Development and Validation of a Novel Variable Selection Technique with Application to Multidimensional Quantitative Structure−Activity Relationship Studies

Chris L. Waller*

OSI Pharmaceuticals, Inc., 4727 University Drive, Suite 400, Durham, North Carolina 27707

Mary P. Bradley[†]

Rhone-Poulenc Agro, 2 T.W. Alexander Drive, Research Triangle Park, North Carolina 27709

Received May 19, 1998

Variable selection is typically a time-consuming and ambiguous procedure in performing quantitative structure−activity relationship (QSAR) studies on overdetermined (regressor-heavy) data sets. A variety of techniques including stepwise and partial least squares/principlal components analysis (PLS/PCA) regression have been applied to this common problem. Other strategies, such as neural networks, cluster significance analysis, nearest neighbor, or genetic (function) or evolutionary algorithms have also evaluated. A simple random selection strategy that implements iterative generation of models, but directly avoids crossover and mutation, has been developed and is implemented herein to rapidly identify from a pool of allowable variables those which are most closely associated with a given response variable. The FRED (fast random elimination of descriptors) algorithm begins with a population of offspring models composed of either a fixed or variable number of randomly selected variables. Iterative elimination of descriptors leads naturally to subsequent generations of more fit offspring models. In contrast to common genetic and evolutionary algorithms, only those descriptors determined to contribute to the genetic makeup of less fit offspring models are eliminated from the descriptor pool. After every generation, a new random increment line search of the remaining descriptors initiates the development of the next generation of randomly constructed models. An optional algorithm that eliminates highly correlated descriptors in a stepwise manner prior to the development of the first generation of offspring greatly enhances the efficiency of the FRED algorithm. A comparison of the results of a FRED analysis of the Selwood data set ($n = 31$ compounds, $k = 53$ descriptors) with those obtained from alternative algorithms reveals that this technique is capable of identifying the same "optimal" solutions in an efficient manner.

INTRODUCTION

Traditionally, quantitative structure−activity relationship (QSAR) models were derived from log *P* (calculated or measured partition coefficients) and steric and electronic parameters (Taft parameters and Hammett σ constants).[1] Descriptors of this type are relatively few and generally describe a measurable molecular property. The use of numerous descriptors that are more indicative of molecular structure and topology is becoming more important for use in QSAR. These types of descriptors are easily calculated from molecular structures and potentially number in the thousands. Selecting the relevant descriptors from among a large pool is therefore necessary if one is to derive useful models. Variable selection in the form of principal components analysis/partial least squares (PCA/PLS)[2] is regularly used in 3D-QSAR models to reduce thousands of comparative molecular field analysis (CoMFA)[3] variables to a manageable handful of latent variables—latent variables being related to the actual variables (i.e., steric and electrostatic interaction energies between commonly aligned molecules with a grid of probe molecules) by a loading matrix. In CoMFA, PLS variable reduction works to our advantage in that the actual latent variables (i.e., components) are essentially transparent to the end-user since the loading matrix will be utilized to allow for visualization of the results on the entire grid or the total set of original variables. While this approach works well for CoMFA variables whose effects can be depicted graphically and thereby interpreted, it is usually preferable to work in the realm of actual variables whose effects can be visualized by chemists in a synthetic chemistry program.

The development of variable selection routines has been the focus of many research efforts over the last several years.[4−9] This area has recently been stimulated by the broad implementation of high-throughput synthetic chemistry and biological screening techniques that have necessitated the advent of computational algorithms for the rapid analysis of large data sets of primary and secondary biological data for structure−activity relationships. To this end, physicochemical parameters that are easily calculable while information-rich are desired. Molecular connectivity and topological indices,[10] two- and three-dimensional molecular fingerprints,[11] and molecular substructure keys[12] are examples of descriptors that are being utilized in QSAR analyses of large data sets. History provides us with an excellent "real-life" example for QSAR practitioners—the Selwood data set.[13] This particular data set was developed as a result of a research project

* To whom correspondence should be addressed: E-mail waller@mindspring.com. Present address: Sphinx Pharmaceuticals, Inc., 20 T.W. Alexander Drive, Research Triangle Park, NC 27709.

[†] E-mail mbradley@rp-agro.com.

10.1021/ci980405r CCC: $18.00
Published on Web 01/30/1999

aimed at the development/discovery of novel antifilarials. While not created for use as a test data set for variable selection algorithms, this data set with its associated variable selection algorithm has served as the benchmark by which variable selection algorithms have been evaluated. A variety of algorithms have been implemented with the goal of obtaining the *best* model(s). Among these are Selwood's original nonlinear mapping algorithm implemented in conjunction with stepwise regression,[13] neural network analysis,[9] cluster significance analysis algorithm,[4] genetic function approximation,[8] and an evolutionary algorithm.[5,6]

In each of the above studies, the result was a *best* model or a population of *best* models. The definition of *best* varies by study and for the genetically based algorithms is highly dependent on the choice and implementation of the fitness function or the criteria against which the acceptability of offspring are measured. Herein, the results of the implementation of a novel evolutionary variable selection routine are contrasted and compared to the previously reported algorithms. The algorithm implements a novel variable elimination philosophy that reduces the chances of finding localized minima while optimizing performance.

METHODS

The program is written entirely in Sybyl programming language (SPL) and runs under the QSAR module.[14] The program presently consists of approximately 800 lines of code. In the interest of simplicity, the fast random elimination of descriptors (FRED) algorithm is presented in pseudocode in the Appendix. The code is available from the authors upon request.

Variable Reduction Strategies. Two variable reduction options are available prior to the generation of the first set of offspring models. The first of these is a variable filter routine that performs simple statistics on each set of descriptor variables. Variables can be eliminated from the available descriptor pool if there is no or little variation in the values within a given set. Elimination of zero variance descriptors is highly recommended and greatly enhances the efficiency of the algorithm. Additionally, it is possible to eliminate interdependent, or collinear, descriptors. The FRED algorithm currently implements a simple stepwise comparison routine that eliminates variables that do not meet a user-defined collinearity threshold. As currently employed, the algorithm begins with two identical lists of the descriptors and compares sequentially each variable in the first list with each variable in the second. If during the comparison process, variables are found to be collinear by a correlation threshold defined by the user, the variable in the second list is eliminated from further consideration. The algorithm terminates once all variables have been compared, and those remaining in the second list are passed to the variable selection routines.

Calculation of Offspring Model Characteristics. Once the available descriptor pool has been culled, the number of allowable descriptors per model is computed. An empirical rule-of-thumb approach was implemented in that for each descriptor (*m*) or independent variable, one must have at least five observations (*n*). This factor varies by opinion; however, it is generally agreed that between five and six observations are required per descriptor in order to minimize the chance of overfitting.[15] Therefore, the maximal number of descriptors per offspring model is calculated as *n*/6. This number can either be fixed so that each offspring in every generation during a run will have exactly the same number of descriptors. Alternatively, it is possible to allow this number to vary with each successive offspring model with the upper limit set to the maximal number of descriptors calculated above. The implications of this latter option with respect to the selection of the fitness function will be discussed.

Calculation of the Number of Offspring To Be Produced per Generation. A simple random number generation algorithm is utilized in the selection of descriptors for the offspring models. An oversampling (progeny) factor is implemented to ensure that all descriptors are selected during this process. This number is variable and affects the algorithm in the following manner: if the number of descriptors to be included in each offspring model is set to the maximal number of allowed descriptors (defined above) so that each descriptor may be chosen only once per offspring and each descriptor is only allowed to be selected once per generation, then the minimal number of models that would be required to sample all the descriptors in the total descriptor pool can be calculated as the total number of descriptors divided by the maximum number of descriptors allowed per model. The algorithm as implemented does not enforce this last restriction, which effectively means that it is possible that one or more descriptors can be either included in or excluded from all models in a generation. An empirical oversampling factor has been utilized that takes this minimal number of models calculated above and factors it upward. It is also possible to allow this oversampling factor value to vary randomly per generation with the minimal and maximal values being set by the user.

