Supervised Self-Organizing Maps in Drug Discovery. 2. Improvements in Descriptor Selection and Model Validation

Yun-De Xiao,*,† Rebecca Harris,† Ersin Bayram,‡,§ Peter Santago II,‡,§ and Jeffrey D. Schmitt†

Molecular Design Group, Targacept Inc., 200 East First Street, Suite 300, Winston−Salem,
North Carolina 27101-4165, Department of Biomedical Engineering, Wake Forest University School of
Medicine, Medical Center Boulevard, Winston−Salem, North Carolina 27157-1022, and Virginia Tech−Wake
Forest University School of Biomedical Engineering and Sciences, Medical Center Boulevard,
Winston−Salem, North Carolina 27157-1022

Received March 11, 2005

The modeling of nonlinear descriptor−target relationships is a topic of considerable interest in drug discovery. We, herein, continue reporting the use of the self-organizing map—a nonlinear, topology-preserving pattern recognition technique that exhibits considerable promise in modeling and decoding these relationships. Since simulated annealing is an efficient tool for solving optimization problems, we combined the supervised self-organizing map with simulated annealing to build high-quality, highly predictive quantitative structure−activity/property relationship models. This technique was applied to six data sets representing a variety of biological endpoints. Since a high statistical correlation in the training set does not indicate a highly predictive model, the quality of all the models was confirmed by withholding a portion of each data set for external validation. Finally, we introduce new cross-validation and dynamic partitioning techniques to address model overfitting and assessment.

INTRODUCTION

With the extraordinary cost and effort required to bring drugs to the market, the pharmaceutical industry continuously strives to find ways to improve the efficiency of the discovery process and to accelerate drug development. Novel computational approaches and methodologies are making drug discovery more efficient, particularly in ligand-based design where the quantitative structure−activity relationship (QSAR) plays a critical role. QSARs are mathematical models approximating the often-complex relationships between molecular descriptors and biological activities. QSAR theory is based on the well-founded tenet that structural descriptors are the basis of observed properties and on the possibility that a structure may be represented by numerical descriptors. *in silico* methodologies built around QSAR theory allow effective prediction of the physical and biological properties of drug candidates, thereby reducing costly laboratory efforts and animal testing. Increasingly sophisticated ways of representing molecular descriptors as well as new development methodologies are allowing investigators to extract clues about which chemical descriptors are likely determinants of biological activity, thereby guiding the design of drugable compounds with fewer side effects. In the face of escalating research and development costs and time to market, researchers rely more and more on QSAR tools not only to find promising candidates faster but also to purge potential failures before they become a drain on scarce time and resources.

Despite their utility, widely used QSAR techniques still exhibit a number of shortcomings related to both the means of molecular representation (i.e., descriptors, fingerprints, etc.) and the statistical methods by which models are developed and validated. One of the major challenges is that widely utilized statistical methods, such as multiple linear regression (MLR) or partial least squares (PLS), are linear techniques that cannot be expected to adequately model the mostly nonlinear feature−target relationships in biological systems. Another is the curse of dimensionality[1] (or overdetermination), in which the dimension of input descriptors is much higher than the number of training samples, and many QSAR models are unable to produce statistically significant decision boundaries. Not only does the predictive ability of a model suffer as a result of having too many descriptors but interpretation is made more difficult, and overfitting of the training data is likely.

The first of these two challenges has been met, in part, by the adaptation of pattern recognition methods to QSAR. One innovative approach to exploring nonlinear relationships, the genetic functional algorithm (GFA), uses a genetic algorithm to build populations of predictive equations while mutations act on the population to introduce nonlinear basis functions.[2] Machine-learning techniques such as artificial neural networks (ANNs) have also been widely applied to QSAR data to achieve nonlinear mapping.[3]

To overcome the second of these challenges, QSAR methodologies typically include a feature selection strategy, which can not only increase the effectiveness of the model but enhance domain interpretability as well. Indeed, feature selection strategies are utilized with many inference models, including supervised and unsupervised machine-learning techniques such as classification, regression, time series

* Corresponding author e-mail: yun-de.xiao@targacept.com.
† Molecular Design Group, Targacept Inc.
‡ Wake Forest University School of Medicine.
§ Virginia Tech−Wake Forest University School of Biomedical Engineering and Sciences.

10.1021/ci0500841 CCC: $33.50
Published on Web 12/10/2005

prediction, clustering, and so forth. As with QSAR, the objective of feature selection is 2-fold:[4] improving prediction performance and enhancing understanding of the underlying concepts in the induction model. Basic linear filters such as elimination based on feature intercorrelation offer computationally inexpensive means of reducing feature space. Bayesian methods have also been commonly used as search engines for the feature selection process.[5−7] More computationally intensive iterative procedures such as stepwise, hill climbing, genetic algorithms, or simulated annealing (SA) often provide more robust selection strategies.

In this study, we address both challenges via a coupling of simulated annealing and the supervised self-organizing map (sSOM). The nonlinear self-organizing map (SOM)[8] attempts to map high dimensional metric space into a low dimensional (in general, 2D) representation space. The SOM trains itself on the basis of weight vector adaptation with respect to the input vectors and is a highly effective tool for clustering and data visualization across a broad spectrum of applications.[9] Such maps have previously been used in the chemical[10,11] and QSAR[12−14] arenas. The sSOM takes class information into account during the learning phase and can, therefore, be utilized to create models for the prediction of class identity.[9] Our previous investigation has determined that sSOMs are well-suited for nonlinear QSAR studies and offer important advantages over existing QSAR methodologies.[15] Although the sSOM is a robust (that is, relatively insensitive to noise and feature redundancy) nonlinear approach,[15] here, we are motivated to couple it with an efficient feature selection algorithm (simulated annealing) with the aim of improving the predictive power and interpretability of the resulting models.

