McQSAR: A Multiconformational Quantitative Structure−Activity Relationship Engine Driven by Genetic Algorithms

Mikko J. Vainio* and Mark S. Johnson

Structural Bioinformatics Laboratory, Department of Biochemistry and Pharmacy, Åbo Akademi University, Tykistökatu 6A, FIN-20520 Turku, Finland

Received May 4, 2005

The generation of quantitative structure−activity relationships (QSARs) under the supervision of a genetic algorithm (GA) is a QSAR modeling approach used for more than a decade. In this paper we present McQSAR, an extension to the traditional GA approach to derive QSARs. McQSAR is able to use descriptors for multiple representations per compound, such as different conformers, tautomers, or protonation forms. Test runs show that the algorithm converges to a set of representations that describe the binding mode of the set of input molecules to a reasonable resolution provided that suitable descriptors—based on the three-dimensional structure—are used. Furthermore, the frequency of chance correlation was measured during multiple runs on a real-life data set using simulated linear relationship functions. The observed frequency of chance correlation, on average $0.3 \pm 0.5\%$, was found independent of the size of the calibration set and the number of terms in the underlying relationship function.

INTRODUCTION

Statistical modeling has many applications throughout the sciences. In the specific niche of computer-assisted drug design, one of the most prominent applications of statistical modeling is the quantitative structure−activity relationship (QSAR) methodology. Since the seminal publication by Hansch and Fujita in 1964,[1] QSAR methodology has evolved into a discipline, and today it is an established tool in the drug discovery and lead optimization pipelines of pharmaceutical companies.

The central paradigm of QSAR is that all properties of a compound depend on the chemical structure of that compound. Thus, if we want to predict the value of a property of a compound, we should be able to do so if we know the chemical structure of the compound and we have a (mathematical) model that tells us exactly how the value of the property is calculated from the structure. The model is derived from a set of compounds for which the value of the desired property is known. A mathematical form of the model is derived by some suitable algorithm (we will come back to this later). Then the coefficients of the model are calibrated on a set of compounds, the "calibration set", with known values of the property. Once the model is calibrated, it can be used to predict the value of the property for other compounds. QSAR methods require that chemical structures are described by numbers, which are suitable input to mathematical functions. Over 3000 different algorithms exist for calculating these variables.[2]

The most persistent problem in statistical modeling is the selection of the most informative variables among those available. Suppose that we have m samples in the calibration set: a vector $\mathbf{y} = (y_i)_m$ of variables of interest (e.g. biological activities) and a matrix $\mathbf{X} = (x_{i,j})_{m \times n}$ of n descriptive variables per sample (the descriptors). We want to find a function $f(\mathbf{X}) = \mathbf{y}' = (\mathbf{y} + \mathbf{e})$, where $\mathbf{e} = (e_i)_m$ is a vector of prediction errors, so that $|\mathbf{e}|$ is as small as possible. If too many variables are added to the regression equation, the expected error of prediction will increase, which motivates the use of a subset of k ($k < n$) regressors to derive the model. The selection of a subset of the most informative variables also is desired when there is an abundance of data of which most is noninformative (noise), or there is redundancy in the information content, i.e., some of the variables correlate. If fewer descriptive variables are used in the model, the interpretation is also simplified: explanations for the underlying phenomena might require only a few variables. For example, the binding of a ligand to a receptor might be adequately described using three variables, one variable for each of the basic interaction types: steric, electrostatic, and hydrophobic.

To select adequate descriptive variables, three algorithmic components are required: a search algorithm, a statistical modeling procedure, and an objective function to guide the search. Tabu search,[3−5] genetic algorithms,[6] simulated annealing,[7,8] ant colony optimization,[9,10] and particle swarm optimization[11−13] are all examples of suitable search algorithms. Some frequently used statistical modeling procedures include multilinear least-squares regression,[14] partial least squares regression,[15] and artificial neural networks.[16] Objective functions in general return an estimate of the error of prediction for a proposed statistical model. A large number of combinations of these algorithmic components has been reported in the literature. Examples of approaches utilizing Bayesian variable selection are reported by George et al.[17] Perkins et al.[18] provide a review on the development of the QSAR methodology over the past decade.

Genetic algorithms (GAs) have been used as a search algorithm for variable selection in chemometrics and QSAR for over a decade.[19−21] GAs are a class of computational

* Corresponding author phone: +358-2-215 4600; fax: +358-2-215 3280; e-mail: mikko.vainio@abo.fi.

10.1021/ci0501847 CCC: $30.25
Published on Web 11/02/2005

problem solving techniques inspired by evolution in nature. The basic idea of a GA is as follows: possible solutions of the problem are encoded in data structures and called *chromosomes*. An *objective function* assigns a numerical *fitness value* to each chromosome according to how good a solution it represents. A *population* of chromosomes is evaluated for their fitness and individuals are *selected* for *reproduction*. Selection favors chromosomes with a high fitness value. Reproduction combines genes of the selected parent chromosomes to produce offspring. This is called *crossover*. The offspring are subject to *mutation* with some probability. The offspring form a new population. The steps of evaluation, selection, and reproduction are applied again to the new population, and the process is repeated many times. Each cycle is called a *generation*. The fitness of individuals gradually improves during the run of the algorithm, analogous to evolution in nature. GAs are excellent for solving problems for which good solutions can be gradually built up from low-order components, e.g. terms of a linear equation.

Druglike molecules can adopt different states under physiological conditions: different protonation states, tautomeric forms, and conformations. It is not always straightforward to decide which representative form of the compound should be used.[22,23] Thus, in QSAR, there is an additional dimension to the problem when using descriptors that depend on the three-dimensional properties of compounds. A GA can derive a meaningful subset of variables for a statistical model. Could a GA also be used to select the representative forms of the compounds used to calibrate the model? This is the question we have sought to address in this study.

METHODS

We have developed software capable of genetic function approximation.[21] In other words, given a calibration data set, the software creates statistical models using a GA. It exceeds being just a variable selection procedure, as the functional form of the QSARs is also subject to evolution. The implementation was designed to be able to use calibration data that contain descriptors for multiple representatives of a single compound. Data for only one conformer per compound are used at a time (for brevity we hereafter use the word conformer to refer to any representative state of a chemical entity). The algorithm was named McQSAR for Multi-conformational QSAR.

