Mining the NCI Anticancer Drug Discovery Databases: Genetic Function Approximation for the QSAR Study of Anticancer Ellipticine Analogues

Leming M. Shi,[†] Yi Fan,[‡,⊥] Timothy G. Myers,[§] Patrick M. O'Connor,[‡] Kenneth D. Paull,[§] Stephen H. Friend,[¶] and John N. Weinstein*[,‡]

Laboratory of Molecular Pharmacology, Division of Basic Sciences, and Information Technology Branch, Division of Cancer Treatment, Diagnosis, and Centers, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, and Fred Hutchinson Cancer Research Center/NCI, 1124 Columbia St., Seattle, Washington 98104

Received September 30, 1997

The U.S. National Cancer Institute (NCI) conducts a drug discovery program in which ∼10 000 compounds are screened every year in vitro against a panel of 60 human cancer cell lines from different organs of origin. Since 1990, ∼63 000 compounds have been tested, and their patterns of activity profiled. Recently, we analyzed the antitumor activity patterns of 112 ellipticine analogues using a hierarchical clustering algorithm. Dramatic coherence between molecular structures and activity patterns was observed qualitatively from the cluster tree. In the present study, we further investigate the quantitative structure−activity relationships (QSAR) of these compounds, in particular with respect to the influence of p53-status and the CNS cell selectivity of the activity patterns. Independent variables (i.e., chemical structural descriptors of the ellipticine analogues) were calculated from the Cerius^2 molecular modeling package. Important structural descriptors, including partial atomic charges on the ellipticine ring-forming atoms, were identified by the recently developed genetic function approximation (GFA) method. For our data set, the GFA method gave better correlation and cross-validation results (R^2 and CVR^2 were usually ∼0.3 higher) than did classical stepwise linear regression. A procedure for improving the performance of GFA is proposed, and the relative advantages and disadvantages of using GFA for QSAR studies are discussed.

INTRODUCTION

The U.S. National Cancer Institute (NCI) conducts an anticancer drug discovery program in which ∼10 000 compounds are screened every year in vitro against a panel of 60 different human cancer cell lines.[1−5] Currently included in the screen are eight melanomas, six leukemias, and eight cancers of breast, two of prostate, nine of lung, seven of colon, six of ovary, eight of kidney, and six of central nervous system (CNS) origin. The purpose of the screen is to provide the initial evaluation of compounds for cytotoxic or growth inhibitory activity against a diverse panel of cancer types. Compounds that show interesting activity patterns in the in vitro screen are selected for additional studies in vitro and in vivo. We can think of this screen as a tool for profiling or "fingerprinting" the tested compounds in terms of their anticancer activity patterns. The activity pattern of a compound is represented by a vector of 60 growth inhibitory activity values, one for each cell line.

Since 1990, ∼63 000 synthetic compounds, plus a larger number of natural product extracts, have been tested. Combinatorial libraries have also been assessed recently. Similarity in activity patterns very often indicates similarity in mechanism of action, mode of drug resistance, and molecular structure of tested compounds.[4−8] Several different algorithms have been introduced to use the activity information for discovery of anticancer drugs and for understanding of the molecular pharmacology of cancer. The COMPARE program[4,6,9,10] has proved very useful for finding agents with activity patterns similar to that of a "seed" compound and for finding compounds with activity patterns that correlate well (positively or negatively) across the 60 cell lines with the expression levels of particular cellular targets. Back-propagation neural networks,[7] Kohonen self-organizing maps,[11] and principal component analysis[12] have been used to predict mechanism of action or to organize compounds into families based on activity patterns. This "information-intensive" approach to the molecular pharmacology of cancer and anticancer drug discovery[7,8,13] has proved useful in identifying subgroups of compounds related to particular biological targets. Growth inhibitory activity for a single cell line is not informative, but activity patterns across the 60 cell lines provide incisive information on the mechanisms of action of screened compounds and also on molecular targets and modulators of activity within the cancer cells.

Our approach to the discovery of anticancer drugs and to the molecular pharmacology of cancer involves three kinds of databases:[8,13] (i) anticancer activity data (*A*) for compounds across the 60 human tumor cell lines; (ii) chemical structure information (*S*) for the tested compounds; and (iii) informa-

* Author to whom correspondence should be addressed at Building 37, Room 5D02, National Cancer Institute, NIH, 9000 Rockville Pike, Bethesda, MD 20892. Tel: (301) 496-9571. Fax: (301) 402-0752. E-mail: weinstein@dtpax2.ncifcrf.gov.

[†] Current address: R. O. W. Sciences, Inc., National Center for Toxicological Research, Mail Code-910, 3900 NCTR Road, Jefferson, AR 72079. E-mail: lshi@nctr.fda.gov.

[‡] Laboratory of Molecular Pharmacology.

[§] Information Technology Branch.

[¶] Fred Hutchinson Cancer Research Center/NCI.

[⊥] Current address: Shionogi Bioresearch Corp., 45 Hartwell Avenue, Lexington, MA 02173. E-mail: yfan@sbrco.com.

S0095-2338(97)00085-1
Published on Web 01/30/1998

tion on possible targets or modulators (*T*) of activity in the 60 cell lines.[14−16] Currently, the size of database *A* is ∼63 000 by 60. The chemical structure (*S*) database can be encoded in terms of any set of 2- (2D) or 3-dimensional (3D) molecular structural descriptors or experimentally measured or theoretically calculated physicochemical properties. The NCI Drug Information System (DIS),[17−20] a major resource for drug discovery, contains structural information for nearly 500 000 molecules, including the 63 000 tested compounds. The NCI DIS 2D database was successfully used to search for multidrug resistance (MDR) reversal agents.[21] In that method, the 2D queries (biophores and biophobes) were automatically generated from the CASE[22] and MULTICASE[23] programs developed by G. Klopman at Case Western Reserve University. The NCI DIS 3D database[24] has been successfully used for the identification of inhibitors for protein kinase C,[25] HIV-1 protease,[26] and HIV-1 integrase.[27,28]

For the analysis and display of these large databases, we have developed the DISCOVERY program set, which maps coherent patterns in the data rather than treating the compounds and targets one pair at a time.[13,29,30]

One cell screen target that we have been studying is the p53 tumor suppressor gene. The p53 gene functions as a transcriptional regulator with the ability to both transactivate and suppress gene transcription. It is activated in response to DNA damage and can orchestrate a number of cellular responses to genotoxic stress, including G1 arrest and apoptosis.[31] The p53 gene is mutated in >50% of human tumors, more than any other gene examined to date.[31,32] The sequence of p53 has been determined for the 60 NCI cell lines.[33] Nineteen of them are p53 wild-type, and 41 are p53 mutant.

Most of the standard clinical anticancer agents are more active on average against p53 wild-type cells than against the p53 mutant ones in the 2-day assay of the NCI screen.[8,33,34] It seemed desirable, therefore, to search for compounds that were more active against p53 mutant cell lines. We termed such agents "p53-inverse". To search for such compounds, we analyzed the database of activity patterns using the COMPARE and DISCOVERY program sets. As part of this process, cluster analysis led to the identification of 1057 agents belonging to 37 cluster families that appeared predominantly p53-inverse (Myers et al., unpublished results). One interesting group of such compounds was the ellipticine family. We have recently analyzed in more detail the activity patterns of 112 ellipticine analogues using a hierarchical clustering method and have observed dramatic coherence between molecular structures and activity patterns.[35]

In the present study, we investigate in detail the quantitative structure−activity relationship (QSAR) of the ellipticine data set in hopes of finding more potent "p53-inverse" agents. We also address the advantages and disadvantages of using genetic function approximation (GFA) for QSAR study and the importance of data quality and "homogeneity" for obtaining meaningful QSAR models.