In general, leave-one-out (LOO) cross-validation[16] is an elegant and straightforward technique for estimating classification error rates and predicted residual sums of squares. The estimated squared correlation coefficient, often denoted q^2, can be utilized in the variable selection procedure as the fitness criterion. However, the computation of the LOO statistic usually entails a large computational expense. LOO also tends to include unnecessary components in models and has proven[17] to be asymptotically incorrect. Furthermore, the method does not work well for data with strong clusterization[18] and underestimates the true predictive error.[19] To reduce the measurement redundancy and to minimize computational intensity, a Monte Carlo approach for cross-validation is herein introduced along with our SA descriptor selection procedure. For a true assessment of model predictive ability, a dynamic partitioning approach was designed and utilized to assess generated models.

MATERIALS AND METHODS

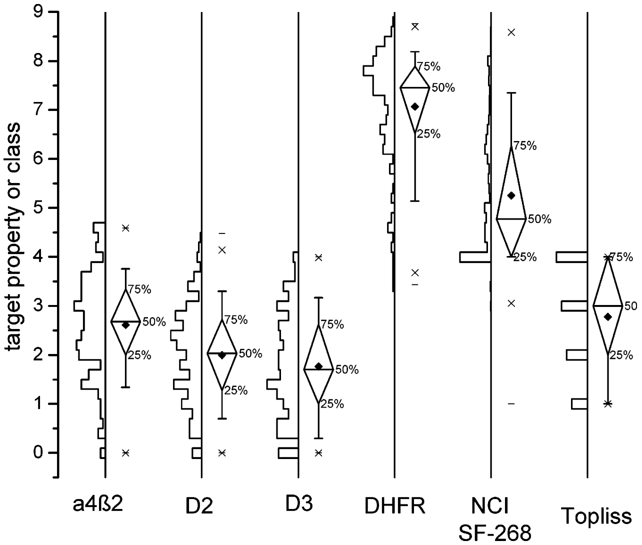

Figure 1. Data set target value distributions. For each data set, a box chart and histogram show the distribution of the target property: (♦) data mean, (×) outliers, (−) minimum and maximum values.

Table 1. Binning of Data Sets by Biological Property

Data Set	Bin 1	Bin 2	Bin 3	Bin 4
α4β2	<200	[200, 1000)	≥1000	
D2	<100	[100, 1000)	≥1000	
D3	<100	[100, 1000)	≥1000	
DHFR	<6.75	[6.75, 7.75)	≥7.75	
Topliss	<1.5	[1.5, 2.5)	[2.5, 3.5)	≥3.5
NCI	<4.2	[4.2, 6.0)	≥6.0	

Data Sets. Six published QSAR data sets of pharmacological interest, representing varied biological endpoints and varied degrees of molecular diversity, were utilized. Three data sets (α4β2 neuronal nicotinic receptor[20] and D2/D3 dopamine receptors[21−26]) represent a moderate degree of diversity and have been shown to evade attempts at linear modeling with generic 1D and 2D descriptors; one data set (Topliss oral bioavailability[27]) is highly diverse, and the remaining two [dihydrofolate reductase (DHFR) inhibition[28] and National Cancer Institute (NCI) Anti-Cancer Screen Database growth inhibition (GI_{50}) data for the SF-268 human CNS cancer cell line (NCI GI_{50})[29]] are inhibition endpoints, with DHFR being a well-studied series of congeneric molecules and NCI GI_{50} being sparse, very large, and diverse. Because SOM is a classification methodology, all data sets were divided according to pharmacologically meaningful classes (herein referred to as bins) of low, moderate, and high biological activity, and a bin number was assigned to each molecule. For two data sets (DHFR and NCI GI_{50}), some molecules were eliminated in order to create evenly distributed bin populations. Box charts and target variable histograms are given in Figure 1, and bin cutoffs are listed in Table 1.

Descriptors. Calculated 1D and 2D descriptors were derived from commercially available software: QSARIS,[30] Cerius2,[31] Volsurf,[32] and Dragon.[33] The number of descriptors was reduced by simulated annealing (see below), leaving only those descriptors most relevant to target prediction in the nonlinear domain.

Data Partitioning. Data sets were partitioned into three sets: training, test, and external validation sets. Although QSAR methodology is almost always reported in terms of a training and test set only, we withheld an external validation set in order to provide an additional rigorous check on model quality. We feel this is necessary since a high statistical correlation on the training and test sets does not necessarily indicate a highly predictive model.[34] To properly partition our data sets so that they each reflect the makeup of the original data set as much as possible, we take into account the distribution of both feature diversity and biological activity as we form our training, test, and external validation

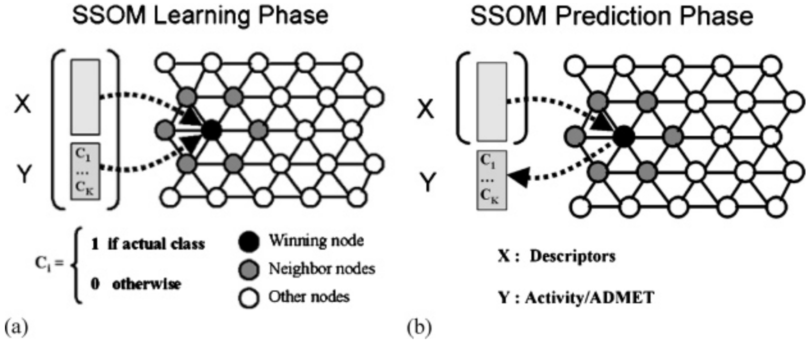

Figure 2. Supervised SOM learning and prediction schemes. A hexagonal neighborhood structure is used for topological connections. (a) During the learning phase, actual class information of each training compound is attached to its descriptor vector in binary format. This combined descriptor vector is fed into the SOM as input to guide the map's organization. (b) During the prediction phase, the map that was created during the learning phase is used to relate the descriptors of the compounds to the unknown bioactivity/property class levels.

sets. In this way, we maintain the original proportions of categorical bins and structural diversity in each of the three sets.

Assessment of Model Quality. In this study, training model quality is simply the percent correct classification (binning) of the target property for the training set (%train). The overall predictive power of a given model is the percent correct classification for the test set (%test) and for the external validation set (%validation), where the external validation set represents naïve holdout data. In our previously published paper, we demonstrated the robustness of this approach using the χ^2 statistic.[15] More extensive model assessment was accomplished by a "dynamic partitioning" procedure (see below), which provides a nonerror rate[35] of the test and external validation sets.