Calculation of the Fitness of Offspring. For each offspring model generated during a generation, a PLS regression analysis with full cross-validation and the number of components set equal to the number of maximum number of allowable descriptors is performed. Two fitness functions are available: (1) the r^2 statistic and (2) the leave-one-out cross-validated r^2, or q^2_{LOO}, statistic. The recommendations for use are discussed below.

Strategy for Variable Reduction. Prior to the generation of the first set of offspring, the user is prompted to supply a kill factor. This factor is used in the algorithm in the following manner. Once all the offspring models have been generated for a given generation, the models are sorted according to fitness. The kill factor is then used to select a given percentage of models from the tails of the distribution of the offspring model population. The descriptors for the models from the lower distribution (i.e., less fit models) are then compared to the descriptors for the models from the upper distribution (i.e., more fit models). At this point, the algorithm implements a modified taboo (aka tabu) search.[16] Those descriptors from the lower distribution not found in the set from the upper distribution are considered detrimental to the fitness of an entire generation of offspring models and are considered suspect. Unlike some previous implementations of taboo search, the variable is not always deselected at this point; however, it is placed on a taboo list. The algorithm is allowed to select this descriptor for incorporation into the makeup of subsequent generations of offspring models on the premise that it may not contribute to a suboptimal model given a second chance in combination with other descriptors. However, once a variable appears on the taboo list in a subsequent but not necessarily sequential generation, it is removed from the allowable descriptor pool. This interpretation of strategic forgetting provides additional insurance that a given descriptor is truly detrimental. A kill factor of 5% is recommended.

Fitness Functions. The fitness function is the primary function in genetic and evolutionary algorithms and the choice of appropriate fitness function has stimulated much debate in the recent literature. A variety of fitness functions are available for use in FRED. In fact, the choice is only limited to the user's imagination. For the purposes of this validation study, the simple q^2_{LOO} value for the offspring model was used. Easily accessible fitness functions also include the simple r^2 and F statistics.

Termination Criteria. Several termination criteria are available for use in the FRED algorithm. The fitness function chosen must be considered when selecting a termination criterion. In this original implementation of the FRED algorithm, a fitness function of q^2_{LOO} was selected as described above. To replicate the results from previous studies in which populations of fit offspring were produced, a termination criterion was designed that would examine the standard deviation of the fitness values (i.e., q^2_{LOO}) for the entire population in a generation. If this value were less than a set threshold value, or minimum σ, then the algorithm would terminate. It is possible to force the algorithm to terminate with a singular fit individual model with a fixed or variable set of characteristics by setting the termination criterion to a smaller minimum σ value. In instances where q^2_{LOO} is implemented as the fitness function and numerous offspring models are desired, a minimum σ value of 0.1 or greater is recommend. A minimum σ value of 0.025 or smaller is generally sufficient to yield a singular offspring model at algorithm termination.

Data Set Composition. Table 1 lists the 53 descriptors for each of the 31 compounds that comprise the training set used in this study. These values as well as the in vitro biological activities (Table 2) for these synthetic derivatives of 3-formamido-2-hydroxysalicylic acid against disease-causing nematodes are used as originally reported.[13]

Table 1. Descriptions of the Variables in the Selwood Data Set

variable	description
ATCH1	partial atomic charge for atom
ATCH2	partial atomic charge for atom
ATCH3	partial atomic charge for atom
ATCH4	partial atomic charge for atom
ATCH5	partial atomic charge for atom
ATCH6	partial atomic charge for atom
ATCH7	partial atomic charge for atom
ATCH8	partial atomic charge for atom
ATCH9	partial atomic charge for atom
ATCH10	partial atomic charge for atom
DIPV_X	dipole vector
DIPV_Y	dipole vector
DIPV_Z	dipole vector
DIPMOM	dipole moment
ESDL1	electrophilic superdelocalizability for atom
ESDL2	electrophilic superdelocalizability for atom
ESDL3	electrophilic superdelocalizability for atom
ESDL4	electrophilic superdelocalizability for atom
ESDL5	electrophilic superdelocalizability for atom
ESDL6	electrophilic superdelocalizability for atom
ESDL7	electrophilic superdelocalizability for atom
ESDL8	electrophilic superdelocalizability for atom
ESDL9	electrophilic superdelocalizability for atom
ESDL10	electrophilic superdelocalizability for atom
NSDL1	nucleophilic superdelocalizability for atom
NSDL2	nucleophilic superdelocalizability for atom
NSDL3	nucleophilic superdelocalizability for atom
NSDL4	nucleophilic superdelocalizability for atom
NSDL5	nucleophilic superdelocalizability for atom
NSDL6	nucleophilic superdelocalizability for atom
NSDL7	nucleophilic superdelocalizability for atom
NSDL8	nucleophilic superdelocalizability for atom
NSDL9	nucleophilic superdelocalizability for atom
NSDL10	nucleophilic superdelocalizability for atom
VDWVOL	van der Waals volume
SURF_A	surface area
MOFI_X	principal moments of inertia
MOFI_Y	principal moments of inertia
MOFI_Z	principal moments of inertia
PEAX_X	principal ellipsoid axes
PEAX_Y	principal ellipsoid axes
PEAX_Z	principal ellipsoid axes
MOL_WT	molecular weight
S8_IDX	substituent dimensions
S8_IDY	substituent dimensions
S8_IDZ	substituent dimensions
S8_ICX	substituent centers
S8_ICY	substituent centers
S8_ICZ	substituent centers
LOGP	partition coefficient
M_PNT	melting point
SUM_F	sum of F substituent constant
SUM_R	sum of R substituent constant

RESULTS

Algorithm Performance (Three-Variable Models). Table 3 summarizes the results of the FRED analysis in which a fixed number of descriptors, in this case three, was desired. An oversampling, or progeny, factor of 30 and a kill factor of 5 were used. A termination criterion of a minimum σ equal to 0.1 was specified. This "loose" criterion allowed for the generation of several optimal models rather than one. These results are more comparable to the techniques in which optimal populations of models are produced rather than a single optimal individual (discussed below). A total of 13 variables were represented in the final population of three-descriptor offspring. In general, these variables were also selected by the MUSEUM algorithm and are, with one exception, a superset of the variables selected by the GFA algorithm. The average fitness values per generation for each of three trial runs are plotted in the upper panel of Figure 1. This plot indicates that the algorithm truly is evolutionary in that the average fitness per generation gradually increases as descriptors are eliminated. The lower panel of Figure 1 indicates that as descriptors are eliminated from the gene pool, fewer offspring models are required to adequately sample the allowable combinations. The total number of model offspring produced per trial run was 2700, 2520, and 2610, with convergence occurring at 21, 20, and 20 generations, respectively.