THE ELLIPTICINE DATA SET

Ellipticine (5,11-dimethyl-6*H*-pyrido [4,3-b]carbazole; Figure 1), is one of the simplest naturally occurring alkaloids

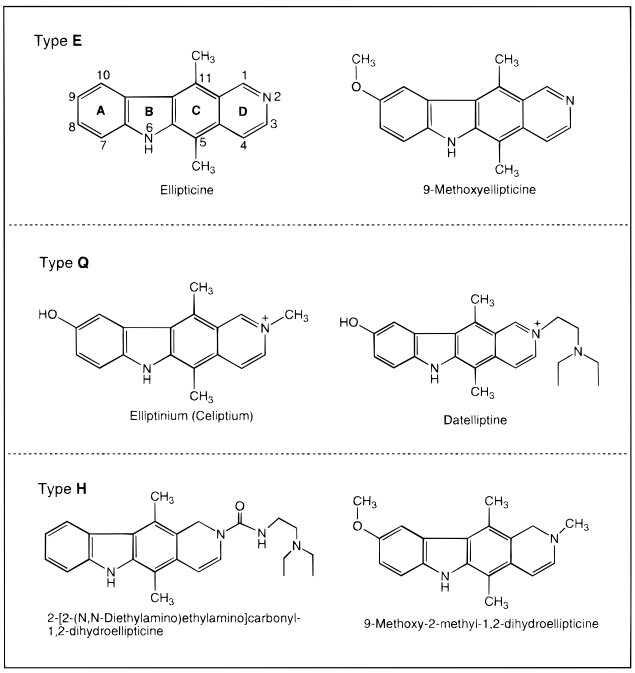

Figure 1. The ellipticine analogues can be classified on the basis of chemical structure into three subgroups: the normal ellipticines (*E*), N^2-substituted ellipticiniums (*Q*), and 1,2-dihydrogenated ellipticines (*H*). Modified from ref 35.

with a planar structure. It was first isolated in 1959 from the leaves of the evergreen tree *Ochrosia elliptica* Labill (Apocynaceae), which grows wild in Oceania,[36] but its biological activities were not recognized then. In 1967, the synthesis and anticancer activity of ellipticine, 9-methoxyellipticine, and other derivatives were reported.[37] Since then, the design, synthesis, and structure−activity relationships of this class of compounds have been studied by a number of laboratories.[38−44] Several publications have reported that some ellipticine analogues, specifically ellipticiniums, are selectively active against cancer cell lines of CNS origin,[39−41] and Boyd and colleagues[45,46] found that cellular uptake was a major factor in this selectivity.

Studies on the mechanism of cytotoxicity and anticancer activity of the ellipticine analogues suggest a complex set of effects,[38,47−49] including: (i) DNA intercalation; (ii) inhibition of topoisomerase 2; (iii) covalent alkylation of macromolecules; and (iv) generation of cytotoxic free radicals.

Because of cardiovascular toxicity and hemolysis observed during preclinical toxicity studies, development of the parent compound ellipticine was halted. Interest then shifted to the 9-substituted derivatives, including 9-hydroxyellipticine, 9-methoxyellipticine, and Elliptinium (Celiptium). Only limited activity was observed in clinical trials with 9-methoxyellipticine and 9-hydroxyellipticine. Phase II clinical trials of Elliptinium yielded moderately promising results.[48] None of the ellipticine derivatives have reached clinical practice, but the ellipticine family may still yield clinically useful anticancer drugs,[38,48] especially if information at the molecular level provides a basis for subsetting potential patients.

For this study we collected information on 112 ellipticine derivatives that have been tested by the NCI anticancer drug discovery program. We were interested in them partly

because of their potent anticancer activity and partly because some of them appeared "p53-inverse,"[8,33] that is, they appeared more active against p53 mutant cell lines than against p53 wild-type ones in the screening assay. In addition to examining their p53-status-related characteristics, we also examined their CNS selectivity.

We initially classified these ellipticine analogues into three subtypes (Figure 1) according to their chemical structures. Normal ellipticines were classified as type E (the D ring is a noncharged pyridyl ring); 1,2-dihydrogenated ellipticines were classified as type H; and N^2-alkyl-substituted ellipticiniums were classified as type Q (the D ring is a quaternized pyridyl ring with a permanent positive charge). In our data set, 30 compounds are type E, 28 are type H (including three 1,2,3,4-tetrahydroellipticine analogues), and 54 are type Q, according to the aforementioned classification criteria.

METHODS

Cell Screen and Activity Data. Details of the NCI cell screening protocols and reporting procedures have been described elsewhere.[1−6,9] Cancer cell growth is determined spectrophotometrically by staining for total cellular protein with sulforhodamine B. Activity is expressed in terms of the quantity -log(GI_{50}), where GI_{50} is the 50% growth inhibitory concentration (as compared with untreated controls). For each compound there are 60 activity values (one for each cell line) that represent the activity pattern or "fingerprint." These are the original entries in the activity matrix, A. The database included 14% missing values, each of which was replaced by the mean value over all remaining cell lines for the compound in question.

In this study, we investigated QSAR for the following seven representative or interesting activity indices of the ellipticine analogues:

1. MOLT-4: activity against leukemia cell line MOLT-4;
2. mean.60: mean activity across all 60 cell lines;
3. mean.CNS: mean activity against the six CNS cell lines;
4. mean.p53W: mean activity against the 19 p53 wild-type cell lines;
5. mean.p53M: mean activity against the 41 p53 mutant cell lines;
6. CNS.sel: (mean.CNS − mean.60), a positive value indicates CNS selectivity;
7. p53.MW: (mean.p53M − mean.p53W), a positive value indicates "p53-inverse" selectivity.

Only the MOLT-4 index is the same as in the original activity database; the other six were derived indices from the original database, as already defined.

Computation of Molecular Descriptors. For computation of molecular descriptors, we used the Cerius^2 molecular modeling package from Molecular Simulations, Inc. (San Diego, CA). The geometries and energies of the ellipticine analogues were optimized using the Cerius^2 Universal Force Field.[50] Partial atomic charges were calculated using the Cerius^2 Charge Equilibration approach.[51] More than 160 molecular descriptors, categorized as: (i) conformational, (ii) electronic, (iii) information theoretic, (iv) quantum mechanical, (v) receptor related, (vi) shape related, (vii) spatial, (viii) thermodynamic, and (ix) topological are available in the Cerius^2 package. Most of the Cerius^2 descriptors were not useful for our QSAR study. In the Cerius^2 package, 49 of the >160 descriptors were identified by MSI as constituting a default descriptor set. Those 49 descriptors were believed, or were demonstrated, by MSI to be more useful. We initially used the 49 Cerius^2 default descriptors (Table 1) plus three indicator variables (*E*, *H*, and *Q*) that reflected our initial classification of the ellipticine derivatives (Figure 1). For additional analyses, we added 17 partial atomic charges on the ellipticine ring-forming atoms to the descriptor set, increasing the number of descriptors to 69.

QSAR and GFA. The QSAR analysis[52−55] relates chemical structural descriptors (x) to a response variable (y) by a mathematical equation, most often in the form

$$y = f(x) = a_0 + \sum a_{ij} B_j(x_i)$$

where a is a (regression) coefficient, i ranges over the set of descriptors, and $B_j(x_i)$ is the jth basis function type for the ith descriptor. The function (operator) B may, for example, be a logarithmic transform, a quadratic transform, a half-space spline, or unity. Given a large number of descriptors and a choice of basis functions, the set of possible terms for fitting the equation is large. There are many statistical and artificial intelligence methods available to fit such equations. Included are multiple linear regression (MLR), stepwise linear regression, principal component regression (PCR), partial least-squares regression (PLS), and artificial neural networks (ANN). Whichever type of algorithm is selected, the QSAR model should be predictive as well as descriptive. Hence, cross-validation and randomization of the dependent variable are often used to test the predictive ability of QSAR models.

