Application of Genetic Function Approximation to Quantitative Structure-Activity Relationships and Quantitative Structure-Property Relationships

David Rogers*

Molecular Simulations Incorporated, 16 New England Executive Park, Burlington, Massachusetts 01803-5297

A. J. Hopfinger

Department of Medicinal Chemistry and Pharmacognosy, College of Pharmacy, University of Illinois at Chicago, Box 6998, Chicago, Illinois 60680

Received November 12, 1993®

The genetic function approximation (GFA) algorithm offers a new approach to the problem of building quantitative structure-activity relationship (QSAR) and quantitative structure-property relationship (QSPR) models. Replacing regression analysis with the GFA algorithm allows the construction of models competitive with, or superior to, standard techniques and makes available additional information not provided by other techniques. Unlike most other analysis algorithms, GFA provides the user with multiple models; the populations of models are created by evolving random initial models using a genetic algorithm. GFA can build models using not only linear polynomials but also higher-order polynomials, splines, and Gaussians. By using spline-based terms, GFA can perform a form of automatic outlier removal and classification. The GFA algorithm has been applied to three published data sets to demonstrate it is an effective tool for doing both QSAR and QSPR.

1. BACKGROUND

Quantitative structure-activity relationship (QSAR) analysis is an area of computational research which builds models of biological activity using physicochemical properties of a series of compounds. The underlying assumption is that the variations of biological activity within a series of similar structures can be correlated with changes in measured or computed molecular features of the molecules. These features could measure, for example, hydrophobic, steric, and electronic properties which may influence biological activity. In this analysis, a data table is formed, each row representing a candidate compound and each column an experimental or computational feature. Regression analysis can be applied to this data table to create a model of activity based on all or some of the features. Quantitative structure-property relationship (QSPR) analysis is a generalization of the QSAR concept. The QSPR philosophy assumes that the behavior of a compound, as expressed by any measured property, can be correlated to a set of molecular features of the compound.

Current QSAR and QSPR methods are limited by the structure of the data: the number of compounds with the requisite behavior measures (e.g. biological activity) is usually small compared with the number of features which can be measured or calculated. This can lead either to models which have low error measure on the training set, but which do not predict well (a phenomena called *overfitting*), or to a complete failure to build a meaningful regression model. Recent findings suggest that features which characterize molecular shape-related properties may be especially useful,[1] but a given compound may have thousands of these features, making the building of a regression model even more problematic. Still, standard regression analysis continues to be the predominant technique for QSAR and QSPR construction, though recent work using partial least-squares (PLS) regression[2] or neural networks[3] show advantages in some situations over standard methods.

The genetic function approximation (GFA) algorithm is derived from Rogers' G/SPLINES algorithm[4-5] and offers a new approach to the construction of QSAR and QSPRs. We propose supplementing standard regression analysis with the GFA algorithm. Application of the GFA algorithm may allow the construction of higher-quality predictive models and make available additional information not provided by standard regression techniques, even for data sets with many features. In particular, the advantages of multiple models made available using GFA will be discussed, as well as the automatic partitioning behavior of GFA when used to build spline models.

2. METHODS

A. QSAR Methodology. QSAR began with the pioneering work of Hansch,[6] who used linear regression to build predictive models of the biological activity of a series of compounds. The general form for the model $F(\bar{\mathbf{X}})$ is as a linear combination of basis functions $\phi_k(\bar{\mathbf{X}})$ of the features $\bar{\mathbf{X}} = \{\mathbf{x}_1, ..., \mathbf{x}_m\}$ in the training data set of size **M**, as given in eq 1.

$$F(\mathbf{X}) = a_0 + \sum_{k=1}^{\mathbf{M}} a_k \phi_k(\mathbf{X}) \qquad (1)$$

The basis functions are functions of one or more features, such as **LOGP**, $(\mathbf{LOGP} - 10.1)^2$, or $\mathbf{DIPV\_X} * (\mathbf{VWDVOL} - 100.0)$. Linear regression takes the list of basis functions and finds a set of coefficients a_k to build a functional model of the activity. The accuracy with which the model describes the training data is termed the *smoothness of fit* of the model. (A *smooth* model represents only the general trends of the data, and not necessarily the details.)

Thirty years later, linear regression remains the major regression technique used to construct QSARs. Unfortunately, a large number of samples is needed, and even moderate numbers of features lead to poor-quality regression models due to *overfitting*. An overfit model can recall the activities of the training set samples but may not accurately predict the

* Author to whom correspondence should be addressed.
® Abstract published in *Advance ACS Abstracts*, April 1, 1994.

0095-2338/94/1634-0854$04.50/0

activity of previously-unseen samples. With linear regression, the amount of fit is determined by the number of basis functions in the model.

To make linear regression suitable when moderate numbers of features are used, it was combined with *principal components analysis* (PCA),[7] which is a technique for selecting the most important set of features from a larger table. Once the most important features are selected, linear regression is used to construct a model. Unfortunately, PCA makes an assumption of independence of the features. If this assumption of independence is not true (as is often the case in real-world applications), the features selected may not be the most predictive set. Still, the reduction in the number of features is vital if overfitting is to be prevented.

More recently, large amounts of three-dimensional (3D) molecular shape data have become available for molecules. For example, in the CoMFA technique,[2] the electrostatic field around a molecule can be calculated on a 3D grid, providing hundreds or thousands of features. However, it is virtually impossible to use so many features in building standard regression models. It was only with the use of partial least-squares (PLS) analysis that model building became possible. These models can show predictiveness, and the CoMFA approach has become a standard for the analysis of 3D field data.

Finally, recently published work suggests that neural networks and genetic algorithms may be useful in data analysis, specifically in the task of reducing the number of features for regression models. Wikel and Dow[3] applied neural networks and Leardi et al.[8] applied genetic algorithms to the feature reduction task, both with some success. Good et al.[9] did a comparison of a neural-network style approach to the PLS approach used in CoMFA.[2] The neural-network analysis was superior in some of the published results as measured by the cross-validated correlation coefficient (r^2) scores. However, the fit of most neural-network models is determined in part by the length of their training cycle, so there is risk of overfitting if trained too long.

B. Genetic Function Approximation. The genetic function approximation algorithm was initially conceived by taking inspiration from two seemingly disparate algorithms: Holland's genetic algorithm[10] and Friedman's multivariate adaptive regression splines (MARS) algorithm.[11]

Genetic algorithms are derived from an analogy with the evolution of DNA. In this analogy, individuals are represented by a one-dimensional string of bits. An initial population is created of individuals, usually with random initial bits. A "fitness function" is used to estimate the quality of an individual, so that the "best" individuals receive the best fitness score. Individuals with the best fitness scores are more likely to be chosen for mating and to propagate their genetic material to offspring through the *crossover* operation, in which pieces of genetic material are taken from each parent and recombined to create the child. After many mating steps, the average fitness of the individuals in the population increases as "good" combinations of genes are discovered and spread through the population. Genetic algorithms are especially good at searching problem spaces with a large number of dimensions, as they conduct a very efficient directed sampling of the large space of possibilities.

Friedman proposed the MARS algorithm as the newest member of a class of well-used statistical modeling algorithms such as CART[12] and *k–d* trees.[13] It uses splines as basis functions to partition data space as it builds its regression models. It was specifically designed to allow the construction of spline-based regression models of data sets with moderate numbers of features. MARS gives high levels of performance and competes well against many neural-network approaches but, unfortunately, is computationally intensive and too expensive to use with more than about 20 features and 1000 input samples. Also, since MARS builds its model incrementally, it may not discover models containing combinations of features that predict well as a group, but poorly individually.

F_1: {LOGP; DIPV_X; (DIPV_Y - 2.0); VDWVOL; (LOGP - 5.1)^2}
F_2: {M_PNT; MOL_WT; LOGP; LOGP^2}
⋮ ⋮
F_K: {(ATCH4 - ATCH6); DIPMOM; DIPV_X; VDWVOL; LOGP}

Figure 1. Examples of a population of models represented for the genetic function approximation algorithm. Each model is represented as a linear string a basis functions. The activity models can be reconstructed by using least-squares regression to regenerate the coefficients $\{a_k\}$. (The sample features are taken from the Selwood data set.)

One of us (D.R.) recognized that Friedman was doing a search over a very large "function space" and that a better search could be done using a genetic algorithm rather than his incremental approach. Replacing the binary strings of Holland with strings of basis functions gave a natural mapping from Holland's genetic approach to the functional models of regression-based approaches. This led to the published work on G/SPLINES, which later evolved into the genetic function approximation (GFA) algorithm.[4,5]

The GFA algorithm approach has a number of important advantages over other techniques: it builds multiple models rather than a single model; it automatically selects which features are to be used in its basis functions and determines the appropriate number of basis functions to be used by testing full-size models rather than incrementally building them; it is better at discovering combinations of basis functions that take advantage of correlations between features; it incorporates the LOF (lack of fit) error measure developed by Friedman[11] that resists overfitting and allows user control over the smoothness of fit; it can use a larger variety of basis functions in construction of its models, for example, splines, Gaussians, or higher-order polynomials; and study of the evolving models provides additional information, not available from standard regression analysis, such as the preferred model length and useful partitions of the data set.

C. Genetic Function Approximation Algorithm. Many techniques, including MARS, CART, and PCA, develop a single regression model by incremental addition or deletion of basis functions. In contrast, the GFA algorithm uses a population of many models and tests only the final, fully-constructed model. Improved models are constructed by performing the genetic crossover operation to recombine the terms of the better-performing models.

A genetic algorithm requires that an individual be represented as a linear string, which plays the role of the DNA for the individual. When using GFA, the string is the series of basis functions, as shown in Figure 1. Using the information in the string, it is possible to reconstruct the activity model by using least-squares regression to regenerate the coefficients $\{a_k\}$. The initial models are generated by randomly selecting some number of features from the training data set, building basis functions from these features using the user-specified basis function types, and then constructing the genetic models from random sequences of these basis functions.