The genetic algorithm used to evolve the QSAR equations is a traditional "simple" GA: it has nonoverlapping generations, i.e., all chromosomes of a generation are offspring of individuals of the previous generation. Elitism is used; the best equation of a generation is carried over to the next generation unchanged. Elitism ensures that the best found solution is not lost between generations. The McQSAR chromosome consists of two parts, a representation of the QSAR equation and a conformer index array (Figure 1). The QSAR equations are represented as rooted tree data structures similar to those used, e.g. in the programs GPQSAR and MoQSAR of Nicolotti et al.[24] A rooted tree has a root node, which has child nodes, which in turn can have child nodes and so forth. The nodes that have children are operators. Operators perform calculations on the value(s) of their child node(s). In this implementation the operator nodes have either one or two child nodes, i.e., there are unary and binary operators. The available operator set is $S_{Op} = \{+, *, /, min, max, average, negation, log, exp, sqrt, power\}$. The operator names should be self-explanatory. The operators *log* and *sqrt* operate on the absolute value of the input. The tree can grow and shrink, which makes it a handy data structure for QSAR as the equation complexity is not fixed a priori to any particular level.

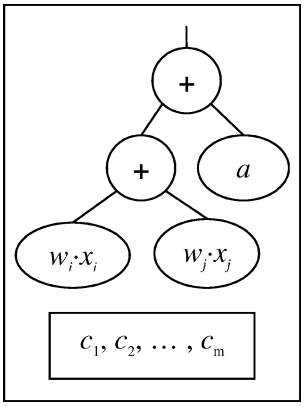

Figure 1. The McQSAR chromosome data structure consists of two parts, the QSAR equation presented as a tree, and a conformation index array. The shown tree structure encodes a linear equation with two variables and a constant, $y = w_i \cdot x_i + w_j \cdot x_j + a$, where w_i and w_j are weights, x_i and x_j are descriptors, and a is a constant. The weights and the constant are fitted to the data for conformers indexed by the array using a least-squares routine. See text for a detailed description.

The terminal (leaf) nodes represent a numerical value. Terminal nodes are from the set $S_{Term} = \{$*descriptor, constant, spline, quadratic spline, Gaussian*$\}$. Each terminal node has a weighting term w (subject to calibration), exponent y (not always used in evaluation), and index to a variable i (not used for constants). The weighting term has different meanings depending on the node type. The terminal nodes evaluate as

descriptor $\quad w \cdot x_i^y$

constant $\quad w$

spline $\quad$ if $w - x_i > 0$, $w - x_i$, otherwise 0

quadratic spline $\quad$ if $w - x_i > 0$, $(w - x_i)^2$, otherwise 0

Gaussian $\frac{1}{2\pi} e^{-(1/2)(x_i - w)^2}$

(standard deviation is fixed to $\sigma = 1$)

where x_i is the value of the descriptive variable with index i. When these basic operations are included at the terminal nodes, the evolution of equations possessing a desired functional form becomes straightforward. Linear equations, for example, are easily evolved using the + operator (minus is implicit as the sign of w can become negative), descriptor and constant terminals, and appropriate mutation operators.

Constant, by the way, is a strange term to use in QSAR equations. Compounds cannot have a "default" property value independent of structure. A constant term is appropriate only in a QSAR model that considers a series of structurally closely related compounds. One can be suspicious about

using linear equations at all, since they might extrapolate wildly outside of the calibration data space. A sum of Gaussians would likely give a zero activity for a compound that does not fall within the calibration data range. Gaussian function approximation is possible with McQSAR; only the standard deviation in the Gaussian terms is not adjustable because of limitations in the underlying data structure. This design flaw occurred to us in a late stage of development of the first version of this software. The next version will correct this flaw instead of calling it a "feature".

Having the building blocks for the tree data structures, genetic operators need to be defined. Equations are combined using the crossover operator, which operates on two parent equations in order to produce two offspring: both parents produce a copy of themselves. A crossover point (node) is randomly selected from both copies, and the subtrees rooted at those nodes are swapped. The newly formed equations are then subject to a reduction operation that ensures their mathematical validity.

The available mutation operators are *insertion*, *deletion* and *swap of a subtree*, *swap of two nodes*, deletion of a leaf node, *random change of the index of a referred variable* (*i*) of a random *descriptor* node, raise/lower exponent (*y*) of a random *descriptor* node, and *shifting of the knot* (*w*) for a random *spline* or *quadratic spline* node. The mutation operator is chosen at random from a user-defined pool. After modification, the resulting equation is again reduced to ensure mathematical validity. As an example, the mutation operator pool for evolving linear equations would minimally consist of two operators, *insertion* and *deletion*, since the swap of subtrees or nodes would not make any difference and the remaining mutation operators are either specializations of the *insertion* and *deletion* operators or inappropriate for the used node types. Polynomial equations can be evolved analogous to linear equations by adding the *raise/lower exponent* operators to the mutator pool.

The second part of the chromosome data structure contains an integer array "gene" that has a locus (an element) for each compound in the training set. Each integer in the array is an index to a conformer of the compound corresponding to that locus. Each array thus represents an "alignment" of the training set compounds and a projection of the three-dimensional training data set to a two-dimensional matrix. The alignment array is initialized with random conformer indices and evolved toward the bioactive conformations by a separate GA as described below.

The statistical modeling procedure that has been used is least-squares regression. The weights *w* of the terminal nodes are fitted to a calibration data set so that descriptors for one conformer per compound are used as indicated by the integer array gene. The fit is done with the Levenberg−Marquardt nonlinear least-squares method[25−27] or by using QR matrix decomposition[14] in the case of a linear equation.