Self-Organizing Maps. A SOM is a neural network method useful for analyzing and visualizing high-dimensional data. In general, unsupervised learning is used to map nonlinear statistical relations between high-dimension input data into a low-dimension lattice, typically in one or two dimensions.[9] The map consists of a regular grid of processing units $m_i(t)$, called "neurons", which can also be understood as prototype vectors.[36] SOMs seek to preserve topology by mapping points that are near each other in the input data space to nearby map units in the lower-dimension map space. The practical result is that each input vector is mapped, according to a distance metric, to the nearest prototype, known as the best matching unit (BMU). Fitting of the model vectors is usually carried out by an iterative optimization process, where $t = 1, 2$, and so forth is the step index. For each observation $x(t)$, the BMU is first identified by the condition

$$\forall_i, ||x(t) - m_c(t)|| \leq ||x(t) - m_i(t)||$$

Then, all model vectors, or a subset centered around the BMU $c = c(x)$, are updated using the Kohonen rule:

$$m(t + 1) = m_i(t) + h_{c(x),i}[x(t) - m_i(t)]$$

Here, $h_{c(x),i}$ is the neighborhood function (or kernel), which is a decreasing function of the distance between the ith and cth unit on the map grid. In our implementation, a hexagonal neighborhood structure is utilized along with a Gaussian neighborhood function. For detailed technical background, see our previous paper[15] or relevant literature.[9]

Supervised Self-Organizing Maps. The most widely used variant of the SOM is the well-known unsupervised SOM, which uses only the independent variables of the data set. In this manuscript, we use the supervised SOM,[37] which takes information about the class identity into account in the learning phase. Because of the self-organizing nature of the system, training is a learning phase during which the class information of each compound ($\mathbf{Y} = [C_1, ..., C_K]$) is appended to its feature vector ($\mathbf{X} = [x_1, ..., x_D]$) to form the manifold ($\mathbf{Z} = [\mathbf{X}^T\mathbf{Y}^T]$). **Y** represents a column binary vector or vectors 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; then, the trained map is used for predicting the unknown **Y** dimension (Figure 2). In this manuscript, Kohonen's batch-training algorithm[38] is used throughout. Training starts with a large neighborhood function to ensure proper topological ordering of the SOM, and the neighborhood kernel is decreased with each iteration to a specified minimum size.

Descriptor Selection based on Simulated Annealing. In this paper, SA is employed to perform feature selection. SA is a global, multivariate optimization technique based on the Metropolis Monte Carlo search algorithm.[39] The concept is inspired by the manner in which liquids freeze or metals recrystallize in the process of annealing. In an annealing process, a real or abstract object is repeatedly heated and cooled, where cooling is conducted such that the system at any time approximates thermodynamic equilibrium. As cooling proceeds, the system becomes more ordered and approaches a "frozen" ground state at $T = 0$. Kirkpatrick and co-workers[40] observed the analogy between finding minimum energy states in a physical system and finding minimum cost configurations in a combinatorial optimization problem. Since then, research on SA has boomed with regards to both theory and applications. The earlier applications of SA in QSAR were used in conjunction with MLR,[41] but a recent report indicates that an improvement also results from its use with the correlation methods such as PLS,[42] adaptations of k-nearest neighbors (kNN),[42−45] or ANNs.[46,47]

In a typical implementation, SA starts from an initial state with a random selection of descriptors and walks through the state space (the feature manifold) by a series of small stochastic steps (descriptor additions or deletions). Using the selected descriptors, an objective function is constructed for each state representing its energy or fitness, which, in our case, is the percentage of correct classification of the training or test set by the sSOM model. While a downhill transition (i.e., where the new model is better than the previous one) is always accepted, an uphill transition is accepted with a probability that is inversely proportional to the energy (negative of the fitness) difference between the two states. This probability is computed using Metropolis' acceptance criterion $p = e^{-\Delta E/(KT)}$, where K is a constant used for scaling purposes and T is an artificial temperature factor that controls the ability of the system to overcome energy barriers. If no downhill transition is found for a given state within a specified number of steps (20, in our case), the temperature is reduced and the search resumes at the beginning point of that state again. The state space exploration continues until no change of the energy is observed or until a specific convergence criterion is met.

Feature Selection based on SA with Cross-Validation. Because overfitting commonly occurs when a correct classification rate based on the training set (%train) is used as the objective function for feature selection, a cross-validation procedure may be used. A simple and commonly used method of cross-validation in chemometrics is the LOO method. This method predicts the target value for each observation in the data set by constructing a model from all other observations. LOO quickly becomes cost-prohibitive with increasing data set size. In addition, this method does not work well for data sets where closely related samples are present[18] and, as a consequence, underestimates the actual predictive error.[19] Because of these shortcomings, we developed a Monte Carlo approach for cross-validation, described as follows:

(1) Randomly select a specified number of descriptors (30, in this case) to generate an initial sSOM model.

(2) Calculate the fitness by our Monte Carlo cross-validation procedure, as follows:

a. Randomly pick 20% of the training set observations to hold out for prediction.

b. Build a sSOM model with the remaining observations and the selected descriptors.

c. Predict the biological properties of the holdout observations using the sSOM model.

d. Repeat steps a−c K times ($K = 20$, in our case).

e. Calculate a fitness score referred to as the nonerror rate (NER),[35] which is the aggregate percentage of correctly classified observations:

$$\mathrm{NER} = 100\sum_{K} m_i / \sum_{K} n_i$$

where m_i and n_i are the number of correctly predicted and the total number of observations in the ith holdout set, respectively.

(3) Change the trial solution descriptor set (from step 1) with a small stochastic perturbation (add and/or delete one or two descriptors using an unbiased random selection process). This selection technique is carried out as follows: two random numbers are generated with a binary allele value, where 1 indicates that a descriptor is randomly selected from the excluded descriptor pool and added to the trial solution descriptor set and 0 indicates a random deletion from the trial solution descriptor set.