The models are scored using Friedman's "lack of fit" (LOF) measure[11] which is given by eq 2. In this equation, c is the number of basis functions (other than the constant term) in

$$\mathrm{LOF} = \frac{\mathrm{LSE}}{\left(1 - \frac{c + dp}{M}\right)^2} \qquad (2)$$

the model; d is the *smoothing parameter* (and is the only parameter adjustable by the user); p is the total number of features contained in all basis functions (some basis functions, such as (**ATCH4** – **ATCH6**), contain more than one feature); and M is the number of samples in the training set. Unlike the more-commonly used least-squares error (LSE), the LOF measure cannot always be reduced by adding more terms to the regression model. While the new term may reduce the LSE, it also increases the values of c and p, which tends to *increase* the LOF score. Thus, adding a new term may reduce the LSE, but actually increase the LOF score. By limiting the tendency to simply add more terms, the LOF measure resists overfitting better than the LSE measure.

Usually, the benefit of new terms for smaller models is enough to offset the penalty from the denominator in the LOF. Hence, the value of LOF decreases as initial terms are added. However, at some point the benefit is no longer enough to offset the penalty, and the value of LOF starts increasing. The location of this minimum is changed by altering the smoothing parameter d. The default value for d is 1.0, and larger values of d shift the minimum toward smoother models (that is, models with fewer terms). In effect, d is the user's estimate of how much detail in the training data set is worth modeling.

Once all models in the population have been rated using the LOF score, the *genetic crossover* operation is repeatedly performed. In this operation, two good models are probabilistically selected as "parents", with the likelihood of being chosen inversely proportional to a model's LOF score. Each parent is randomly "cut" into two pieces, and a new model is created using a piece from each parent, as shown in Figure 2. The coefficients of the new model are determined using least-squares regression.

Next, mutation operators may alter the newly-created model. Two mutation operators are possible by default: *new* alters by appending a new random basis function, and *shift* moves the knot of a spline basis function. These two mutation operators have a default 50% probability of being applied to the newly-created model.

Finally, if a duplicate of the resulting model does not already exist in the population, the model with the worst LOF score is replaced by the new child.

The overall process is ended when the average LOF score of the models in the population stops significantly improving. For a population of 300 models, 3000–10 000 genetic operations are usually sufficient to achieve "convergence". For typical data sets, this process takes between 10 min and 1 h on a Macintosh-IIfx computer.

Upon completion, one can simply select the model from the population with the lowest score, though it is usually preferable to inspect the different models and select on the basis of the appropriateness of the features, the basis functions, and the feature combinations.

Selecting a single model is not always desirable; the population can be studied for information on feature use, and predictions can often be improved by averaging the results of multiple models rather than relying on an individual model. Different models may have different regions in which they predict well, and, by averaging, the effect of models which are extrapolating beyond their predictive region is reduced.

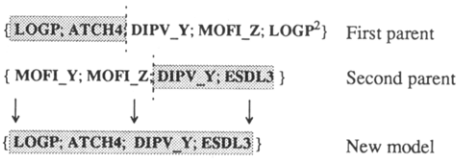

Figure 2. The crossover operation. Each parent is cut at a random point, and a piece from each parent is used to construct the new model which now uses some basis functions from each parent.

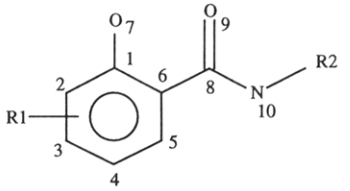

Figure 3. Shared structure of the antimycin analogous in the Selwood data set.

- **LOGP**: Partition coefficient
- **M_PNT**: Melting point
- **DIPMOM**: Dipole moment
- **VDWVOL**: van der Waals volume
- **SURF_A**: Surface area
- **MOL_WT**: Molecular weight)
- **DIPV_X, DIPV_Y, DIPV_Z**: Dipole vector components in X, Y, and Z
- **MOFI_X, MOFI_Y, MOFI_Z**: Principal moments of inertia in X, Y, and Z
- **PEAX_X, PEAX_Y, PEAX_Z**: Principal ellipsoid axes in X, Y, and Z
- **S8_1DX, S8_1DY, S8_1DZ**: Substituent on atom 8 dimensions in X, Y, and Z
- **S8_1CX, S8_1CY, S8_1CZ**: Substituent on atom 8 center in X, Y, and Z
- **ATCH1** to **ATCH10**: Partial atomic charges for atoms 1-10
- **ESDL1** to **ESDL10**: Electrophilic superdelocalizability for atoms 1-10
- **NSDL1** to **NSDL10**: Nucleophilic superdelocalizability for atoms 1-10
- **SUM_F** and **SUM_R**: Sums of the F and R substituent constants

Figure 4. Features contained in the Selwood data set.

3. RESULTS

Three data sets were chosen to illustrate application of the GFA algorithm. The Selwood data set[15] illustrates feature selection and the utility of exploring multiple models. The Cardozo/Hopfinger data set[16] demonstrates the automatic partitioning behavior of spline models. The Koehler/Hopfinger data set[17] shows the applicability of the GFA algorithm in polymer modeling. The only difference in the application of the GFA algorithm to these three problems was in the types of basis functions used. In the Selwood application only linear polynomials were considered, while, in the Cardozo/Hopfinger application, linear and quadratic polynomials and splines were considered. For the Koehler/Hopfinger application, linear polynomials and splines were considered.

A. Selwood Data Set. The Selwood data set[15] contains 31 compounds, 53 features, and a set of corresponding antifilarial antimycin activities. In order to save space, the complete data set is not presented here. Reference 15 should be consulted for details on the structure–activity data. The series of analogs are of the general form shown in Figure 3.

This data set was of particular interest because it contains a large number of features relative to the number of compounds. The list of features is shown in Figure 4.

Selwood et al. used multivariate regression to develop a QSAR. Later, this same data set was studied by Wikel and Dow,[3] who used a neural network to select features for their QSAR model. Other groups have studied this same data set.[18–22] In this study, a comparison is made to one of the models of Selwood et al. and the model of Wikel and Dow. Both groups developed models using 3 of the 53 features to predict the activity as measured by $-[\log(IC_{50})]$, where IC_{50} refers to the concentration of an analog needed to reduce the

biological activity by 50%. Their judgement was that the relatively small number of compounds allows only a few features in the final regression model.

Selwood proposed a QSAR model of three features: **LOGP**, the partition coefficient; **M_PNT**, the melting point; the **ESDL10**, the electrophilic superdelocalizability at atom 10. The model derived using all 31 samples is given by eq 3. (This is eq 4 in the Selwood reference.[15]

$$-\log(IC_{50}) = -3.93 + 0.44 * \mathbf{LOGP} + 0.008 * \mathbf{M\_PNT} - 0.30 * \mathbf{ESDL10} \quad (3)$$

LOF: 0.487
r: 0.737
F: 13.29

Selwood used a technique which incrementally adds features to a model based on maximizing the correlation between the new feature (after decorrelation with previously-selected features) and the activity. This technique, known as *forward-stepping regression analysis*, is a common method of feature selection. Its success requires that the information of interest is contained in the correlation of individual features with the response. However, information requiring the combined effect of sets of features may not be discovered. Still, it is exactly such a combined effect that will likely be operative in a chemical system. This means that models which reflect such a combined effect may exist and may outperform models discovered with the forward-stepping technique. (Selwood was able to improve this model by performing *outlier removal* on the data set. We did not perform outlier removal. The next section demonstrates how splines can be used to perform automatic outlier removal.)

Wikel and Dow[3] circumvented the problem of incremental feature selection by using a neutral network to select the appropriate combination of variables in place of forward-stepping regression analysis. While the network does not directly select the features, after training it can be analyzed to reveal the features of interest. Regression can then be used to build the corresponding QSAR model.

A QSAR model of three features has been proposed by Wikel and Dow: **LOGP**, the partition coefficient; **ATCH4**, the partial atomic charge on atom 4; and **MOFI_X**, the principal moment of inertia in the *X* dimension. This QSAR is given by eq 4. It shows better correlation and a lower LOF score than the QSAR model of Selwood.

$$-\log(IC_{50}) = -1.63 + 0.231 * \mathbf{LOGP} + 4.415 * \mathbf{ATCH4} + 0.000659 * \mathbf{MOFI\_X} \quad (4)$$

LOF: 0.525
r: 0.774
F: 13.47

That two different techniques generate two significantly different QSAR models raises some questions: is the discovered model the "best" model, that is, does it minimize the error (versus models of the same size) over the training set? Does a single "best" model even exist, or instead is there a collection of models of the same performance quality? Is the number of terms in the model appropriate, or should larger, or smaller, models be considered? Can the differences between models which score well due to predictiveness, and models which perform well due to chance correlations between features in the training set, be identified?

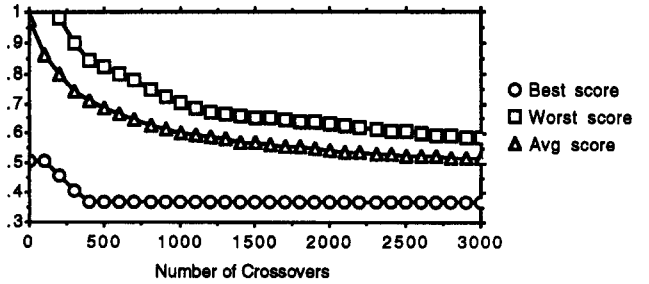

Figure 5. Number of crossover operations versus the best, worst, and average LOF scores in the population of models.