To fit the equations and thereby build QSAR models we used the recently developed GFA method of D. Rogers.[56,57] GFA combines J. H. Friedman's multivariate adaptive regression splines (MARS) algorithm[58,59] with J. Holland's genetic algorithm (GA).[60] The MARS algorithm uses "truncated power spline" terms[56] to build regression models. The spline term can be in the form $\langle x - t \rangle$ or $\langle t - x \rangle$, where x is the value of the original variable and t is the "knot" of the spline. In the case of an $\langle x - t \rangle$ expression, the value of the spline term is 0 for $x \leqslant t$, and, otherwise, $x - t$. In the case of a $\langle t - x \rangle$ expression, the value of the spline term is 0 for $x \geqslant t$, and, otherwise, $t - x$. The spline basis functions introduce nonlinearity into the regression model. MARS was designed to allow the construction of spline-based regression models with moderate numbers of descriptors, usually <20. This algorithm can provide high levels of performance and often competes well with neural network approaches. The utility of spline terms for dealing with nonlinearity and outliers in multivariate regression has been well documented.[58,59] However, MARS is computationally intensive and too expensive to run when the number of descriptors becomes large (e.g., >20). Furthermore, because MARS builds its model incrementally (like forward stepwise regression), it may not be able to find models containing combinations of features that predict well as a group but poorly individually.[56,57]

Briefly, GFA uses a GA to search the MARS problem space to evolve a population of equations (combination of descriptors and coefficients) that best fit the training data

Table 1. Table of All Descriptors Used in This Study

No.	Descriptor	Family	Description
		49 default descriptors in Cerius[2]	
1	Apol	electronic	sum of atomic polarizabilities
2	BIC	information	bonding information content
3	CIC	information	complementary information content
4	E_ADJ_equ	information	edge adjacency information index (equility)
5	E_ADJ_mag	information	edge adjacency information index (magnitude)
6	E_DIST_equ	information	edge distance information index (equality)
7	E_DIST_mag	information	edge distance information index (magnitude)
8	IAC-Mean	information	mean atomic composition information
9	IC	information	multigraph information content
10	SIC	information	structural information content
11	V_ADJ_equ	information	vertex adjacency information index (equality)
12	V_ADJ_mag	information	vertex adjacency information index (magnitude)
13	V_DIST_equ	information	vertex distance information index (equality)
14	V_DIST_mag	information	vertex distance information index (magnitude)
15	Area	spatial	molecular surface area
16	MW	spatial	molecular weight
17	PMI	spatial	principal moment of inertia
18	PMIX	spatial	principal moment of inertia − *X* component
19	PMIY	spatial	principal moment of inertia − *Y* component
20	PMIZ	spatial	principal moment of inertia − *Z* component
21	Rotlbonds	spatial	number of rotatable bonds
22	Vm	spatial	molecular volume
23	AlogP	thermodynamic	Ghose and Crippen log *P*
24	MolRef	thermodynamic	Ghose and Crippen molar refractivity
25	CHI-0	topological	molecular connectivity index 0
26	CHI-1	topological	molecular connectivity index 1
27	CHI-2	topological	molecular connectivity index 2
28	CHI-3_C	topological	molecular connectivity index 3 (cluster)
29	CHI-3_P	topological	molecular connectivity index 3 (path)
30	CHI-V-0	topological	molecular connectivity index V0
31	CHI-V-1	topological	molecular connectivity index V1
32	CHI-V-2	topological	molecular connectivity index V2
33	CHI-V-3_C	topological	molecular connectivity index V3 (Cluster)
34	CHI-V-3_P	topological	molecular connectivity index V3 (path)
35	JX	topological	Balaban index *X*
36	Kappa-1	topological	molecular shape index order 1
37	Kappa-1-AM	topological	molecular shape index order 1 with alpha-modified atom count
38	Kappa-2	topological	molecular shape index order 2
39	Kappa-2-AM	topological	molecular shape index order 2 with alpha-modified atom count
40	Kappa-3	topological	molecular shape index order 3
41	Kappa-3-AM	topological	molecular shape index order 3 with alpha-modified atom count
42	PHI	topological	molecular flexibility index
43	SC-0	topological	subgraph count order 0
44	SC-1	topological	subgraph count order 1
45	SC-2	topological	subgraph count order 2
46	SC-3_C	topological	subgraph count order 3 (cluster)
47	SC-3_P	topological	subgraph count order 3 (path)
48	Wiener	topological	Wiener index
49	Zagreb	topological	Zagreb index
50−66		17 partial atomic charges on the ellipticine core structure	
67−69		3 indicator descriptors (*E, H,* or *Q*)	

(i.e., calculate the activity index in question). The algorithm is as follows:

(i) An initial population of equations or individuals (e.g., 100) is generated by a random choice of descriptors (e.g., the initial set of 52 or the extended set of 69 in our case) and basis functions (linear, quadratic, or spline in our study). The knot (cut-point) of a spline term is randomly initialized (and later optimized during GFA evolution). The "fitness" of each initial equation for calculating an activity index over the training data set is scored by using Friedman's lack of fit (LOF) measure:[58,59]

$$\mathrm{LOF} = \mathrm{LSE}/\{1 - (c + d^{*}p)/m\}^2$$

where LSE is the least-squares error (calculated from the difference between actual and calculated values for the activity index over data set), c is the number of basis functions in the model, d is a smoothing parameter that controls the number of terms in the model equation (a larger value of d leads to fewer terms), p is the number of features contained in all terms of the model, and m is the number of samples (compounds) in the training set. The LOF measure penalizes appropriately for the addition of terms to the equation (and consequent loss of degrees of freedom) in such a way as to resist overfitting. In all our calculations, the smoothing parameter d is set to the default value, 1.

(ii) Pairs from the population of equations are chosen at random (equations with lower LOF scores have a greater probability of being chosen), and "crossovers" are done at randomly chosen points within the equations to produce "progeny" equations combining the characteristics of both "parents." The crossovers are allowed to take place at different points in the two parents (cf., unequal crossover in

biology), so that the progeny may have fewer or more terms than their parents.

(iii) The fitness of the progeny equation is assessed by calculating the LOF measure.

(iv) If the fitness of the new equation is in the top 100, it is kept, and equation number 100 is dropped; otherwise, the progeny equation is discarded. Steps (ii) to (iv) are repeated until a preset number of crossovers has been performed. The final population of equations is examined for important information.

A distinctive feature of GFA is that instead of generating a single model, as do most other statistical methods, it produces a population of models (e.g., 100). The range of variation in this population gives added information on the quality of fit and importance of descriptors. By examining these models, additional information can be discerned. For example, the frequency of use of a particular descriptor in the population of equations may indicate how relevant the descriptor is to the prediction of activity. For comparison with GFA, we also performed classical stepwise linear regression analyses on the same data set. In preliminary studies,[61,62] GFA performed better than did the stepwise regression.

Cluster Analysis. To cluster compounds in terms of their GI_{50} activity patterns across the 60 cell lines we used the "hclust" (hierarchical clustering) function in the S−Plus statistical package (StatSci Division, MathSoft, Inc., Seattle, WA).[63] Compounds with similar activity patterns appear together, and dissimilar compounds appear distant from each other in the cluster tree. For this study, we used the "average linkage" clustering algorithm and a distance metric of ($1 - r$), where r is the average Pearson correlation coefficient between the activity patterns of two merging groups of compounds.

RESULTS

Cluster Analysis Based on Cell Screen Activity Patterns. For a better understanding of the current work, it will be helpful to summarize our previous cluster analyses of and mechanistic studies on the ellipticine data set.[35] The average linkage hierarchical cluster tree for 112 ellipticine analogues (Figure 2) indicates a remarkable separation of the data set into two subgroups. The first subgroup (compounds **1**−**66**) consists principally of normal ellipticines (*E*), whereas the second subgroup (compounds **67**−**112**) consists principally of N^2-alkyl substituted ellipticiniums (*Q*). The *H*-type compounds cluster in the first subgroup (*E*), with a few exceptions that fall in the second subgroup (*Q*). Apparent discrepancies were explainable on the basis of chemical transformations in the incubation medium, leading to the *EE* (**1**−**66**) and *QQ* (**67**−**112**) classifications.[35]

GFA Results Without Considering Partial Atomic Charges as Descriptors. After some preliminary calculations on different data sets, we decided to set the number of GFA crossovers to 200 000 to achieve reasonable convergence. The fitness (LOF and R^2) and predictivity (cross-validated R^2 or CVR^2) of the best model for each activity index are shown in Table 2. When the whole data set (*EHQ*, 112 compounds) was taken into account for QSAR analysis with the initial set of 52 descriptors, the GFA models for most of the dependent variables were not good, except for those obtained for activity index CNS.sel, and to a lesser extent for index p53.MW. In fact, reasonably acceptable QSAR models for CNS.sel and p53.MW were obtained only after inclusion of the three indicator variables (*E*, *H*, and *Q*). We found that CNS.sel and p53.MW were highly positively correlated with *Q*. That is, compounds of the *Q*-type were not only more CNS selective but also more active against p53 mutant cell lines. CNS.sel for the *Q* subset was ∼0.5 log units higher than those for the *E* and *H* subsets, and p53.MW for the *Q* subset was ∼0.18 log units higher than those for the *E* and *H* subsets. The QSAR models obtained here are mainly distinguishing the differences between type *Q* and type *E* or *H* compounds.