The objective function for the GA returns a correlation index for the input equation. In this implementation, the score given by the objective function is used as the fitness value without any scaling. First, the standard correlation index r^2 is calculated and compared to a threshold value T (a parameter of the algorithm). If r^2 is smaller than T and smaller than the fitness of the best equation in the population, r^2 is returned. If r^2 is greater than T or greater than the fitness of the current best equation in the population, a nested GA is invoked to evolve that portion of the chromosome that encodes the alignment. The output of this "inner" GA is the cross-validated correlation index q^2 for the input equation. The role of the parameter T is thus to reduce the need to run the cross-validation test and to ensure that a correlation value higher than T is always a cross-validated one.

The individuals of a population are sorted in decreasing order according to the raw values of the objective function. Selection of individuals for reproduction is based on the rank of the individuals in the sorted population: an individual becomes selected if a biased "coin flip" test returns true. The bias (probability for a single flip to return true) is equal to the reciprocal of the size of the population. Testing starts from the highest ranking chromosome and proceeds successively to lower ranking ones until a selection is made. Should the end of the population be reached, a random individual is selected.

The GA used to evolve the conformer index arrays (alignments) is also a traditional simple GA with elitism of one individual. The conformer index array portion of the equation chromosome is copied to initialize a population of arrays, which is then evolved. The *random flip* mutation operator changes the value at a random locus into a random index valid for that locus. Selection of arrays for reproduction is done with the same routine as is used for the equations above. The one-point crossover copies two parent arrays, splits the copies into two halves at random loci, and swaps the tails. The objective function returns a cross-validated correlation index q^2 (see below) for a least-squares fit to the data for conformers indexed by the evaluated array.

We have followed the principle of Occam's razor or the principle of parsimony, which is widely considered as good scientific practice: of two equally valid explanations, go for the simpler one (as Albert Einstein put it "Make it as simple as possible, but not any simpler."). Occam's razor in QSAR refers to preferring the simplest equation of those that fit the data equally well. Complex equations that contain a term for each observation in the calibration set can fit the data perfectly; r^2 becomes one. This is called overfitting. To protect against overfitting, cross-validation tests are made during the conformer sampling GA run. A portion of the observations is put aside from the calibration set, the equation is recalibrated on the remaining samples, and the dependent variable values for the omitted samples are predicted. This is repeated leaving out another set of observations until all observations have a predicted dependent value. The cross-validated correlation index q^2 is calculated from the errors of the predictions and used as the fitness value instead of r^2. Equations that overfit tend to give a high error of prediction for the omitted observations, which leads to a low q^2, a low fitness, and eventual extinction of the overly "flexible" model. Cross-validation (CV) thus controls the model complexity when it comes to overfitting, but it does not explicitly prefer simpler equations.

The objective function is the most critical component of a variable selection procedure. If the search is guided astray, it will certainly go astray no matter how elaborate the search algorithm itself is. McQSAR uses leave-*d*%-out-averaged-over-*B*-repetitions CV, an equivalent to leave-multiple-out cross-validation (LMO-CV);[28,29] *d* and *B* are user configurable parameters. In the literature, this method is also known as balanced incomplete CV[30] and the repeated learning-

testing method.[31] Note that Baumann[28,29] uses d to denote the number of observations to leave out, whereas we use it to denote the percentage of observations to leave out. The traditional leave-one-out CV (LOO-CV) is not recommended because it returns an overestimated figure of merit (q^2). LMO-CV underestimates the figure of merit but, on the other hand, has a higher theoretical probability for selecting the correct (linear) QSAR model[30] and produces models with better external predictivity than LOO-CV.[28]

The balanced incomplete CV requires the following "balance" conditions on the leave-out subsets:

1. Every compound appears in the same number of subsets, and

2. Every pair of compounds appears in the same number of subsets.

Additionally,

3. All subsets should be of the same size.[31]

In McQSAR, the algorithm for splitting the calibration data set into leave-out subsets explicitly ensures that condition 3 holds, which might require that the last subset is padded with some randomly selected compounds from the other subsets. This obviously does not conform to conditions 1 and 2. However, since the compounds used for padding the last subset are selected randomly and the prediction error is averaged over B repetitions, conditions 1 and 2 are asymptotically met as $B \rightarrow \infty$. Recommended values of d and B to start with are $d = m^{3/4}/m = m^{-1/4}$ (the theoretical optimum)[30] and approximately $2m \leq B \leq 3m$,[29] where m is the number of observations (compounds) in the calibration set.

A conformer combination that represents an incoherent view of the training data (erroneous alignment), ill-fitting the equation, produces a low q^2 value. This gives rise to evolutionary pressure in the conformer sampling GA run.

Medicinal chemists are interested in models that reliably predict activities of compounds outside the calibration set. While the use of q^2 to guide evolution might prevent equations from becoming overfitted, a high q^2 value does not imply that an equation has high predictive ability[32] nor does it rule out the possibility of chance correlation[33] even if variable selection is guided by a stringent cross-validation objective function. An external test set needs to be used. The following criteria have been proposed for a model to be highly predictive:[32]

1. A high q^2 value.

2. A high correlation coefficient value from a linear regression analysis of experimental versus predicted activity values of compounds from an external test set.

3. Linear regression of the experimental versus the predicted activity values of compounds from an external test set, forced through the origin (i.e. intercept fixed to zero), should give a slope close to one and a correlation coefficient close to q^2. This should hold at least for one of the two possible regressions r^2_O, experimental versus predicted, and r'^2_O, predicted versus experimental (No, they do not give the same results. We strongly recommend reading the paper by Golbraikh and Tropsha.[32]).

Criterion 3 above is emphasized also by Hawkins:[34] "An extreme form of inappropriate measure is the correlation coefficient between the predicted and observed values in the validation sample.". He also states that a consideration of the raw deviations between the predicted and observed activities is the only correct way to use a validation set. Indeed, the correlation coefficient alone gives an erroneous impression of the predictive ability since it only reports on the *extent of correlation* between predicted and observed data points but does not imply whether the predictions are *accurate* or not. Correlation coefficients of linear regressions via the origin, r^2_O and r'^2_O, however, are measures of the accuracy of the predictions and readily comprehensible to the user unlike (a sum of) the raw deviations. During the evolution of equations, McQSAR can use a validation set to calculate the predictive-r^2, r^2_O, and r'^2_O for the best equation of each generation and report their values to the output console so that the user can follow the progress of the evolutionary process in terms of the external predictive ability of the best model. Predictive-r^2, r^2_O, and r'^2_O cannot be used to guide the evolutionary process since they would become optimistically biased, but after each run these external figures of merit can guide the selection of a model for further consideration.

