Multiobjective Optimization in Quantitative Structure−Activity Relationships: Deriving Accurate and Interpretable QSARs

Orazio Nicolotti,[†] Valerie J. Gillet,*[,†] Peter J. Fleming,[‡] and Darren V. S. Green[§]

Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield, Western Bank, Sheffield S10 2TN, United Kingdom, Department of Automatic Control and Systems Engineering, University of Sheffield, Western Bank, Sheffield S10 2TN, United Kingdom, and GlaxoSmithKline, Gunnels Wood Road, Stevenage SG1 2NY, United Kingdom

Received April 25, 2002

Deriving quantitative structure−activity relationship (QSAR) models that are accurate, reliable, and easily interpretable is a difficult task. In this study, two new methods have been developed that aim to find useful QSAR models that represent an appropriate balance between model accuracy and complexity. Both methods are based on genetic programming (GP). The first method, referred to as genetic QSAR (or GPQSAR), uses a penalty function to control model complexity. GPQSAR is designed to derive a single linear model that represents an appropriate balance between the variance and the number of descriptors selected for the model. The second method, referred to as multiobjective genetic QSAR (MoQSAR), is based on multiobjective GP and represents a new way of thinking of QSAR. Specifically, QSAR is considered as a multiobjective optimization problem that comprises a number of competitive objectives. Typical objectives include model fitting, the total number of terms, and the occurrence of nonlinear terms. MoQSAR results in a family of equivalent QSAR models where each QSAR represents a different tradeoff in the objectives. A practical consideration often overlooked in QSAR studies is the need for the model to promote an understanding of the biochemical response under investigation. To accomplish this, chemically intuitive descriptors are needed but do not always give rise to statistically robust models. This problem is addressed by the addition of a further objective, called chemical desirability, that aims to reward models that consist of descriptors that are easily interpretable by chemists. GPQSAR and MoQSAR have been tested on various data sets including the Selwood data set and two different solubility data sets. The study demonstrates that the MoQSAR method is able to find models that are at least as good as models derived using standard statistical approaches and also yields models that allow a medicinal chemist to trade statistical robustness for chemical interpretability.

Introduction

Quantitative structure−activity relationships (QSARs) attempt to relate a numerical description of molecular structure to known biological activity. Hansch pioneered the approach by demonstrating that biological activity could be correlated to a few simple thermodynamic or electronic variables using a simple regression equation. Since this first analysis, two significant developments have been made.[1] The first is that a wide range of easily computable molecular descriptors is now available, and the second is that many sophisticated techniques have emerged that are a significant improvement over the original linear regression analysis.

Despite the developments that have taken place in QSAR, deriving models that are accurate, reliable, and easily interpretable remains a difficult task. While the availability of large numbers of easily computable descriptors such as topological indices, substructural keys, and two-dimensional (2D) and three-dimensional (3D) fingerprints can help in providing a variety of different ways of describing structures, it can also make the task of deriving accurate and easily interpretable QSAR models harder. Complexity in QSAR can be due to a number of different factors including the number of terms included, the mathematical operators and functions used to combine the terms, the inclusion of nonlinear and cross-terms, and the inclusion of descriptors that are hard to interpret. Often these factors conflict with model accuracy, with more accurate models tending to be more complex and hence harder to interpret and vice versa.

When there are more descriptors available than data points, the use of inappropriate analysis methods can lead to overfitting of the data with the generation of models that have poor predictive ability. In these cases, the number of descriptors should be reduced in order to develop a model that is predictive and easier to interpret. A systematic search for the best subset of features is generally not possible since there are a total of $2^N - 1$ possible subsets of features for a data set consisting of N descriptors. For example, as McFarland and Gans[2] have noted, there are 9×10^{15} possible combinations of descriptors for the well-known Selwood

* To whom correspondence should be addressed. Tel. +44 1142 222 652. Fax: +44 1142 780 300. E-mail. v.gillet@sheffield.ac.uk.
[†] Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield.
[‡] Department of Automatic Control and Systems Engineering, University of Sheffield.
[§] GlaxoSmithKline.

10.1021/jm020919o CCC: $22.00
Published on Web 10/15/2002

data set[3] that is typically characterized by 53 descriptors. The computational cost associated with feature selection has resulted in a number of different algorithms being developed for feature selection and QSAR generation, such as principal component analysis,[4] nonlinear mapping,[5] partial least squares,[6] neural networks,[7] and evolutionary algorithms.[8]

The Selwood data set has been well-studied in QSAR and has become a standard against which new methods are tested. Selwood's[3] initial approach involved using forward-stepping multivariate regression analysis to obtain a three descriptor model. However, the stepwise nature of this procedure fails to take into account any coupled effects between the descriptors, and subsequently, Wikel and Dow[7] derived an improved three descriptor model using a neural network as the descriptor selection method.

Rogers and Hopfinger developed the genetic function approximation (GFA)[9] method where descriptor selection is performed using a genetic algorithm (GA) and QSAR models are obtained by performing least squares regression to regenerate the coefficients. The models are scored using Friedman's lack of fit (LOF) measure, which is based on the least squares error combined with a user definable smoothing parameter that penalizes the effect of including additional terms in a model. They found improved three descriptor models in addition to good two and four descriptor models. Reducing the smoothing parameter led to 4−6 descriptor models with modest improvements in prediction, estimated using cross-validation.

The mutation and selection uncover models (MUSEUM) algorithm[10,11] developed by Kubinyi is based on an evolutionary algorithm involving mutation only (that is, there is no crossover operator). It avoids the need for a user-defined parameter by using the FIT value as the fitness criterion. The FIT value is based on the Fischer significance value adjusted with respect to the number of independent variables selected in each model. In the related evolutionary programming (EP) method developed by Luke,[8] fitness is defined using a three term function. The first term is the root mean square (RMS) between predicted and measured values, the second term is used to drive the solution toward a given number of descriptors, and the final term is used to weight the descriptors according to their exponent values; for example, quadratic terms are penalized relative to linear terms. Both Kubinyi's and Luke's methods were able to find three descriptor models that had not been found previously.

The previous methods are limited to finding linear models. So and Karplus[12] developed a hybrid method that combines a GA for descriptor selection with an artificial neural network for model building. They found improved models for the Selwood data set where the improvement appears to be due to the selection of nonlinear descriptors. The neural network was able to explore nonlinear relationships without the need to examine each possible nonlinearity. More recently, a novel algorithm based on the fast random elimination of descriptors (FRED)[13] has been proposed that was able to find the same solutions for the Selwood data sets as the previous methods.

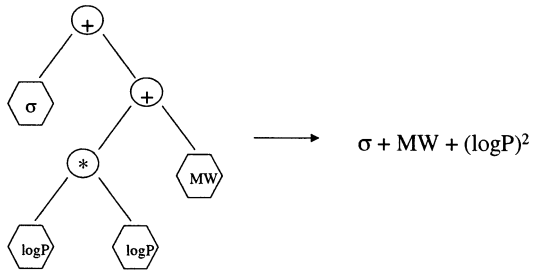

Figure 1. Example of the tree representation used in GP together with the decoded mathematical expression that it represents.

We have developed two approaches to deriving QSARs that seek to balance model accuracy with complexity. Both approaches are based on genetic programming (GP). In the first, the balance between the accuracy and the number of terms in the model is controlled via the use of a penalty function. The second approach is based on a multiobjective GP (MOGP) method in which a family of equivalent models is found, where each model represents one particular compromise between accuracy and complexity. In the latter approach, several objectives are used to control complexity including the number of terms, the number of nonlinear terms, and a knowledge-based objective that is able to drive the solutions toward descriptors that are easily interpretable. The method is able to find models that are at least as good as models derived using standard statistical approaches and also yields models that allow a medicinal chemist to trade statistical robustness for chemical interpretability. In the following section, we give a brief overview of GP before describing the approaches that we have developed.

Computational Methods

GP. GP[14] is a branch of GAs and is based on the principles of Darwinian evolution and survival of the fittest. The main difference between GP and GAs is the representation of potential solutions. In GAs, an individual is usually represented as a fixed length linear string. In GP, however, an individual is represented as a tree, which can vary in shape and size as the population undergoes evolution. Thus, the complexity of the representation is increased relative to a GA. GP was originally developed to evolve computer programs, or mathematical expressions, where an individual is represented as a parse tree. The internal nodes of the tree represent mathematical operators or mathematical functions, and the terminal nodes represent variables or constant values. An individual is evaluated by converting the tree into the corresponding mathematical expression. The process is illustrated in Figure 1 where the internal nodes are mathematical functions and the terminal nodes are molecular descriptors.

GP has been applied to many problems such as automated design, pattern recognition, robot control, symbolic regression, music generation, image compression, and image analysis. Despite the widespread application of GAs to problems in computer-aided molecular design[15] such as ligand docking, pharmacophore detection, and variable selection in QSAR, the applications of GP in the field have been more limited. Examples include the use of GP to evolve molecules to

fit a QSAR or quantitative structure−property relationship (QSPR)[16] and the use of GP to design molecules based on 2D similarity to a target compound.[17] Both of these examples exploit the relationship between 2D chemical structures and graph theory where the GP manipulates the structures directly.