The failure to obtain good QSAR equations for the whole set seemed to be consistent with the hypothesis that different subgroups of ellipticine analogues may have different mechanisms of action or resistance. We therefore decided to divide the whole data set into three subsets (*E*, *H*, and *Q*) according to their chemical structures, in hopes of separating these compounds by mechanism of action or cellular handling. The results of GFA calculations for the seven activity indices for each subset are also summarized in Table 2. Generally speaking, better results were obtained by separating the whole data set into three smaller subsets.

One might argue that generation of better QSAR models could simply be the result of decreased size of the data sets when the same number of descriptors was retained for GFA, but that did not seem to be the case here. After separation of the whole data set into *E*, *H*, and *Q* portions, we got significantly improved models for the *E* subset (except for the CNS.sel and p53.MW indices) and *Q* subset. However, there was no appreciable improvement for the *H* subset (except for its CNS.sel and p53.MW indices; as will be demonstrated later, QSAR equations for the CNS.sel and p53.MW indices of the H-subset were not stable). It appeared that the three types of ellipticine analogues inhibit cancer cell growth by different mechanisms or are handled by the tumor cells in different ways. Consequently, separation of the whole data set into three subgroups had the effect of separating them by mechanism of action or cellular handling (e.g., transport), thus leading to much improved QSAR models. The reasonably good results for the *E* and *Q* subsets suggested that there might be a unique mechanism of action or cellular handling within each subgroup of *E* and *Q*. Poor QSAR models obtained for the *H* subset suggested, however, that there was no unique structure−activity relationship among them and that the *H* subset might be heterogeneous with respect to mechanism of action, cellular handling, or metabolism in the cell culture medium.

The poor QSAR models obtained for the CNS.sel and p53.MW indices with the *E* subset might be explained by the fact that this subset was neither CNS selective nor more active against p53 mutant cell lines. In other words, differences in these indices among compounds are at the level of experimental error. This explanation was supported by examining the statistics of these activity indices. For the *E* subset, CNS.sel and p53.MW had a very small range of values. Apparently, the *H* and *Q* subsets covered a much larger range of activity values than did the *E* subset, and indeed, for these two activity indices, much better QSAR models were obtained for both *H* and *Q* subsets.

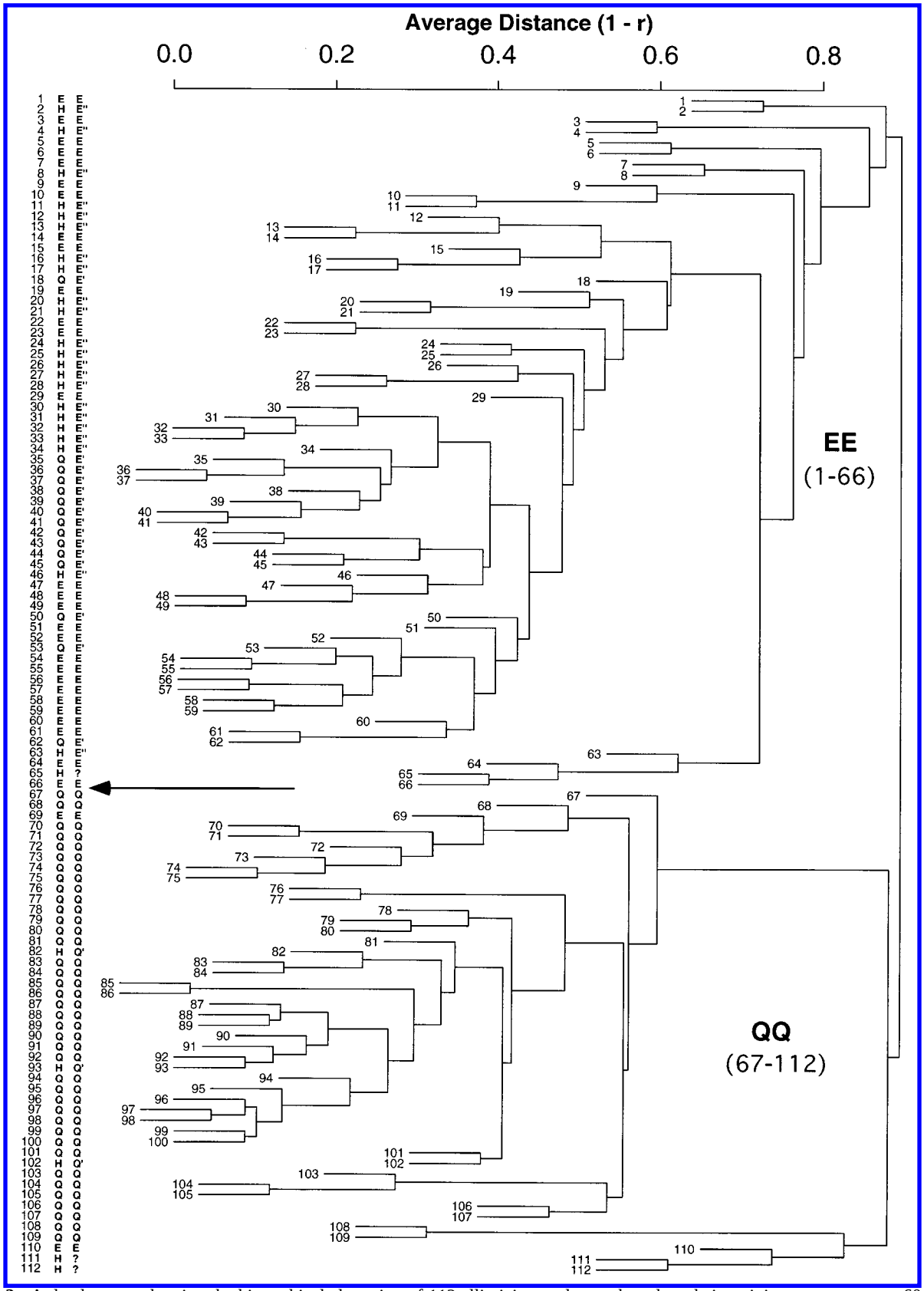

Figure 2. A dendrogram showing the hierarchical clustering of 112 ellipticine analogues based on their activity patterns across 60 human tumor cell lines. The first column shows compound numbers in cluster order; the second column (symbol *E*, *H*, or *Q*) specifies the original structural type of the compound; the third column (symbol *E*, *E*′, *E*″, *Q*, or *Q*′) indicates the active structural type to which the compound was probably transformed in the culture medium: *E*′ was transformed from *Q*; *E*″ was transformed from *H*; and *Q*′ was transformed from *H*. For compounds with a question mark (?), we are not sure of their active forms. There is a major gap between compounds **66** and **67**, dividing the data set into two subgroups (*EE* and *QQ*). Modified from ref 35.

GFA Results With Partial Atomic Charges as Descriptors. Because the default Cerius[2] descriptors did not produce very good QSAR models, we looked for additional descriptors. Because substitutions to the ellipticine core structure

Table 2. Performance of GFA Models as Assessed by Three Different Scores (LOF, R^2, and CVR^2) Based on 52 Descriptors (49 Cerius^2 Descriptors and 3 Indicator Variables)

	LOF				R^2				CVR^2			
activity index	*EHQ*	*E*	*H*	*Q*	*EHQ*	*E*	*H*	*Q*	*EHQ*	*E*	*H*	*Q*
MOLT-4	0.684	0.401	0.577	0.339	0.396	0.821	0.485	0.800	0.342	0.741	0.377	0.750
mean.60	0.361	0.229	0.342	0.171	0.096	0.412	0.431	0.741	0.263	0.584	0.326	0.672
mean.CNS	0.388	0.210	0.406	0.191	0.273	0.765	0.328	0.742	0.301	0.651	0.232	0.676
mean.p53W	0.398	0.226	0.373	0.204	0.290	0.732	0.445	0.768	0.234	0.598	0.341	0.696
mean.p53M	0.374	0.227	0.325	0.190	0.240	0.686	0.436	0.667	0.217	0.528	0.336	0.622
CNS.sel	0.053	0.006	0.014	0.043	0.682	0.156	0.907	0.842	0.640	0.407	0.833	0.802
p53.MW	0.020	0.005	0.010	0.015	0.511	0.546	0.676	0.741	0.462	0.374	0.592	0.708
Average	0.325	0.186	0.292	0.165	0.355	0.588	0.530	0.757	0.351	0.555	0.434	0.704