GFA analysis attempts to both generate models which are competitive with or superior to models generated by other techniques and also answer crucial questions such as those above. The answer to all these questions is based upon the construction and analysis of multiple models, rather than optimization of a single model.

B. GFA Applied to the Selwood Data Set. The GFA algorithm was applied to the Selwood data set with an intention to illustrate the advantages and uses of multiple models in QSAR analysis.

QSAR analysis with GFA begins by generating a population of random models. These models are generated by randomly selecting some number of features from the file and using regression to generate the coefficients of the models. For the Selwood data set, a population of 300 models was used, and the terms of the models were limited to linear polynomials. (The limitation to linear polynomials was done for easier comparison to literature models).

The generic operator was applied until the average LOF score showed little improvement over a period of 100 crossover operations. This convergence criteria was met after 3,000 operations. The evolution took approximately 15 min on a Macintosh–IIfx. Figure 5 shows a graph of the evolution of the LOF scores. Some preliminary runs suggested that for an average model length of three features, the parameter *d* (the smoothing parameter in Friedman's LOF function) should be set to 2.0. Provided the value of *d* is set appropriately, there is little risk of overfitting even if the crossover operations are continued beyond convergence.

After convergence, the population was sorted in order of LOF score. The top 20 models in the population are shown in Table 1.

The model of Wikel and Dow was discovered and was rated 131 out of 300. The model of Selwood was not discovered, though similar models were generated. It is possible it may have been discovered if the algorithm was allowed to run further, as some similar runs with a different random seed did discover it. Had it been discovered in this run, it would have been rated about 200 out of 300.

All of the top 20 models, and many of the additional 280 models, approximately match or exceed the correlation scores of the models built by either Selwood or Wikel and Dow. The large number of high-scoring models, and the large amount of variation, appears to refute the supposition that there may be a single best model using this data. The mixture of models with 2–4 features suggests that both smaller and larger models may be appropriate for consideration. Thus, the population of models allows a deeper understanding of the range of possible models from a data set.

Table 1. Top 20 Models Generated Using the Full 31-Sample Selwood Data Set[a]

1: $-\log(IC_{50}) = -2.501 + 0.584 * \mathbf{LOGP} + 1.513 * \mathbf{SUM\_F} - 0.000075 * \mathbf{MOFI\_Y}$
LOF: 0.366
r: 0.849
F: 23.27

2: $-\log(IC_{50}) = 2.871 + 0.568 * \mathbf{LOGP} - 0.013 * \mathbf{SURF\_A} + 0.810 * \mathbf{ESDL3}$
LOF: 0.368
r: 0.848
F: 23.04

3: $-\log(IC_{50}) = -0.805 + 0.589 * \mathbf{LOGP} + 0.736 * \mathbf{ESDL3} - 0.000077 * \mathbf{MOFI\_Y}$
LOF: 0.392
r: 0.838
F: 21.15

4: $-\log(IC_{50}) = 1.791 + 0.500 * \mathbf{LOGP} + 0.842 * \mathbf{ESDL3} - 0.200 * \mathbf{PEAX\_X} + 2.807 * \mathbf{ATCH4}$
LOF: 0.397
r: 0.880
F: 22.31

5: $-\log(IC_{50}) = -2.148 + 0.694 * \mathbf{LOGP} - 0.000084 * \mathbf{MOFI\_Y}$
LOF: 0.405
r: 0.776
F: 21.15

6: $-\log(IC_{50}) = -1.749 + 0.486 * \mathbf{LOGP} - 0.000055 * \mathbf{MOFI\_Y} + 10.124 * \mathbf{ATCH5} + 3.444 * \mathbf{ATCH4}$
LOF: 0.407
r: 0.776
F: 21.57

7: $-\log(IC_{50}) = -0.226 + 0.608 * \mathbf{LOGP} - 0.206 * \mathbf{PEAX\_X}$
LOF: 0.396
r: 0.781
F: 20.81

8: $-\log(IC_{50}) = -0.777 + 0.503 * \mathbf{LOGP} + 1.345 * \mathbf{SUM\_F} - 0.177 * \mathbf{PEAX\_X}$
LOF: 0.409
r: 0.830
F: 19.86

9: $-\log(IC_{50}) = 1.643 + 0.666 * \mathbf{LOGP} - 0.0138 * \mathbf{SURF\_A}$
LOF: 0.410
r: 0.772
F: 20.69

10: $-\log(IC_{50}) = 0.823 + 0.553 * \mathbf{LOGP} + 1.347 * \mathbf{SUM\_F} - 0.0118 * \mathbf{SURF\_A}$
LOF: 0.410
r: 0.829
F: 19.77

11: $-\log(IC_{50}) = 0.849 + 0.510 * \mathbf{LOGP} + 0.686 * \mathbf{ESDL3} - 0.185 * \mathbf{PEAX\_X}$
LOF: 0.416
r: 0.827
F: 19.41

12: $-\log(IC_{50}) = -3.314 + 0.568 * \mathbf{LOGP} + 6.852 * \mathbf{ATCH1} - 0.000071 * \mathbf{MOFI\_Y}$
LOF: 0.430
r: 0.820
F: 18.44

13: $-\log(IC_{50}) = -2.169 + 0.576 * \mathbf{LOGP} + 6.154 * \mathbf{ATCH3} - 0.000070 * \mathbf{MOFI\_Y}$
LOF: 0.432
r: 0.819
F: 18.31

14: $-\log(IC_{50}) = -2.956 + 0.553 * \mathbf{LOGP} + 0.0030 * \mathbf{M\_PNT} + 1.286 * \mathbf{SUM\_F} - 0.000061 * \mathbf{MOFI\_Y}$
LOF: 0.437
r: 0.867
F: 19.44

15: $-\log(IC_{50}) = -0.147 + 0.561 * \mathbf{LOGP} + 0.864 * \mathbf{ESDL3} + 2.030 * \mathbf{ATCH4} - 0.000076 * \mathbf{MOFI\_Y}$
LOF: 0.440
r: 0.866
F: 19.50

16: $-\log(IC_{50}) = -2.322 + 0.590 * \mathbf{LOGP} + 5.507 * \mathbf{ATCH5} - 0.000068 * \mathbf{MOFI\_Y}$
LOF: 0.441
r: 0.815
F: 17.76

17: $-\log(IC_{50}) = 3.280 + 0.536 * \mathbf{LOGP} + 0.911 * \mathbf{ESDL3} + 1.555 * \mathbf{ATCH4} - 0.0126 * \mathbf{SURF\_A}$
LOF: 0.444
r: 0.864
F: 19.21

18: $-\log(IC_{50}) = 0.960 + 0.552 * \mathbf{LOGP} + 6.011 * \mathbf{ATCH3} - 0.0114 * \mathbf{SURF\_A}$
LOF: 0.445
r: 0.813
F: 17.54

19: $-\log(IC_{50}) = -0.091 + 0.544 * \mathbf{LOGP} + 6.565 * \mathbf{ATCH1} - 0.0115 * \mathbf{SURF\_A}$
LOF: 0.446
r: 0.812
F: 17.43

20: $-\log(IC_{50}) = -5.571 + 0.560 * \mathbf{LOGP} - 13.758 * \mathbf{ATCH6} - 0.000065 * \mathbf{MOFI\_Y}$
LOF: 0.447
r: 0.812
F: 17.43

[a] In this study, 300 random models were created and then evolved with 3000 genetic crossover operations. All of these models, and many of the additional 280 models, approximately match or exceed the scores of the models built with the variables used by either Selwood et al. or Wikel or Dow. The large number of high-scoring models appears to refute the supposition that there may be a single "best" model using the data.

QSARs are developed to make predictions, not merely for the ability to reproduce the results in the training set. Predictiveness can be estimated using cross-validation. Each sample is systematically removed from the data set, and new regression coefficients are generated for a given model. This newly-regressed model is used to predict the removed sample. This procedure is performed on each sample in sequence. The series of predictions is used to calculate a new value for r, called the *cross-validated r*. Figure 6 shows the values of cross-validated *r* using the features of Selwood et al., Wikel and Dow, and the top four models discovered by GFA.

As measured by the cross-validated *r*, the combinations of features discovered by the GFA algorithm yield models which are more predictive than the combinations of features discovered by either the forward-stepping regression of neural-network techniques. This is because the genetic algorithm specifically searches for combinations of features which score well, rather than trying to identify individual features.

Cross-validation can also be used to confirm or refute hypotheses about the appropriate number of terms in a model. We decided that models with three features were the most desirable from the reported QSARs. Preliminary runs of the program with different values for *d* indicated a value 2.0 as being most favorable to generate three-feature models. However, without this prior knowledge, we could have decided to use the default value $d = 1.0$, which favors a population with an average model length of five features. Figure 7 shows the top four models (rated by LOF score), and the model with the highest cross-validated-*r* score (which was rated 6 out of 300 by LOF score) for a run using the default value $d = 1.0$. (Interestingly, the second model is identical to the model recently developed by McFarland and Gans using cluster significance analysis (CSA).[22]) Whether the improvement in cross-validated *r* is worth the extra terms may depend on whether the terms suggest a plausible underlying mechanism for their predictiveness. Otherwise, it may be best to stay

$-\log(IC_{50})_{Selwood} = -3.93 + 0.44 * LOGP + 0.008 * M\_PNT - 0.30 * ESDL10$
r : 0.737
F: 13.29
crossvalidated-r: 0.667

$-\log(IC_{50})_{Wikel+Dow} = -1.63 + 0.231 * LOGP + 4.415 * ATCH4 + 0.000659 * MOFI\_X$
r: 0.774
F: 13.47
crossvalidated-r: 0.679

$-\log(IC_{50}) = -2.501 + 0.584 * LOGP + 1.513 * SUM\_F - 0.000075 * MOFI\_Y$
r: 0.849
F: 23.27
crossvalidated-r: 0.804

$-\log(IC_{50}) = 2.871 + 0.568 * LOGP - 0.013 * SURF\_A + 0.810 * ESDL3$
r: 0.848
F: 23.04
crossvalidated-r: 0.803

$-\log(IC_{50}) = -0.805 + 0.589 * LOGP + 0.736 * ESDL3 - 0.000077 * MOFI\_Y$
r: 0.838
F: 21.15
crossvalidated-r: 0.777

$-\log(IC_{50}) = 1.791 + 0.500 * LOGP + 0.842 * ESDL3 - 0.200 * PEAX\_X + 2.807 * ATCH4$
r: 0.880
F: 22.31
crossvalidated-r: 0.798

Figure 6. Correlation coefficient r, cross-validated r, and F for models using the features of Selwood et al., Wikel and Dow, and the top four models discovered by the GFA algorithm using $d = 2.0$ for the smoothing parameter.