GP begins with the definition of a set of functions and terminals that are appropriate for the domain. An initial population of trees is then generated consisting of random compositions of nodes. Each tree represents a mathematical expression, or computer program, and the fitness function consists of executing the computer program and assigning a fitness value according to how well it solves the problem in question. GP then enters an iterative cycle where in each iteration a new population is generated by applying the genetic operators reproduction, mutation, and crossover.

Each of the genetic operators involves the selection of one or more parent chromosomes where the probability of a chromosome being selected is proportional to its fitness. The reproduction operator involves selecting one parent chromosome, which is copied unchanged into the next generation. In mutation, a single parent is selected and a mutation point is chosen at random. The subtree at the mutation point is deleted, and a new subtree is grown. In crossover, two parents are selected, which are usually of different shape and size, a crossover point is chosen at random in each parent, and subtrees are exchanged. Crossover is the predominant operator in GP and is performed with a high probability relative to mutation and reproduction. The iterations continue until some convergence criterion has been reached when the best solution found is designated to be the result of the GP.

Applying GP to QSAR. QSAR is a regression problem where an attempt is made to relate a numerical description of molecular structure or properties to a known biological activity. An example QSAR model is shown in eq 1

$$y_{\text{pred}} = ax_1 + bx_2 + ...cx_n + d \qquad (1)$$

where y_{pred} is the predicted or calculated activity; x_i represents the variables or molecular descriptors used in the model; a, b, and c are coefficients; and d is a constant.

GP can be used to solve this problem by defining the function set to be a set of mathematical operators, for example, $F = \{+, -, *, \sin, \cos, \exp, \log\}$, and the terminal set to be the independent variables and the coefficients and constants. The fitness function then involves converting the tree representation into the corresponding mathematical expression, applying the model to the known data points, and measuring how well it is able to predict the known activities.

A GP approach to deriving QSAR has been implemented in the program GPQSAR. Here, the terminal set is limited to the set of molecular descriptors available for a data set, and coefficients and constants for the model are calculated during the fitness function itself, as described below. The function set has been restricted to the sum operator, i.e., $F = \{+\}$ to allow the method to be compared with existing published methods. (The minus operator is implicit as will be seen later.)

One potential problem with GP is that there can be a tendency to generate large and complex trees that result in overfitting of the data. This can cause difficulties for QSAR since, as already mentioned, it is desirable to achieve a balance between model accuracy and model complexity in order that the model can be used to make predictions about previously unknown compounds. The usual way of evaluating a model in QSAR is to use r^2, which is the squared correlation between the y response (the activity) and a set of variables (descriptors). It is well-known that r^2 tends to increase as the number of variables increases, so that more complex models tend to be more accurate. However, a good QSAR model is considered to be one with a small number of terms and a high value of r^2. Thus, using r^2 as the fitness function is inappropriate since this would naturally favor more complex models.

Many approaches to QSAR have tackled this difficulty by specifying the exact number of terms required; however, this is often difficult since it is not easy to know beforehand the number of terms that is likely to give rise to a good model and often several runs will be performed with varying numbers of terms.

In GP, one way in which tree complexity can be controlled is to restrict the maximum number of nodes in a tree or the maximum depth allowed. However, appropriate limits may vary from one problem to another depending on the particular relationship that exists between accuracy and model complexity. An alternative approach is used here with model complexity being controlled through the use of a penalty function. This approach has the advantage that it is not necessary to set arbitrary limits on the size of the trees.

In GPQSAR, model complexity is controlled through the use of the Akaike Information Criterion (AIC).[18] The AIC function is used to identify an appropriate model structure when choosing between models. Specifically, when two or more competing models can explain data, then the model with the smallest number of parameters should be chosen. This principle is known as Occam's razor or the law of parsimony.

The AIC penalty function is shown below (eq 2)

$$\text{AIC} = \log(\sigma_\theta) + kp/N \qquad (2)$$

where σ_θ, is the variance of the residuals and is a measure of the performance of the model; N is the number of fitness cases or data points; p is the number of terms in the model; and k is a penalty factor. It can be seen that the inclusion of additional terms in the model increases the complexity of the model and is therefore penalized according to the value of k. Calibration of k is required in order that an appropriate balance is found between the residual variance σ_θ and the number of terms p. (The calibration of k is described in the Experimental Section.) Thus, because the AIC function is used to control model complexity, there is no need to set limits on the number of nodes or the depth of tree allowed.

The fitness function of GPQSAR involves the following three steps. (i) The expression encoded in an individual is extracted to determine the descriptors that will be used in the QSAR model, i.e., the x_i values in eq 1. (ii) Optimum values for coefficients and constants ($a-d$) in the model are calculated by applying the least squares

method (LSM). Positive coefficients indicate a favorable relationship between a given descriptor, x_i, and the response, y_{pred}, and negative coefficients indicate the reverse. (The production of negative coefficients during the LSM means that it is not necessary to include the minus operator in the function set.) (iii) Fitness is measured using the AIC function described above. Note that the number of terms in the model, p, is not necessarily the same as the number of terminal nodes in an individual since a given descriptor can appear more than once; for example, a subtree containing the expression $x_1 + x_1$ consists of two terminal nodes but the expression unfolds to a single term, $2x_1$.

Various parameters of GPQSAR are user definable, for example, population size; the maximum number of generations; and the relative probabilities of the breeding operators. In the runs described in the Experimental Section, the population size was set to 200 and the maximum number of generations was set to 5000. Crossover, mutation, and reproduction were assigned with relative probabilities of 0.7, 0.2, and 0.1, respectively, and the trees were free to grow with no limits on the number of nodes or depth.

Multiobjective Optimization. The aim of GPQSAR is to derive a single model that represents an appropriate balance between the variance and the number of terms. However, in general, there will be a family of equivalent models, where each model represents a different compromise in the objectives. Thus, QSAR is an example of a multiobjective optimization problem that typically comprises two (or more) competitive objectives.

In GPQSAR, the multiobjective QSAR problem is effectively reduced to a single objective optimization problem through the use of the AIC function, which involves summing the different objectives to give a single value. The result of running GPQSAR is a single model only, with the particular compromise between the objectives being determined by the value of k used in the AIC function. A further disadvantage of the GPQSAR method is the need to calibrate k for each data set.

Evolutionary algorithms such as GAs and GP are well-suited to the true optimization of multiobjective problems.[19] Both GAs and GP have been adapted for multiobjective optimization in the development of the multiobjective GA (MOGA)[20] and the MOGP,[18] respectively.

Both MOGAs and MOGP are based on the idea of Pareto optimality where a Pareto optimal, or nondominated, solution is one where another solution does not exist in the population that is better than it in all of the objectives. As a result, one solution dominates another if it is either equivalent, or better, in all of the objectives and strictly, it is better in at least one objective. For simplicity, consider a two objective problem f_1 and f_2 where the aim is to minimize both objectives, as shown in Figure 2. Each point in the graph represents a pair of values, which are the objectives f_1 and f_2. A solution is nondominated if the square area bounded by the axes and lines drawn parallel with the axes from the point does not include any other point.

Pareto frontiers have been used in many applications of multiobjective optimization; see, for example, the review article by Coello Coello.[21] The first application

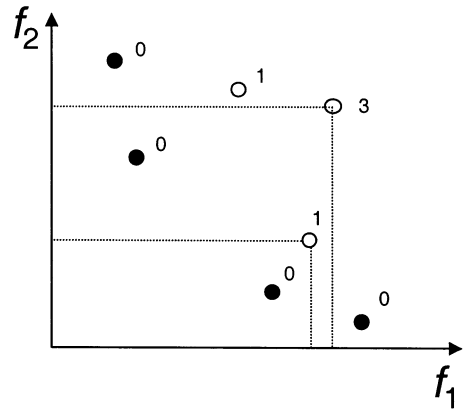

Figure 2. Potential solutions to a two objective (f_1 and f_2) problem. The solid circles are nondominated solutions and fall on the Pareto frontier. Dominated solutions are shown as unfilled circles. In MoQSAR, individuals are ranked according to the number of times they are dominated; thus, nondominated solutions are given rank zero and the dominated solutions are given ranks as shown.

in chemoinformatics of which we are aware is the work of Handschuh et al.[22] who used Pareto optimization in the GA they developed for the flexible superposition of 3D structures. Their method finds the maximum common substructures, MCSS, between two molecules. The search for the MCSS involves two criteria: the number of atoms in the substructure and the fit of the matching atoms. These are conflicting criteria since a larger MCSS will by definition have a larger deviation in the coordinates of the superimposed atoms when the larger MCSS is a superset of the smaller. Rather than attempting to combine the different criteria into a single weighted sum fitness function, a set of Pareto solutions is obtained at the end of each run whereby an optimal geometric fit is found for each possible size of MCSS.