Algorithm Performance (Six-Variable Models). The algorithm was also utilized to identify the best six-variable model. This was accomplished by adjusting the minimum σ value to 0.025. This relatively tight criterion forced the algorithm to terminate with a singular model. In the upper panel of Figure 2, the average fitness values per generation for three trial runs along with the standard deviation of the fitness values in a generation are plotted for three separate

Table 2. Biological Activity and Parameter Values

	activity	random	ATCH1	ATCH2	ATCH3	ATCH4	ATCH5	ATCH6	ATCH7	ATCH8	ATCH9	ATCH10	DIPV_X	DIPV_Y	DIPV_Z	DIPMOM
K17	−1	0	0.17	0.04	−0.01	−0.1	0	−0.24	−0.25	0.42	−0.41	−0.4	1	−2.67	−0.03	2.85
D30	−1	−0.88	0.26	−0.15	0.09	−0.16	0.11	−0.29	−0.21	0.43	−0.4	−0.32	3.13	−1.34	0.19	3.41
J19	−0.9	0.92	0.17	0.04	−0.01	−0.1	0	−0.24	−0.25	0.42	−0.41	−0.4	0.95	−2.4	0.06	2.58
A5	−0.88	1.03	0.25	−0.14	0.09	−0.47	0.11	−0.29	−0.22	0.43	−0.4	−0.33	0.03	−2.28	3.68	4.33
J1	−0.85	1.55	0.17	0.04	−0.01	−0.1	0	−0.24	−0.25	0.42	−0.41	−0.4	2.2	−1.78	−0.02	2.83
K18	−0.41	1.07	0.17	0.04	−0.01	−0.1	0	−0.24	−0.25	0.42	−0.41	−0.4	0.99	−2.65	−0.02	2.83
G2	−0.38	1.4	0.17	0.04	−0.01	−0.1	0	−0.25	−0.25	0.43	−0.4	−0.33	−1.25	−1.49	0.39	1.98
L25	−0.04	1.84	0.28	−0.17	0.1	−0.15	0.08	−0.29	−0.19	0.42	−0.4	−0.4	−1.12	−8.92	−0.04	8.99
A10	0	1.02	0.26	−0.14	0.1	−0.51	0.15	−0.29	−0.22	0.43	−0.4	−0.32	−3.16	1.58	−2.7	4.45
C11	0.1	0.1	0.26	−0.15	0.09	−0.16	0.11	−0.29	−0.21	0.43	−0.39	−0.33	1.15	0.07	0.77	1.38
D23	0.23	1.13	0.26	−0.15	0.09	−0.16	0.11	−0.29	−0.21	0.43	−0.4	−0.32	3.43	0.18	0.25	3.45
F15	0.3	1.36	0.29	−0.52	0.13	−0.14	0.08	−0.29	−0.23	0.43	−0.39	−0.33	0.96	−6.13	−3.3	7.03
G4	0.32	0.23	0.23	−0.14	0.05	−0.26	0.08	−0.28	−0.23	0.43	−0.4	−0.33	−3.1	−2.05	−1	3.85
G9	0.42	−1	0.23	−0.25	0.06	−0.13	0.04	−0.28	−0.23	0.43	−0.4	−0.33	−2.37	−3.76	−0.82	4.52
I26	0.43	0.89	0.26	−0.15	0.09	−0.16	0.11	−0.29	−0.21	0.42	−0.39	−0.32	2.27	1.35	0.66	2.72
N31	0.48	1.41	0.25	−0.15	0.09	−0.16	0.12	−0.29	−0.21	0.35	−0.33	−0.15	3.58	1.79	0.7	4.07
H14	0.77	0.3	0.26	−0.15	0.09	−0.16	0.11	−0.29	−0.21	0.43	−0.4	−0.32	2.9	−0.28	0	2.91
M6	0.82	−0.38	0.28	−0.16	0.1	−0.15	0.08	−0.29	−0.19	0.43	−0.39	−0.32	0.38	−8.07	0.39	8.09
L21	0.82	0.32	0.28	−0.17	0.1	−0.15	0.08	−0.29	−0.19	0.42	−0.4	−0.4	−1.1	−8.93	−0.03	9
E20	0.89	0.42	0.26	−0.15	0.09	−0.16	0.11	−0.28	−0.21	0.42	−0.41	−0.4	3.42	−0.38	−0.02	3.44
B13	0.92	0.77	0.26	−0.15	0.09	−0.16	0.11	−0.29	−0.21	0.43	−0.4	−0.32	−2.35	5.16	1.35	5.83
B8	1.02	0.43	0.26	−0.15	0.09	−0.16	0.11	−0.29	−0.21	0.43	−0.4	−0.32	−0.04	3.71	1.24	3.92
B27	1.03	−0.85	0.26	−0.15	0.09	−0.16	0.11	−0.29	−0.21	0.43	−0.4	−0.32	0.44	2.61	−1.53	3.06
B29	1.07	−0.9	0.26	−0.15	0.09	−0.16	0.11	−0.29	−0.21	0.43	−0.39	−0.32	−1.08	1.99	−1.56	2.75
C12	1.13	−1	0.26	−0.15	0.09	−0.16	0.11	−0.29	−0.21	0.43	−0.39	−0.33	−0.73	0.77	0.25	1.1
C16	1.36	−0.41	0.26	−0.15	0.09	−0.16	0.11	−0.29	−0.21	0.43	−0.39	−0.33	0.28	1.08	−0.21	1.14
L22	1.36	0.82	0.28	−0.16	0.11	−0.09	0.1	−0.28	−0.18	0.42	−0.4	−0.4	0.42	−8.08	−0.03	8.09
B3	1.4	1.36	0.26	−0.15	0.09	−0.16	0.11	−0.29	−0.21	0.43	−0.39	−0.32	0.11	1.46	0.45	1.53
E24	1.41	−0.04	0.26	−0.15	0.09	−0.16	0.11	−0.28	−0.21	0.42	−0.41	−0.4	3.42	−0.38	−0.02	3.44
B28	1.55	0.82	0.26	−0.15	0.09	−0.16	0.11	−0.29	−0.21	0.43	−0.39	−0.32	−0.3	2.56	0.46	2.62
B7	1.84	0.48	0.24	−0.14	0.06	−0.08	0.07	−0.28	−0.22	0.43	−0.4	−0.33	−1.52	−0.54	0.44	1.67

Table 2. (Continued)