Table 3. Performance of GFA Models as Assessed by Three Different Scores (LOF, R^2, and CVR^2) Based on 69 Descriptors (49 Cerius^2 Descriptors, 3 Indicator Variables, and 17 Partial Atomic Charges)

	LOF				R^2				CVR^2			
activity index	*EHQ*	*E*	*H*	*Q*	*EHQ*	*E*	*H*	*Q*	*EHQ*	*E*	*H*	*Q*
MOLT−4	0.555	0.267	0.324	0.269	0.539	0.831	0.837	0.857	0.482	0.745	0.707	0.804
mean.60	0.318	0.092	0.280	0.169	0.474	0.890	0.609	0.755	0.423	0.820	0.515	0.700
mean.CNS	0.308	0.080	0.370	0.155	0.543	0.925	0.486	0.762	0.486	0.868	0.378	0.707
mean.p53W	0.332	0.104	0.263	0.177	0.477	0.919	0.729	0.767	0.425	0.861	0.660	0.720
mean.p53M	0.350	0.098	0.266	0.154	0.320	0.898	0.680	0.765	0.262	0.855	0.167	0.697
CNS.sel	0.052	0.007	0.011	0.041	0.664	0.543	0.873	0.808	0.636	0.458	0.295	0.766
p53.MW	0.018	0.003	0.005	0.015	0.583	0.220	0.862	0.764	0.529	0.309	0.571	0.727
Average	0.276	0.093	0.217	0.140	0.514	0.747	0.725	0.783	0.463	0.702	0.470	0.732
						0.893[a]				0.830[a]		

[a] Average calculated excluding activity indices CNS.sel and p53.MW.

Table 4. Performance of Stepwise Regression Models as Assessed by Two Different Scores (R^2 and CVR^2) with 69 Descriptors (49 Cerius^2 Descriptors, 3 Indicator Variables, and 17 Partial Atomic Charges)

	EHQ (112)		*E* (30)		*H* (28)		*Q* (54)	
activity index	R^2	CVR^2	R^2	CVR^2	R^2	CVR^2	R^2	CVR^2
MOLT−4	0.422	0.347	0.278	0.161	0.589	0.485	0.592	0.519
mean.60	0.345	0.232	0.782	0.644	0.317	0.167	0.558	0.463
mean.CNS	0.232	0.189	0.776	0.678	0.324	0.203	0.450	0.363
mean.p53W	0.279	0.208	0.792	0.651	0.686	0.561	0.613	0.533
mean.p53M	0.258	0.203	0.808	0.710	0.294	0.141	0.511	0.424
CNS.sel	0.657	0.609	0.322	0.253	0.142	0.055	0.732	0.678
p53.MW	0.510	0.433	0.197	−0.005	0.622	−14.22	0.366	0.321
Average	0.386	0.317	0.565	0.442	0.425	−0.180	0.546	0.472

should affect the electron distribution on 17 ring-forming atoms, we decided to add partial atomic charges on the 17 atoms to the descriptor set. Therefore, the augmented set contained 49 (from Cerius^2) + 3 (*E*, *H*, *Q* indicators) + 17 (partial atomic charges) = 69 descriptors. Results of GFA calculations are shown in Table 3. Generally speaking, QSAR models were appreciably better with inclusion of partial atomic charges. In most cases, partial atomic charges were picked up by GFA as the most descriptive variables. The results in Table 3 showed the same trends as those in Table 2; that is, (i) results for the whole data set were still not good; (ii) good results were obtained for the *E* subset (except for CNS.sel and p53.MW); (iii) good results were obtained for the *Q* subset; and (iv) the results for the *H* subset were improved but still not good.

Comparison With Stepwise Linear Regression. For comparison, we used a cross-validated forward stepwise regression procedure with $F = 4.0$ as the threshold value for adding additional variables. The performance of stepwise linear regression (Table 4) was much worse than that of the GFA algorithm: R^2 and CVR^2 values were ~0.3 lower than those from GFA. We believe that the superior performance of GFA was due to its ability to search more possible combinations of variables and to the inclusion of spline terms in building QSAR models. The spline terms permit modeling of nonlinearity.[56−59]

Randomization Tests. The *H* subset seemed to provide GFA models with very high R^2 values (see Table 3), but the predictivity (CVR^2) of the models varied considerably, suggesting that the models for the *H* subset were not robust. A randomization (permutation) test further confirmed this suggestion. The test was done by (i) repeatedly permuting the activity values of the data set, (ii) using the permuted values to generate QSAR models, and (iii) comparing the resulting scores with the score of the original QSAR model generated from nonrandomized activity values. If the original QSAR model is statistically significant, its score should be significantly better than those from permuted data. When we did the randomizations with the CNS.sel index for the *H* subset, two of the 19 random trials had better scores than that of the original model, and several other trials yielded R^2 values very close to that of the original model (see Figure 3).

GFA Calculations on the Transformed *QQ′* Subset. Because the *E*, *H*, and *Q* subsets showed apparently different patterns of activity by clustering and in our previous analyses

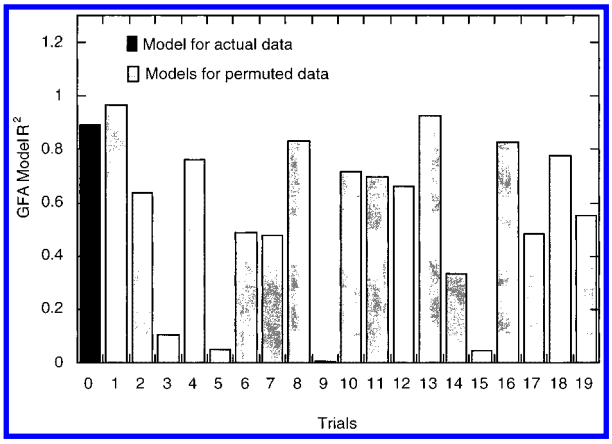

Figure 3. GFA randomization tests for the CNS.sel index of the *H* subset. The first bar (solid) shows the R^2 value for the model based on actual data; the other 19 bars (open) show the R^2 for 19 models based on permuted data. This figure demonstrates that the model for actual data for the *H* subset was not robust. See text for details.

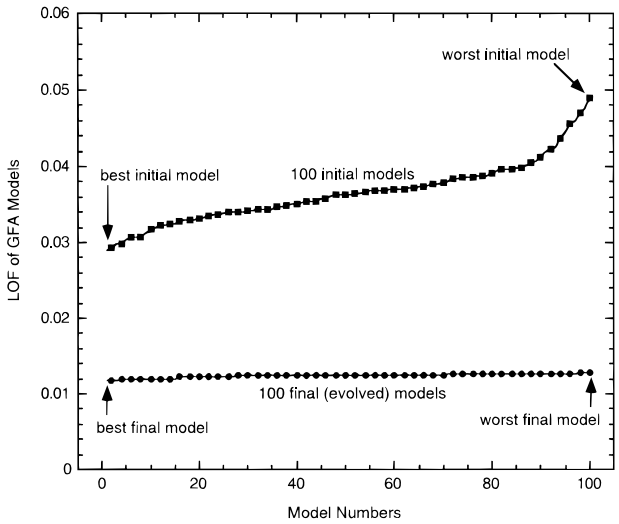

Figure 4. LOF values for the 100 evolved GFA equations for the p53.MW index of the *QQ′* subset, compared with those for the initial (not evolved) population of models.

on the interconversions among them, we focused our calculations on a subset of 42 compounds (*QQ′*; **67**, **68**, **70**−**109**), which presumably possess the same ellipticinium active form (see Figure 1) under cell culture conditions. Compounds **69** and **110**−**112** were excluded from the calculations because they do not have the ellipticinium active form. The *QQ′* subset was composed of 39 original ellipticiniums (54 ellipticiniums, excluding 15 prodrugs that showed activity patterns closer to those of normal ellipticines) and three 1,2-dihydroellipticines (**82**, **93**, and **102**), which were likely to be converted to the ellipticinium forms by a mechanism of oxidation. For these three "pseudo-*Q*" compounds, we used the corresponding ellipticinium forms to calculate their molecular descriptors. The major chemical difference between ellipticiniums and normal ellipticines is the permanent charge on the D-ring of ellipticiniums. This structural difference appeared to be the major correlate of p53-inverse character. Also interesting, however, were the significant differences in p53.MW index within the *QQ′* subset. Therefore, this subset of 42 compounds was subjected to extensive GFA calculations. We focused our calculations on the most interesting activity index, p53.MW.