$-\log(IC_{50})_{GFA-1} = -1.277 + 0.402 * LOGP + 4.824 * ATCH4 + 12.017 * ATCH5 - 0.114 * DIPV\_X - 0.000050 * MOFI\_Z$
r: 0.909
F: 23.82
crossvalidated-r: 0.834

$-\log(IC_{50})_{GFA-2} = -1.268 + 0.406 * LOGP + 4.712 * ATCH4 + 12.406 * ATCH5 - 0.118 * DIPV\_X - 0.000050 * MOFI\_Y$
r: 0.909
F: 23.78
crossvalidated-r: 0.834

$-\log(IC_{50})_{GFA-3} = 2.618 + 0.442 * LOGP + 3.095 * ATCH4 + 0.766 * ESDL3 - 0.0137 * VDWVOL + 0.000433 * MOFI\_X$
r: 0905
F: 22.71
crossvalidated-r: 0.822

$-\log(IC_{50})_{GFA-4} = -1.815 + 0.490 * LOGP + 2.609 * ATCH4 + 1.972 * SUM\_F - 0.125 * DIPV\_X - 0.000073 * MOFI\_Z$
r: 0.904
F: 22.48
crossvalidated-r: 0.836

$-\log(IC_{50})_{GFA-6} = 3.301 + 0.435 * LOGP + 5.480 * ATCH4 + 21.025 * ATCH5 + 22.636 * ATCH6 - 0.153 * DIPV\_X - 0.000056 * MOFI\_Z$
r: 0.920
F: 22.02
crossvalidated-r: 0.849

Figure 7. Top four models when the evolution is repeated using the smaller value $d = 1.0$ for the smoothing parameter; along with the 6th model, which had the highest cross-validated r score in the population.

with smaller models of nearly equivalent predictiveness.

Both Selwood and Wikel and Dow give short lists of about 10 features which they felt were the most useful for building activity models. GFA can provide similar information by counting the number of times each feature is used in the population and ranking the features by that value. (We counted feature use using the original value $d = 2.0$ for the smoothing parameter). Table 2 shows the top ten features selected by Selwood, Wikel and Dow, and GFA.

The different techniques show little overlap, agreeing only on **LOGP** (the partition coefficient). This is likely caused by different selection pressures under each technique. For forward-stepping regression, the appearance of a feature reflects high correlation with the response after decorrelation with all previously-selected variables. For the neural network, the appearance reflects an ability to work in concert with all of the other features, though the final feature set used in the activity model will be much smaller. For the GFA algorithm, the appearance reflects a feature's utility in many different combinations toward building high-scoring models.

The usage of features in the population using GFA changes over the evolution of model evaluation. Graphing the feature usage as it changes is a dramatic way to watch the evolution of the population of models, to estimate when the population has converged and to quickly judge the relative utility of different features. Such a graph is shown in figure 8 for the Selwood data set.

Only the feature **LOGP** stands out as highly significant. The remainder of the features, those both shown in the graph of Figure 6 and unshown, are not well-distinguished from each other by usage. This suggests that the information in the data set is duplicated over many of the features. The continuing change in the usage suggests that further evolution of the population may be warranted. However, it should be noted that the genetic algorithm searches for combinations of features which score well, rather than identifying individual features. Hence, a count of use may not necessarily be the best measure of whether a feature was useful in making the best-scoring models, though in this example the top ten features accounted for nearly all the features in the top twenty models.

The above results demonstrate that GFA discovers models that are, at the least, important candidates that must be considered in the search for the best predictive model. However, selection of a single best model and the discarding of the remaining models may not be the most advantageous course. It is proposed that the outputs of the multiple models can be averaged to gain additional predictivity.

Averaging the predictions of some number of the higher-scoring models often gave better predictions than any of the individual models. This behavior can be seen if the results predicted by some number of the top-rated 20 models are averaged. Figure 9 shows the effect of this on the cross-validation coefficient r. The top model predicted the training set with $r = 0.849$; by averaging its output with the second-rated model, the correlation coefficient climbs to 0.86, and by averaging the result with the second and third rated models, the correlation coefficient is greater than 0.89. Further additions do not increase the correlation coefficient, but the

Table 2. Preferred Variables for the Selwood Data Set from the Three Studies[a]

	Selwood	Wikel	GFA
ATCH1			*
ATCH2	*	*	
ATCH4		*	*
ATCH			*
ATCH6			*
DIPV_X		*	
DIPV_Y	*		
DIPV_Z	*		
ESDL3			*
ESDL5	*		
ESDL10	*		
LOGP	*	*	*
M_PNT	*	*	
MOFI_X		*	
MOFI_Y		*	*
NSDL2	*		
PEAX_X			*
PEAX_Y		*	
S8_1CZ	*		
SUM_F			*
SUM_R	*		
SURF_A			*
VDWVOL		*	

[a] (Variables not selected by any of the three techniques are not shown.) Selwood used a technique that selected the best correlated variable after decorrelation with the previous selections; Wikel and Dow used as neural network, selecting variables which had large hidden-unit weight in the trained network; this study used a population of initially-random models trained using a genetic algorithm and shows the variables used in 10% or more of the models. The genetic algorithm searches for combinations of features which score well, rather than identifying individual features. Thus a count of use is not the best measure of whether a feature useful in making the best-scoring models. The different techniques show little overlap, agreeing only on LOGP. This is likely caused by different selection pressures under each technique.

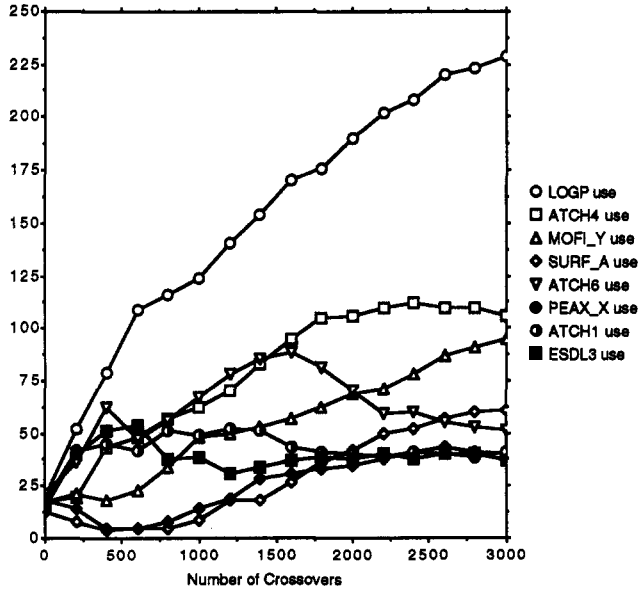

Figure 8. Change in variable use as the evolution proceeds. (The graph only shows varaibles that are used in 15% or more of the 300 models). The feature **LOGP** is the only one whose use is widespread and is used in more than 80% of the models. The next most used feature, **ATCH4**, is used in about 35% of the models, followed closely by **MOFI_Y**. The remainder of the features, both shown in the graph and unshown, are not well-distinguished from each other by usage. This suggests that the information in the data set is duplicated over many of the features. The continuing change in the usage suggests that further evolution of the population may be warranted.

average is remarkably robust, remaining above 0.88 even if we average the output of all 300 models in the population.

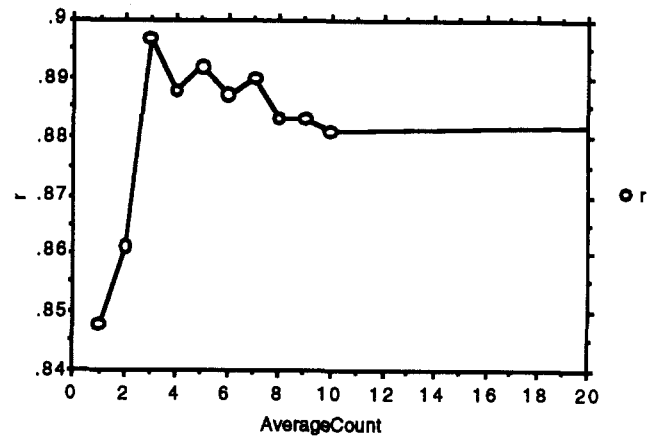

Figure 9. Number of models in the average versus the correlation coefficient. As the number of models in the average increases, the correlation coefficient increases to 0.88, which is higher than any of the individual models.

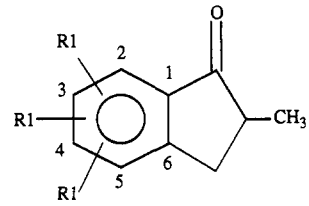

Figure 10. Shared structure of the acetylcholinesterase inhibitor analogs in the Cardozo/Hopfinger data set.