(4) Calculate the new fitness ($\mathrm{NER}_{\mathrm{new}}$) as in step 2 for the new model.

(5) Apply the optimization criterion: if $\mathrm{NER}_{\mathrm{new}} > \mathrm{NER}$, the new trial is accepted; if $\mathrm{NER}_{\mathrm{new}} < \mathrm{NER}$, then the new solution is accepted with a probability, $p = e^{-\Delta E/T}$, where $\Delta E = \mathrm{NER} - \mathrm{NER}_{\mathrm{new}}$ and T is the current artificial temperature.

(6) Repeat steps 3−5 until the termination condition is satisfied. With each step, lower the artificial temperature[48] whenever a new solution is accepted (see step 5) or after a preset number of successive steps (20 in our case). The new temperature is calculated as $T = C_d T_{\mathrm{old}}$, where C_d is the rate of temperature decrement ($C_d = 0.9$ in these experiments). The process is terminated if the temperature is lower than a preset value (at a resolution of 10^{-6}).

Model Assessment via "Dynamic Partitioning". Conventionally, QSAR model quality is almost always reported in terms of training and test statistics, such as model, or training, r^2 (in our case, %train); q^2 (NER); and test r^2 (%test). However, it has been demonstrated that leave-one-out q^2 is not necessarily adequate to assess the predictive ability of models[49] and may be biased toward overestimating the general ability of the model.[50] It is also recognized that large random variability is frequently encountered when using a locked test set.[51] The true assessment of a model is how well it predicts target values from a representative population that has not been used in the model construction itself. For this reason, we designed a "dynamic partitioning" procedure, a bias-free assessment of the model quality, to evaluate derived models. The detailed implementation is as follows:

(1) Randomly divide the entire data set (along with the selected descriptor set) into training (60%), test (20%), and external validation (20%) sets.

(2) Build a sSOM model on the training set.

(3) Predict the test set target values using the sSOM model, and record the size of the test set (n_i^t) and the number of correctly classified observations (m_i^t) by the model.

(4) Merge the training and test sets and build a new sSOM model on the combined set.

(5) Predict the external validation set using the sSOM model from step 4. Record the size of the external validation set n_i^v and the number of correct classification observations m_i^v by the new model.

(6) Repeat steps 1−5 K times (20, in our case).

(7) Calculate test NER ($100 \times \sum_{i=1}^{K} m_i^t / \sum_{i=1}^{K} n_i^t$) and external validation NER ($100 \times \sum_{i=1}^{K} m_i^v / \sum_{i=1}^{K} n_i^v$); the higher they are, the better the model.

The sSOM-QSAR Algorithm. In general, our sSOM-QSAR method employs the sSOM combined with descriptor selection procedures that seek to find the optimal subset of descriptors from the original descriptor manifold. Partitioning data sets into training, test, and external validation sets rigorously assesses model quality. We extend this methodology by implementation of the dynamic model assessment described above. A flowchart of the sSOM-QSAR algorithm is provided in Figure 3, which involves the following steps:

(1) Divide each data set into two parts: one used to build models (80% of the observations), the other (20%) to validate models (external validation set); in our imple-

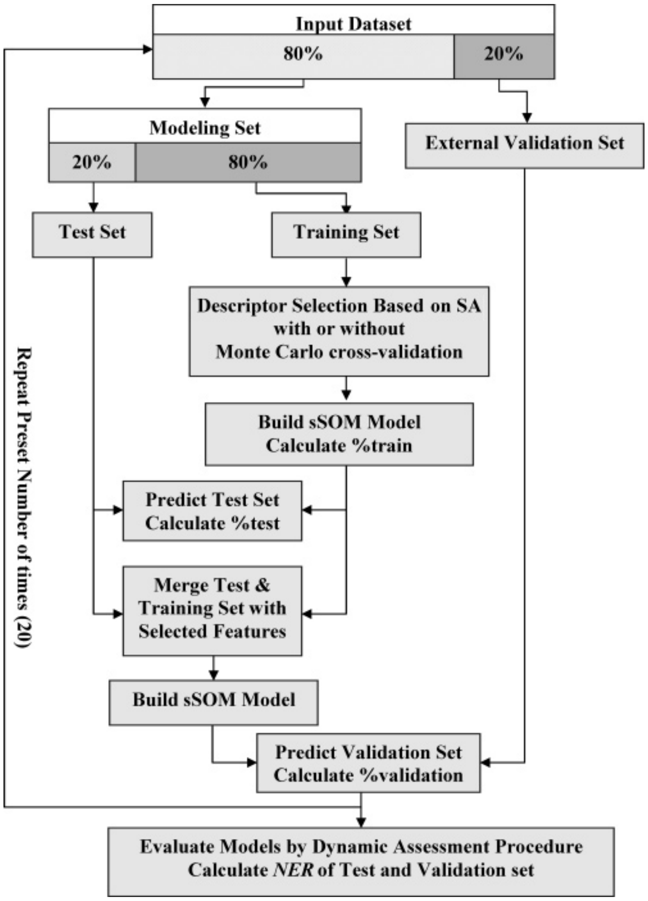
Input Dataset
80%
20%
Modeling Set
20%
80%
External Validation Set
Test Set
Training Set
Descriptor Selection Based on SA with or without Monte Carlo cross-validation
Build sSOM Model Calculate %train
Predict Test Set Calculate %test
Merge Test & Training Set with Selected Features
Build sSOM Model
Predict Validation Set Calculate %validation
Repeat Preset Number of times (20)
Evaluate Models by Dynamic Assessment Procedure Calculate *NER* of Test and Validation set

Figure 3. Flowchart of the sSOM-QSAR Algorithm

mentation, the external validation set is selected to have a high level of diversity (see discussion of data set partitioning, above).

(2) Further partition the 80% identified for model building to form two more sets: training (80%) and test (20%) sets.

(3) Select an optimized subset of descriptors using a stochastic technique (in this case, SA) based on the training set with or without the Monte Carlo cross-validation procedure (described above); then, train the sSOM.

(4) Use the sSOM model from the previous step to calculate the training set NER and percent correct classification (%train).