	ESDL1	ESDL2	ESDL3	ESDL4	ESDL5	ESDL6	ESDL7	ESDL8	ESDL9	ESDL10	NSDL1	NSDL2	NSDL3	NSDL4	NSDL5	NSDL6
K17	−0.55	−0.8	−1.7	−0.29	−1.29	−1.42	−0.4	−0.47	−0.61	−0.39	7.53	4.16	1.1	5.97	4	1.03
D30	−0.66	−0.95	−0.99	−0.42	−1.02	−0.7	−0.41	−0.93	−0.78	−0.34	6.04	3.43	1.08	4.04	1.29	1.68
J19	−0.56	−0.84	−1.78	−0.29	−1.36	−1.49	−0.41	−0.49	−0.62	−0.39	6.55	3.65	1.06	5.18	3.47	1
A5	−1.21	−0.87	−1.12	−1.07	−1.53	−0.84	−0.51	−0.45	−0.55	−0.38	1.16	1.15	1.04	2.17	0.94	1.07
J1	−0.55	−0.81	−1.7	−0.29	−1.29	−1.43	−0.4	−0.47	−0.61	−0.39	7.67	4.24	1.1	6.07	4.07	1.03
K18	−0.55	−0.8	−1.7	−0.29	−1.29	−1.43	−0.4	−0.46	−0.61	−0.39	7.74	4.28	1.1	6.13	4.11	1.03
G2	−7.2	−3.75	−1.6	−5.52	−4.04	−1.32	−1.35	−0.51	−0.6	−0.47	0.8	0.99	0.91	0.88	0.82	0.94
L25	−1.27	−0.43	−0.84	−0.66	−0.6	−0.66	−0.53	−0.25	−0.45	−0.33	0.99	1.97	1.26	2.6	2.78	1.3
A10	−0.87	−0.76	−0.95	−1.03	−1.21	−0.71	−0.45	−0.55	−0.59	−0.35	1.14	1.21	1.17	1.18	0.95	1.1
C11	−18.52	−10.23	−0.94	−8.61	−2.13	−2.34	−2.97	−1.24	−0.98	−0.79	0.99	1.05	1.14	1.66	1.01	1.14
D23	−0.74	−1.07	−1.01	−0.43	−1.04	−0.69	−0.42	−1.08	−0.85	−0.34	6.6	3.56	1.08	4.39	1.3	1.85
F15	−1.36	−1	−0.86	−0.86	−0.88	−0.68	−0.51	−0.37	−0.5	−0.34	0.99	1.11	1.23	1.17	1.06	1.13
G4	−2.07	−1.35	−1.38	−1.84	−2.36	−1	−0.64	−0.46	−0.56	−0.41	0.91	0.96	0.95	1.05	0.8	0.96
G9	−5.28	−2	−1.62	−3.53	−2.52	−1.2	−1.12	−0.5	−0.59	−0.45	0.82	1.06	0.95	0.89	0.84	0.95
I26	−34.2	−14	−1.09	−18.54	−2.43	−7.08	−5.16	−1.53	−1.02	−1.58	1	1.06	1.14	1.68	1.02	1.14
N31	−41.81	−27.23	−2.09	−16.57	−4.02	−3.35	−6.08	−5.71	−3.43	−2.77	0.97	1.03	1.17	1.67	0.97	1.15
H14	−51.02	−29.87	−1.52	−22.46	−4.89	−4.21	−7.72	−4.53	−2.71	−1.49	3.51	2.2	1.14	2.9	1.15	1.48
M6	−1.09	−0.41	−0.79	−0.66	−0.58	−0.6	−0.5	−0.39	−0.51	−0.33	1.14	5.48	3.77	17.26	18.24	3.59
L21	−1.27	−0.43	−0.84	−0.67	−0.6	−0.66	−0.53	−0.25	−0.45	−0.33	0.99	1.96	1.26	2.59	2.76	1.3
E20	−0.51	−0.55	−0.81	−0.4	−0.95	−0.65	−0.39	−0.24	−0.44	−0.33	5.88	3.8	1.13	3.71	1.34	1.42
B13	−3.63	−2.31	−0.79	−1.77	−1.05	−0.86	−0.83	−0.63	−0.62	−0.41	1.02	1.11	1.22	1.8	1.05	1.23
B8	−8.44	−4.91	−0.86	−3.96	−1.43	−1.34	−1.53	−0.82	−0.74	−0.54	1	1.06	1.15	1.69	1.02	1.16
B27	−9.83	−5.53	−0.87	−4.67	−1.52	−1.54	−1.73	−0.86	−0.76	−0.58	0.99	1.06	1.15	1.68	1.02	1.15
B29	−4.11	−2.63	−0.79	−1.96	−1.11	−0.86	−0.91	−0.62	−0.63	−0.42	1.02	1.1	1.21	1.78	1.04	1.22
C12	−3.63	−2.4	−0.78	−1.72	−1.07	−0.79	−0.83	−0.59	−0.61	−0.41	1.02	1.11	1.23	1.81	1.05	1.24
C16	−5.58	−3.47	−0.82	−2.6	−1.23	−0.99	−1.12	−0.7	−0.67	−0.46	1	1.08	1.18	1.73	1.03	1.19
L22	−0.95	−0.74	−1.09	−2.48	−3.08	−1.12	−0.47	−0.79	−0.74	−0.43	1.13	1.94	1.23	4.69	1.04	1.25
B3	−5.84	−3.58	−0.83	−2.75	−1.25	−1.03	−1.16	−0.74	−0.69	−0.47	1	1.08	1.18	1.73	1.03	1.19
E24	−0.51	−0.55	−0.81	−0.4	−0.95	−0.65	−0.39	−0.24	−0.44	−0.33	5.88	3.8	1.13	3.71	1.34	1.42
B28	−4.15	−2.66	−0.79	−1.98	−1.12	−0.86	−0.91	−0.62	−0.62	−0.42	1.02	1.1	1.21	1.78	1.04	1.21
B7	−1.08	−0.82	−1.03	−0.93	−1.39	−0.78	−0.49	−0.41	−0.53	−0.36	1.04	1.1	1.05	0.91	0.88	1.03

Table 2. (Continued)

	NSDL7	NSDL8	NSDL9	NSDL10	VDWVOL	SURF_A	MOFI_X	MOFI_Y	MOFI_Z	PEAX_X	PEAX_Y	PEAX_Z	MOL_WT	S8_1DX	S8_1DY	S8_1DZ	S8_1CX
K17	1.26	1.46	0.49	0.68	250.4	314.23	574.2	9189.65	8711.2	12.81	0.95	3.16	264.33	6.9	9.71	5.36	5.24
D30	1.09	1.88	0.6	1.04	307.8	364.03	1837.91	15123.22	13701.33	13.61	1.69	4.73	364.36	9.43	10.58	6.82	6.48
J19	1.12	1.44	0.48	0.67	396.2	506.68	793.23	40830.82	39243.73	22.54	2.26	3.91	390.57	10.82	18.74	6.98	7.18
A5	0.37	1.9	0.6	0.94	342.3	404.43	3099.53	21450.85	18866.43	14.66	1.72	5.73	436.31	11.77	10.66	5.53	8.54
J1	1.28	1.46	0.49	0.68	380	488.07	766.61	37378.3	36572.98	22.06	0.51	3.23	376.54	17.09	13.62	7.08	11.2
K18	1.29	1.46	0.49	0.68	282.8	358.11	623.6	13796.1	13431.95	15.09	1.49	2.91	292.38	7.96	11.89	5.95	5.75
G2	0.31	1.65	0.51	0.85	328	386.59	1686.44	21047.63	19971.37	15.41	1.92	4.08	417.25	11.8	9.96	6.26	8.53
L25	0.36	2.11	0.76	0.85	405	520.97	808.41	49091.14	46927.39	24.21	2.89	4.28	406.57	11.52	21.11	7.54	33.79
A10	0.38	2.05	0.65	0.97	353.6	413.68	3400.44	22357.97	19442.15	14.62	1.64	5.93	452.31	11.73	10.64	5.6	8.51
C11	0.37	2.04	0.64	0.96	304.8	355.8	1894.86	15368.85	13607.31	13.68	0.96	5.03	362.34	11.79	9.56	4.83	8.52
D23	1.17	1.82	0.58	1.06	280.4	330.79	1422.58	10025.04	8799.78	11.77	1.25	4.59	314.34	9.26	8.39	5.28	6.34
F15	0.37	2.24	0.72	1.05	353.6	411.79	1993.35	23754.71	22894.08	15.77	2.51	3.99	452.31	11.69	9.61	6.44	8.41
G4	0.33	1.71	0.54	0.86	328.8	395.37	2776.7	19569.54	17354.98	14.31	1.84	5.47	420.31	11.76	10.1	6.22	8.52
G9	0.32	1.69	0.53	0.86	328.8	396.11	1822.81	20390.84	19440.97	15.1	2.29	4.08	420.31	11.78	10	6.26	8.53
I26	0.37	2.01	0.63	0.97	328.3	380.38	3064.93	19326.95	16851.72	14.12	1.88	5.78	418.24	8.62	11.7	7.88	6
N31	0.38	2.22	0.8	1.05	216.2	259.56	1631.94	5819.5	4289.13	8.76	0.96	5.35	277.66	5.61	7.78	4.45	3.87
H14	0.73	2.1	0.67	1.05	328.5	394.02	2529.68	19250	17192.37	14.33	1.69	5.27	414.8	12.95	10.08	6.41	9.12
M6	0.38	5.78	2.47	2.16	320.6	378.27	1726.75	20579.16	19403.1	15.16	1.82	4.18	419.22	11.8	9.96	6.27	8.53
L21	0.36	2.12	0.76	0.85	372.6	477.43	765.72	37097.85	36119.48	21.89	1.19	3.4	378.51	10.76	18.89	7.27	30.98
E20	1.06	2.19	0.8	0.86	340.2	436.41	1650.21	25060.27	23662.13	18.34	1.34	4.67	350.46	10.03	16.48	7.11	6.76
B13	0.38	2.4	0.76	1.1	332	392.1	2622.25	1989.53	20498.27	15.26	2.57	4.9	429.77	12.5	9.28	8.44	8.69
B8	0.37	2.12	0.66	1	321.6	386.11	2267.29	22328.79	20831.54	15.7	2.15	4.76	418.33	12.71	9.9	7.28	8.82
B27	0.37	2.08	0.65	0.99	321.6	385.27	1846.98	22716.93	21426.62	15.96	1.83	4.35	418.33	11.12	11.96	6.8	7.15
B29	0.37	2.39	0.75	1.1	336.1	401.83	2447.47	23459.07	21598.62	15.42	1.81	4.9	452.77	11.13	11.67	6.23	7.14
C12	0.38	2.48	0.8	1.11	333.8	385.53	3188.15	20791.55	17782.86	14.38	1.02	6.02	431.23	11.82	11.62	4.71	8.54
C16	0.37	2.25	0.71	1.04	332.3	391.06	3615.87	20164.71	17081.31	13.95	1.76	6.23	435.28	10.18	11.91	6.95	7.37
L22	0.39	2.03	0.66	0.85	387.1	493.84	1628.62	39861.4	38461.7	21.6	1.18	4.29	412.96	10.78	18.95	7.22	313.32
B3	0.37	2.27	0.71	1.06	320.6	379.91	2651.73	19371.61	17267.4	14.29	1.81	5.35	419.22	11.77	10.01	6.23	8.53
E24	1.06	2.19	0.8	0.86	340.2	436.41	1650.21	25060.27	23662.13	18.34	1.34	4.67	350.46	10.03	16.48	7.11	6.76
B28	0.37	2.37	0.75	1.09	354	421.54	3010.16	28906.99	26566.66	16.53	1.87	5.28	484.83	10.57	13.45	6.9	6.86
B7	0.35	1.92	0.61	0.94	313.8	373.31	2345.07	17645.04	15850.81	14.02	1.86	5.11	399.23	11.78	9.97	6.26	8.53