Table 3. Cardozo/Hopfinger Data Set

compd no.	−[log(IC_{50})]	C_4	U_t (D)	HOMO energy (eV)
1	8.88	0.356	3.918	−9.305
2	8.28	0.454	2.301	−9.533
3	8.20	0.505	2.855	−9.631
4	(data sample missing from published data set and not used by GFA)			
5	8.15	0.234	2.966	−9.593
6	8.05	0.449	3.332	−9.451
7	7.92	0.249	3.127	−9.545
8	7.88	0.468	2.765	−9.615
9	7.70	0.452	2.629	−9.461
10	7.64	0.409	2.785	−9.314
11	7.60	0.364	2.845	−9.581
12	7.44	0.251	3.248	−9.443
13	7.16	0.168	3.004	−9.615
14	7.09	0.247	2.980	−9.577
15	7.06	0.027	3.192	−9.376
16	6.89	0.021	3.009	−9.716
17	6.70	0.226	3.467	−9.856
18	6.42	0.463	1.946	−9.508

C. Cardozo/Hopfinger Data Set. The Cardozo/Hopfinger data set[16] contains 17 analogs, 3 features, and a set of corresponding acetylcholinesterase inhibitor activities. The series of analogs have the structure shown in Figure 10. (The activity models used in the original publication used 18 compounds, but the data for one compound was left out of the publication, and so a reduced set of 17 compounds was used in this study.) The data set is given in Table 3.

The compounds were sorted in order of decreasing activity, so that compound **1** is the most active and compound **18** the least active. This data set describes a series of acetylcholinesterase inhibitors with activity measured by −[log(IC_{50})], where IC_{50} is the concentration of the analog needed to inhibit the enzyme by 50%. Unlike the Selwood data set, this data set was already reduced using a technique called molecular decomposition–recomposition (MDR),[16] so it contained a small

number of features relative to the number of compounds. Hence there was no need for a further reduction in the number of features. The three features selected for this QSAR were C_4, the out-of-plane π orbital coefficient of ring carbon 4; U_t, the total dipole moment, and **HOMO**, the energy of the highest occupied molecular orbital.

Cardozo et al.[16] proposed a model of three features and six terms. This QSAR is given by eq 5. Most of the error in the

$$\begin{aligned} -\log(IC_{50})_{full} = &-740.93 \\ &+ 2.73 * C_4 \\ &+ 1.86 * U_t \\ &- 0.14 * (U_t)^2 \\ &- 156.7 * \mathbf{HOMO} \\ &- 8.25 * \mathbf{HOMO}^2 \end{aligned} \quad (5)$$

N: 18
r: 0.804
F: 3.66

QSAR is due to the contributions of two compounds in the data set, analogs **5** and **18**. Removal of these two compounds yields another QSAR, given by eq 6. The elimination of these

$$\begin{aligned} -\log(IC_{50})_{reduced} = &-757.52 \\ &+ 2.21 * C_4 \\ &- 6.65 * U_t \\ &- 1.18 * (U_t)^2 \\ &- 162.9 * \mathbf{HOMO} \\ &- 8.58 * \mathbf{HOMO}^2 \end{aligned} \quad (6)$$

N: 16
r: 0.939
F: 13.95

two compounds yields a QSAR with a much higher correlation score. However, explicit user intervention was required to identify and remove the outliers. Moreover, in most cases no classification process is used that would assist in identifying whether a given test compound should be treated as an outlier. In the next section, we will show how GFA uses spline-based terms to automatically partition the compounds in the data set, and give models over the full data set which are superior to the QSAR expressed by eq 5.

D. GFA Applied to the Cardozo/Hopfinger Data Set. The GFA algorithm was applied to the Cardozo/Hopfinger data set to illustrate the automatic partitioning behavior of spline-based models in QSAR analysis.

QSAR analysis with GFA began with a population of 300 random models. The terms of the models were linear polynomials, quadratic polynomials, linear splines, and quadratic splines. Because there were only three features in the data set, feature selection was not a critical issue. Instead, we included spline basis functions to explore partitions of the data set. The population was evolved for 5,000 crossover operations. By that point there was little continued improvement in the average score of the models in the population.

The splines used are *truncated power splines* and are denoted with angle brackets. For example, $\langle f(\mathbf{x}) - a \rangle$ is equal to zero if the value of $(f(\mathbf{x}) - a)$ is negative, else it is equal to $(f(\mathbf{x}) - a)$. For example, $\langle U_t - 2.765 \rangle$ is zero when $U_t < 2.765$, and equal to $(U_t - 2.765)$ otherwise, as shown in Figure 11. The constant a is called the *knot* of the spline. When a spline term is created, the knot is set using the value of the feature in a random data sample.

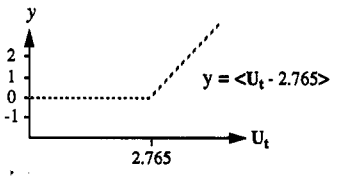

Figure 11. Graph of the truncated power spline $\langle U_t - 2.765 \rangle$.

A spline partitions the data samples into two classes, depending on the value of some feature. The value of the spline is zero for one of the classes and nonzero for the other classes. When a spline basis function is used in a linear sum, the contribution of members of the first class can be adjusted independently of the members of the second class. Linear regression assumes that the effect of a feature on the response is linear over its range. Regression with splines allows the incorporation of features which do not have the linear effect over their entire range.

Splines are interpreted as performing either *range identification* or *outlier removal*. If there are many members in the nonzero partition, then the spline is identifying a range of effect. For example, the interpretation of the term $\langle U_t - 2.765 \rangle$ in a model is that only high values for U_t have an effect on the response. If there are only a few members of the nonzero set, the spline is identifying outliers. Regression can use the spline term to fit these members independently of the other terms of the model by, effectively, making them "special cases" based on the extreme value of a feature.

The top ten functions discovered by the GFA algorithm are reported in Table 4. The list of the ten most frequently used basis functions is shown in Figure 12.

An example of outlier removal is the most frequently used spline, $\langle 2.301 - U_t \rangle$. The only compound for which this item is nonzero is compound 18, which was identified as one of the two outliers in the original study. All of the top ten models isolated compound **18** in this manner. Since the LOF score for the models extracts a cost for each added term, there was clearly a significant reduction in error attributable to the use of this term.

An example of range identification is the use of spline terms based on **HOMO**. For example, model 3 contains the term $\langle -9.545 - \mathbf{HOMO} \rangle^2$, which is nonzero for approximately half of the compounds: {**3, 5, 8, 11, 13, 14, 16, 17**}. The interpretation is that **HOMO** has an effect on activity only after it achieves a more negative value than −9.545. Four of the top ten most frequently used basis functions are of this form.

Splines identify possible partition points in the range of a feature which may be related to an underlying mechanism of activity. However, the correlation may be an artifact of the small size of the data set; this risk is accentuated since the spline term is nonzero only for some reduced number of compounds in the data set. Thus, it is important to consider whether the partitions may have physical meaning before accepting their predictiveness.

An interesting effect in the selection of splines for the top 10 models was the isolation of the least active compounds. If a histogram is constructed of the number of times each spline term is nonzero for each of the compounds in the data set, a strong skew toward the least active compounds is observed, as shown in Figure 13. The system is achieving the highest scores by building general models of the most-active compounds and taking the least-active as special cases. This may reflect different modes of activity of the least-active and most-active compounds.

Table 4. Top 10 Models Derived for the Cardozo/Hopfinger Data Set[a]

1: $-\log(IC_{50}) = 6.950$
$+ 2.046 * C_4$
$- 6.523 * \langle 2.301 - U_t \rangle$ {18}
$+ 1.037 * (U_t - 2.845)^2$
$- 22.090 * \langle -9.631 - \mathbf{HOMO} \rangle^2$ {16, 17}
LOF: 0.209
r: 0.923
F: 17.16

2: $-\log(IC_{50}) = 7.101$
$+ 2.040 * C_4$
$- 4.581 * \langle 2.301 - U_t \rangle$ {18}
$+ 2.988 * \langle U_t - 3.467 \rangle$ {1}
$- 7.178 * (\mathbf{HOMO} + 9.508)^2$
LOF: 0.216
r: 0.920
F: 16.55

3: $-\log(IC_{50}) = 7.053$
$+ 1.983 * \langle C_4 - 0.027 \rangle$ {15 16}
$- 6.468 * \langle 2.301 - U_t \rangle$ {18}
$+ 0.987 * (U_t - 2.845)^2$
$- 11.185 * \langle -9.545 - \mathbf{HOMO} \rangle^2$ {3, 5, 8, 11, 13, 14, 16, 17}
LOF: 0.221
r:= 0.918
F: 16.10

4: $-\log(IC_{50}) = 7.035$
$+ 1.771 * C_4$
$- 7.683 * \langle 2.301 - U_t \rangle$ {18}
$+ 1.241 * (U_t - 2.966)^2$
$- 18.192 * \langle -9.615 - \mathbf{HOMO} \rangle^2$ {3, 16, 17}
LOF: 0.224
r: 0.917
F: 15.84

5: $-\log(IC_{50}) = 7.080$
$+ 1.778 * \langle C_4 - 0.027 \rangle$ {15, 16}
$- 7.686 * \langle 2.301 - U_t \rangle$ {18}
$+ 1.242 * (U_t - 2.966)^2$
$- 18.196 * \langle -9.615 - \mathbf{HOMO} \rangle^2$ {3, 16, 17}
LOF: 0.224
r: 0.917
F: 15.82

6: $-\log(IC_{50}) = 7.138$
$+ 1.615 * \langle C_4 - 0.027 \rangle$ {15, 16}
$- 7.619 * \langle 2.301 - U_t \rangle$ {18}
$+ 1.231 * (U_t - 2.966)^2$
$- 4.616 * \langle -9.631 - \mathbf{HOMO} \rangle$ {16, 17}
LOF: 0.225
r: 0.917
F: 15.80