(5) Predict the test set target values using the sSOM model and calculate the percent correct classification of the test set (%test).

(6) Merge the training and test sets, and build a new sSOM model.

(7) Predict external validation set target values using the new model, and calculate the percent correct classification of the external validation set (%validation).

(8) Repeat steps 1−8 a preset number of times (in our case, 20).

(9) Assess each model (i.e., its selected descriptors) by the dynamic partitioning process described above, and generate test and external validation NERs.

RESULTS AND DISCUSSION

It is well-known that the quality of QSAR models is strongly affected by data set partitioning. Careful partitioning of data sets allowed us not only to build highly predictive models but also to rigorously assess them. To properly partition our data sets into training, test, and external validation sets, their biological activity and descriptor diversity were taken into account so that each subset had almost the same distribution as the data set as a whole (see Figure 1). Furthermore, to account for sampling error, models were generated from 20 different training/test/external validation set partitions.

To set a baseline for comparison, conventional PLS and sSOM models were built on training sets using the entire descriptor space (described above) for all six published data sets of pharmacological interest. Predictions of test and external validation sets were obtained from the relevant models (Table 2). When compared with the PLS baseline, sSOMs consistently outperformed PLS based on the conventional criterion of predictive ability on test sets (see classification results of the bold-faced values for the test set in Table 2). For example, the %test of the best D2 model for PLS and sSOMs are 42.4% and 66.7%, respectively. The results reported herein are consistent with those of our previous study.[15] It is important to note that the PLS models are trained on higher-resolution information (i.e., actual bioactivity values rather than bin numbers) than sSOM models for all data sets except Topliss, for which actual bioactivity values were not available. For five of the six data sets, sSOMs also demonstrated superior performance on external validation sets (see the external validation column and bold-faced values in Table 2).

For ease of interpretation, a good model should be as parsimonious as possible; that is, it should contain the minimum number of components (original descriptors) necessary to maintain accurate predictive ability. Since an exhaustive evaluation of all possible descriptor subsets is usually computationally intractable, some method of descriptor selection is typically employed. We utilized the SA strategy for descriptor selection to build QSAR models for the six data sets. To demonstrate the effectiveness of this methodology, two experiments with different protocols, both with and without Monte Carlo cross-validation, were conducted (i.e., MC-SA-sSOM and SA-sSOM, respectively). The results of training sSOMs under both conditions are shown in Table 2. We found a striking difference in model complexity between the two protocols, on the basis of the number of descriptors in each model. SA without Monte Carlo cross-validation selected 17−33% of the total descriptors, but when coupled with the Monte Carlo cross-validation, SA selects only 8−25%. Despite the descriptor space reduction, there is very little diminution in predictive ability as indicated by %test and %validation. For instance, with the D3 data set, the average number of selected descriptors is 37.8 without cross-validation and 27.2 with Monte Carlo cross-validation, while the mean %test values are 66.7% and 68.0%, the mean %validation values are 58.5% and 59.8%, the best %test values are 84.8% and 78.8%, and the best %validation values are 63.4% and 68.3%, respectively. This indicates that SA with Monte Carlo cross-validation is a highly efficient descriptor selection approach and can greatly help to simplify sSOM and perhaps other models as well.

It is well-known that large, overdetermined data sets are subject to overfitting, also known as the "curse of dimensionality"[1]. Motivated by this fact and because locked test and external validation sets are subject to random vari-

Table 2. Classification Results for PLS, sSOMs, SA-sSOMs, and SA-sSOMs with Monte Carlo Cross-Validation[a]

	training		test		external validation		
data set/method	av. %correct	best	av. %test	best	av. %validation	best	NER
$\alpha4\beta2$ (169)[b] Three-Way							
PLS	48.9 ± 6.6 (4.3)[c]	56.2	42.8 ± 7.1	**56.5**	53.3 ± 6.1	**65.5**	
sSOM	84 ± 3.3	89.6	38.7 ± 11.1	**60.9**	45.7 ± 6.5	**58.6**	
SA-sSOM	97.1 ± 1.9 (39.1)[d]	99	46.8 ± 10.5	**60.9**	47.1 ± 7.8	**62.1**	
MC-SA-sSOM	85.6 ± 3.2 (27.5)[d]	90.6	43.9 ± 8.6	**60.9**	61.8 ± 4.2	**65.5**	63.7
D2 (207)[b] Three-Way							
PLS	34.7 ± 2.4 (1.3)[c]	40.6	31.1 ± 6.1	**42.4**	32.8 ± 4.4	**43.9**	
sSOM	80 ± 2.1	84.2	54.4 ± 7.3	**66.7**	50.6 ± 5.4	**61**	
SA-sSOM	90 ± 1.7 (37.2)[d]	94	55.9 ± 8.5	**75.8**	48.8 ± 4	**56.1**	
MC-SA-sSOM	79.4 ± 2.7 (31.4)[d]	86.5	55.3 ± 6.9	**69.7**	62.8 ± 2.2	**63.4**	62.7
D3 (207)[b] Three-Way							
PLS	24.3 ± 5.8 (1.3)[c]	36.8	20 ± 6	**33.3**	25.7 ± 3.2	**36.6**	
sSOM	85.5 ± 1.9	88.7	64.8 ± 6.9	**81.8**	59.9 ± 6.2	**73.2**	
SA-sSOM	92.7 ± 0.8 (37.8)[d]	94	66.7 ± 7.7	**84.8**	58.5 ± 3.5	**63.4**	
MC-SA-sSOM	84.7 ± 2.7 (27.2)[d]	88.7	68 ± 6.2	**78.8**	59.8 ± 5	**68.3**	72.7
DHFR (135)[b] three-way							
PLS	58 ± 4.2 (3.5)[c]	67.8	49.3 ± 6.4	**61.9**	50.4 ± 3.5	**55.6**	
sSOM	91 ± 1.8	95.4	61 ± 8	**76.2**	64.8 ± 4.6	**70.4**	
SA-sSOM	98.5 ± 1.1 (63.5)[d]	100	59.3 ± 7.9	**71.4**	62.8 ± 6.2	**70.4**	
MC-SA-sSOM	88.5 ± 3.4 (30.3)[d]	94.3	58.8 ± 9.4	**81**	60.6 ± 6.5	**70.4**	63.5
Topliss (272)[b] Four-Way							
PLS	35.9 ± 3.6 (2.8)[c]	41.4	32.9 ± 5.1	**41.9**	30.7 ± 2.8	**35.2**	
sSOM	74 ± 2.2	77.6	39.7 ± 5.8	**51.2**	38.8 ± 5	**48.1**	
SA-sSOM	94.2 ± 1.7 (74.5)[d]	96.6	39.7 ± 6.2	**55.8**	39.7 ± 4.4	**46.3**	
MC-SA-sSOM	84.3 ± 4.3 (32.4)[d]	89.1	40 ± 8.4	**55.8**	37.7 ± 5.4	**50**	51.2
NCI SF-268 (2400)[b] Three-Way							
PLS	54.8 ± 0.7 (5)[c]	56.8	53.4 ± 2.5	**57**	52.9 ± 0.8	**53.7**	
sSOM	71.9 ± 0.8	73	54.9 ± 2.3	**59.6**	56.8 ± 1.9	**60.4**	
SA-sSOM	83.9 ± 1.4 (79.9)[d]	85.5	53.1 ± 2.3	**57**	57 ± 1.7	**60.6**	
MC-SA-sSOM	76.8 ± 1.8 (41.8)[d]	81.7	53.7 ± 2.6	**58.9**	56.2 ± 2.5	**60.8**	55.9