Table 2. (Continued)

	S8_1CY	S8_1CZ	LOGP	M_PNT	SUM_F	SUM_R
K17	−5.06	−0.57	3.01	62	0.25	−0.23
D30	−5.42	−0.91	3.69	178	0.67	0.16
J19	−9.58	−1.5	7.23	71	0.28	−0.26
A5	−1.62	−1.14	5.73	165	0.52	0.01
J1	−4.96	−1.57	7.24	81	0.25	−0.23
K18	−6.15	−0.93	4.07	78	0.25	−0.23
G2	−1.27	−1.23	5.96	183	0.25	−0.23
L25	18.3	0	9.52	79	0.67	0.16
A10	−1.67	−1.42	5.67	195	0.54	0.22
C11	−2.02	1.25	4.89	212	0.67	−0.23
D23	−4.37	1.2	5.35	227	0.67	0.16
F15	−1.29	−1.73	5.68	178	0.54	0.23
G4	−1.44	−1.03	7.37	143	0.2	−0.18
G9	−1.34	−1.22	7.37	151	0.2	−0.18
I26	−5.75	−1.62	6.81	173	0.67	0.16
N31	−3.8	0.22	4.65	170	0.67	0.16
H14	−1.77	−1.23	6.18	159	0.67	−0.23
M6	−1.27	−1.24	6.99	192	0.67	0.16
L21	16.46	−1.67	8.47	67	0.67	0.16
E20	−8.46	−1.46	8.47	90	0.67	0.16
B13	−1.57	−1.17	6.11	208	0.67	−0.23
B8	−1.92	−1.58	6.7	199	0.67	0.16
B27	−5.93	−1.2	6.7	176	0.67	0.16
B29	−5.79	−1.28	7.27	192	0.67	0.16
C12	−2.26	1.2	6.2	246	0.67	−0.23
C16	−2.47	2.28	6.84	222	0.67	−0.23
L22	16.5	−1.63	9.3	81	1.08	0.01
B3	−1.36	−1.28	6.99	207	0.67	0.16
E24	−8.46	−1.46	7.41	85	0.67	0.16
B28	−6.68	−1.29	7.87	195	0.67	0.16
B7	−1.32	−1.28	6.76	256	0.51	0.19

Table 3. Comparison of Variables Selected by Different Algorithms[a]

	NLM/SR	NN	CSA	GFA	MUSEUM	FRED
ATCH1				X	X	X
ATCH2	X	X	X			
ATCH3					X	X
ATCH4		X[b]	X[b]	X	X[b]	X[c]
ATCH5			X[b]	X	X[b]	
ATCH6				X	X	
ATCH7						X
DIPV_X		X	X[b]		X[b]	
DIPV_Y	X					
DIPV_Z	X					
ESDL3				X	X[b]	X[c]
ESDL5	X					
ESDL8						X
ESDL10	X					
NSDL2	X					
VDWVOL		X	X		X[b]	X[c]
SURF_A				X	X[b]	X
MOFI_X		X[b]				X[c]
MOFI_Y		X		X	X[b]	
MOFI_Z					X[b]	X
PEAX_X				X	X[b]	X
PEAX_Y		X				
LOGP	X	X[b]	X	X	X[b]	X[c]
M_PNT	X		X		X	
SUM_F				X	X[b]	X[c]
SUM_R	X					

[a] NLM/SR, nonlinear mapping/stepwise regression; NN, neural network; CSA, cluster significance analysis; GFA, genetic function approximation; MUSEUM, mutation and selection uncover models. [b] Primary descriptors selected by algorithm. [c] Descriptors selected in six-variable model (progeny factor = 40, kill factor = 5, minimum σ = 0.025).

analyses of the Selwood data set. The lower panel, once again, indicates the efficiency of the sampling algorithm. The final results of these three analyses were identical in that each analysis selected ATCH_4, ESDL_3, VDWVOL, LOGP, and SUM_F as the primary determinants of biological activity. In Table 3, the variables identified in the three analyses are signified by footnote *b*. These offspring models express a q^2 fitness value of 0.683 with a corresponding r^2 value of 0.829. The total number of generations required for each trial run to reach this solution was 32, 33, and 20 with a total number of models generated per trial run of 5240, 4880, and 3400, respectively.

Algorithm Performance (Timing). Since the algorithm is implemented in SPL (an interpreted language), a large amount of input/output (I/O) traffic is required. This effectively makes the algorithm much less efficient than an algorithm written purely in a compiled language (i.e., C). However, the SPL implementation does allow the algorithm to take advantage of the statistical routines (i.e., cross-validation, PLS) of Sybyl as well as the expression generators and variables that are unique to the Sybyl program. Given all this, a typical run time on an Indigo II (R10000) is less than 15 min for the six-variable models described above. This time increases linearly with the total number of models to be generated (as a function of the progeny factor and the number of descriptors), as the rate-limiting factor is the amount of time required to perform a single cross-validated analysis on the data set. In addition, program status and diagnostics are displayed in real time and written to a log file.