Sixty-nine descriptors were included in the descriptor set, and the linear, quadratic, and spline basis functions were allowed for each descriptor. The GFA population size was set to 100, with initial equation length of 4. Theoretically, because of the inherent "randomness" of the genetic algorithm incorporated into the GFA approach, it is unlikely that the same results will be obtained repeatedly for different runs of the GFA procedure. However, there did appear to be considerable consistency with regard to the types of descriptors that most frequently appeared in the final population of equations. The average LOF, R^2, and CVR^2 for six separate runs of GFA were 0.0135, 0.782, and 0.696, respectively, indicating a relatively high structure−activity correlation. Note that, as mentioned earlier in the *Methods* section, the GFA searching process was not driven by R^2 or LSE (fitness measures with no penalty for addition of terms to the equation). Instead, it was guided by LOF, as defined by Friedman.[58,59]

A list of the 10 most frequently used descriptors in a population of 100 evolved equations was furnished after each run of GFA. Seventeen descriptors appeared in the top-10-descriptors list for at least one of the six separate GFA runs. This subset of 17 descriptors was then used as the working descriptor set to ask the following questions: (i) Did GFA find a good global optimum from the much larger descriptor set (69 descriptors) compared with that from the smaller one (17 descriptors)?; (ii) How robust were the variable selection and fitting processes?; (iii) How similar to each other were the 100 evolved equations?; and (iv) How did the final model relate to the biology?

Table 5. Summary of the Best Models for Prediction of p53.MW from 20 Separate Runs (from Different Starting Random Numbers) of GFA for the *QQ′* Subset with 17 Preselected Descriptors and 200 000 Crossovers (see also Figure 5)[a]

model number	best model (from a population size of 100)	number of times the equation was picked up as "best" (percentage)
1	p53.MW = −1.265 + 26.203*C8Ch2 − 16.871*⟨−0.38 − N2Ch⟩ + 0.0641*IC2 + 0.284*⟨1 − RB⟩ + 9.320*⟨C9Ch − 0.287⟩ + 8.652*C14Ch LOF = 0.0118, R^2 = 0.812, CVR^2 = 0.744 (6 descriptors)	16 (80%)
2	p53.MW = −1.299 + 0.282*⟨1 − RB⟩ + 9.211*⟨C9Ch − 0.287⟩ + 8.761*C14Ch + 0.0658*IC2 + 26.399*C8Ch2 − 17.613*⟨−0.38 − N2Ch⟩ LOF = 0.0120, R^2 = 0.813, CVR^2 = 0.744 (6 descriptors)	3 (15%)
3	p53.MW = −0.7078 − 1.275*N6Ch − 3.596*C8Ch + 10.378*⟨C9Ch − 0.287⟩ − 9.783*⟨−0.379 − N2Ch⟩ LOF = 0.0154, R^2 = 0.659, CVR^2 = 0.530 (4 descriptors)	1 (5%)

[a] The charges are the partial atomic charges at the specified atomic positions calculated by using the Cerius[2] Charge Equilibration method; IC, information content; RB, number of rotatable bonds in a molecule.

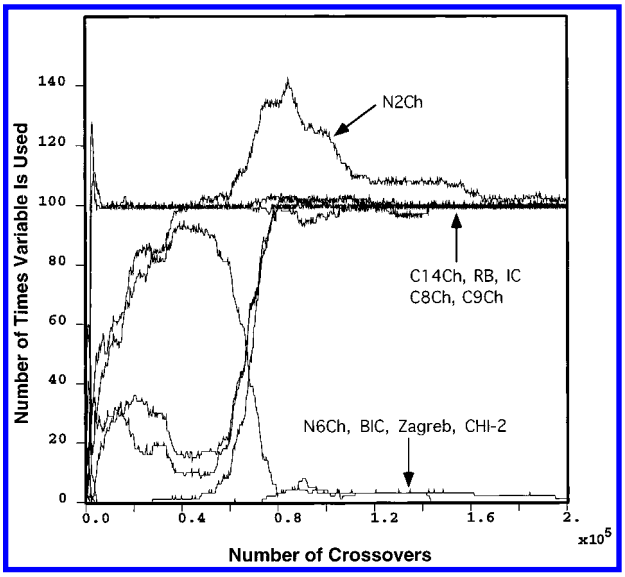

Figure 5. Frequency of descriptor use for the p53.MW index of the *QQ′* subset. The six descriptors used in the best model in Table 5 were used in almost every model (~100 usages per 100 models), and other descriptors were rarely used.

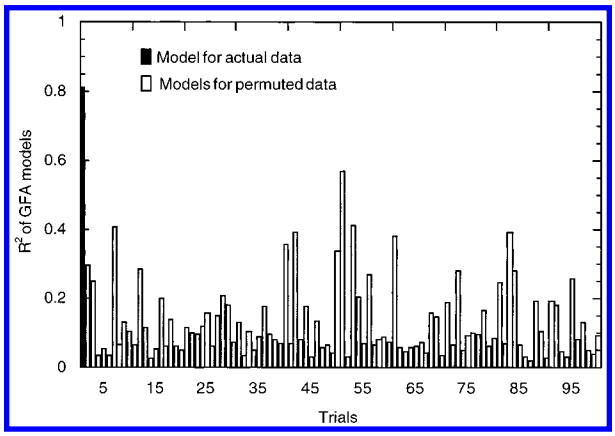

Figure 6. GFA randomization tests for the p53.MW index of the *QQ′* subset. The first bar (solid) shows the R^2 value for the model based on actual data; the other 99 bars (open) show the R^2 values for 99 models based on permuted data. This figure demonstrates that the model for actual data for the *QQ′* subset was statistically robust and reliable. See text for details.

We ran GFA 20 times, starting from the same 17 descriptors and GFA parameter settings, but with different initializations of the random number generator. A summary of the results is shown in Table 5. Sixteen of the 20 runs generated the same best equation, with LOF, R^2, and CVR^2 values of 0.0118, 0.812, and 0.744, respectively. Although we have not assessed the statistical significance, there appeared to be a moderate improvement over the values obtained using the larger descriptor set (i.e., 0.0135, 0.782, and 0.696, respectively). The second model in Table 5 was essentially the same as the first. The third model apparently represented a local optimum; it was quite different from the other two. The results showed that by using a pre-selected smaller subset of descriptors (down from 69 to 17), the GFA was able to obtain more consistent results from run to run. With the smaller subset of descriptors it seemed less likely that the GFA would be trapped in local optima, because the chance for an important descriptor or combination of descriptors to be lost during the crossover process was lower.

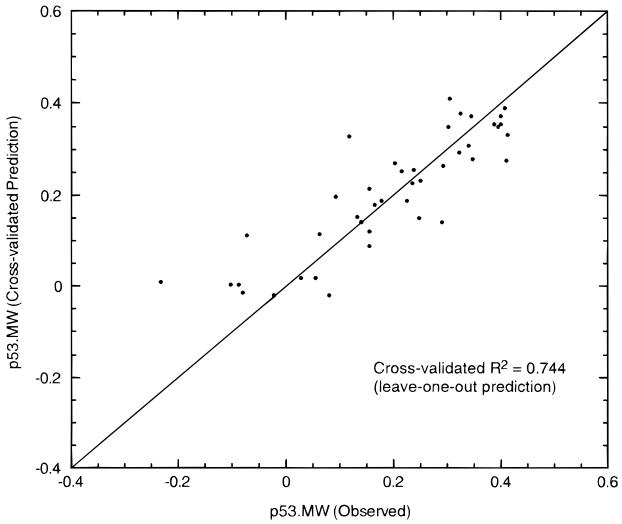

Figure 7. Observed and GFA cross-validation-predicted p53.MW values for the 42 *QQ′* type compounds.

It is important to note that once an important descriptor is dropped during the crossover process, it can never be recovered.

At the end of GFA evolution, the 100 models in the equation population were quite similar in terms of their LOF values, as shown in Figure 4. Shown in Figure 5 are the frequencies at which the top 10 descriptors were used in the equation population (size 100) during the GFA evolution process from one representative run. These figures clearly demonstrate the stability of the final GFA models.