More recently, the concept of Pareto optimality has been applied to combinatorial library design in the program MoSELECT, which is based on a MOGA.[23−25] In MoSELECT, a typical library design scenario would be to design a library that is simultaneously diverse, cheap to synthesize, and has druglike physicochemical properties.

Here, the multiobjective approach to deriving QSARs has the aim of generating QSARs that are simultaneously accurate, reliable, and easy to interpret. To this end, the GP in GPQSAR has been adapted to a MOGP in the program MoQSAR. The definition of the individuals and their representation as parse trees in MoQSAR are the same as in GPQSAR. The main difference to GPQSAR is that the AIC fitness function is no longer used and fitness is now calculated on the basis of Pareto ranking determined from the values of the individual objectives, without the need for summation. Each time a new individual is generated, the expression encoded in the tree is extracted, LSM is run to calculate the best coefficients for the model, and the objectives are calculated independently and stored. Because each objective is treated independently, accuracy can now be measured using r^2. Using the number of terms p as a second objective allows the relationship between r^2 and p to be explored directly.

After each generation, the individuals are ranked according to the number of individuals in the population

by which they are dominated. Thus, nondominated individuals are assigned rank zero, individuals that are dominated by one individual are assigned rank one, and so on. The fitness of an individual is then determined by its rank, or dominance, with individuals having low dominance being preferred to individuals with higher dominance.

The result of MoQSAR is a family of models where each model represents a different compromise in the objectives. A further advantage as compared to the AIC function used in GPQSAR is that calibration is no longer required to find a balance between the accuracy and the number of terms. In fact, for a two objective problem where the objectives are r^2 and p, there will be one solution for each distinct number of terms and that solution will be the one with the best value of r^2 generated throughout the GP for the given number of terms.

MOGP (and MOGAs) can be applied to include any number of objectives, not just two, and the following section describes objectives additional to r^2 and p that have been investigated so far in the QSAR context. In MoQSAR, the function set has been extended to include quadratic and cubic terms and the objective s is used to penalize models that contain higher order terms over linear models. Thus, s can be thought of as a measure of the internal complexity of a model. Models with linear terms only are assigned $s = 1$; s is incremented by one for each quadratic term and by two for each cubic term contained within the model.

The final objective, called desirability, relates to the chemical interpretability of the descriptors used in the QSAR model. So far, we have described QSAR as a compromise between different objectives that are essentially statistical parameters. However, QSAR is not only a mathematical problem aimed at relating a numerical description of molecular structure or properties to known biological activity; it is also important that the models can be used predictively, to suggest new compounds for testing. To be useful for prediction purposes, models should be constructed from descriptors that are easily interpretable by chemists. However, analyses involving large numbers of descriptors often include some descriptors whose meaning can be obscure. In such cases, models having poorer statistics may be preferable to more accurate models that contain hard to interpret descriptors.

The chemical desirability objective, D, is a knowledge-based function designed to drive the selection of descriptors in QSAR models toward those that have been flagged as desirable. Each descriptor is assigned a desirability value, d, of 3, 2, or 1, which corresponds to the categories: excellent, fair, or poor, respectively. The assignments are user definable and should be based on a priori knowledge or chemical intuition. The chemical desirability of the model is then calculated as

$$D = (\text{avg} \times \text{geomean} + 1)/(n \times \text{geomean}) \quad (3)$$

where avg is the arithmetic mean, $\sum_{i=1}^{p} d_i/p$, and geomean is the geometric mean, $\sqrt[p]{d_1.d_2...d_p}$, respectively; d_i is the desirability of the ith descriptor in the model; and n is the number of unique descriptors included in the model. For a model consisting of linear terms only, n is the same as p; however, it is possible for the same descriptor to appear as both a linear term and a power term in which case it is counted only once and $n < p$. Thus, D is used to reward the presence of more desirable descriptors.

Table 1. Descriptors for the Selwood Data Set[a]

ID	descriptor	*d*	ID	descriptor	*d*	ID	descriptor	*d*
X1	ATCH1	2	X19	ESDL5	1	X37	MOFI_X	1
X2	ATCH2	2	X20	ESDL6	1	X38	MOFI_Y	1
X3	ATCH3	2	X21	ESDL7	1	X39	MOFI_Z	1
X4	ATCH4	2	X22	ESDL8	1	X40	PEAX_X	1
X5	ATCH5	2	X23	ESDL9	1	X41	PEAX_Y	1
X6	ATCH6	2	X24	ESDL10	1	X42	PEAX_Z	1
X7	ATCH7	2	X25	NSDL1	1	X43	MOL_WT	3
X8	ATCH8	2	X26	NSDL2	1	X44	S8_1DX	3
X9	ATCH9	2	X27	NSDL3	1	X45	S8_1DY	3
X10	ATCH10	2	X28	NSDL4	1	X46	S8_1DZ	3
X11	DIPV_X	2	X29	NDSL5	1	X47	S8_1CX	1
X12	DIPV_Y	2	X30	NDSL6	1	X48	S8_1CY	1
X13	DIPV_Z	2	X31	NDSL7	1	X49	S8_1CZ	1
X14	DIPMOM	2	X32	NDSL8	1	X50	LOGP	3
X15	ESDL1	1	X33	NDSL9	1	X51	M_PNT	1
X16	ESDL2	1	X34	NDSL10	1	X52	SUM_F	3
X17	ESDL3	1	X35	VDVWOL	3	X53	SUM_R	3
X18	ESDL4	1	X36	SURF_A	3			

[a] The descriptors are as follows: partial atomic charges for atoms 1−10 (ATCH1−ATCH10); dipole vector (DIPV_X, DIPV_X, DIPV_X); dipole moment (DIPMOM); electrophilic superdelocalizability for atoms 1−10 (ESDL1−ESDL10); nucleophilic superdelocalizability for atoms 1−10 (NSDL1−NSDL10); van der Waals volume (VDWVOL); surface area (SURF_A); principal moments of inertia (MOFI_X, MOFI_Y, MOFI_Z); principal elipsoid axes (PEAX_X, PEAX_Y, PEAX_Z); molecular weight (MOL_WT); substituent dimensions (S8_1DX, S8_1DY, S8_1DZ); substituent centers (S8_1CX, S8_1CY, S8_1CZ); partition coefficient (LOGP); melting point (M_PNT); sums of the F and R substituent constants (SUM_F, SUM_R). The columns headed ID give codes assigned to the descriptors in the subsequent tables. The columns headed *d* represent user-defined chemical desirability values.

As with GPQSAR, various parameters are configurable, and for the runs described in the Experimental Section, MoQSAR was run for 2000 generations with a population size of 200, with the breeding parameters unchanged from the GPQSAR runs. The maximum number of nodes allowed in a tree was limited to 14, which means that the maximum number of terms that can be included in a model is seven.

Results and Discussion

(1) Data Sets. The GPQSAR approach has been tested on the Selwood data set,[3] which consists of 31 compounds, 53 descriptors, and a set of corresponding antifilarial antimycin activities, expressed as $-\log(IC_{50})$. The molecular descriptors are listed in Table 1.

The MoQSAR approach has been tested on three data sets. The first is the Selwood data set already described. The other two data sets are solubility data sets where MoQSAR is used to find QSPRs. The first solubility data set was supplied by Huuskonen[26] and consists of 1272 structures with logP values and with solubility as the dependent variable, given as logS. The solubility values are in the range of −11.62 to +1.58 log units. In addition, 72 Molconn-Z parameters were calculated as molecular descriptors.[27] These include simple and valence molecular connectivity indices, simple and valence difference connectivity indices, shape indices, electrotopological state indices, and hydrogen bond donor indices. The full list of molecular descriptors is reported

Table 2. Descriptors for the Aquax Data Set[a]

ID	descriptor	*d*	ID	descriptor	*d*	ID	descriptor	*d*
X1	LOGP	3	X26	dx1	2	X51	ka3	3
X2	x0	2	X27	dx2	2	X52	si	1
X3	x1	2	X28	dxp3	2	X53	Totop	2
X4	x2	2	X29	dxp4	2	X54	sumI	3
X5	xp3	2	X30	dxp5	2	X55	sumdelI	3
X6	xp4	2	X31	dxp6	2	X56	tets2	2
X7	xp5	2	X32	dxp7	2	X57	Phia	3
X8	xp6	2	X33	dxp8	2	X58	SHsOH	3
X9	xp7	2	X34	dxp9	2	X59	SHdNH	3
X10	xp8	2	X35	dxv0	2	X60	SHsSH	3
X11	xp9	2	X36	dxv1	2	X61	SHsNH2	3
X12	xv0	3	X37	dxv2	2	X62	SHssNH	3
X13	xv1	3	X38	dxvp3	2	X63	SHtCH	3
X14	xv2	3	X39	dxvp4	2	X64	SHother	3
X15	xvp3	3	X40	dxvp5	2	X65	SHCHnX	3
X16	xvp4	3	X41	dxvp6	2	X66	Hmax	3
X17	xvp5	3	X42	dxvp7	2	X67	Gmax	2
X18	xvp6	3	X43	dxvp8	2	X68	Hmin	3
X19	xvp7	3	X44	dxvp9	2	X69	Gmin	2
X20	xvp8	3	X45	k0	1	X70	Hmaxpos	1
X21	xvp9	3	X46	k1	1	X71	SHHBD	3
X22	xc3	2	X47	k2	1	X72	SHHBA	3
X23	xc4	2	X48	k3	1	X73	Qv	3
X24	xpc4	2	X49	ka1	3			
X25	dx0	2	X50	ka2	3			

[a] The reader is referred to the Molconn-Z manual for descriptions of properties 2−73. The columns headed ID give codes assigned to the descriptors in the subsequent tables. The columns headed *d* represent user-defined chemical desirability values.