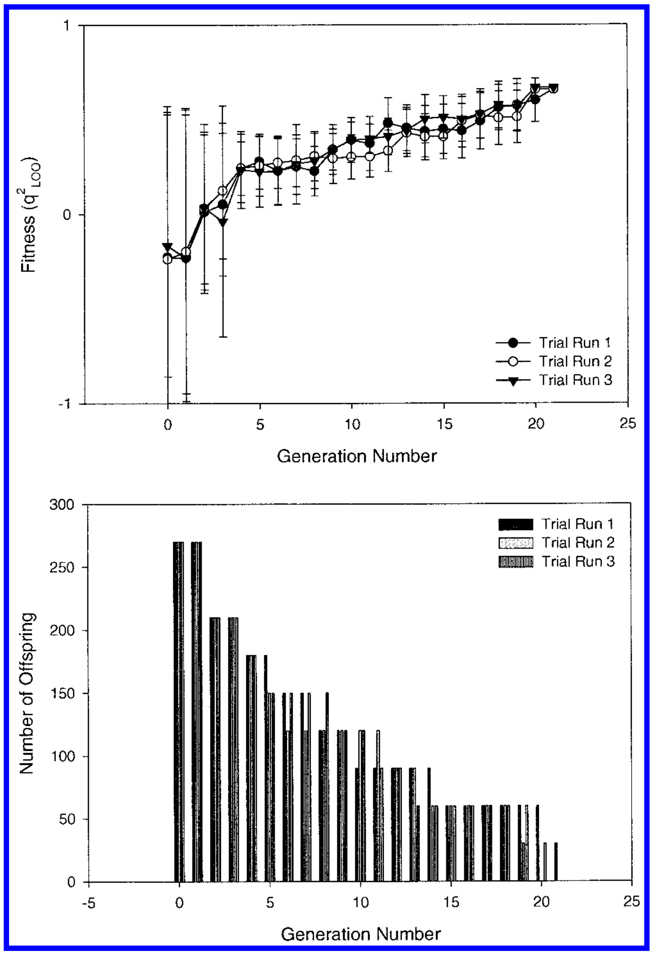


Figure 1. Evolution of three-variable FRED models.

Algorithm Validation (Randomized Data Trials). To validate that the algorithm was actually capable of identifying optimal solutions to problems rather than just spuriously happening upon them, the actual biological data for the Selwood data set were randomized (see Table 2) and three trial runs were performed. The algorithm was asked to find the best six-variable model using the same parameters as the above six-variable analyses (i.e., progeny factor = 40, kill factor = 5, minimum σ = 0.025). These trial runs required 26, 20, and 24 generations to converge and resulted in average q^2 fitness values of 0.244, 0.108, and 0.193 with r^2 values of 0.416, 0.246, and 0.400, respectively. That these trial runs did not converge to the same model and descriptor set is not unexpected due to the low signal-to-noise ratio inherent in the random data set. It is possible that a greater progeny factor is required in order to force the algorithm to converge to a single solution. Regardless, implementation of the algorithm on the randomized data set using the same parameters as the six-variable analyses on the original data set resulted in the discovery of statistically insignificant and inferior models.

In the upper panel of Figure 3, the average fitness value per generation is plotted for each of the trial runs on the randomized data set. In comparison with the upper panels of Figures 1 and 2, the average fitness values for early generations of models were comparable while the standard deviation values for models developed from the randomized data were smaller than those for the trial runs on the original data set. The larger standard deviation of the latter examples

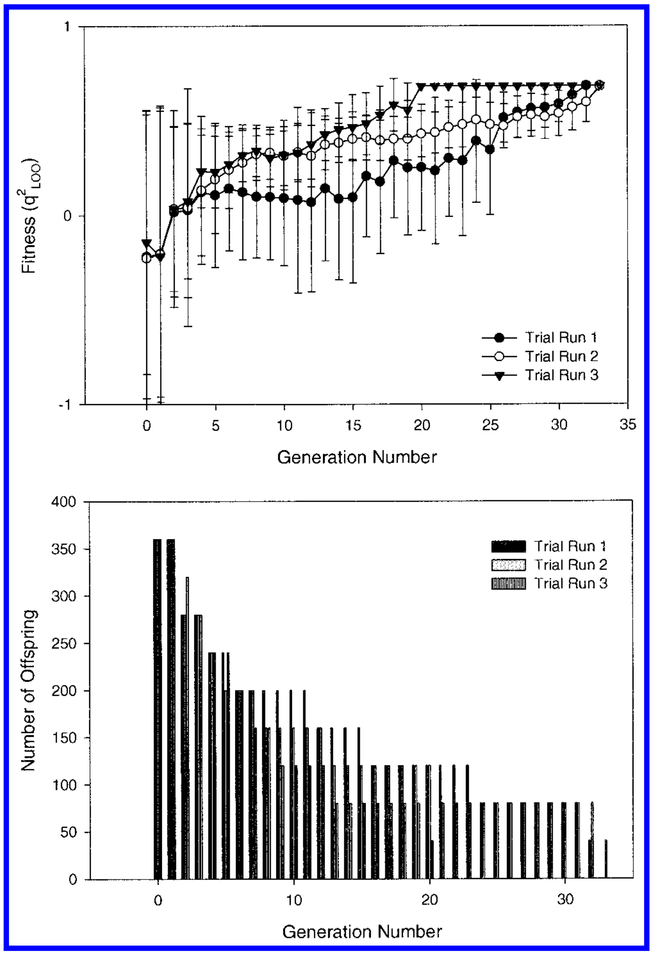


Figure 2. Evolution of six-variable FRED models.

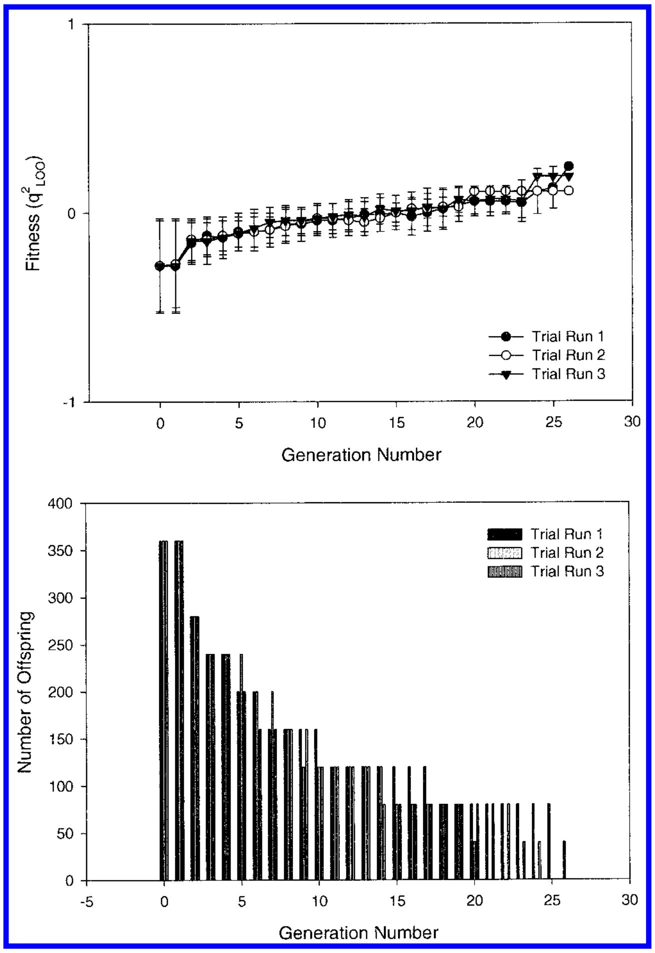


Figure 3. Evolution of six-variable FRED models using randomized biological data.

is indicative of the existence and discovery of very fit model offspring in the early generations. As the algorithm continues into the latter generations, the average fitness of both the three-variable and six-variable analyses on the original data set (Figures 1 and 2) increases dramatically, while the analysis on the randomized data set reveals only a slight increase in overall fitness.