Todeschini et al.[35] have recently reported a set of criteria, combined within a fitness function called RQK, that can be used to detect models with "pathological" conditions, i.e., models exhibiting chance correlation, lack of predictive ability, or containing redundant variables. To use multiple criteria, McQSAR would require substituting the underlying GA with a multiobjective GA such as the Elitist Non-Dominated Sorting GA II (NSGA-II),[36] which is an improvement worth consideration.

Parallelization. The extra dimension of sampling the conformers makes the algorithm very calculation-intensive. Therefore, a parallel implementation was written to run on a Linux cluster of Intel x86 PCs. A master-slave configuration was an intuitive choice for the parallelization scheme due to its straightforward implementation. A master process handles the initialization, selection, crossover, and mutation and sends equations to slave processes for fitting and conformer sampling GA runs. This scheme parallelizes well for a few processors, provided that the conformer sampling GA is frequently initiated. Otherwise the interprocess communication overhead outweighs the benefit of parallel execution. There is also an overhead associated with the use of a simple GA: toward the end of a generation cycle, all processes but one stand idle waiting for the one to finish the conformer sampling GA on the last remaining equation before the evaluation of the next generation begins. This idle time might take a good while if the parameters cause a lengthy conformer sampling GA run. A steady-state GA constantly dispatching equations to slave processes, receiving results, and resorting the population would utilize the CPUs more efficiently.

Predicting Activity for a Compound with Multiple Conformers. Once a QSAR model is derived, one wishes to use it to predict activities for compounds outside the calibration set. Calculating the predicted activity value for a compound with several different low-energy conformers using a 3D dependent QSAR model is problematic. Obviously there is no way of telling which conformer should be used. McQSAR applies a straightforward means of taking the average value of all conformers of a compound as the predicted activity for the compound. If the (relative) energies for the conformers are known, they can be used to calculate a weighted average according to the Boltzmann distribution, which states that a compound is found in a high-energy

conformation with a lower probability than in a low-energy conformation. Since the free ligand and the ligand−receptor complex are two different thermodynamical systems (the energy of a conformer of the ligand in the gas phase (as in the conformer generation procedure) is different from the net energy of a system containing the same conformer in contact with a receptor), weighting by the conformer ensemble probabilities of the free ligand is not physically valid. It is, however, the best approximation we can make about the state distribution of the receptor−ligand system without running a full-scale docking/free energy perturbation test.

Working Hypothesis. McQSAR grew out of the interest to test whether a GA could be used to select the representative state for the compounds in the calibration set, based on the following hypothesis:

We are operating on descriptors which depend on the 3D properties (conformation, tautomeric form, protonation state) of the compounds. We have a set of compounds to be used as the calibration set for a QSAR study. Suppose all the compounds in the calibration set bind the biological target in the same binding mode (which may or may not be true for a real-life set), and the set contains data for several low-energy conformers per compound. The bioactive conformer most probably closely resembles one of the low-energy conformers, so we can assume that the calibration set contains the bioactive conformer for each compound. Therefore, if we randomly select one conformer for each compound, the bioactive conformer is selected for some compounds. If equation *f* adequately describes the correct binding mode, the least-squares fit to the data for the selected conformers should give a moderate r^2 because some of the observations satisfy *f*. The conformer sampling GA should then find the bioactive conformation for most of the compounds and return a high cross-validated correlation value q^2.

If equation *g* does not represent the binding mode, it is likely that not all compounds in the calibration set have a conformer that fits the equation. Therefore, it is not likely for the conformer sampling GA, if even initiated in the first place, to find an alignment that would give a high q^2.

Would the above hypothesis hold in real life? One readily can think of several factors that affect the quality of the solution for such an experiment: Since McQSAR not only screens different variables but also considers multiple alternative values for the screened variables, the risk of the occurrence of chance correlation[33] should be high. Thus, what is the probability for a "wrong" equation to correlate by chance (chance correlation, part I)? What is the probability for the conformer sampling GA to find a wrong alignment that ranks better than the correct one (chance correlation, part II)? If we know the bioactive conformation for some of the compounds from X-ray crystal structures, do these fixed points guide the conformer sampling GA toward the correct alignment of compounds? We can find qualitative, and partially quantitative, answers to these questions by observing the behavior of McQSAR (and its subroutines) on benchmark data sets.

There did not seem to be any readily suitable benchmark data sets available. Random numbers could be used to test a QSAR method but that would require assumptions to be made about the distribution and range of the descriptor values that might not reflect the properties of a typical "production run" data set. In order not to make such assumptions, data on existing chemical entities seemed a more appealing test case. The data set of 405 benzodiazepine ligands of Sutherland et al.[37] was taken as an example. The data generation procedure follows: Compound collection was run through the conformer ensemble generation procedure of the program CATALYST version 4.8 (Accelrys, Inc., San Diego, CA). The program DRAGON[2] version 3.0 (Talete srl, Milano, Italy) was run on the conformers to produce the 3D-dependent geometrical, charge, RDF, 3D-MoRSE, WHIM, and GETAWAY descriptors, aromaticity indices, and Randic molecular profiles (see ref 2 for individual references for these descriptors).

This data set was input to a routine that generates linear equations with variables randomly chosen from the input data set. The selected variables were checked for 3D dependencies—according to the hypothesis, the descriptor values must vary among the conformers. A random "activity" value between −10 and 10 was assigned to each compound and the coefficients of the equation were fit. This was necessary in order to keep the scale of the dependent variable values in a numerically stabile range in the next step: for each compound, a conformer was picked randomly, and the equation was evaluated on the data for that conformer. The resulting value for the dependent variable was stored as the activity of the compound, and the conformer was stored as the "bioactive" conformer. The generated linear equation with randomly selected terms, the combination of randomly selected conformers, and the "activity" values calculated for those conformers constitute a simulated "correct" relationship. Therefore, the equation generated in this process is hereafter referred to as the "correct equation", and the combination of selected conformers is referred to as the "correct alignment" of compounds.