Table 3. Descriptors for the Nitrogen Data Set[a]

ID	descriptor	*d*	ID	descriptor	*d*	ID	descriptor	*d*
X1	MDE_14	3	X15	x2	2	X29	k1	1
X2	GEOM_3	3	X16	xv0	3	X30	k2	1
X3	PPSA_1	3	X17	xv1	3	X31	k3	1
X4	FPSA_1	3	X18	xv2	3	X32	ka1	3
X5	SCDH_2	3	X19	xvp3	3	X33	ka2	3
X6	NN	3	X20	dx0	2	X34	ka3	3
X7	NSB	3	X21	dx1	2	X35	Si	1
X8	WTPT_2	3	X22	dx2	2	X36	Totop	2
X9	EAVE_2	3	X23	dxp3	2	X37	SumI	3
X10	GEOM_1	3	X24	dxv0	2	X38	SumdelI	3
X11	FPSA_2	3	X25	dxv1	2	X39	tets2	2
X12	CTDH	3	X26	dxv2	2	X40	Phia	3
X13	x0	2	X27	dxvp3	2			
X14	x1	2	X28	k0	1			

[a] The descriptors are as follows: MDE_14 is the molecular distance edge between all primary and quaternary carbons; GEOM_3 is the third geometric moment; PPSA_1 is the summation of the positive surface area; FPSA_1 is the positive surface area divided by the total surface area; SCDH_2 is the average surface area times charge on donatable hydrogens; NN is the number of nitrogens; NSB is the number of single bonds; WTPT_2 is the sum of unique weighted paths divided by the total number of atoms; EAVE_2 is the average E-state value over all heteroatoms; GEOM_1 is the first geometric moment; FPSA_2 is the fractional charged partial surface area; CTDH is the number of donatable hydrogens. The reader is referred to the Molconn-Z manual for descriptions of X13−X40. The columns headed *d* represent user-defined chemical desirability values.

in Table 2. This data set is referred to as the Aquax data set.

The second QSPR data set was supplied by McElroy and Jurs[28] and consists of 176 organic compounds with associated solubility values ranging from −7.41 to 0.96 log units and 12 descriptors. An additional 28 descriptors were calculated using Molconn-Z. The full list of molecular descriptors is reported in Table 3. The compounds contain a minimum of one nitrogen atom, zero or more oxygen atoms, and zero or more halogens per molecule. This data set is referred to as the Nitrogen data set.

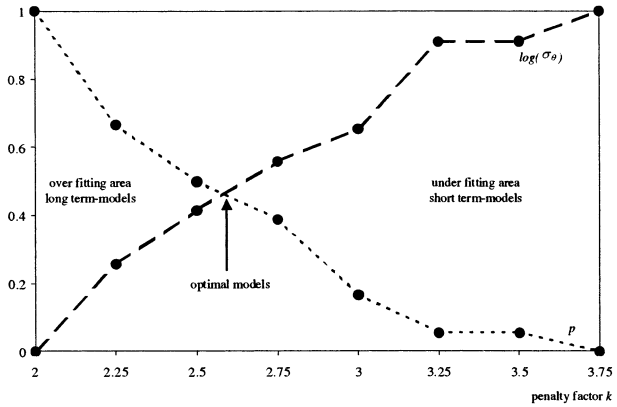

Figure 3. Calibrating the penalty factor *k* in the AIC function for the Selwood data set. The dashed line shows the normalized residual variance (log(σ_θ)), and the dotted line shows the normalized number of terms. In both cases, the results are averaged over five runs at different values of *k*. The optimum value of *k* is chosen as the intersection point.

(2) Application of GPQSAR to the Selwood Data Set. The program was parametrized to combine the 53 physical property descriptors (Table 1) using only the "+" operator (note that this can result in both + and "−" arithmetic in the QSAR equation, according to the signs of the coefficients generated during the GP).

For GPQSAR to function properly, an appropriate value of *k*, the penalty factor, must be determined, to balance the accuracy of the model, i.e., how well the model fits the training data measured by the residual variance, log(σ_θ), with the number of terms, *p*. This was achieved by carrying out several runs of GPQSAR for different values of *k*. Figure 3 shows the relationship between log(σ_θ) and *p* observed during the calibration runs. As can be seen, high values of *k* lead to models with relatively high residual variance and a low number of terms. These models are called short term QSAR models and represent a poor fit to the data points. Conversely, low values of *k* lead to models with low residual variance and a high number of terms. These models are called long term models since they represent a good fit to the data points, but to do so, they employ complex QSARs. Optimal models are expected to be those that represent a compromise between short and long term models, that is, models that are neither too simple nor too complex. The intersection point in Figure 3 between the log(σ_θ) and the *p* curves is taken as the optimal *k* value; hence, *k* was set at 2.58.

Ten runs of GPQSAR were then carried out. Table 4 lists the best models found (some of the models were found more than once as indicated) where q^2, calculated using leave-one-out (LOO) cross-validation, is also reported. As can be seen, the results consist of models with between three and six molecular descriptors. The method has been successful in finding literature models,[2,3,7−11,13] with five of the models corresponding to the top five models reported by Kubinyi[11] for this data set.

Table 4. Models Found by Applying GPQSAR to the Selwood Data Set Using AIC as the Fitness Function with the Penalty Factor k Set to 2.58[a]

descriptors	r^2	q^2	AIC	N
$4.824\ X4 + 12.018\ X5 - 0.114\ X11 - 5.019 \times 10^{-5} X39 + 0.402\ X50 - 1.277$	0.826	0.696	0.510	2
$4.712\ X4 + 12.406\ X5 - 0.118\ X11 - 5.027 \times 10^{-5} X38 + 0.406\ X50 - 1.268$	0.826	0.696	0.510	3
$2.669\ X4 - 0.182\ X11 - 6.150 \times 10^{-5} X38 - 0.035\ X48 + 0.471\ X50 - 2.577\ X52 - 2.209$	0.853	0.754	0.499	1
$2.886\ X4 - 0.176\ X11 - 6.102 \times 10^{-5} X39 + 0.033\ X48 + 0.462\ X50 + 2.488\ X52 - 2.179$	0.853	0.751	0.499	1
$2.328\ X4 - 0.130\ X11 - 7.415 \times 10^{-5} X38 + 0.499\ X50 + 2.049\ X52 - 1.821$	0.813	0.692	0.478	2
$-7.488 \times 10^{-5} X38 + 0.584\ X50 + 1.514\ X52 - 2.501$	0.721	0.647	0.470	1

[a] The solutions are sorted by their AIC values. N is the number of times each model was found. q^2 is also reported for each model.

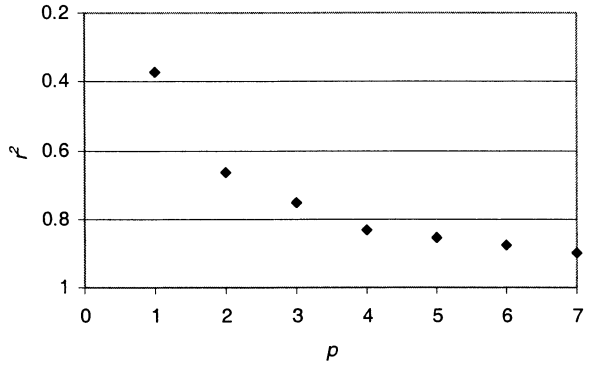

Figure 4. Models found for the Selwood data set using MoQSAR parametrized to optimize the two objectives, r^2 and the number of terms p.

The use of the AIC penalty function as the fitness function in GPQSAR has therefore proved to be useful in finding good QSAR models and has the advantage over existing methods that the user does not need to specify the number of terms required in the model. Varying sized models are explored during the search process, and the solution found is the model that reflects the best compromise between the residual variance and the number of terms. However, as already stated, disadvantages of the approach are that a single model only is found and that k must be calibrated for each data set. Hence, the next experiments were carried out using MoQSAR, which does not suffer from these limitations.

(3) Application of MoQSAR to the Selwood Data Set. Initially, MoQSAR was used to find models for the Selwood data set based on optimizing two objectives only, namely, the correlation of predicted vs observed response, r^2, and the number of terms, p. The goal was to maximize r^2 while minimizing p. This time, the program was allowed to use +, quadratic, and cubic functions with which to combine the 53 descriptors. Thus, nonlinear terms were allowed when building QSAR models (although cross-terms were not allowed).