Algorithm Advantages. The single largest problem with optimization algorithms, in general, is the chance of finding local optimal, or minimal, solutions. Monte Carlo-based algorithms, such evolutionary algorithms, provide a means to disrupt the process of settling on local solutions by randomizing the state of system, thus putting it either closer to, or farther from, the global optimum state. Purely genetic algorithms, such as GFA and MUSEUM, require that individuals (QSAR models in these cases) be crossed-over with other individuals, similar to reproduction, or mutated. In crossing-over, portions of the genetic makeup of one individual are combined with portions of the generic makeup of another (one or more) individual, thus creating a hybrid. In mutation, one or more of the individual genes of an individual is randomly changed to something else. These techniques promise to disrupt the state of the system with the result ideally being a state closer to the global solution.

In the FRED algorithm, these techniques are modified somewhat in that while individuals are assessed for fitness, it is not the entire individual that is allowed to pass to the next generation. In fact, not even intact portions of the individuals are allowed to pass. Instead, only the genes (traits) that have proven not to be detrimental to the fitness of the individual are allowed to pass to the next generation. The remaining genes (traits) are then randomly reassembled into individuals whose fitness is judged. The FRED algorithm ensures that those genes (traits) that contribute to poorly fit individuals are culled from the available gene pool, unlike the competing algorithms that would not necessarily identify these traits due to the ability of a potentially detrimental gene (trait) to pass to the next generation when assessed in combination with a particularly beneficial gene (trait). In this manner, the algorithm converges rapidly to optimal, statistically indifferent, models following a few generations of culling of traits (Figure 1).

DISCUSSION

In the publication in which this data set was first introduced,[13] the authors' intent was to develop QSAR models for the prediction of biological activity for untested compounds. To this end, the authors split the 31 compound data set into a training set composed of 16 molecules while the balance of the data set made up an external test set. The process implemented to reduce the 53 descriptor set to an allowable set of three descriptors proceeded in a very logical manner with the creation of a Pearson's correlation matrix. The removal of highly interrelated ($r > 0.75$) descriptors reduced the data set to 23 parameters. These remaining descriptors were then analyzed by a stepwise regression technique to identify those most determinant of biological activity. The first 10 descriptors selected by this technique were M_PNT, LOGP, ESDL10, DIPV_Z, ESDL5, SUM_R, S8_ICZ, DIPV_Y, NSDL2, and ATCH2, listed in order of selection. These results are presented here primarily for historical purposes since the results of this analysis are not directly comparable to analyses on the full data set.

Wikel and Dow[9] presented the first results of an analysis in which the entire data set was used as the training set. In this study, a neural network was utilized to identify the top nine candidate descriptors (ATCH2, ATCH4, DIPV_X, VDWVOL, MOFI_X, MOFI_Y, LOGP, and M_PNT). These descriptors were then subjected to multiple regression analysis (MRA) to identify the primary three determinants (ATCH4, MOFI_X, and LOGP). These results were quickly followed by those of McFarland and Gans.[4] In this latter study, the authors utilized a sequential implementation of a random cluster significance analysis (CSA). In a manner similar to that of Selwood, the highly correlated ($r > 0.7$ in this case) descriptors were eliminated. Unlike Selwood's approach, prior to this step, each descriptor was evaluated for its contribution to model based on its p-value when used as the sole descriptor. The p-values were then used to rank-sort (lowest to highest) prior to a stepwise variable elimination routine in which the codependent variable with the lowest p-value was retained. Additionally, descriptors with high p-values ($p > 0.3$) were eliminated. The biological activities of the compounds were also transformed from a continuous response variable to a binary variable where the first 15 compounds (ranked according to potency) were assigned to the active group while the remaining 16 were deemed inactive. CSA then identified six candidate descriptors (ATCH2, ATCH4, ATCH5, DIPV_X, VDWVOL, and S8_IDX) that were subsequently subjected to stepwise MRA to reduce the set to three primary determinants (ATCH4, ATCH5, and DIPV_X).

More recently, Rogers et al.[7,8] applied a genetic function approximation (GFA) to this multiple minima problem. In this implementation, a population of random models with random numbers of descriptors was generated. The models were then allowed to reproduce (cross over) and mutate. The offspring were evaluated with a fitness function, and in subsequent generations the fitness function served to weight the chances of crossover (i.e., more fit models reproduced more often). After a given number of iterations, the fitness function was used to select to "best models". Rogers identified a population of models in which the descriptors ATCH1, ATCH4, ATCH5, ATCH6, ESDL3, SURF_A, MOFI_Y, PEAX_X, LOGP, and SUM_F were represented. An extension of the GFA technique has been proposed by Kubinyi[5,6] as the MUSEUM (mutation and selection uncover models) algorithm. This technique differs from GFA in that only the fittest model survives from one generation to the next and, as the name implies, is mutated (addition or elimination of descriptors) numerous times. The MUSEUM algorithm identified the variables ATCH4, ATCH5, DIPV_X, ESDL3, VDWVOL, SURF_A, MOFI_Y, MOFI_Z, PEAX_X, LOGP, and SUM_F in the top 10 models from an analysis in which only three variable models were produced. The variables ATCH1, ATCH3, ATCH6, ATCH7, and M_PNT were additionally selected if the top 25 models were considered.

The FRED algorithm borrows characteristics from the techniques discussed above; however, certain fundamental differences exist. Like the above techniques, the FRED program is primarily an evolutionary algorithm in the sense that the descriptors represent traits, a fitness is assessed, and several generations are produced in the course of the analysis. In a manner similar to the generation of the first set of offspring models in the GFA approach, a large number of randomly generated models are produced. This is distinctly different from the MUSEUM approach in which only one model is produced and then mutated to produce a generation of offspring. The primary difference between the FRED algorithm and the genetic-type evolutionary algorithms is in the philosophy that is implemented to eliminate or select variables. In the other evolutionary algorithms, either a population of fit models or a singular fit individual model are allowed to pass on to subsequent generations. In the GFA implementation, the probability of reproduction is set proportional to the fitness from the previous generation. In this implementation, an elimination routine simply identifies the undesirable traits present in the less fit models and eliminates them from the available descriptor pool. In the FRED algorithm, weighting functions are not utilized and all remaining (nontaboo) descriptors are given an equal opportunity to contribute to offspring models in subsequent generations. In this manner it is more difficult for poorer descriptors to be masked by exceptionally good descriptors and be passed on to subsequent generations, as they are eliminated from the allowable descriptor pool.

It is interesting to note the differences between the variables selected by the previously published algorithms. The extremely high degree of interrelated variables makes this a data set an ideal choice for variable reduction algorithm validation. These relationships have been noted in previous publications and have served as the primary basis for any disagreements between the results. Given this, it is clear that certain variables, including LOGP, SUM_F, and others identified in Table 3, are generally selected by all algorithms; therefore, any novel technique must be capable of identifying these variables at a minimum. The results from the FRED analysis of the Selwood data set are consistent with the results of the genetic and evolutionary algorithms. These results further indicate that genetic algorithm-derived variable reduction routines can be utilized on data sets containing highly interrelated variables without the need for preprocessing the data by variable elimination based on pairwise correlation analyses.