The conformer sampling GA, using the correct equation, was then evolved on the data set, starting from a population of random conformer combinations. After the evolution of alignments had terminated, the best individual of the run was compared against the correct "bioactive" conformers. The percentage of correctly assigned "bioactive" conformers in the best individual tells us how correct the solution is. For the incorrect conformers, the average root-mean-square deviation (RMSD) to the "bioactive" conformer gives a metric for how close to the correct binding mode the solution is. The RMSD values have a lower limit above zero, since most conformer generation procedures apply an RMSD cutoff: if two conformers have RMSD less than the cutoff, the higher energy conformer is discarded.

If we know the correct equation and the correct alignment of compounds, the occurrence of chance correlation can be detected by comparing the q^2 values of alignments produced by the conformer sampling GA to the q^2 value of the correct alignment. The relative frequency (approximating probability) of chance correlation is then obtained as the number of incorrect individuals with q^2 equal to or higher than the q^2 of the correct alignment divided by the total number of alignments tested. Baumann used an analogous definition for the probability of chance correlation in permutation tests for variable selection.[29]

The equation length was varied from three to eight terms in one term increments, and the size of the calibration set

was increased from 20 compounds to the full size of the set (241) in steps of 50 compounds (the last step being less), taking random subsets of the whole set. Statistics on the performance indicators were collected over 50 runs per each combination of equation length and set size, each run with a new random equation. This procedure tests the first part of our hypothesis: "the correct equation *f* will end up with a relatively high fitness value (the cross-validated correlation index q^2) and correct bioactive conformers".

The conformer sampling GA consisted of a population size of 30 individuals, mutation probability 0.02, crossover probability 0.99, and termination of the procedure after 1000 successive generations if no improvement has taken place or after a maximum of 10 000 generations. The cross-validation settings were leave-25%-out averaged over 10 times. These parameter values are not optimal by any means, since 10 repetitions are probably too few to average over. Statistically, a more ensuring number would be two or three times the size of the calibration set, but the computation time for evaluating an alignment scales linearly with the number of repetitions and becomes a significant issue when the number of repetitions is increased 10-fold or 100-fold.

To study the second part of the hypothesis—"wrong equations do not score well" —the whole McQSAR procedure was run on a real data set for which a number of random correct linear equations were generated as described above. Five runs were performed for each combination of equation length and calibration set size. Equation length was again increased from three to eight in one term steps, and the calibration set size was increased from 20 to 241 in steps of 50. Each run used a new random correct equation (i.e. an equation generated by a simulation procedure as described earlier). During each run, the GA generates a large number of equations that try to model the relationship given by the correct equation. Chance correlation was detected by comparing the q^2 values of the models proposed by the GA to that of the correct equation. (The q^2 value of the correct model was updated during the run whenever the GA produced the correct equation with a q^2 value higher than the previous value.) The frequency of occurrence of chance correlation was recorded. Distinction was made among chance correlating models according to how many variables of the correct equation they contained: incorrect models contain none, partially correct equations contain some, and overfitting models contain all the terms of the correct equation and some unnecessary terms. Settings used for the equation generation GA were as follows: mutation probability 0.02, crossover probability 0.99, population size of 100 equations, and termination after 10 elapsed generations. The conformer sampling GA used the same probabilities for mutation and crossover, a population size of 30 alignments, and a maximum of 200 generations. Settings used for the CV procedure were as follows: leave out 25% of observations and average over 10 repetitions.

RESULTS AND DISCUSSION

Statistics on the performance of the conformer sampling GA were collected and averaged over 50 random equations per each combination of equation length (3 to 8 terms) and calibration set size (starting from 20 in steps of 50 compounds) on the benzodiazepine ligand data set. A summary of the results is presented in Table 1. The original data set contains 405 ligands. Due to the 300 atom limit for compound sizes in the program DRAGON used to calculate the descriptors, the final data set size was reduced to 241 compounds. The number of descriptors was 692 after discarding variables with a standard deviation equal to zero (no information content), missing values, and variables with a correlation coefficient $R^2 > 0.9$ to some other variable. As anticipated, the conformer sampling GA did not converge to the global optimum—the nature of a GA is to find a good solution, not the optimal one. The final q^2 values were high, with an average of 0.86 and average standard deviation of 0.08, and did not depend on equation length ($R^2 = 0.1$) nor calibration set size ($R^2 = 0.04$). This finding corroborates the rather trivial part of the working hypothesis*: the algorithm will return a high cross-validated correlation index value q^2 for the correct model.*

Table 1. Statistics on the Performance of the Conformer Sampling GA[c]

calibration set size	average q^2	average % correct conformers	average RMSD [Å][a]	sample size [equations, alignments][b]
20	0.85 ± 0.09	31 ± 26	1.2 ± 0.9	300
70	0.87 ± 0.08	38 ± 19	1.2 ± 0.9	300
120	0.86 ± 0.07	44 ± 15	1.2 ± 0.9	300
170	0.86 ± 0.08	49 ± 11	1.2 ± 0.9	300
220	0.85 ± 0.08	54 ± 8	1.2 ± 0.9	300
241	0.85 ± 0.08	57 ± 8	1.2 ± 0.9	300

[a] The values differ in decimals but round to 1.2 ± 0.9 Å. [b] Each run uses a different equation and produces one alignment. [c] A summary of statistics collected for the final best alignment given by runs of the conformer sampling GA, each time using a different simulated correct equation. The results are averaged over all equation lengths. See text for details.

What is the probability for the conformer sampling GA to find the correct alignment of compounds for a "correct" equation?

The size of the alignment search space increases from $\sim 10^{100}$ to 10^{217} when the calibration set size increases from 20 to 241 compounds. One would expect to find less correctly selected conformations with an increasing number of possible choices, but the percentage of correctly selected conformers increases from about 30 ± 30% for a set of size 20 to ~55 ± 10% for a set of size 241 compounds (Figure 2). The dependency is linear ($R^2 = 0.96$) with an increase (slope) of 0.1% per added compound. It appears that the number of *feasible* alignments *decreases* as the number of compounds increases. Each compound poses a restriction on the alignment; even if two compounds would fit the equation nicely but in the wrong binding mode, three compounds cannot necessarily be aligned in the wrong binding mode while maintaining a high correlation index. This reinforces our intuition that knowing the correct bioactive conformation for some of the ligands (e.g. from the X-ray structure of a complex) will guide the conformer sampling GA toward the correct alignment.