The models resulting from a single run of MoQSAR are shown in Figure 4 where r^2 is plotted against p. Seven solutions were found with each solution being the best model found for a given number of terms, for example, the best model consisting of a single term, the best model consisting of two terms, up to the best model consisting of seven terms. As expected, the graph shows that accuracy is in conflict with model complexity with r^2 increasing as the number of terms increases. Each of the solutions found represents a different compromise between r^2 and p. The models corresponding to these solutions are shown in Table 5, where q^2, the squared correlation coefficient of prediction calculated using the LOO procedure, is also reported. Improved values of r^2 and q^2 are achieved relative to GPQSAR optimization, due to the presence of quadratic and cubic terms, which were permitted in the function set. The best four term QSAR model found includes a cubic term (SURF_A) and a quadratic term (SUM_R) and is shown in eq 4.

$$-\log(EC_{50}) = 0.67183\ LOGP - 2.8519 \times 10^{-8}\,(SURF\_A)^3 + 1.8824\ SUM\_F + 17.485\,(SUM\_R)^2 + 3.68537$$

$$r^2 = 0.830;\ q^2 = 0.782 \qquad (4)$$

In general, the presence of higher order relationships in QSAR models is undesirable; hence, the next experiment investigated the effect of including an additional objective in the search, namely, the number of nonlinear terms, s. The goal of the optimization was to maximize r^2 and to minimize p and s. Results are shown in Figure 5 where a parallel coordinates graph is used to illustrate the relationship between the three objectives. In this representation, each line in the graph represents a nondominant solution to the problem, indicating the achieved objective values for that solution. The competing nature of the objectives is shown clearly by the crossing lines with the more accurate models consisting of larger numbers of terms and also containing terms that are nonlinear.

Including the third objective, s, results in more than one QSAR model for a given number of terms with the total number of solutions increasing from seven for the two objective case to 16 for three objectives. Statistical details of the models are given in Table 6. All linear 3−6 term models are included in Kubinyi's best list. Additional models are also found, which contain nonlinear terms. Three four term models were identified as follows: one consisting of linear terms only (with $s = 1$), one including a quadratic term (with $s = 2$), and one including a cubic term (with $s = 3$). All of these models represent simpler models than the previously reported four term model, when optimizing r^2 and p only. The linear model is shown in eq 5 where it can be seen that the simpler model is achieved at the expense of some loss in fit to the data points relative to eq 4.

$$-\log(EC_{50}) = 0.49984\ LOGP + 2.8075\ ATCH4 + 0.84222\ ESDL3 - 0.19960\ PEAK\_X + 1.7908$$

$$r^2 = 0.774;\ q^2 = 0.636 \qquad (5)$$

The model includes the well-understood descriptor logP and a descriptor based on atomic charge (ATCH4), which is also readily interpretable. However, relationships between electrophilic superdelocalizability (ESDL3) and biological response and between principle ellipsoid axis (PEAK_X) and biological response are not

Table 5. Models Shown in Figure 4 that Were Found by Applying MoQSAR to the Selwood Data Set Optimized on r^2 and p^a

descriptors	r^2	p	q^2
$177.467(X7)^3 + 2.338$	0.371	1	0.303
$-2.827 \times 10^{-8}(X36)^3 + 0.676\ X50 - 2.037$	0.663	2	0.606
$-2.493 \times 10^{-8}(X36)^3 + 0.572\ X50 + 1.336\ X52 - 2.344$	0.754	3	0.692
$-2.852 \times 10^{-8}(X36)^3 + 0.672\ X50 + 1.882\ X52 + 17.485(X53)^2 - 3.685$	0.830	4	0.782
$0.101\ X20 - 2.986 \times 10^{-8}(X36)^3 + 0.679\ X50 + 1.970\ X52 + 18.637(X53)^2 - 3.590$	0.855	5	0.805
$0.119\ X20 - 3.178 \times 10^{-8}(X36)^3 - 0.138\ X49 + 0.682\ X50 + 2.239\ X52 + 22.223(X53)^2 - 3.852$	0.876	6	0.826
$0.035\ X18 - 0.285\ X30 - 3.453 \times 10^{-8}(X36)^3 - 0.144\ X49 + 0.710\ X50 + 2.612\ X52 + 23.347(X53)^2 - 3.789$	0.897	7	0.835

[a] q^2 is also reported for each model.

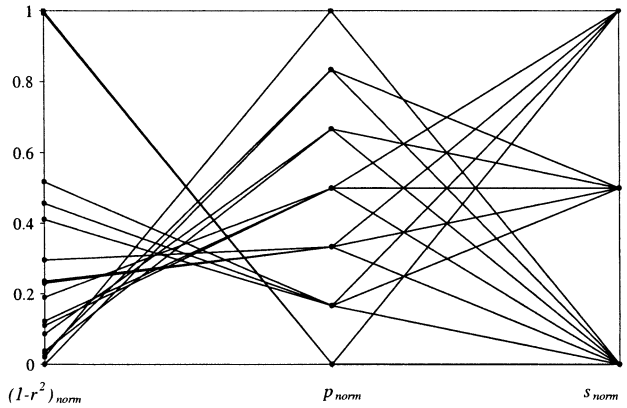

Figure 5. Parallel coordinates graph representation is shown of models found for the Selwood data set optimized on three objectives (r^2, p, and s). Normalized values of the objectives are plotted on the y axis with r^2 plotted as $1 - r^2$ so that the direction of improvement in all objectives is toward zero. Each continuous line in the graph represents one solution.

Table 6. Descriptors of the Models Shown in Figure 5 that Were Found by Applying MoQSAR to the Selwood Data Set Optimized on r^2, p, and s^a

descriptors	r^2	p	s	q^2
X6	0.367	1	1	0.309
$(X7)^3$	0.371	1	3	0.303
X39 X50	0.610	2	1	0.533
$(X36)^2$ X50	0.641	2	2	0.573
$(X36)^3$ X50	0.663	2	3	0.606
X38 X50 X52	0.721	3	1	0.647
$(X38)^2$ X50 X52	0.751	3	2	0.692
$(X36)^3$ X50 X52	0.754	3	3	0.692
X4 X17 X40 X50	0.774	4	1	0.636
X4 X5 $(X39)^2$ X50	0.809	4	2	0.682
X4 X5 $(X39)^3$ X50	0.814	4	3	0.697
X4 X5 X11 X39 X50	0.826	5	1	0.696
X4 X5 X11 $(X39)^2$ X50	0.851	5	2	0.733
X4 X11 X39 X48 X50 X52	0.853	6	1	0.751
X4 X5 X6 X11 $(X39)^2$ X50	0.870	6	2	0.767
X4 X5 X6 X11 X33 X39 X50	0.859	7	1	0.392

[a] The coefficients have not been included for reasons of brevity. q^2 is also reported for each model.

particularly intuitive. Thus, although the model is statistically sound, it does not promise to engage the chemists' imagination.

MoQSAR was then used to investigate the effect of including chemical desirability, D, as an objective. The molecular descriptors were assigned desirability values as shown in the columns headed d in Table 1. Examples of descriptors rated as excellent are dipole moment (DIPMOM), molecular weight (MOL_WT), van der Waals volume (VDWVOL), surface area (SURF_A), and LOGP. These are all descriptors that are intuitive to the chemist. MoQSAR was run with the four objectives r^2, p, s, and D_{norm}, where D is calculated from the individual values of the descriptors (see Table 1 and eq 3) and D_{norm} is D normalized by the best and worst values achievable for a given number of descriptors.

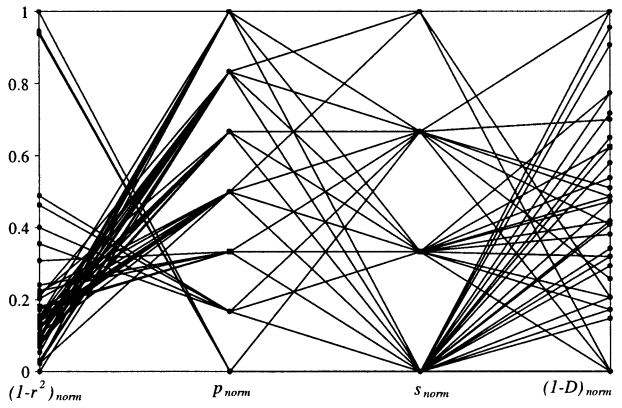

Figure 6. Parallel coordinates graph representation is shown of the solutions found for the Selwood data set optimized on four objectives (r^2, p, s, and D_{norm}). The normalized values of the objectives are plotted, and r^2 and D_{norm} are plotted as $1 - r^2$ and $1 - D_{norm}$, respectively, so that the direction of improvement is toward zero on the y axis.