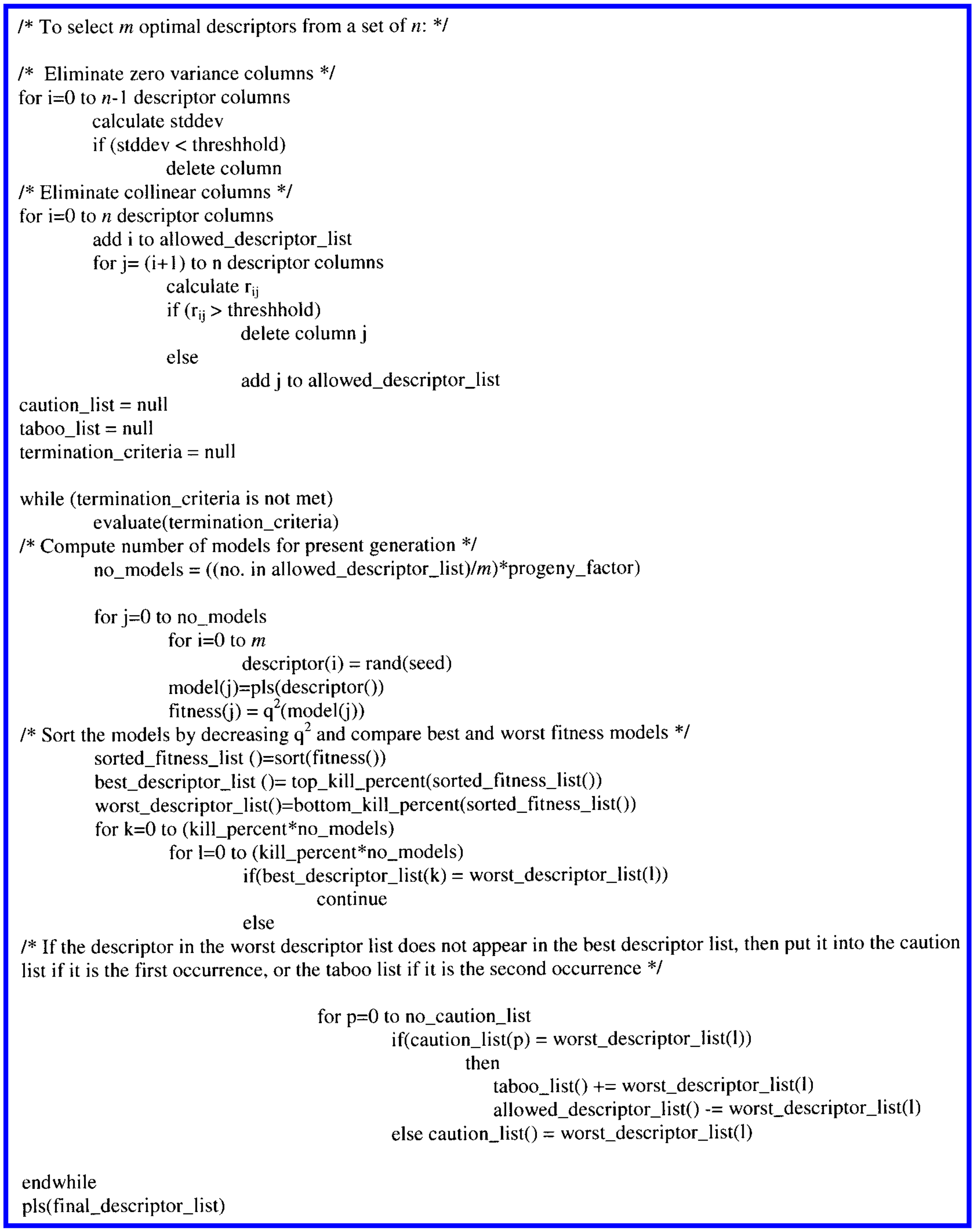
CONCLUSION

A novel technique for variable selection with potential application to QSAR analysis has been developed and validated against several existing algorithms. In the present study, the algorithm was demonstrated to be capable of rapidly identifying the *best* determinants of biological activity from a multidimensional data matrix of physicochemical parameters.

ACKNOWLEDGMENT

We gratefully acknowledge Dr. Hugo Kubinyi for "lively" e-mail discussions and for providing the Selwood, and other, sample data sets.

APPENDIX 1. PSEUDOCODE FOR FRED ALGORITHM



```
/* To select $m$ optimal descriptors from a set of $n$: */

/*  Eliminate zero variance columns */
for i=0 to $n$-1 descriptor columns
        calculate stddev
        if (stddev < threshhold)
                delete column
/* Eliminate collinear columns */
for i=0 to $n$ descriptor columns
        add i to allowed_descriptor_list
        for j= (i+1) to n descriptor columns
                calculate $r_{ij}$
                if ($r_{ij}$ > threshhold)
                        delete column j
                else
                        add j to allowed_descriptor_list
caution_list = null
taboo_list = null
termination_criteria = null

while (termination_criteria is not met)
        evaluate(termination_criteria)
/* Compute number of models for present generation */
        no_models = ((no. in allowed_descriptor_list)/$m$)*progeny_factor)

        for j=0 to no_models
                for i=0 to $m$
                        descriptor(i) = rand(seed)
                model(j)=pls(descriptor())
                fitness(j) = $q^2$(model(j))
/* Sort the models by decreasing $q^2$ and compare best and worst fitness models */
        sorted_fitness_list ()=sort(fitness())
        best_descriptor_list ()= top_kill_percent(sorted_fitness_list())
        worst_descriptor_list()=bottom_kill_percent(sorted_fitness_list())
        for k=0 to (kill_percent*no_models)
                for l=0 to (kill_percent*no_models)
                        if(best_descriptor_list(k) = worst_descriptor_list(l))
                                continue
                        else
/* If the descriptor in the worst descriptor list does not appear in the best descriptor list, then put it into the caution
list if it is the first occurrence, or the taboo list if it is the second occurrence */

                                for p=0 to no_caution_list
                                        if(caution_list(p) = worst_descriptor_list(l))
                                                then
                                                   taboo_list() += worst_descriptor_list(l)
                                                   allowed_descriptor_list() -= worst_descriptor_list(l)
                                        else caution_list() = worst_descriptor_list(l)

endwhile
pls(final_descriptor_list)
```

REFERENCES AND NOTES

(1) Hansch, C. *Drug Dev. Res.* **1981**, *1*, 267−309.
(2) Wold, S.; Ruhe, A.; Wold, H.; Dunn, W. J. *SIAM J. Sci. Stat. Comput.* **1984**, *5*, 735−743.
(3) Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.
(4) McFarland, J. W.; Gans, D. J. *Quant. Struct.-Act. Relat.* **1994**, *13*, 11−17.
(5) Kubinyi, H. *Quant. Struct.-Act. Relat.* **1994**, *13*, 393−401.
(6) Kubinyi, H. *Quant. Struct.-Act. Relat.* **1994**, *13*, 285−294.
(7) Rogers, D. *Genetic Function Approximation: A Genetic Approach to Building Quantitative Structure−Activity Relationship Models*; Rogers, D., Ed.: Prous, Barcelo, Spain, 1994.
(8) Rogers, D.; Hopfinger, A. J. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854−866.
(9) Wikel, J.; Dow, E. *Bioorg. Med. Chem. Lett.* **1993**, *3*, 645−651.
(10) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; Research Studies Press Ltd.: Letchworth, Hertfordshire, England, 1986.
(11) *Unity Chemical Information Software*; Version 2.3; Tripos, Inc.: St. Louis, MO, 1998.
(12) *ISIS*, Version 2.1; MDL Information Systems, Inc.: San Leandro, CA, 1998.
(13) Selwood, D. L.; Livingstone, D. J.; Comley, J. C. W.; O'Dowd, A. B.; Hudson, A. B.; Jackson, P.; Jandu, K. S.; Rose, V. S.; Stables, J. N. *J. Med. Chem.* **1990**, *33*, 136−142.
(14) Sybyl; Version 6.4; Tripos, Inc.: St. Louis, MO, 1998.
(15) Topliss, J. G.; Edwards, R. P. *J. Med. Chem.* **1979**, *22*, 1238−1244.
(16) Cvijovic, D.; Klinowski, J. *Science* **1995**, *267*, 664−666.

CI980405R