The statistical significance of the relationship between p53.MW and chemical structural descriptors was further demonstrated by a randomization procedure of the type discussed earlier. The R^2 values for 99 trials based on permuted activity data are shown in Figure 6. In contrast to the results obtained for the CNS.sel index of the *H* subset (shown in Figure 3), the R^2 value of the original model for p53.MW of the *QQ′* subset was much higher than that of any of the trials using permuted data.

A full cross-validation was performed on the p53.MW index for the *QQ′* subset. At each step, one of the 42 compounds was left out in the entire GFA training process, and its value was predicted by the model obtained from the remaining 41 compounds. This process was repeated 42 times until every compound had been left out and predicted once. The observed and cross-validation-predicted p53.MW values are shown in Figure 7. A CVR^2 of 0.744 was obtained, indicating the predictive ability of the GFA model (i.e., $p \ll 0.01$ with respect to the null hypothesis that $R^2 = 0$).

DISCUSSION

Cluster analyses were performed for the ellipticine analogues based on their in vitro anticancer activity patterns. The results from this work were in good agreement with previous observations that in vitro activity patterns could distinguish structurally diverse compounds by their mechanisms of action.[6−9] In this study, we demonstrate at a "micro-level" (for a more homogeneous data set) that the NCI in vitro cell screen program generates rich information about the mechanisms of action and selective cytotoxicity

of tested compounds. The p53- and CNS-related activity indices were fairly well predicted by GFA on the basis of chemical structural descriptors. The GFA algorithm performed much better than did classical stepwise linear regression. However, QSAR models were not particularly successful for the whole data set with either method. Satisfactory QSAR results were obtained only after we divided the set into three subsets according to their chemical structures, presumably separating them (though not perfectly) by mechanism of action and/or cellular handling.

Extensive GFA calculations were performed on a subset of 42 ellipticiniums in the *QQ′* set to search for structural features that predict the p53.MW and CNS.sel indices. We found that partial atomic charges were the descriptors most frequently picked up by GFA. This finding might reflect the way in which the ellipticine compounds interact with their biological targets or it might reflect differences in transport or metabolism. Although the GFA models appeared robust and significant, their relevance to biological problems demands further clarification and explanation. More molecular biology and molecular modeling studies will be required to explain the details of these drug−target(s) interactions.

To obtain stable and consistent results from different GFA runs, we proposed and used a new procedure to select a subset of descriptors from a much larger pool of descriptors. GFA was run several times from different random number starting points to compile a list of the most frequently used descriptors. The smaller subset of descriptors was then subjected to further GFA calculations. Good descriptors appeared less likely to be dropped from the candidate descriptor set during the GFA evolutionary calculations if they were not buried in a huge number of irrelevant descriptors. Despite its simplicity, this procedure seemed to work well in this study. Consistent and stable "best models" were generated using the pre-selected subset of descriptors.

ACKNOWLEDGMENT

We are grateful to Drs. Kurt Kohn, Yves Pommier, and Mark Waltham of the NCI for helpful discussions on the biology involved in this study. We thank Dr. David Rogers of MSI for helpful discussions of GFA. We are also very grateful to members of the Developmental Therapeutics Program, NCI, for the anticancer activity data used in this study. Their extraordinary efforts in developing and maintaining the cancer cell screen have made these theoretical studies possible.

REFERENCES AND NOTES

(1) Boyd, M. R. In *Cancer: Principles and Practice of Oncology Update*; DeVita, V. T.; Hellman, S.; Rosenberg, S. A.; Eds.; J. B. Lippicott: Philadelphia, 1989; Vol. 3.

(2) Alley, M. C.; Scudiero, D. A.; Monks, A.; Hursey, M. L.; Czerwinski, M. J.; Fine, D. L.; Abbott, B. J.; Mayo, J. G.; Shoemaker, R. H.; Boyd, M. R. Feasibility of drug screening with panels of human tumor cell lines using a microculture tetrazolium assay. *Cancer Res.* **1988**, *48*, 589−601.

(3) Monks, A.; Scudiero, D. A.; Shoemaker, R. H.; Paull, K. D.; Vistica, D.; Hose, C.; Langley, J.; Cronise, P.; Vaigro-Wolff, A.; Gray-Goodrich, M.; Campell, H.; Mayo, J.; Boyd, M. R. Feasibility of a high-flux anticancer screen using a diverse panel of cultured human tumor lines. *J. Natl. Cancer Inst.* **1991**, *83*, 757−766.

(4) Boyd, M. R.; Paull, K. D. Some practical considerations and applications of the National Cancer Institute in vitro anticancer drug discovery screen. *Drug Devel. Res.* **1995**, *34*, 91−109.

(5) Boyd, M. R. In *Anticancer Drug Development Guide: Preclinical Screening, Clinical Trials, and Approval*; Teicher, B. A., Ed.; Humana: Totowa, NJ, 1997.

(6) Paull, K. D.; Shoemaker, R. H.; Hodes, L.; Monks, A.; Scudiero, D. A.; Rubinstein, L.; Plowman, J.; Boyd, M. R. Display and analysis of patterns of differential activity of drugs against human tumor cell lines: development of mean graph and COMPARE algorithm. *J. Natl. Cancer Inst.* **1989**, *81*, 1088−1092.

(7) Weinstein, J. N.; Kohn, K. W.; Grever, M. R.; Viswanadhan, V. N.; Rubinstein, L. V.; Monks, A. P.; Scudiero, D. A.; Welch, L.; Koutsoukos, A. D.; Chiausa, A. J.; Paull, K. D. Neural computing in cancer drug development: Predicting mechanism of action. *Science* **1992**, *258*, 447−451.

(8) Weinstein, J. N.; Myers, T. G.; O'Connor, P. M.; Friend, S. H.; Fornace, A. J., Jr.; Kohn, K. W.; Fojo, T.; Bates, S. E.; Rubinstein, L. V.; Anderson, N. L.; Buolamwini, J. K.; van Osdol, W. W.; Monks, A. P.; Scudiero, D. A.; Sausville, E. A.; Zaharevitz, D. W.; Bunow, B.; Viswanadhan, V. N.; Johnson, G. S.; Wittes, R. E.; Paull, K. D. An information-intensive approach to the molecular pharmacology of cancer. *Science* **1997**, *275*, 343−349.

(9) Paull, K. D.; Hamel, E.; Malspeis, L. In *Cancer Chemotherapeutic Agents*; Foye, W. O., Ed.; American Chemical Society: Washington, D. C., 1995.

(10) Koo, H.-M.; Monks, A.; Mikheev, A.; Rubinstein, L. V.; Gray-Goodrich, M.; McWilliams, M. J.; Alvord, W. G.; Oie, H. K.; Gazdar, A. F.; Paull, K. D.; Zarbl, H.; Vande Woude, G. F. Enhanced sensitivity to 1-beta-D-arabinofuranosylcytosine and topoisomerase II inhibitors in tumor cell lines harboring activated ras oncogenes. *Cancer Res.* **1996**, *56*, 5211−5216.

(11) van Osdol, W. W.; Myers, T. G.; Paull, K. D.; Kohn, K. W.; Weinstein, J. N. Use of the Kohonen self-organizing map to study the mechanisms of action of chemotherapeutic agents. *J. Natl. Cancer Inst.* **1994**, *86*, 1853−1859.

(12) Koutsoukos, A. D.; Rubinstein, L. V.; Faraggi, D.; Simon, R. M.; Kalyandrug, S.; Weinstein, J. N.; Kohn, K. W.; Paull, K. D. Discrimination techniques applied to the NCI in vitro anti-tumour drug screen: predicting biochemical mechanism of action. *Stat. Med.* **1994**, *13*, 719−730.

(13) Weinstein, J. N.; Myers, T.; Buolamwini, J.; Raghavan, K.; van Osdol, W.; Licht, J.; Viswanadhan, V. N.; Kohn, K. W.; Rubinstein, L. V.; Koutsoukos, A. D.; Monks, A.; Scudiero, D. A.; Anderson, N. L.; Zaharevitz, D.; Chabner, B. A.; Grever, M. R.; Paull, K. D. Predictive statistics and artificial intelligence in the U.S. National Cancer Institute's Drug Discovery Program for Cancer and AIDS. *Stem Cells* **1994**, *12*, 13−22.

(14) Bates, S. E.; Fojo, A. T.; Weinstein, J. N.; Myers, T. G.; Alvarez, M.; Paull, K. D.; Chabner, B. A. Molecular targets in the National Cancer Institute drug screen. *J. Cancer Res. Clin. Oncol.* **1995**, *121*, 495−500.