The self-restrictive effect of the alignments depends on the number of compounds in the calibration set. Access to biological activity data for hundreds of compounds, needed to obtain more reliable results, is not a problem for large pharmaceutical companies but university groups with a

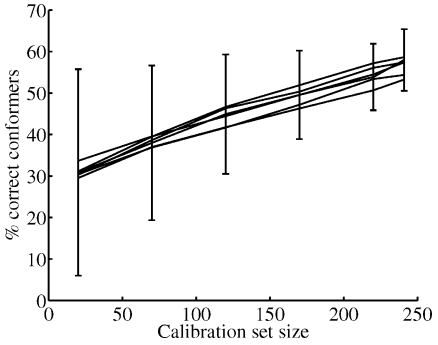

Figure 2. Percentage of correctly selected conformers in the best alignment of compounds produced by the conformer sampling GA averaged over 50 runs. The lines represent the results for equations of lengths of 3 to 8 terms; there is no correlation between the equation length and the percentage of correctly selected conformers. The error bars are for equations with five terms and show the general trend in uncertainty associated with the GA method to find the correct alignment of compounds in the calibration set.

Table 2. Occurrence of Chance Correlating Compound Alignments[a]

calibration set size	average RMSD [Å]	frequency of chance correlation	sample size [alignments]
20	0.1 ± 0.1	0.147	22 183 080
70	0.4 ± 0.2	0.067	23 294 130
120	0.6 ± 0.2	0.027	23 020 830
170	0.8 ± 0.1	0.012	24 080 070
220	1.06 ± 0.05	0.020	25 096 170
241	1.08 ± 0.04	0.006	23 865 180

[a] Summary of statistics on the frequency of occurrence of chance correlation (i.e. compound alignments that give better q^2 than the "correct" alignment) as observed during multiple runs of the conformer sampling GA. For each calibration set size, the results were accumulated over 300 runs, each run using a different simulated equation, and averaged over all equation lengths. See text for details.

modest budget can hardly afford such a screening effort, which may hinder these groups from exploiting the full potential of McQSAR. On the other hand, the standard deviation of the percentage of correctly selected conformers is large for small sets (Table 1; error bars in Figure 2), implying that even with a small calibration set one might get lucky and find an alignment describing the correct binding mode.

The average of the RMSDs of the incorrect conformers to the "bioactive" conformers measures the distance of an alignment to the correct binding mode. The average of the RMSDs was 1.2 ± 0.9 Å for all measurements (Table 1). Thus, the average RMSD of 1.2 Å can be taken as the higher limit of resolution of the conformer selection procedure for the benzodiazepine data set. If we consider that a typical resolution of an X-ray crystal structure of a biomolecule is approximately 2 Å, the first part of the working hypothesis, *a correct model will always rank well and converge to an alignment that describes the common binding mode of the compounds in the calibration set*, seems to hold within a moderate error limit in this particular example case.

What is the probability of a wrong alignment to rank better than the correct one?

An instance of chance correlation was defined in the context of alignments as an alignment that gives a q^2 value better than or equal to that of the correct alignment. The frequency of chance correlation was measured during

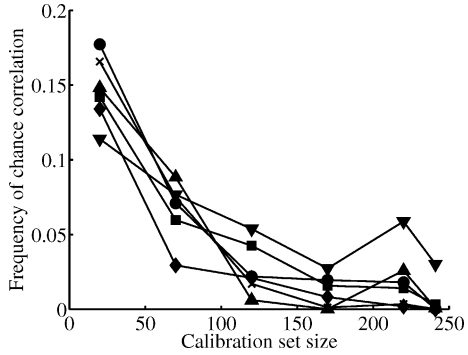

Figure 3. The observed frequency of occurrence of chance correlation for the conformer sampling GA. Each data point represents results accumulated over 50 runs. The lines are for equations containing 3 (down-triangle), 4 (square), 5 (circle), 6 (diamond), 7 (up-triangle), and 8 (cross) terms.

multiple runs of the conformer sampling GA using simulated correct equations. A summary of the results is presented in Table 2. Chance correlation does occur. Its frequency depends ($R^2 = 0.7$) on the calibration set size ranging from 15% (set size 20) to 0.6% (set size 241) (Figure 3). The frequency of chance correlation did not depend linearly on the equation length ($R^2 = 0.01$). On average, $4 \cdot 10^6$ alignments were tested for each combination of equation length and set size, resulting in a total of $141 \cdot 10^6$ alignments. The average RMSD to the correct conformers for a chance correlating alignment increases with set size, from 0.1 ± 0.1 Å (set size 20) to 1.08 ± 0.04 Å (set size 241). This means that the bigger the calibration set, the less probable the occurrence of chance correlation, but when it occurs, it is more severe than in the case of a small calibration set.

What is the probability for a "wrong" equation to correlate by chance?

The traditional concept of chance correlation is defined in the context of equations. The occurrence of chance correlation is known to become more probable as the number of screened variables increases;[33] buying 10 lottery tickets gives a better chance to win than one ticket would. In this study, we did not explicitly keep track of the number of screened variables. The number of screened variables is a function of the equation population size, the number of elapsed generations, and the mutation probability. As generations of the equation generating GA elapse, more and more variables are selected from the pool of the 692 available variables. The frequency of chance correlation was recorded over five runs of McQSAR per each combination of equation length and calibration set size, each run using a new simulated correct equation.