The results are shown in Figure 6 as a parallel coordinates graph representation where it can be seen that the introduction of desirability as a fourth objective has resulted in a further increase in the number of solutions. A total of 44 QSARs were identified during the MoQSAR run. The 3−5 term models are reported in Table 7. The four term model with the highest r^2 was found previously and has also been reported by Kubinyi; however, this model has a relatively poor desirability rating. Three additional four term models, not reported by Kubinyi, were also found in which accuracy has been traded for desirability. The linear four term model with the best desirability rating is shown in eq 6.

$$-\log(EC_{50}) = 0.46825\ LOGP - 1.9043 \times 10^{-2}\ VDVWOL + 7.0068 \times 10^{-3}\ MOL\_WT + 1.3645\ SUM\_F + 0.14079$$

$$r^2 = 0.730;\ q^2 = 0.616 \qquad (6)$$

All of the descriptors are rated as excellent. The model has slightly poorer statistics than the model found previously (eq 5); however, it represents a much more intuitive model and hence may represent a better compromise in the objectives. Thus, MoQSAR has yielded a model in which statistical robustness has been traded for chemical interpretability. Such models will not be found by traditional optimization methods that are based on optimizing model accuracy alone.

Table 7. Descriptors for the Three, Four, and Five Term Models Shown in Figure 6 that Were Found by Applying MoQSAR to the Selwood Data Set Optimized on r^2, p, s, and D[a]

descriptors	r^2	p	s	D_{norm}	q^2
X38 X50 X52	0.721	3	1	0.611	0.647
X36 X50 X52	0.687	3	1	1.000	0.586
$(X38)^2$ X50 X52	0.751	3	2	0.611	0.692
$(X36)^2$ X50 X52	0.729	3	2	1.000	0.657
$(X36)^3$ X50 X52	0.754	3	3	1.000	0.692
X4 X5 X39 X50	0.772	4	1	0.403	0.624
X4 X39 X50 X52	0.756	4	1	0.552	0.619
X38 X43 X50 X52	0.740	4	1	0.704	0.646
X35 X43 X50 X52	0.730	4	1	1.000	0.616
X39 X50 X52 $(X53)^2$	0.789	4	2	0.704	0.720
X36 $(X36)^2$ X50 X52	0.766	4	2	1.000	0.682
X13 $(X36)^3$ X50 X52	0.777	4	3	0.839	0.696
$(X35)^3$ X43 X50 X52	0.768	4	3	1.000	0.688
$(X36)^3$ X50 X52 $(X53)^2$	0.830	4	4	1.000	0.782
X4 X5 X38 X43 X50	0.800	5	1	0.516	0.640
X4 X36 X37 X50 X52	0.782	5	1	0.638	0.630
X11 X14 X36 X50 X52	0.765	5	1	0.744	0.655
X9 X35 X43 X50 X52	0.764	5	1	0.871	0.668
X35 X43 X44 X50 X52	0.735	5	1	1.000	0.591
X4 X10 $(X39)^2$ X50 X52	0.804	5	2	0.516	0.673
X36 X38 X50 X52 $(X53)^2$	0.802	5	2	0.761	0.705
X8 X36 X50 X52 $(X53)^2$	0.799	5	2	0.871	0.729
X35 X36 $(X36)^2$ X50 X52	0.773	5	2	1.000	0.648
X4 X11 $(X36)^3$ X50 X52	0.802	5	3	0.744	0.668
X4 X36 $(X36)^3$ X50 X52	0.800	5	3	0.839	0.610
$(X35)^3$ X36 X43 X50 X52	0.780	5	3	1.000	0.680

[a] The coefficients have not been included for reasons of brevity. q^2 is also reported for each model.

Table 8. Descriptors for the Four Term Models Found by Applying MoQSAR to the Selwood Data Set Optimized on r^2, p, s, D, and q^{2a}

descriptors	r^2	p	s	D_{norm}	q^2
X4 X5 X39 X50	0.772	4	1	0.403	0.624
X4 X39 X50 X52	0.756	4	1	0.552	0.619
X11 X38 X50 X52	0.755	4	1	0.552	0.660
X12 X38 X50 X52	0.745	4	1	0.552	0.665
X38 X43 X50 X52	0.740	4	1	0.704	0.646
X35 X43 X50 X52	0.730	4	1	1.000	0.616
X8 X36 X50 X52	0.730	4	1	0.839	0.641
X4 $(X39)^2$ X50 X52	0.803	4	2	0.552	0.685
X38 X50 X52 $(X53)^2$	0.799	4	2	0.704	0.740
X36 $(X36)^2$ X50 X52	0.766	4	2	1.000	0.682
$(X36)^2$ X50 X52 $(X53)^2$	0.807	4	3	1.000	0.751

[a] The coefficients have not been included for reasons of brevity.

The final run on the Selwood data set was based on optimizing q^2, calculated using LOO cross-validation, in addition to the previous objectives. The four term models found are reported in Table 8. Two extra linear models are found as compared to those found in the previous run. One of these models is shown in eq 7.

$$-\log(EC_{50}) = 0.57740\ \mathrm{LOGP} - 1.3356 \times 10^{-2}\ \mathrm{SURF\_A} + 1.3728\ \mathrm{SUM\_F} + 12.977\ \mathrm{ATCH8} - 4.2123$$
$$r^2 = 0.730;\ q^2 = 0.641 \qquad (7)$$

The model has the same value of r^2 as the previous model (eq 6) but has a higher q^2. However, this has been achieved at the expense of some chemical desirability; hence, in this case, improved prediction has been traded with interpretability of descriptors. The descriptors van der Waals volume and molecular weight, both rated as excellent, have been replaced by surface area, also rated excellent, and an atomic charge descriptor, which is rated as fair only.

Table 9. Descriptors for Models Found by Applying MoQSAR to the Aquax Data Set Optimized on r^2, p, s, and D[a]

descriptors	r^2	p	s	D_{norm}	q^2
X1	0.687	1	1	1.000	0.684
X1 X54	0.798	2	1	1.000	0.796
X1 X54 X67	0.830	3	1	0.786	0.828
X1 X12 X68	0.818	3	1	1.000	0.815
X1 $(X1)^2$ X4	0.831	3	2	0.681	0.828
X1 $(X1)^2$ X54	0.827	3	2	1.000	0.824
X1 X54 X61 X67	0.838	4	1	0.839	0.836
X1 X12 X61 X68	0.827	4	1	1.000	0.825
X1 $(X1)^2$ X54 X67	0.844	4	2	0.786	0.841
X1 $(X1)^2$ X12 X68	0.837	4	2	1.000	0.835
X1 X36 X54 X55 X67	0.843	5	1	0.744	0.840
X1 X54 X58 X61 X67	0.842	5	1	0.871	0.840
X1 X12 X61 X66 X68	0.834	5	1	1.000	0.831
X1 $(X1)^2$ X54 X58 X67	0.853	5	2	0.839	0.850
X1 $(X1)^2$ X12 X66 X68	0.844	5	2	1.000	0.841
X1 X36 X54 X58 X67 X68	0.846	6	1	0.786	0.843
X1 X54 X58 X61 X67 X68	0.845	6	1	0.892	0.842
X1 X12 X18 X61 X66 X68	0.834	6	1	1.000	0.831
X1 $(X1)^2$ X54 X58 X67 X71	0.856	6	2	0.871	0.853
X1 $(X1)^2$ X12 X61 X66 X68	0.847	6	2	1.000	0.844
X1 $(X1)^2$ X54 X58 X68 $(X73)^2$	0.848	6	3	1.000	0.845
X1 X15 X25 X35 X54 X66 X73	0.85	7	1	0.816	0.847
X1 X12 X54 X61 X66 X67 X68	0.849	7	1	0.908	0.846
$(X1)^2$ X1 X12 X54 X58 X67 X68	0.858	7	2	0.892	0.855

[a] The coefficients have not been included for reasons of brevity. q^2 is also reported for each model. The two term model was used to predict logS for the 272 structures in the test set.

(3) Application of MoQSAR to the Aquax Data Set. The Aquax data set was partitioned at random into a training and a test set made up of 1000 and 272 structures, respectively. The program was allowed to use +, quadratic, and cubic functions with which to combine the 73 molecular descriptors reported in Table 2. Desirability values were assigned to each of the descriptors on the basis of ranks indicated in the Molconn-Z manual, and the values are shown in the columns headed *d* in Table 2.

Table 9 summarizes the results of running MoQSAR with the four objectives, r^2, p, s, and D_{norm}. A family of 22 models was found consisting of up to seven variables. LOGP occurs in all of the models, which is not surprising since it is known to be highly relevant in explaining molecular solubility; hence, this is consistent with chemical intuition.

The two term linear model reported in Table 9 is discussed here. The full QSAR notation for the model is given in eq 8.

$$\mathrm{logS} = -0.795\ \mathrm{LOGP} - 0.041\ \mathrm{SUML} + 0.555$$
$$r^2 = 0.798;\ q^2 = 0.796 \qquad (8)$$

The model derived using the training set was subsequently validated on the test set. Figure 7 shows logS predicted by the model plotted against observed logS for the test set. The correlation coefficient ($r^2_{trn/tst}$ = 0.784) between the predicted and the experimental logS of the training and test set as well as the $q^2 = 0.796$ calculated using LOO cross-validation of the training set proves the statistical reliability of this model.