[a] PLS and sSOM without descriptor reduction. %correct: Average percent correct classification of 20 different runs with different training, test, and external validation partition. best: The most accurate classification percentage. [b] The number of observations. [c] The average number of principal components of 20 runs for PLS. [d] The average number of selected descriptors of 20 runs for SA-sSOM and MC-SA-sSOM.

ability,[51] we chose to further analyze our results. It is apparent that the predictive results of a locked test set (% test) are an unreliable criterion for assessing and selecting the best model (see bold-faced values in Table 2). It is also clear that there is little correlation between the Monte Carlo cross-validation NER and the predictive quality of the model with regard to test and external validation sets (see last column of results for each data set in Table 2). A high value of cross-validation q^2 (NER), therefore, is an insufficient indicator of the predictive ability of a model. However, our results show that a very large and diverse test set may yield an acceptable model assessment criterion (see results for NCI data set, Table 2). To circumvent the problems inherent in using cross-validation and a single test set to assess the predictive ability of a model, we implemented a dynamic partitioning process on the six study data sets using both PLS and sSOM protocols. This procedure was performed with all descriptors as well as with a selected subset of descriptors (Table 3). Results showed no significant difference between the NER of the dynamic test set and external validation set for all six data sets. These results indicate that the nonerror rate of the dynamic test set can be used to assess model quality and to select the best model. Because this process provides a robust estimation of model predictive ability, we were able to utilize it as a powerful investigative tool for the analysis and comparison of models under different protocols and conditions:

1. We compared PLS and sSOM (see Table 3). Our results show that, as with other commonly utilized QSAR methodologies, performance varies with the data set. In our experiments, dynamic partitioning of sSOMs significantly improved the classification power of the model for the D2, D3, DHFR, and Topliss data sets and showed marginal improvement for the $\alpha4\beta2$ and NCI data sets. Overall, these encouraging results demonstrate that sSOM has the ability to produce predictions that are more accurate than traditional linear QSAR methods.

2. We compared sSOM results with and without SA as a descriptor selection tool. Our results confirm those of our earlier study that showed sSOM to be a very robust methodology with the ability to quickly identify relevant information in even a large input manifold[15] (see sSOM results in Table 3), but our results with SA show that it can significantly reduce the descriptor space while providing a model with at least as good a predictive quality as one with all the descriptors. For instance, comparing the external validation NER for the D2 data set under our three protocols (sSOM, SA-sSOM, and MC-SA-sSOM) gives 54.0, 54.9, and 56.5, respectively.

3. We compared the results of SA-sSOMs and MC-SA-sSOMs using dynamic test NER in Table 3. Models whose optimization was driven by Monte Carlo cross-validation were equally as robust in predictive power as those driven by %train, but MC-SA-sSOM was able to reduce the

Table 3. Nonerror-Rate Results of "Dynamic Partitioning" for PLS, SA-sSOMs, and MC-SA-sSOMs with Different Model Descriptor Sets[a]

	%train		%test		%validation	
data set/method	mean	best	mean	best	mean	best
α4β2 Three-Way						
PLS		45.3		**42.3**		**48.4**
sSOM		85.9		**42.5**		**45.3**
SA-sSOM	90.4	94	46	**52.5**	47.1	**52.2**
MC-SA-sSOM	85.4	89.2	48.8	**53.3**	49.7	**56.9**
D2 Three-Way						
PLS		33		**34.1**		**30.2**
sSOM		80.9		**50.3**		**54**
SA-sSOM	82.5	84	52.5	**56.8**	53	**58.4**
MC-SA-sSOM	80	81.6	53.4	**56.5**	54.8	**57.2**
D3 Three-Way						
PLS		29.8		**28.4**		**26.8**
sSOM		85.1		**63.7**		**63.8**
SA-sSOM	85.9	87.6	63.1	**66.6**	63.7	**67.3**
MC-SA-sSOM	83.6	86	65	**68.4**	66.1	**70.6**
DHFR Three-Way						
PLS		52.7		**51.4**		**51.5**
sSOM		90.8		**63.4**		**58**
SA-sSOM	93.1	94.3	58.6	**63.4**	60.4	**63.9**
MC-SA-sSOM	90.1	93.5	60	**64.3**	62.1	**67**
Topliss Four-Way						
PLS		35.6		**30.8**		**30.8**
sSOM		74.5		**40.3**		**46.1**
SA-sSOM	86.4	87.9	39.4	**42**	39.8	**44**
MC-SA-sSOM	83.5	87.7	40.7	**43.5**	41.9	**46.9**
NCI SF-268 Three-Way						
PLS		54.9		**52.8**		**53.4**
sSOM		72.3		**55.1**		**56.1**
SA-sSOM	80.9	83.3	54.2	**55.9**	54.6	**56.6**
MC-SA-sSOM	76.8	79.5	54.7	**55.9**	54.8	**56.2**

[a] mean: Average NER of 20 different descriptor subset models. best: The NER value for PLS and sSOM without descriptor reduction and the highest NER of 20 models for SA-sSOM and MC-SA-sSOM.

descriptor space even more. In general, NER {dynamic test, all descriptors} < NER {dynamic test, SA-sSOM} < NER {dynamic test, MC−SA-sSOM}.