The observed "wrong" equations were divided into three classes: incorrect, partially correct, and overfitting models (see Methods). A summary of the observed frequencies of occurrence of equations of each class is presented in Table 3. Of the total of 198 000 tested equations, four models (0.002%) were overfitting, all of which occurred within a single run, which implies that the CV procedure adequately restricts the degree of complexity of the evolved models. The frequency of occurrence of incorrect and partially correct models did not depend on the calibration set size ($R^2 \approx 0.006$ for both model classes) nor on the equation length ($R^2 \approx 0.2$ for both classes). Incorrect models had an average frequency of 0.2 ± 0.5%. Partially correct models, which

contain the building blocks needed by the GA to construct the correct model, occurred more often, with an average of $0.8 \pm 2\%$ over all test runs. The Schema Theorem of GAs[6] implies that the number of samples (instances) of low-order building blocks that have fitness greater than the average in the population will increase exponentially over the elapsed generations. Therefore, one can expect the GA to find the correct model before a high-correlating incorrect model is developed. Thus, a naïve conclusion is obtained: the second part of our working hypothesis (*equations that do not represent the correct binding mode are not likely to give a high* q^2) holds with the apparent probability of one minus the frequency of incorrect equations ≈0.99.

DERIVATION OF PHARMACOPHORE MODELS

The alignment of ligand molecules obtained from the McQSAR algorithm could be used to derive pharmacophore models that describe the binding mode of the ligands. The alignment inherently contains errors (incorrect conformers for some compounds) which give rise to incorrect pharmacophores. If the alignment contains about 50% correct conformers, can it be used to derive a pharmacophore model that accurately describes the binding mode?

Pharmacophores are commonly derived using some consensus method on a few of the most active compounds of the calibration set (see, e.g., ref 38). Assume that we use two active ligands at a time to derive consensus pharmacophore models. The likelihood of one compound to have the correct conformer is denoted *f*(correct). The probability of two compounds to have the correct conformer is the product of the likelihoods for the individual ligands: $P(\text{correct}) = f(\text{correct})^2$. We coarsely assume that if both conformers are correct, then a useful pharmacophore model will also be derived. The likelihood of the complement event, that one or both of the ligands have a wrong conformer and all pharmacophores based on that pair of ligands are incorrect, is $P(\text{incorrect}) = 1 - P(\text{correct})$. For k active ligands, there are $k!/2!(k-2)!$ possible different pairs (combinations of two) of ligands. If consensus pharmacophores are derived using all the possible different pairs of active ligands, the probability for a useful pharmacophore model to be created becomes

$$P(\text{useful pharmacophore created}) = 1 - P(\text{all pharmacophores incorrect})$$

$$= 1 - P(\text{incorrect})\frac{k!}{2!(k-2)!}$$

$$= 1 - (1 - P(\text{correct}))\frac{k!}{2!(k-2)!}$$

$$= 1 - (1 - f(\text{correct})^2)\frac{k!}{2!(k-2)!}$$

In our situation *f*(correct) = 0.57 (Table 1, calibration set of size of 241). Substituting the value for *f*(correct) to the above equation and varying the number of active ligands k yields probabilities 0.3249 ($k = 2$), 0.6923 ($k = 3$), 0.9053 ($k = 4$), 0.9803 ($k = 5$), and 0.9972 ($k = 6$). The probability for a useful pharmacophore to be created increases heavily with the number of active ligands used for derivation of the candidate pharmacophores, and thus the probability of finding a useful pharmacophore depends on the number of ligands we define as "active". The activity values must vary between the ligands of the calibration set in order to be useful for QSAR. Therefore one can always pick, say, five most active compounds as basis for the pharmacophore generation process. Setting $k = 5$ in the above equation and varying *f*(correct) yields probabilities 0.0956 (*f*(correct) = 0.1), 0.3352 (0.2), 0.6106 (0.3), and 0.8251 (0.4). The odds turn to favor the creation of a useful pharmacophore when 30% or more of the conformers are correct (30% is the average for a calibration set of size of 20; Table 1), and five active ligands are used to generate the set of candidate pharmacophore models.

Table 3. Statistics on the Occurrence of Chance Correlating Equations[b]

calibration set size	frequency of occurrence of "wrong" models: incorrect	partially correct	overfitting	sample size [equations]
20	0.0054	0.0103	0.0	33 000
70	0.0010	0.0060	0.0	33 000
120	0.0008	0.0035	0.0	33 000
170	0.0017	0.0026	0.00012	33 000
220	0.0025	0.0046	0.0	33 000
241	0.0031	0.0192	0.0	33 000
average[a]	0.002 ± 0.005	0.008 ± 0.020		

[a] Averaged over all combinations of calibration set size and equation length. [b] Summary of statistics on the occurrence of chance correlating equations during multiple runs of the McQSAR algorithm. The results for each calibration set size are averaged over all equation lengths. See text for details.

CONCLUSIONS

The objective of this study was to find out whether a GA could select the representative forms of the compounds as well as the QSAR model. The answer is yes, but you need a large calibration set and/or a regular serving of serendipity to achieve a good result. Furthermore, test runs show that chance correlation is not more of an issue with the McQSAR algorithm than with any other variable selection technique used for statistical modeling.

The method addresses the interpretability of QSAR models by suggesting a set of bioactive conformations for the ligands in the calibration set. The suggested conformers can be used to derive a pharmacophore model describing the common binding mode of the compounds.

ACKNOWLEDGMENT

We thank Dr. Pauli Saarenketo, Dr. Anna-Marja Hoffrén, Kurt Kokko, Dr. Ville-Veikko Rantanen, and Mikko Huhtala for suggestions and inspiring discussions. Juvantia Pharma Ltd. is acknowledged for the cooperation in generating conformer databases. McQSAR uses portions of the GAlib genetic algorithm package written by Matthew Wall at the Massachusetts Institute of Technology. The authors gratefully acknowledge funding from the Academy of Finland and from the Sigrid Jusélius Foundation.

REFERENCES AND NOTES

(1) Hansch, C.; Fujita, T. Rho-Sigma-Pi Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* **1964**, *86*, 1616−1626.