7: $-\log(IC_{50}) = 7.092$
$+ 1.756 * \langle C_4 - 0.027 \rangle$ {15, 16}
$- 7.667 * \langle 2.301 - U_t \rangle$ {18}
$+ 1.234 * (U_t - 2.966)^2$
$- 15.250 * \langle -9.593 - \mathbf{HOMO} \rangle^2$ {3, 8, 13, 16, 17}
LOF: 0.225
r: 0.917
F: 15.79

8: $-\log(IC_{50}) = 7.098$
$+ 1.607 * C_4$
$- 7.615 * \langle 2.301 - U_t \rangle$ {18}
$+ 1.231 * (U_t - 2.966)^2$
$- 4.610 * \langle -9.631 - \mathbf{HOMO} \rangle$ {16, 17}
LOF: 0.225
r: 0.917
F: 15.78

9: $-\log(IC_{50}) = 8.076$
$- 2.110 * \langle 0.468 - C_4 \rangle$ {all but 3 and 8}
$+ 3.354 * \langle U_t - 3.467 \rangle$ {1}
$- 12.972 * \langle 2.301 - U_t \rangle^2$ {18}
$- 8.219 * (\mathbf{HOMO} + 9.545)^2$
LOF: 0.225
r: 0.916
F: 15.74

10: $-\log(IC_{50}) = 7.056$
$+ 1.739 * C_4$
$- 21.539 * \langle 2.301 - U_t \rangle^2$ {18}
$+ 1.223 * (U_t - 2.966)^2$
$- 13.468 * \langle -9.577 - \mathbf{HOMO} \rangle^2$ {3, 5, 8, 11, 13, 16, 17}
LOF: 0.226
r: 0.916
F: 15.65

[a] Angle brackets are used to denote splines terms, zero if the contents are negative, otherwise the value of the contents. Curly brackets list the compound numbers for which that term is nonzero. Terms explored were linear and quadratic polynomials and linear and quadratic splines.

Basis function	# models	Nonzero samples	Comments
$\langle 2.301 - U_t \rangle$	153	{18}	Outlier removal
C_4	61	all	
$(U_t - 2.966)^2$	58	all but {5}	
$\langle 2.301 - U_t \rangle^2$	44	{18}	Outlier removal
$\langle U_t - 2.980 \rangle^2$	44	{1 6 7 12 13 15 16 17}	High values of U_t
$\langle -9.615 - \mathbf{HOMO} \rangle$	26	{3 16 17}	Highly negative HOMO
$\langle U_t - 3.467 \rangle$	21	{1}	Outlier removal
$\langle -9.545 - \mathbf{HOMO} \rangle^2$	20	{3 5 8 11 13 14 16 17}	Highly negative HOMO
$\langle -9.615 - \mathbf{HOMO} \rangle^2$	19	{3 16 17}	Highly negative HOMO
$\langle -9.577 - \mathbf{HOMO} \rangle^2$	19	{5 8 11 13 14 16 17}	Highly negative HOMO

Figure 12. Ten most frequently used basis functions in the population of 300 models. The first column contains the basis function; the second, the number of models which use that basis function; the third, the set of samples for which the function is nonzero; and the last column gives comments on the role the spline term is playing in the model.

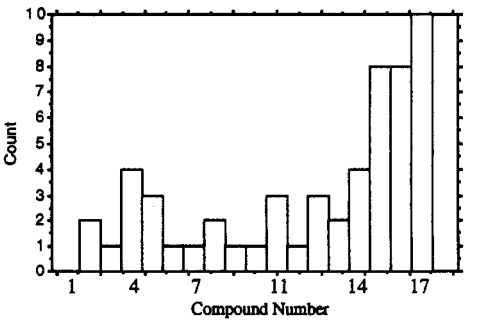

Figure 13. Histogram of the number of times a given compound was made a special case in the top ten models. The histogram shows a skew toward the highest-numbered compounds, which are also the least active. In effect, the GFA algorithm is making special cases of the least-active compounds and modeling the patterns it found in the most active analogs.

E. QSPR Data Sets. The QSPR data sets from Koehler and Hopfinger[17] contain seven features and either 35 or 30 compounds of structurally diverse polymers. The former data set was used to predict the property T_g, the glass transition temperature, and the latter data set was used to predict the property T_m, the melt transition temperature. The seven features were S_B and S_S, the backbone and side-chain contributions to the monomer conformational entropy; M_B and M_S, the backbone and side-chain mass moments; $\bar{E}_D$, $\bar{E}_+$, $\bar{E}_-$, the dispersion, positive electrostatic, and negative electrostatic intermolecular energies for the complete monomer unit. The data set for T_g is given in Table 5. The data set for T_m is given in Table 6.

Table 5. Koehler and Hopfinger Data Set for T_g, the Glass Transition Temperature[a]

compd no.	S_B	M_B	S_S	M_S	E_D	E_+	E_-	$T_g(1)$	$T_g(2)$	T_g (used)
1	3.33	15	0	0	−0.86	−1.53	0.69	188	243	243
2	3.65	14.7	0	0	−1.04	−0.87	0.37	206	246	206
3	3.82	14.5	0	0	−1.13	−0.54	0.22	195	228	228
4	3.91	14.4	0	0	−1.18	−0.34	0.12	185	194	194
5	4.3	14	0	0	−1.39	0.45	0.26	143	250	143
6	1.93	14	0	0	−1.51	0.51	−0.37	238	299	238
7	1	34.3	0	0	−3.26	−1.09	−0.59	353	380	353
8	1.7	24	0	0	−1.52	−0.81	−0.38	253	314	314
9	1.48	32.5	0	0	−1.69	−1.27	−0.16	238	286	286
10	1.06	32	0	0	−1.36	−1.03	−0.5	323	371	323
11	0.8	31	0	0	−1.93	−0.9	−0.46	247	354	354
12	1.12	58	0	0	−2.18	−1.92	−1.08	318	373	373
13	1.96	38.4	0	0	−2.59	−2.1	−0.22	346	346	346
14	0.58	56.5	0	0	−3.77	−2	−0.36	393	420	393
15	0.53	63.5	0	0	−4.13	−1.26	−0.53	414	423	414
16	0.76	28.3	1.7	28.5	−2.38	−2.29	−0.41	378	378	378
17	0.76	28.3	2.29	23.7	−2.15	−1.79	−0.56	338	338	338
18	0.76	28.3	3.95	15.7	−1.59	−0.14	−0.34	288	288	288
19	0.76	28.3	0.89	39	−2.72	−1.93	−0.42	380	380	380
20	0.76	28.3	0.49	43.7	−2.71	−2.41	−0.69	385	385	385
21	0.76	28	0	0	−2.5	0.77	−0.66	198	243	243
22	1.22	22	0	0	−1.83	−0.8	−1.88	343	372	343
23	1.48	12	2.89	14	−1.36	0.4	−0.32	228	249	228
24	2.92	12	3.36	14	−1.37	0.41	−0.31	221	287	221
25	2.92	14.5	3.45	15	−0.91	−1.58	0.36	242	260	242
26	3.29	14.5	3.28	14.7	−0.97	−1.2	0.27	231	254	231
27	3.29	23.6	1.92	29	−1.91	−2.49	0.35	279	282	282
28	2.38	23.6	2.44	24	−1.77	−1.95	0.05	251	251	251
29	2.38	23.6	1.1	39.5	−2.25	−2.13	0.33	314	314	314
30	2.38	23.6	3.19	20	−1.66	−1.26	−0.04	219	219	219
31	3.4	23.4	0	0	−1.84	−0.54	−0.31	268	298	268
32	0.59	59.5	0	0	−3.97	−2.34	−2.25	411	428	428
33	3.27	18.8	0	0	−1.65	−1.16	−1.58	318	330	318
34	2.92	12	3.83	14	−1.38	0.43	−0.29	208	228	208
35	2.92	12	3.86	14.3	−1.28	0.02	−0.21	196	223	223

[a] S_B and S_S are the backbone and side-chain contributions to the monomer conformational entropy; M_B and M_S are the backbone and side-chain mass moments; $\bar{E}_D$, $\bar{E}_+$, $\bar{E}_-$ are the dispersion, positive electrostatic, and negative electrostatic probe energies for the monomer unit. Some of the compounds have two experimental values for T_g. The final column is the observed value for T_g which Koehler and Hopfinger compared their predictions against, and the one used to train the GFA models. The polymers corresponding to each row in the table are reported in ref 17.

Koehler and Hopfinger proposed separate models of five features and six terms for T_g and T_m. These models are shown in Figure 14.

F. GFA Applied to the QSPR Data Sets. Genetic Function Approximation was applied to the QSPR data set to illustrate the applicability of the analysis process to QSPR problems.

Two separate analysis were conducted for the two variables T_g and T_m. Each analysis with GFA began with a population of 300 random models. The terms of the models were linear polynomials and linear splines. The population was evolved for 5000 crossover operations.

The top 10 models discovered by the GFA algorithm for T_g are shown in Table 7; the top 10 models for T_m are shown in Table 8.

The T_g models discovered and rated best have a small improvement in the correlation coefficient and fewer terms than the QSPR of Koehler and Hopfinger. The feature use appears different in the GFA models as compared to the original QSPR. For example, S_S, M_B, and M_S are not used in any of the top 10 models for T_g, while E_D (which was not used in the original QSPR) was used in 9 out of 10. In fact, the ability to build models which are competitive with the original QSPR, but which contain only features relating to energy is an interesting result, and parallels a discussion in the original paper,[17] which suggests (and rejects) the possibility of not considering mass moments or side chain conformational energy.