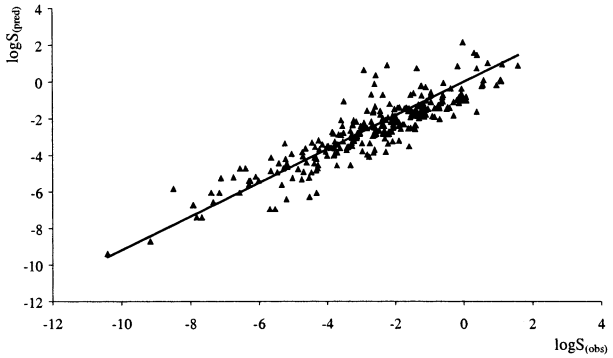

Figure 7. Correlation between the experimental and the predicted logS of the test set of 272 structures for the Aquax data set. The line represents the ideal correlation between experimental and predicted logS.

Table 10. Statistics Are Reported for 50 Runs of MoQSAR on the Aquax Data Set with Scrambled Data[a]

p	r^2	r^2_{MAX}	q^2	q^2_{MAX}
1	0.687	0.019	0.684	0.013
2	0.798	0.026	0.796	0.021
3	0.831	0.031	0.828	0.024
4	0.844	0.033	0.841	0.023
5	0.853	0.034	0.850	0.022
6	0.856	0.035	0.853	0.02
7	0.858	0.04	0.855	0.026

[a] The runs were designed to optimize r^2 and p. r^2 and q^2 represent the values found for unscrambled data. r^2_{MAX} and q^2_{MAX} represent the best values of r^2 and q^2 found for each value of p over 50 runs using the scrambled data.

The effect of removing LOGP as a descriptor was also investigated. MoQSAR was run with the terminal set reduced by one element, which was LOGP, and a family of 28 equivalent QSAR was collected. Despite the removal of LOGP, MoQSAR resulted in r^2 and q^2 statistics that are still acceptable, with their values falling in the ranges of 0.420−0.762 and 0.418−0.758, respectively. The three two term models identified when optimizing r^2, p, s, and D_{norm} are given in eqs 9−11.

$$\log S = -0.565\ \mathrm{xv0} + 0.134\ \mathrm{sumdell} + 0.772$$
$$r^2 = 0.555;\ q^2 = 0.551 \qquad (9)$$

$$\log S = -0.447\ \mathrm{xv0} + 0.267\ \mathrm{Gmax} - 1.304$$
$$r^2 = 0.590;\ q^2 = 0.587 \qquad (10)$$

$$\log S = -0.843 \times 1 - 0.851\ \mathrm{dxv0} + 0.317$$
$$r^2 = 0.604;\ q^2 = 0.601 \qquad (11)$$

Finally, randomization studies were performed to verify that the models found by MoQSAR were not due to chance correlations. For ease of interpretation, the randomization experiments were performed for the two objective case with MoQSAR configured to optimize r^2 and p. The activity data for the training set were scrambled 50 times, and MoQSAR was applied to each randomized data set. The results for the scrambled training sets are presented in Table 10 as the maximum values of $r^2$and q^2 found for each value of p. The values of r^2 and q^2 for the original (unscrambled data) are also shown. The much lower values of r^2 and q^2 for the scrambled data for any given number of terms confirm that the correlations found previously are not due to chance correlations.

Table 11. Descriptors for Models Found by Applying MoQSAR to the Nitrogen Data Set Optimized on r^2, p, s, and D_{norm}[a]

descriptors	r^2	p	s	D_{norm}	q^2
X17	0.461	1	1	1.000	0.436
X1 X16	0.609	2	1	1.000	0.592
X14 $(X29)^3$	0.610	2	3	0.155	0.590
X1 X2 X16	0.666	3	1	1.000	0.646
X1 X2 X5 X16	0.699	4	1	1.000	0.676
X1 X2 X5 X16 X22	0.726	5	1	0.871	0.703
X1 X2 X5 X9 X16	0.716	5	1	1.000	0.689
X1 X2 X12 $X16(X22)^2$	0.735	5	2	0.871	0.714
X1 X2 X5 $(X9)^2$ X16	0.719	5	2	1.000	0.693
X1 X2 X5 $(X9)^3$ X16	0.719	5	3	1.000	0.694
X1 X2 X5 X9 X15 X38	0.741	6	1	0.892	0.710
X1 X2 X5 X8 X9 X16	0.738	6	1	1.000	0.709
X1 X2 X5 $(X9)^2$ X15 X38	0.748	6	2	0.892	0.721
X1 X2 X5 X8 $(X9)^2$ X16	0.740	6	2	1.000	0.712
X1 X2 X5 $(X8)^3$ X9 X16	0.740	6	3	1.000	0.711
X1 X2 X5 X9 X16 X36 X38	0.747	7	1	0.908	0.709
X1 X2 X3 X5 X8 $(X9)^2$ X16	0.742	7	2	1.000	0.708
X1 X2 X5 $(X8)^3$ X9 X16 X37	0.743	7	3	1.000	0.707

[a] The coefficients have not been included for reasons of brevity. q^2 is also reported for each model. The four term model was used to predict logS for the 35 structures in the test set.

The results presented here are not directly comparable with those obtained by Huuskonen for several reasons. First, the data set was supplied as a whole and was subsequently partitioned at random; thus, the test set and training set used here will differ from those used by Huuskonen. Second, there are some differences in the descriptors used in the two studies. Third, MoQSAR was used to attempt to find models that balanced accuracy with complexity; hence, in general, the aim was to find models consisting of a relatively small number of easily interpretable descriptors. On the basis of multilinear regression and artificial neural networks, Huuskonen was able to derive a model with impressive statistics ($n = 884$, $r^2 = 0.89$, $q^2 = 0.88$); however, the model consists of 30 different parameters. Conversely, MoQSAR resulted in a large number of equivalent QSARs, which are much simpler in terms of their structural complexity and consequently more interpretable.

(4) Application of MoQSAR to the Nitrogen Data Set. Finally, the Nitrogen data set was analyzed. The 176 structures were divided at random into a training set of 141 structures and a test set of 35 structures. The terminal set consists of the 40 molecular descriptors listed in Table 3. The desirability values assigned to each descriptor are shown in the column headed d.

MoQSAR was run to optimize four objectives (r^2, s, p, and D_{norm}), and the results are collected in Table 11. As can be seen, five of the descriptors provided by McElroy and Jurs occur with high frequency, namely, WTPT-2 and MDE-14 that code topological information, GEOM-3 that represents the third geometric moment, and SDCH-2 and CTDH that represent the average surface area times the charge on donatable hydrogens and the number of donatable hydrogens, respectively. Another frequently occurring descriptor is the Molconn-Z parameter, xv_0, which is a connectivity valence path index. The following four term model found by

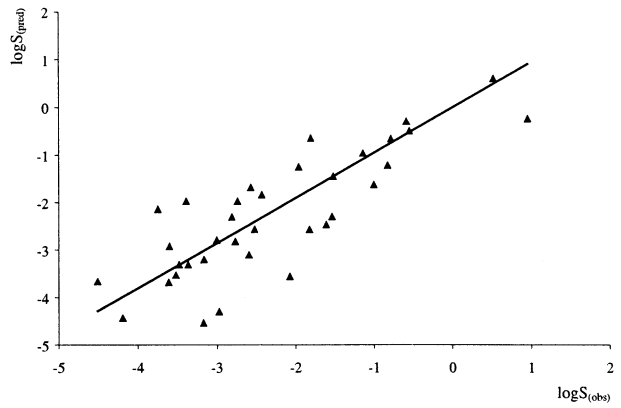

Figure 8. Correlation between the experimental and the predicted logS of the test set of 35 structures for the Nitrogen data set. The line represents the ideal correlation between experimental and predicted logS.

Table 12. Statistics Are Reported for 50 Runs of MoQSAR on the Nitrogen Data Set with Scrambled Data[a]

p	r^2	r^2_{MAX}	q^2	q^2_{MAX}
1	0.461	0.108	0.436	0.081
2	0.614	0.128	0.595	0.086
3	0.666	0.165	0.646	0.124
4	0.699	0.191	0.676	0.140
5	0.735	0.210	0.714	0.133
6	0.747	0.223	0.724	0.142
7	0.760	0.219	0.732	0.134

[a] The runs were designed to optimize r^2 and p. r^2 and q^2 represent the values found for unscrambled data. r^2_{MAX} and q^2_{MAX} represent the best values of r^2 and q^2 found for each value of p over 50 runs using the scrambled data.

running MoQSAR on the training set is presented here

$$\text{logS} = 0.137\ \text{MDE\_14} + 1.259\ \text{GEOM\_3} + 0.102\ \text{SCDH\_2} - 0.549\ \text{xv0} + 1.278$$

$$r^2 = 0.699;\ q^2 = 0.676 \qquad (12)$$

As for the Aquax data set, a validation study was performed to predict the 35 structures in the test set. Figure 8 shows the predicted logS plotted against observed logS when the model is applied to the test set. The correlation coefficient ($r^2_{trn/tst} = 0.658$) between the predicted and the experimental logS of the training and test set, as well as the $q^2 = 0.676$ of LOO cross-validation of the training set proves the statistical reliability of this model.