(2) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, 2000; Vol. 11, p 667.
(3) Glover, F.; Laguna, M. *Tabu Search*; Kluwer Academic Publishers: 1997; p 408.
(4) Glover, F. Tabu Search- - Part I. *INFORMS J. Comput.* **1989**, *1*, 190−206.
(5) Glover, F. Tabu Search- - Part II. *INFORMS J. Comput.* **1990**, *2*, 4−32.
(6) Holland, J. H. *Adaptation in Natural and Artificial Systems: And Introductory Analysis With Applications to Biology, Control, and Artificial Intelligence*, 1st MIT Press edition ed.; The MIT Press: Cambridge, 1975; p 211.
(7) Brooks, S. P.; Friel, N.; King, R. Classical model selection via simulated annealing. *J. R. Stat. Soc. Ser. B-Stat. Methodol.* **2003**, *65*, 503−520.
(8) Kirkpatrick, S.; Gelatt, C. D.; Vecchi, M. P. Optimization by Simulated Annealing. *Science* **1983**, *220*, 671−680.
(9) Dorigo, M. Optimization, Learning and Natural Algorithms (in Italian), Ph.D. Thesis, Politecnico di Milano, Italy, Milano, 1992.
(10) Dorigo, M.; Stützle, T. *Ant Colony Optimization*; The MIT Press: 2004; p 328.
(11) Eberhart, R.; Shi, Y.; Kennedy, J. *Swarm Intelligence*; Morgan Kaufmann: 2001; p 512.
(12) Eberhart, R. C.; Kennedy, J. A. A new optimizer using particle swarm theory. *Proceedings of the Sixth International Symposium on Micro-machine and Human Science, Nagoya, Japan;* 1995; pp 39−43.
(13) Kennedy, J. A.; Eberhart, R. C. Particle swarm optimization. *Proceedings of IEEE International Conference on Neural Networks, Piscataway, NJ;* 1995; pp 1942−1948.
(14) Press: W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipies in C: the art of scientific computing*, 2nd ed.; Cambridge University Press: Cambridge, 1992; p 994.
(15) Wold, S.; Sjostrom, M.; Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109−130.
(16) Haykin, S. *Neural Networks: A Comprehensive Foundation*, 2nd ed.; Prentice Hall: 1999; p 842.
(17) George, E. I.; McCulloch, R. E. Approaches for Bayesian variable selection. *Stat. Sin.* **1997**, *7*, 339−373.
(18) Perkins, R.; Fang, H.; Tong, W. D.; Welsh, W. J. Quantitative structure−activity relationship methods: Perspectives on drug discovery and toxicology. *Environ. Toxicol. Chem.* **2003**, *22*, 1666−1679.
(19) Leardi, R.; González, A. L. Genetic algorithms applied to feature selection in PLS regression: how and when to use them. *Chemom. Intell. Lab. Syst.* **1998**, *41*, 195−207.
(20) Jouan-Rimbaud, D.; Massart, D. L.; de Noord, O. E. Random correlation in variable selection for multivariate calibration with a genetic algorithm. *Chemom. Intell. Lab. Syst.* **1996**, *35*, 213−220.
(21) Rogers, D.; Hopfinger, A. J. Application of Genetic Function Approximation to Quantitative Structure−Activity-Relationships and Quantitative Structure−Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854−866.
(22) Mekenyan, O.; Nikolova, N.; Schmieder, P.; Veith, G. COREPA-M: A multidimensional formulation of COREPA. *QSAR Comb. Sci.* **2004**, *23*, 5−18.
(23) Mekenyan, O. Dynamic QSAR Techniques: Applications in Drug Design and Toxicology. *Curr. Pharm. Des.* **2002**, *8*, 1605−1621.
(24) Nicolotti, O.; Gillet, V. J.; Fleming, P. J.; Green, D. V. S. Multiobjective optimization in quantitative structure−activity relationships: Deriving accurate and interpretable QSARs. *J. Med. Chem.* **2002**, *45*, 5069−5080.
(25) Gill, P. R. M., W.; Wright, M. H. The Levenberg−Marquardt Method. In *Practical Optimization*; Academic Press: London, 1981; pp 136−137.
(26) Levenberg, K. A Method for the Solution of Certain Problems in Least Squares. *Quart. Appl. Math.* **1944**, *2*, 164−168.
(27) Marquardt, D. An Algorithm for Least-Squares Estimation of Nonlinear Parameters. *SIAM J. Appl. Math.* **1963**, *11*, 431−441.
(28) Baumann, K.; von Korff, M.; Albert, H. A systematic evaluation of the benefits and hazards of variable selection in latent variable regression. Part II. Practical applications. *J. Chemom.* **2002**, *16*, 351−360.
(29) Baumann, K. Cross-validation as the objective function for variable-selection techniques. *TRAC, Trends Anal. Chem.* **2003**, *22*, 395−406.
(30) Shao, J. Linear-Model Selection by Cross-Validation. *J. Am. Stat. Assoc.* **1993**, *88*, 486-494.
(31) Zhang, P. Model Selection Via Multifold Cross-Validation. *Ann. Stat.* **1993**, *21*, 299−313.
(32) Golbraikh, A.; Tropsha, A. Beware of q^2! *J. Mol. Graph. Model.* **2002**, *20*, 269−276.
(33) Topliss, J. G.; Edwards, R. P. Chance Factors in Studies of Quantitative Structure−Activity-Relationships. *J. Med. Chem.* **1979**, *22*, 1238−1244.
(34) Hawkins, D. M. The problem of overfitting. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1−12.
(35) Todeschini, R.; Consonni, V.; Mauri, A.; Pavan, M. Detecting "bad" regression models: multicriteria fitness functions in regression analysis. *Anal. Chim. Acta* **2004**, *515*, 199−208.
(36) Deb, K.; Pratap, A.; Agarwal, S.; Meyarivan, T. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comput.* **2002**, *6*, 182−197.
(37) Sutherland, J. J.; O'Brien, L. A.; Weaver, D. F. Spline-fitting with a genetic algorithm: A method for developing classification structure−activity relationships. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1906−1915.
(38) Li, H.; Sutter, J.; Hoffmann, R., HypoGen: An Automated System for Generating 3D Predictive Pharmacophore Models. In *Pharmacophore Perception, Development, and Use in Drug Design*, 2nd ed.; Güner, O. F., Ed.; International University Line: La Jolla, CA, 2000; pp 171−187.

CI0501847