Other patterns emerge that may be of note. For example, a pair of spline terms based on $\bar{E}$, but with opposite signs, appears in 8 out of 10 of the top models. These pairs appear to be isolating a central region of values for $\bar{E}_-$ that has the most effect on the value of T_g. The relatively large coefficients of these terms suggests caution, as we may be seeing the amplification of a chance pattern in the data set, but it is certainly worth presenting to the researcher for consideration.

The T_m models discovered and rated best also show improvement in the correlation coefficient over the original QSPR, and have fewer terms. The feature use is different, both from the original QSPR and from the patterns of the T_g models. Seven of the 10 top models use only $\bar{E}_D$, $\bar{E}_+$, $\bar{E}_-$, and S_S. There were two models of only four terms, and one model of six terms. No example of the pairing of dual $\bar{E}_-$ spline terms was seen. None of the spline terms were found to isolate one or two outliers in the data set. Instead, the splines seem to be performing identification of ranges of the variables that may be of interest. For example, 8 out of the top 10 models use the spline term $\langle E_D + 1.660 \rangle$, which separates out the 14 data compounds with $E_D > -1.660$. Whether this is due to any underlying mechanism that would make only that range of the feature important needs to be determined.

The differences in feature use between the T_g models and the T_m models is best illustrated by graphing the use of features in the population of models as evolution proceeds. This is shown in Figure 15. Some features, such as $\bar{E}_+$ and $\bar{E}_-$ are similar in their use. Others, such as S_B, are quite different, being greatly used to predict one but not both of the transition temperatures. Again, it can be seen that studying the populations of models, and comparing populations, can be a

Table 6. Koehler and Hopfinger Data Set for T_m, the Melt Transition Temperature[a]

compd no.	S_B	M_B	S_S	M_S	E_D	E_+	E_-	$T_m(1)$	$T_m(2)$	T_m (used)
1	3.33	15	0	0	−0.86	−1.53	0.69	333	473	333
2	3.65	14.7	0	0	−1.04	−0.87	0.37	335	349	335
3	3.82	14.5	0	0	−1.13	−0.54	0.22	308	308	308
4	3.91	14.4	0	0	−1.18	−0.34	0.12	308	333	333
5	4.3	14	0	0	−1.39	0.45	−0.26	410	410	410
6	1.93	14	0	0	−1.51	0.51	−0.37	385	481	385
7	1	34.3	0	0	−3.26	−1.09	−0.59	498	523	498
8	1.7	24	0	0	−1.52	−0.81	−0.38	473	473	473
9	1.48	32.5	0	0	−1.69	−1.27	−0.16	410	511	410
10	0.76	28	0	0	−2.5	0.77	−0.66	275	317	317
11	0.8	31	0	0	−1.93	−0.9	−0.46	485	583	485
12	1.12	58	0	0	−2.18	−1.92	−1.08	483	533	533
13	0.76	28.3	1.7	28.5	−2.38	−2.29	−0.41	433	473	473
14	2.92	14.5	3.45	15	−0.91	−1.58	0.36	417	423	417
15	1.48	12	2.89	14	−1.36	0.4	−0.32	359	359	359
16	2.38	23.6	3.19	20	−1.66	−1.26	−0.04	275	317	317
17	3.4	23.4	0	0	−1.84	−0.54	−0.31	396	411	396
18	0.84	48	0	0	−3	−1.36	−0.38	463	483	463
19	0.91	31	0	0	−1.96	−2.13	−0.8	292	672	672
20	2.92	12	2.89	14	−1.36	0.4	−0.32	379	415	379
21	2.92	12	3.83	14	−1.38	0.43	−0.29	235	235	235
22	2.38	23.6	2.91	21.5	−1.71	−1.55	0	388	435	388
23	1.56	38	0	0	−2.72	−2.47	−2.59	728	728	728
24	3.43	21.3	0	0	−1.72	−0.48	−0.1	338	338	338
25	3.29	14.5	4.02	14.2	−1.23	−0.18	−0.05	280	280	280
26	2.35	34.3	0	0	−2.39	−1.67	−0.23	533	537	533
27	3.24	21.5	0	0	−1.75	−1.1	−0.24	332	332	332
28	3.27	18.8	0	0	−1.65	−1.16	−1.58	523	545	545
29	2.39	29.2	0	0	−2.3	−1.2	−1.4	606	613	606
30	3.87	17	0	0	−1.53	−0.17	−0.27	344	358	358

[a] S_B and S_S are the backbone and side-chain contributions to the monomer conformational entropy; M_B and M_S are the backbone and side-chain mass moments; $\bar{E}_D$, $\bar{E}_+$, $\bar{E}_-$ are the dispersion, positive electrostatic, and negative electrostatic probe energies for the monomer unit. Some of the compounds have two experimental values for T_m. The final column is the observed value for T_m which Koehler and Hopfinger compared their predictions against, and the one used to train the GFA models. The polymers corresponding to each row in the table are reported in ref 17.

$$T_g = 288.83 - 27.3 * S_B - 10.1 * S_S + 1.07 * M_B - 29.3 * \bar{E}_+ - 15.1 * \bar{E}_-$$

n: 35
r: 0.954
F: 60.51

$$T_m = 493.7 - 32.6 * S_B - 22.1 * S_S - 2.51 * M_B - 50.5 * \bar{E}_+ - 109.8 * \bar{E}_-$$

n: 30
r: 0.907
F: 23.91

Figure 14. Proposed QSPR models for T_g and T_m.

powerful tool in analyzing a data set.

4. CONCLUSION

The genetic function approximation (GFA) algorithm offers a new approach to the problem of building activity models. Replacing standard regression analysis with the GFA algorithm allows the construction of models competitive with, or superior to, "standard" techniques and makes available additional information not provided by other techniques.

A fundamental difference between GFA and other techniques is the creation and use of multiple models rather than a single model. While one can simply select the model from the population with the lowest LOF score, it is usually preferable to inspect the different models and select, with the aid of scientific intuition, using the appropriateness of the features, the basis functions, and the combinations. The population can be studied for information on feature use, and predictions can often be improved by averaging the results of multiple models rather than relying on an individual model.

The method of model construction also has important consequences. Most techniques choose features incrementally and may not find combinations whose components are not

Table 7. Top 10 Models for the QSPR T_g Data Set

1: T_g = 332.836
+ -46.059 * <$\bar{E}_+$ + 2.130>
+ -21.538 * <$\bar{E}_D$ + 3.970>
+ 221.511 * <-0.040 - $\bar{E}_-$>
+ -219.826 * <-0.460 - $\bar{E}_-$>
LOF: 840.801
r: 0.964
F: 97.34

2: T_g = 269.490
+ -48.654 * <$\bar{E}_+$ + 1.920>
+ 25.335 * <-1.660 - $\bar{E}_D$>
+ 232.012 * <-0.040 - $\bar{E}_-$>
+ -229.866 * <-0.460 - $\bar{E}_-$>
LOF: 844.545
r: 0.963
F: 96.88

3: T_g = 333.626
+ -48.684 * <$\bar{E}_+$ + 2.130>
+ 212.309 * <-0.040 - $\bar{E}_-$>
+ -196.174 * <-0.370 - $\bar{E}_-$>
+ -17.642 * S_B
LOF: 859.638
r: 0.963
F: 95.04

4: T_g = 330.184
+ -40.268 * <$\bar{E}_+$ + 2.130>
+ -23.916 * $\bar{E}_D$
+ 438.901 * <$\bar{E}_-$ + 0.260>
+ -466.898 * <$\bar{E}_-$ + 0.410>
LOF: 874.026
r: 0.962
F: 93.36

5: T_g = 375.022
+ -44.262 * <$\bar{E}_+$ + 2.130>
+ 27.015 * <-1.910 - $\bar{E}_D$>
+ 422.346 * <$\bar{E}_-$ + 0.260>
+ -469.224 * <$\bar{E}_-$ + 0.410>
LOF: 874.603
r: 0.962
F: 93.29

6: T_g = 306.729
+ -44.767 * <$\bar{E}_+$ + 1.950>
+ 19.550 * <-1.520 - $\bar{E}_D$>
+ 196.267 * <-0.040 - $\bar{E}_-$>
+ -182.987 * <-0.370 - $\bar{E}_-$>
+ -12.993 * S_B
LOF: 876.948
r: 0.971
F: 97.01

7: T_g = 177.068
+ -44.893 * $\bar{E}_+$
+ 23.826 * <-1.660 - $\bar{E}_D$>
+ 234.735 * <-0.040 - $\bar{E}_-$>
+ -234.500 * <-0.460 - $\bar{E}_-$>
LOF: 878.854
r: 0.962
F: 92.80

8: T_g = 425.653
+ -40.164 * <$\bar{E}_+$ + 2.130>
+ -24.153 * <$\bar{E}_D$ + 3.970>
+ 440.733 * <$\bar{E}_-$ + 0.260>
+ -468.315 * <$\bar{E}_-$ + 0.410>
LOF: 885.056
r: 0.962
F: 92.10

9: T_g = 290.656
+ -72.806 * <$\bar{E}_+$ + 0.540>
+ -28.495 * $\bar{E}_D$
+ 24.502 * <0.120 - $\bar{E}_-$>
+ -25.612 * S_B
LOF: 885.998
r: 0.962
F: 91.99

10: T_g = 294.253
+ -71.257 * <$\bar{E}_+$ + 0.540>
+ -28.585 * $\bar{E}_D$
+ 25.333 * <-0.040 - $\bar{E}_-$>
+ -26.336 * S_B
LOF: 888.349
r: 0.961
F: 91.73

Table 8. Top 10 Models for the QSPR T_m Data Set

1: T_m = 473.471
+ -79.526 * E_+
+ 206.516 * <E_D + 1.660>
+ -275.659 * <E_- + 0.800>
+ -95.585 * <S_S - 2.910>
LOF: 3090.626
r: 0.934
F: 42.45