Again, 50 randomization studies were conducted by scrambling the logS data in the training set to verify that the models found were not due to chance correlations. MoQSAR was applied to each randomized data set to optimize the two objectives, r^2 and p, and the results for the scrambled training sets are presented in Table 12. The much lower values of r^2 and q^2 found for the scrambled data for any given number of terms relative to the unscrambled data indicate that the correlations reported in Table 11 are not due to chance correlations.

(5) Robustness and Efficiency. GP is a nondeterministic search method, and so, a final experiment was carried out to test the robustness of MoQSAR. Ten runs were carried out on the Selwood data set with the aim of optimizing r^2 and p. As can be seen in Table 13, the low standard deviations over the 10 runs demonstrate the robustness of the method.

Table 13. Ten Runs of MoQSAR Were Carried out on the Selwood Data Set to Evaluate the Robustness of the Approach[a]

p	r^2_{MEAN}	r^2_{SD}	p	r^2_{MEAN}	r^2_{SD}
1	0.371	0	5	0.844	0.014
2	0.639	2E-08	6	0.869	0.008
3	0.735	0.004	7	0.875	0.015
4	0.812	0.019			

[a] The runs were designed to optimize r^2 and p; hence, each run generated seven solutions, one for each number of terms (1−7). The column headed r^2_{MEAN} shows the results averaged for each value of p. Standard deviations are shown in the column headed r^2_{SD}.

A typical GPQSAR run on the Selwood data set with a population size of 200 and 2000 iterations takes on average 60 min to find a single solution whereas a single run of MoQSAR with the same GP parameters finds a whole family of solutions in an average of 45 min (SGI R10K workstation at 195 MHz).

Conclusions

Two novel methods have been developed that aim to derive QSAR models that explore the tradeoff between model accuracy and complexity. Both methods are based on GP, which is a branch of GAs with the main difference being that the individuals in the population can vary in shape and size. When applied to the derivation of QSARs, this allows models of varying complexity to be explored. In GPQSAR, a single solution is evolved with the balance between model complexity and accuracy being controlled by a penalty function. The advantage of this approach over existing approaches to deriving QSARs is that the number of terms required does not have to be specified. A disadvantage is that the penalty function has to be calibrated for each data set. A further disadvantage is that a single solution is found, which represents one particular compromise solution when typically a family of different compromise solutions exists.

The second approach, MoQSAR, is based on a MOGP and exploits the population nature of the GP to optimize a family of solutions in parallel. It is no longer necessary to calibrate the method nor is it necessary to specify the number of terms required. A family of QSAR models is obtained where each model represents a different compromise in the objectives. Model complexity has been measured using a number of different objectives including the total number of terms, the number of nonlinear terms, and a knowledge-based measure of the chemical interpretability of the descriptors used in the model.

The method has been applied to several different data sets, and in each case, a variety of different models were found. In the case of the Selwood data set, these models include "best" models previously reported in the literature. Additional models are also found where accuracy is traded for improved interpretability. The full range of models can be presented to the user who is then able to select a model that represents the best compromise in the objectives.

Acknowledgment. We thank Jarmo Huuskonen for providing the Aquax data set; Nathan McElroy and Peter Jurs for providing the Nitrogen data set; Katya Rodriguez-Vazquez for help with GP; Iain McLay for

helpful discussion; and Peter Willett for careful reading of this manuscript. We thank the GARAGe group at Michigan State University for making their lilgp code available (http://garage.cps.msu.edu). We thank GlaxoSmithKline for funding. The Krebs Institute for Biomolecular Research is a designated Biomolecular Sciences Centre of the Biotechnology and Biological Sciences Research Council.

References

(1) So, S.-S. Quantitative Structure−Activity Relationships. In *Evolutionary Algorithms in Molecular Design*; Clark, D. E., Ed.; Wiley-VCH: Weinheim, 2000; pp 71−97.
(2) McFarland, J. W.; Gans, D. J. On identifying likely determinants of biological activity in high dimensional QSAR problem. *Quant. Struct.-Act. Relat.* **1994**, *13*, 11−17.
(3) Selwood, D. L.; Livingstone, D. J.; Comley, J. C.; O'Dowd, B. A.; Hudson, A. T.; Jackson, P.; Jandu, K. S.; Rose, V. S.; Stables, J. N. Structure−activity relationships of antifilarial antimycin analogues: a multivariate pattern recognition study. *J. Med. Chem.* **1990**, *33*, 136−142.
(4) Camilleri, P.; Livingstone, D. J.; Murphy, J. A.; Manallack, D. T. Chiral chromatography and multivariate quantitative structure−property relationships relationships of bezoimidazole sulphoxides. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 61−69.
(5) Domine, D.; Devillers, J.; Chastrette, M. A non linear map of substituent constants for selecting test series and deriving structure−activity relationships. II. Aliphatic series. *J. Med. Chem.* **1994**, *37*, 981−987.
(6) Norinder, U. A. A PLS QSAR analysis using 3D generated aromatic descriptors of principal property type: Application of some dopamine D2 benzamide antagonists. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 671−682.
(7) Wikel, J. H.; Dow, E. R. The use of neural networks for variable selection in QSAR. *Bioorg. Med. Chem. Lett.* **1993**, *3*, 645−651.
(8) Luke, B. T. Comparison of different data set screening methods for use in QSAR/QSPR generation studies. *J. Mol. Struct. (THEOCHEM)* **2000**, *507*, 229−238.
(9) Rogers, D. R.; 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*, 854−866.
(10) Kubinyi, H. Variable selection in QSAR studies. I. An evolutionary algorithm. *Quant. Struct.-Act. Relat.* **1994**, *13*, 285−294.
(11) Kubinyi, H. Variable selection in QSAR studies. II. A highly efficient combination of systematic search and evolution. *Quant. Struct.-Act. Relat.* **1994**, *13*, 393−401.
(12) So, S.; Karplus, M. Evolutionary optimisation in Quantitative Structure−Activity Relationship: an application of genetic neural networks. *J. Med. Chem.* **1996**, *39*, 1521−1530.
(13) Waller, C. L.; Bradley, M. P. Development and validation of a novel variable selection technique with application to multidimensional quantitative structure−activity relationship studies. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 345−355.
(14) Koza, J. R. *Genetic Programming*; The MIT Press: Cambridge, Massachusetts, 1993.
(15) Clark, D. E., Ed. *Evolutionary Algorithms in Molecular Design*; Wiley-VCH: Weinheim, 2000.
(16) Nachbar, R. B. Molecular evolution: A hierarchical representation for chemical topology and its automated manipulation. In *Proceedings of the Third Annual Genetic Programming Conference*, University of Wisconsin, Madison, Wisconsin, 22−25 July, 1998; pp 246−253.
(17) Globus, A.; Lawton, J.; Wipke, T. Automatic molecular design using evolutionary techniques. *Nanotechnology* **1999**, *10*, 290−299.
(18) Rodríguez-Vázquez, K.; Fleming, P. J. Multi-objective genetic programming for nonlinear system identification. *Electron. Lett.* **1998**, *34*, 930−931.
(19) Fonseca, C. M.; Fleming, P. J. An overview of evolutionary algorithms in multiobjective optimisation. In *Evolutionary Computation*; De Jong, K., Ed.; The Massachusetts Institute of Technology: Cambridge, MA, 1995; Vol. 3, No. 1, pp. 1−16.
(20) Fonseca, C. M.; Fleming, P. J. Genetic algorithms for multiobjective optimization: formulation, discussion and generalisation. In *Genetic Algorithms*: *Proceedings of the Fifth International Conference*; Forrest, S., Ed.; Morgan Kaufmann: San Mateo, CA, 1993; pp 416−423.
(21) Coello Coello, C. A. An updated survey of GA-based multiobjective optimization techniques. *ACM Computing Surveys* **2000**, *32*, 109−143.
(22) Handschuh, S.; Wagener, M.; Gasteiger, J. Superposition of three-dimensional chemical structures allowing for conformational flexibility by a hybrid method. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 220−232.
(23) Gillet, V. J.; Willett, P.; Fleming, P.; Green, D. V. S. Designing focused libraries using MoSELECT. *J. Mol. Graphics Modell.* **2002**, *20*, 491−498.
(24) Gillet, V. J.; Khatib, W.; Willett, P.; Fleming, P. J.; Green D. V. S. Combinatorial library design using a multiobjective genetic algorithm. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 375−385.
(25) International Patent Application No. PCT/GB01/05347.
(26) Huuskonen, J. Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 773−777.
(27) Molconn-Z is available from eduSoft, LC., P.O. Box 1811, Ashland, VA 23005.
(28) McElroy, N. R.; Jurs, P. C. Prediction of aqueous solubility of heteroatom-containing organic compounds from molecular structure. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1237−1247.

JM020919O