CONCLUSIONS

This paper further demonstrates the utility of sSOMs in QSAR and quantitative structure−property relationship modeling. By coupling SA to sSOMs, we were able to create more predictive classification models (with the exception of the Topliss data set) and significantly reduce model complexity. A Monte Carlo descriptor selection technique is described as a means to minimize overfitting while also providing simpler final models in the cases studied. Finally, a dynamic partitioning process for model development is described which overcomes the problem of deleterious sampling error frequently encountered using small locked test sets. Our results indicate that the need for dynamic partitioning depends on data set size (number of observations) and complexity.

ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under SBIR Grant #0319787. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The authors would like to thank the SOM Toolbox Team from Helsinki University of Technology for making the SOM Toolbox freely available for research.

REFERENCES AND NOTES

(1) Bellman, R. *Adaptive Control Processes: A Guided Tour*; Princeton University Press: Princeton, NJ, 1961.

(2) Rogers, D.; Hopfinger, A. J. Application of genetic function approximation to quantitative structure−activity relationships and quantitative structure−property relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 4−866.

(3) Gasteiger, J.; Zupan, J. Neural Networks in Chemistry. *Angew. Chem., Int. Ed. Engl.* **1993**, *32*, 503−527.

(4) Kittler, J. Feature selection and extraction. In *Handbook of Pattern Recognition and Image Processing*; Young, T. Y., Fu, K.-S., Eds.; Academic Press: New York, 1986.

(5) Inza, I.; Merino, M.; Larranaga, P.; Quiroga, J.; Sierra, B.; Girala, M. *Feature Subset Selection by Population-Based Incremental Learning*; Technical Report no. EHU−KZAA-IK-1/99; University of the Basque Country: Spain, 1999.

(6) Yang, J.; Honavar, V. Feature subset selection using a genetic algorithm. In *Genetic Programming 1997: Proceedings of Second Annual Conference*; Koza, J., Ed.; Morgan Kaufmann: San Fransisco, CA, 1997; p 380.

(7) Kudo, M.; Somol, P.; Pudil, P.; Shimbo, M.; Sklansky, J. Comparison of classifier-specific feature selection algorighms. In SSPP/SPR; 2000; pp 677−686.

(8) Kohonen, T. Self-organized formation of topologically correct feature maps. *Biol. Cybernetics* **1982**, 43.

(9) Kohonen, T. *Self-organizing Maps*; Springer-Verlag: Berlin, 1995.

(10) Gasteiger, J.; Zupan, J. *Neural Networks for Chemists An Introduction*. VCH: Weinheim, Germany, 1993; pp 277−291.

(11) Gasteiger, J.; Li, X. *Angew. Chem.* **1994**, *106*, 671−674.

(12) Rose, V. S.; Croall, I. F.; MacFie, H. J. H. An Application of Unsupervised Neural Network Methodology Kohonen Topology-Preserving Mapping to QSAR Analysis. *Quant. Struct.-Act. Relat.* **1991**, *10* (1), 6−15.

(13) Polanski, J. Self-organizing neural network for modeling 3D QSAR of colchicinoids. *Acta Biochim. Pol.* **2000**, *47* (1), 37−45.

(14) Espinosa, G.; Arenas, A.; Giralt, F. An integrated SOM-fuzzy ARTMAP neural system for the evaluation of toxicity. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 343−59.

(15) Xiao, Y.-D.; Clauset, A.; Harris, R.; Bayram, E.; Santago, P., II; Schmitt, J. D. Supervised Self-Organizing Maps in Drug Discovery. 1. Robust Behavior with Overdetermined Data Sets. *J. Chem. Inf. Model.* **2005**, *45*, 1749−1758.

(16) Lachenbrush, P. A.; Mickey, M. Estimation of error rates in discriminant analysis. *Technometrics* **1968**, *10*, 1−11.

(17) Stone, M. An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion. *J. R. Stat. Soc., B* **1977**, *38*, 44−47.

(18) Eriksson, L.; Johansson, E.; Muller, M.; Wold, S. On the selection of the training set in environmental QSAR analysis when compounds are clustered. *J. Chemom.* **2000**, *14*, 599−616.

(19) Martens, H. A.; Dardenne, P. Validation and verification of regression in small data sets. **1998**, *44*, 99−121.

(20) Schmitt, J. D. Exploring the Nature of Molecular Recognition in Nicotinic Acetylcholine Receptors. *Curr. Med. Chem.* **2000**, *7*, 749−800.

(21) Huang, Y.; Hammond, P. S.; Whirrett, B. R.; Kuhner, R. J.; Ross, J.; Wu, L.; Childers, S. R.; Mach, R. H. Synthesis and quantitative structure−activity relationships of N-(1-benzylpiperidin-4-yl)-phenylacetamides and related analogues as potent and selective s_1 receptor ligands. *J. Med. Chem.* **1998**, *41* (13), 2361−2370.

(22) Yang, B.; Johnston, D. E., Jr.; Luedtke, R. R.; Hammond, P. S.; Mach, R. H. Synthesis and in vitro binding of N-alkyl-2,3-dimethoxy[3.3.1]-azabicyclononane benzamides at dopamine D_2 and D_3 receptors. *Med. Chem. Res.* **1998**, *8* (3), 115−131.

