoopers, 1999.
2. Hansch, C., Acc. Chem. Res., 2 (1969) 232.
3. Draper, N.R. and Smith, H., Applied Regression Analysis. Wiley, New York, 1998.
4. Lindberg, W., Persson, J.A. and Wold, S., Anal. Chem., 55 (1983) 643.
5. Geladi, P. and Kowalski, B.R., Anal. Chim. Acta, 185 (1986) 1.
6. Rogers, D.R. and Hopfinger, A.J., J. Chem. Inf. Comput. Sci., 34 (1994) 854.
7. Simon, V., Gasteiger, J. and Zupan, J., J. Am. Chem. Soc., 115(20) (1993) 9148.
8. Kohonen, T., Self-Organizing Maps, 3rd edn., Springer, Berlin, 2001.
9. Polanski, J., Acta Biochim. Pol., 47 (2000) 37.
10. Kovalishyn, V.V., Tetko, I.V., Luik, A.I., Ivakhnenko, A.G. and Livingstone, D.J., Proceedings of the 12th European Symposium on Quantitative Structure–Activity Relationships: Molecular Modeling and Prediction of Bioactivity, August 23–28, 1998, pp. 444–445, 2000.
11. Agrafiotis, D.K. and Lobanov, V.S., J. Chem. Inf. Comput. Sci., 40 (2000) 1356.
12. Espinosa, G., Yaffe, D., Arenas, A., Cohen, Y. and Giralt, F., Ind. Eng. Chem. Res., 40 (2001) 2757.
13. Rose, V.S., Macfie, H.J.H. and Croall, I.F., QSAR: Ration. Approaches Des. Bioact. Compd., 16 (1991) 213.
14. Anzali, S., Gasteiger, J., Holzgrabe, U., Polanski, J., Sadowski, J., Teckentrup, A. and Wagener, M., Pers. Drug Discov. Design, 9 (1998) 273.
15. Bernard, P., Golbraikh, A., Kireev, D., Chretien, J.R. and Rozhkova, N., Analusis, 26 (1998) 333.
16. Pintore, M., Taboureau, O., Ros, F. and Chretien, J., Eur. J. Med. Chem., 36 (2001) 349.
17. Leardi, R., Boggia, R. and Terrile, M., J. Chemom., 6 (1992) 267.
18. Luke, B.T., J. Chem. Inf. Comput. Sci., 34 (1994) 1279.
19. Kubinyi, H., Quant. Struct.-Act. Relat., 13(3) (1994) 285.
20. So, S.S. and Karplus, M., J. Med. Chem., 39 (1996) 1521.
21. Li, T., Mei, H. and Cong, P., Chemometr. Intell. Lab. Syst., 45 (1991) 177.
22. Tang, K. and Li, T., Chemometr. Intell. Lab. Syst., 64 (2002) 55.
23. Vesanto, J., Himberg, J., Alhoniemi, E. and Parhankangas, J., In Proceedings of the Matlab DSP Conference 1999. pp. 35–40, Espoo, Finland, 1999.
24. Gao, H., J. Chem. Inf. Comput. Sci., 41 (2001) 402.
25. Schmitt, J.D., Curr. Med. Chem., 7 (2000) 749.
26. Hammond, P.S., Cheney, J.T., Johnston, D.E., Ehrenkaufer, R.L., Luedtke, R.R. and Mach, R.H., Med. Chem. Res., 9 (1999) 35.
27. Hansch, C., Silipo, C. and Steller, E.E., J. Pharm. Sci., 64 (1975) 1186.
28. Andrea, T.A. and Kalayeh, H., J. Med Chem., 34 (1991) 2824.
29. Yoshida, F. and Topliss, J.G., J. Med. Chem., 43 (2000) 2375.
30. National Cancer Institute Anti-cancer Screen Database, http://dtp.nci.nih.gov/docs/cancer/cancer_data.html
31. Golbraikh, A. and Tropsha, A., J. Comput.-Aided Mol. Des., 17 (2003) 241.
32. Tropsha, A., Gramatica, P. and Gombar, V.K., QSAR Comb. Sci., 22 (2003) 69.