2: T_m = 473.216
+ -79.625 * E_+
+ 205.330 * <E_D + 1.660>
+ -274.773 * <E_- + 0.800>
+ -93.295 * <S_S - 2.890>
LOF: 3092.823
r: 0.934
F: 42.42

3: T_m = 476.864
+ -76.994 * E_+
+ 195.629 * <E_D + 1.660>
+ -279.549 * <E_- + 0.800>
+ -201.405 * <S_S - 3.450>
LOF: 3180.766
r: 0.932
F: 41.07

4: T_m = 657.982
+ -80.012 * <E_+ + 2.290>
+ 209.310 * <E_D + 1.660>
+ -278.816 * <E_- + 0.800>
+ -95.818 * <S_S - 2.910>
LOF: 3204.616
r: 0.931
F: 40.72

5: T_m = 488.296
+ 104.078 * <-0.340 - E_+>
+ -115.934 * <E_- + 1.400>
+ -3.261 * M_S
LOF: 3210.991
r: 0.917
F: 45.95

6: T_m = 486.084
+ 95.324 * <-0.180 - E_+>
+ -120.007 * <E_- + 1.400>
+ -2.973 * M_S
LOF: 3249.374
r: 0.916
F: 45.31

7: T_m = 445.224
+ -88.847 * E_+
+ 245.339 * <E_D + 1.660>
+ -319.856 * <E_- + 0.660>
+ -98.103 * <S_S - 2.910>
LOF: 3272.867
r: 0.930
F: 39.74

8: T_m = 468.177
+ -85.367 * E_+
+ 168.978 * <E_D + 1.660>
+ -254.480 * <E_- + 0.800>
+ -89.452 * <S_S - 2.910>
+ -5.986 * <M_S - 15.000>
LOF: 3273.182
r: 0.942
F: 37.94

9: T_m = 657.682
+ -78.982 * <E_+ + 2.290>
+ 213.580 * <E_D + 1.660>
+ -285.085 * <E_- + 0.800>
+ -133.628 * <S_S - 3.190>
LOF: 3273.205
r: 0.930
F: 39.74

10: T_m = 435.377
+ 81.740 * <0.400 - E_+>
+ 205.992 * <E_D + 1.660>
+ -271.999 * <E_- + 0.800>
+ -94.093 * <S_S - 2.910>
LOF: 3278.962
r: 0.929
F: 39.66

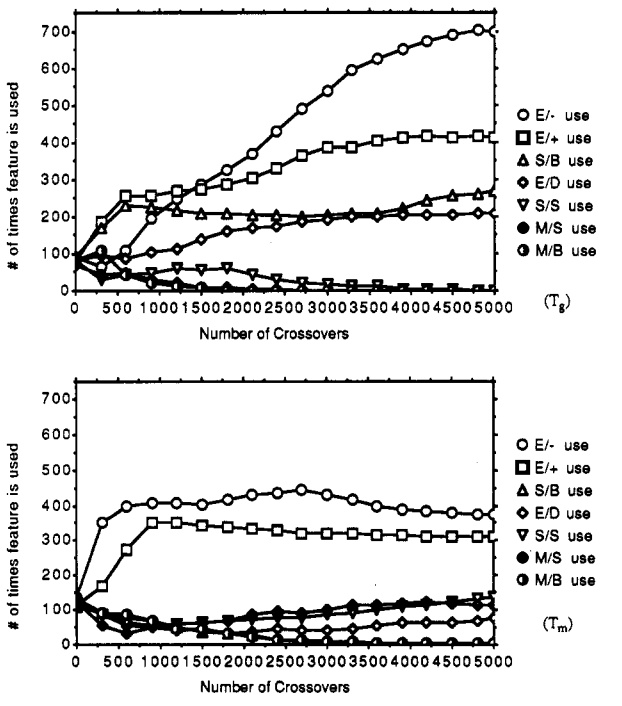

Figure 15. Feature use versus the number of crossovers for T_g and T_m. The top graph shows the information for T_g; the bottom graph shows the information for T_m. The largest difference is in the use of S_B, which is often used for building T_g models but not T_m models.

predictive individually. In contrast, by testing full-size models, rather than incrementally building them, the GFA algorithm is better at discovering combinations of basis functions that take advantage of correlations only available in combination.

Limitations of some other algorithms are not present in the GFA algorithm. For example, GFA can build models using only linear polynomials, but also higher-order polynomials, splines, and Gaussians. By choosing the appropriate basis function types for a given data set, models which will be composed of the most appropriate terms can be constructed.

A form of automatic outlier removal is performed by GFA if spline-based terms are included in the model construction. This is a consequence of the partitioning behavior of spline-based terms. For features with limited ranges of effect, the use of splines provide important information about the location of each range of effect and assists in the avoidance of terms which extrapolate poorly due to a false assumption of a linear effect.

The above results demonstrate that GFA discovers models that are, at the least, important candidates that must be considered in a search for the "best" predictive model. Of itself this would be significant. Combined with the other attributes of the algorithm, and other information that the algorithm makes available, we believe that this confirms the GFA algorithm as a valuable technique for quantitative structure–activity relationship and quantitative structure–property relationship analyses.

ACKNOWLEDGMENT

D.R. wishes to acknowledge his parents, Cecilia and Philip, who wanted him to become a scientist; Pentti Kanerva, who supported this work in its early stages; Molecular Simulations Inc. and Mick Savage, who saw the potential of the theoretical work and helped it become a practical tool; and Doug Brockman, his domestic partner, who made him work harder on this than he had ever worked before.

REFERENCES AND NOTES

(1) Hopfinger, A. J.; Burke, B. J. In *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, C. A., Eds.; Wiley: New York, 1990, p 173.

(2) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.*, **1988**, *110*, 5959–5967.

(3) Wikel, J.; Dow, E. The Use of Neural-Networks for Variable Selection in QSAR. *Bioorg. Med. Chem. Lett.* **1993**, *3*, 645–651.

(4) Rogers, D. G/SPLINES: A Hybrid of Friedman's Multivariate Adaptive Regression Splines (MARS) Algorithm with Holland's Genetic Algorithm. *The Proceedings of the Fourth International Conference on Genetic Algorithms*, San Diego, July 1991.

(5) Rogers, D. Data Analysis using G/SPLINES. *Advances in Neural Processing Systems 4*; Kaufmann; San Mateo, CA, 1992.

(6) Hansch, C.; Fujita, T. ρ–σ–π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* **1964**, *86*, 1616.

(7) Glen, W. D.; Dunn, W. J.; Scott, R. D. Principal Components Analysis and Partial Least Squares Regression. *Tetrahedron Comput. Methodol.* **1989**, *2*, 349–376.

(8) Leardi, R.; Boggia, R.; Terrile, M. Genetic Algorithms as a Strategy for Feature Selection. *J. Chemom.* **1992**, *6*, 267–281.

(9) Good, A. C.; So, S.; Richards, W. G. Structure–Activity Relationships from Molecular Similarity Matrices. *J. Med. Chem.* **1993**, *36*, 433–438.

(10) Holland, J. *Adaptation in Artificial and Natural Systems*; University of Michigan Press: Ann Arbor, MI, 1975.

(11) Friedman, J. Multivariate Adaptive Regression Splines, Technical Report No. 102, Laboratory for Computational Statistics, Department of Statistics, Stanford University; Stanford, CA, Nov 1988 (revised Aug 1990).

(12) Breiman, L.; Friedman, J.; Olshen, R.; Stone, C. *Classification and Regression Trees*; Wadsworth: Belmont, CA, 1984.

(13) Bentley, J. Multidimensional Binary Search Trees used for Associative Searching. *Commun. ACM* **1975**, *18*, 509–517.

(14) Dunn, W. J., III; Greenberg, M. J.; Callejas, S. S. Use of Cluster Analysis in the Development of Structure–Activity Relations for Antitumor Triazines. *J. Med. Chem.* **1976**, *19*, 1299–1301.

(15) Selwood, D. L.; Livingtone, D. J.; Comley, J. C.; O'Dowd, A. B.; Hudson, A. T.; Jackson, P.; Jandu, K. S.; Rose, V. S.; Stables, J. N. *J. Med. Chem.* **1990**, *33*, 136.
(16) Cardozo, M. G.; Iimura, Y.; Sugimoto, H.; Yamanishi, Y.; Hopfinger, A. J. QSAR Analysis of the Substituted Indanone and Benzylpiperidine Rings of a Series of Indanone–Benzylpiperidine Inhibitors of Acetylcholinesterase. *J. Med. Chem.* **1992**, *35*, 584–589.
(17) Koehler, M. G.; Hopfinger, A. J. Molecular modelling of polymers: 5. Inclusion of intermolecular energetics in estimating glass and crystal-melt transition temperatures. *Polymer*, **1989**, *30*, 116–126.
(18) Livingstone, D. J.; Hesketh, G.; Clayworth, D. Novel Method for the Display of Multivariate Data Using Neural Networks. *J. Mol. Graphics* **1991**, *9*, 115–118.
(19) 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*, 6–15.
(20) Rose, V. S.; Wood, J.; MacFie, H. J. H. Single Class Discrimination Using Principal Component Analysis (SCD–PCA). *Quant. Struct.-Act. Relat.* **1991**, *10*, 359–368.
(21) Rose, V. S.; Wood, J.; MacFie, H. J. H. Generalized Single Class Discrimination (GSCD). A New Method for the Analysis of Embedded Structure-Activity Relationships. *Quant. Struct.-Act. Relat.* **1992**, *11*, 492–504.
(22) McFarland, J. W.; Gans, D. J. On Identifying Likely Determinants of Biological Activity in High-Dimensional QSAR Problems. *Quant. Struct.-Act. Relat.*, in press.