(23) Hammond, P. S.; Cheney, J. T.; Johnston, D. E.; Ehrenkaufer, R. L.; Luedtke, R. R.; Mach, R. H. Synthesis, in vitro dopamine D_2 and D_3 receptor binding and quantitative structure−activity studies on substituted 2,3-dimethoxy-N-1-benzyl-4-piperidnyl)benzamides and related compounds. *Med. Chem. Res.* **1999**, *9* (1), 35−49.

(24) Mach, R. H.; Hammond, P. S.; Huang, Y.; Yang, B.; Xu. Y.; Cheney, J. T.; Freeman, R.; Luedtke, R. R. Structure−activity relationship studies of N-(9-benzyl)-9-azabicyclo[3.3.1]nonan-3b-yl benzamide analogues for dopamine D2 and D3 receptors. *Med. Chem. Res.* **1999**, *9* (6), 355−373.

(25) Huang, Y.; Luedtke, R. R.; Freeman, R. A.; Wu, L.; Mach, R. H. Synthesis of 2-(2, 3,-dimethoxyphenyl)-4-(aminomethyl)imidazole

analogues and their binding affinities for dopamine D2 and D3 receptors. *Bioorg. Med. Chem.* **2001**, *9*, 3113−3122.

(26) Huang, Y.; Luedtke, R. R.; Freeman, R. A.; Wu, L.; Mach, R. H. Synthesis and structure−activity relationships of naphthamides as dopamine D3 receptor ligands. *J. Med. Chem.* **2001**, *44*, 1815−1826.

(27) Yoshida, F.; Topliss, J. G. QSAR model for drug human oral bioavailability. *J. Med. Chem.* **2000**, *43*, 2575−85.

(28) Hansch, C. A Quantitative Approach to Biochemical Structure−Activity Relationships. *Acc. Chem. Res*. **1969**, *2*, 232−39.

(29) Developmental Therapeutics Program NCI/NIH. http://dtp.nci.nih.gov/.

(30) *QSARIS*, 1.1; MDL Information Systems Inc.: San Leandro, CA.

(31) *Cerius^2 Modeling Environment*, release 4.9; Accelrys Inc.: San Diego, CA, 2003.

(32) Cruciani, G.; Pastor, M.; Mecucci, S. *Volsurf*, 3.0.9; Molecular Discovery Ltd.: Pinner, Middlesex, U. K.

(33) *Dragon*, 4.0; Milano Chemometrics and QSAR Research Group, University of Milano-Bicocca: Milano, Italy. http://www.telemacus.it/talete/.

(34) Golbraikh, A.; Shen, M.; Xiao, Z.; Xiao, Y.-D.; Lee, K.-H.; Tropsha, A. Rational selection of training and test sets for the development of validated QSAR models.

(35) Frank, I. E.; Todeschini, R. *The Data Analysis Handbook.* Elsevier: Amsterdam, 1994.

(36) Van der Putten, P. Utilizing the Topology Preserving Property of Self-Organizing Maps for Classification. MSc Thesis, Cognitive Artificial Intelligence, Utrecht University, NL, 1996.

(37) Kohonen, T.; Makisara, K.; Saramaki, T. Phonotopic Maps − Insightful Representation of Phonological Features for Speech Recognition. *Proceedings of the IEEE Seventh International Conference on Pattern Recognition*, Montreal, Canada, July 30−August 2, 1984; pp 182−185.

(38) Kohonen, T. The self-organizing map. *Neurocomputing* **1998**, *21* (1−3), 1−6.

(39) Metroplis, N.; Rosenbluth, M. N.; Teller, A. H. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys*. **1953**, *21*, 1087−1092.

(40) Kirkpatrick, S.; Gelatt, C. D., Jr.; Vecchi, M. P. Optimization by Simulated Annealing. *Science* **1983**, *220*, 671−680.

(41) Sutter, J. M.; Kalivas, J. H. Comparison of forward selection, backward elimination, and generalized simulated annealing for variable selection. *Microchem. J.* **1993**, *47*, 60−66.

(42) Shen, M.; LeTiran, A.; Xiao, Y.; Golbraikh, A.; Kohn, H.; Tropsha, A. Quantitative Structure−Activity Relationship Analysis of Functionalized Amino Acid Anticonvulsant Agents Using *k* Nearest Neighbor and Simulated Annealing PLS Methods. *J. Med. Chem.* **2002**, *45* (13), 2811−2823.

(43) Zheng, W.; Tropsha, A. A novel variable selection QSAR approach based on the K-nearest neighbor principle. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 185−194.

(44) Xiao, Z.; Xiao, Y.-D.; Feng, J.; Golbraikh, A.; Tropsha, A.; Lee, K.-H. Antitumor Agents. 213. Modeling of Epipodophyllotoxin Derivatives Using Variable Selection *k* Nearest Neighbor QSAR Method. *J. Med. Chem.* **2002**, *45* (11), 2294−2309.

(45) Shen, M.; Xiao, Y.; Golbraikh, A.; Gombar, V. K.; Tropsha, A. Development and Validation of *k*-Nearest-Neighbor QSPR Models of Metabolic Stability of Drug Candidates. *J. Med. Chem.* **2003**, *46* (14), 3013−3020.

(46) Sutter, J. M.; Dixon, S. L.; Jurs, P. C. Automated descriptor selection for quantitative structure−activity relationships using generalized simulated annealing. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 77−84.

(47) So, S.-S.; van Helden, S. P.; van Geerestein, V. J.; Karplus, M. Quantitative Structure−Activity Relationship Studies of Progesterone Receptor Binding Steroids. *J. Chem. Inf. Comput. Sci.* **2000**, *40* (3), 762−772.

(48) Sun, L.; Xie, Y.; Song, X.; Wang, J.; Yu, R. Cluster Analysis By Simulated Annealing. *Comput. Chem.* **1994**, *18*, 103−108.

(49) Golbraikh, A.; Tropsha, A. Beware of q2! *J. Mol. Graphics Modell.* **2002**, *20*, 269−276.

(50) Stone, M. Cross validatory choice and the assessment of statistical prediction. *J. R. Stat. Soc*., *B* **1974**, *36*, 111−133.

(51) Hawkins, D. M. The Problem of Overfitting. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1−12.

CI0500841