(15) Li, G.; Waltham, M.; Unsworth, E.; Treston, A.; Mushine, J.; Anderson, N. L.; Kohn, K. W.; Weinstein, J. N. Rapid protein identification from two-dimensional polyacrylamide gels by MALDI mass spectrometry. *Electrophoresis* **1997**, *18*, 391−402.

(16) Myers, T. G.; Waltham, M.; Li, G.; Buolamwini, J. K.; Scudiero, D. A.; Rubinstein, L. V.; Paull, K. D.; Sausville, E. A.; Anderson, N. L.; Weinstein, J. N. A protein expression database for the molecular pharmacology of cancer. *Electrophoresis* **1997**, *18*, 647−653.

(17) Milne, G. W. A.; Miller, J. A. The NCI Drug Information System. 1. System overview. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 154−159.

(18) Milne, G. W. A.; Feldman, A.; Miller, J. A.; Daly, G. P.; Hammel, M. J. The NCI Drug Information System. 2. DIS pre-registery. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 159−168.

(19) Milne, G. W. A.; Feldman, A.; Miller, J. A.; Daly, G. P. The NCI Drug Information System. 3. The DIS chemistry module. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 168−179.

(20) Milne, G. W. A.; Miller, J. A.; Hoover, J. R. The NCI Drug Information System. 4. Inventory and shipping modules. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 179−185.

(21) Klopman, G.; Shi, L. M.; Ramu, A. Quantitative structure−activity relationship of multidrug resistance reversal agents. *Mol. Pharmacol.* **1997**, *52*, 323−334.

(22) Klopman, G. Artificial intelligence approach to structure−activity studies. Computer automated structure evaluation of biological activity of organic molecules. *J. Am. Chem. Soc.* **1984**, *106*, 7315−7320.

(23) Klopman, G. MULTICASE: a hierarchical computer automated structure evaluation program. *QSAR* **1992**, *11*, 176−184.

(24) Milne, G. W. A.; Nicklaus, M. C.; Driscoll, J. S.; Wang, S.; Zaharevitz, D. National Cancer Institute Drug Information System 3D database. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1219−1224.

(25) Wang, S.; Zaharevitz, D. W.; Sharma, R.; Marquez, V. E.; Lewin, N. E.; Du, L.; Blumberg, P. M.; Milne, G. W. A. The discovery of novel, structurally diverse protein kinase C agonists through computer 3D-database search. Molecular modeling studies. *J. Med. Chem.* **1994**, *37*, 4479−4489.
(26) Wang, S.; Milne, G. W. A.; Yan, X.; Posey, I.; Nicklaus, M. C.; Graham, L.; Rice, W. G. Discovery of novel, non-peptide HIV-1 protease inhibitors by pharmacophore searching. *J. Med. Chem.* **1996**, *39*, 2047−2054.
(27) Nicklaus, M. C.; Neamati, N.; Hong, H.; Mazumder, A.; Sunder, S.; Chen, J.; Milne, G. W. A.; Pommier, Y. HIV-1 integrase pharmacophore: Discovery of inhibitors through three-dimensional database searching. *J. Med. Chem.* **1997**, *40*, 920−929.
(28) Hong, H.; Neamati, N.; Wang, S.; Nicklaus, M. C.; Mazumder, A.; Zhao, H.; Burke, T. R.; Pommier, Y.; Milne, G. W. A. Discovery of HIV-1 integrase inhibitors by pharmacophore searching. *J. Med. Chem.* **1997**, *40*, 930−936.
(29) Myers, T. G.; Weinstein, J. N.; Raghavan, K.; Buolamwini, J.; Anderson, N. L.; O'Connor, P.; Kohn, K. W.; Scudiero, D. A.; Monks, A. P.; Friend, S. An 'information-intensive' strategy for drug discovery in cancer and AIDS: Relating cell cycle factors to patterns of drug activity. *Proc. Ann. Meeting Am. Assoc. Cancer Res.* **1995**, *36*, 305.
(30) Myers, T. G.; Paull, K. D.; Kohn, K. W.; Fojo, A. T.; Bates, S.; Weinstein, J. N. Molecular determinants of in vitro cytotoxicity in the NCI drug screen. *Proc. Ann. Meeting Am. Assoc. Cancer Res.* **1996**, *37*, 299.
(31) Harris, C. C. p53 tumor suppressor gene: from the basic research laboratory to the clinic − an abridged historical perspective. *Carcinogenesis* **1996**, *17*, 1187−1198.
(32) Hollstein, M.; Sidransky, D.; Vogelstein, B.; Harris, C. C. p53 mutations in human cancers. *Science* **1991**, *253*, 49−53.
(33) O'Connor, P. M.; Jackman, J.; Bae, I.; Myers, T. G.; Fan, S.; Mutoh, M.; Scudiero, D. A.; Monks, A.; Sausville, E. A.; Weinstein, J. N.; Friend, S.; Fornance, J., A. J.; Kohn, K. W. Characterization of the p53-tumor suppressor pathway in cell lines of the National Cancer Institute anticancer drug screen and correlations with the growth-inhibitory potency of 123 anticancer agents. *Cancer Res.* **1997**, *57*, 4285−4300.
(34) Benhattar, J.; Cerottini, J. P.; Saraga, E.; Metthez, G.; Givel, J. C. p53 mutations as a possible predictor of response to chemotherapy in metastatic colorectal carcinomas. *Int. J. Cancer (Pred. Oncol.)* **1996**, *69*, 190−2.
(35) Shi, L. M.; Myers, T. G.; Fan, Y.; O'Conner, P. M.; Paull, K. D.; Friend, S. H.; Weinstein, J. N. Mining the NCI anticancer drug discovery database: cluster analysis of ellipticine analogs with p53-inverse and CNS-selective patterns of activity. *Mol. Pharmacol.* in press.
(36) Goodwin, S.; Smith, A. F.; Horning, E. C. Alkaloids of Ochrosia elliptica Labill. *J. Am. Chem. Soc.* **1959**, *81*, 1903−1908.
(37) Dalton, L. K.; Demerac, S.; Elmes, B. C.; Lorder, J. W.; Swan, J. M.; Teitel, T. Synthesis of the tumor-inhibitory alkaloids, ellipticine, 9-methoxyellipticine, and related pyrido[4,3-b]carbazoles. *Aust. J. Chem.* **1967**, *20*, 2715−2727.
(38) Gribble, G. W. In *The Alkaloids: Chemistry and Pharmacology*; Brossi, A., Ed.; Academic: San Diego, 1990; Vol. 39.
(39) Acton, E. M.; Narayanan, V. L.; Risbood, P. A.; Shoemaker, R. H.; Vistica, D. T.; Boyd, M. R. Anticancer specificity of some ellipticinium salts against human tumors in vitro. *J. Med. Chem.* **1994**, *37*, 2185−2189.
(40) Anderson, W. K.; Gopalsamy, A.; Reddy, P. S. Design, synthesis and study of 9-substituted ellipticine and 2-methylellipticinium analogues as potential CNS-selective antitumor agents. *J. Med. Chem.* **1994**, *37*, 1955−1963.
(41) Jurayi, J.; Haugwitz, D.; Varma, R. K.; Paull, K. D.; Barrett, J. F.; Cushman, M. Design and synthesis ellipticinium salts and 1,2-dihydroellipticines with high selectivities against human CNS cancers in vitro. *J. Med. Chem.* **1994**, *37*, 2190−2197.
(42) Devraj, R.; Jurayj, J.; Fernandez, J. A.; Barrett, J. F.; Cushman, M. Synthesis of a series of cytotoxic 2-acyl-1,2-dihydroellipticines which inhibit topoisomerase II. *Anti-Cancer Drug Design* **1996**, *11*, 311−324.
(43) Devraj, R.; Barrett, J. F.; Fernandez, J. A.; Katzenellenbogen, J. A.; Cushman, M. Design, synthesis, and biological evaluation of ellipticine-estradiol conjugates. *J. Med. Chem.* **1996**, *39*, 3367−3374.
(44) Shimamoto, T.; Imajo, S.; Honda, T.; Yoshimura, S.; Ishiguro, M. Structure−activity relationship study on N-glycosyl moieties through model building of DNA and ellipticine N-glycoside complex. *Bioorg. Med. Chem. Lett.* **1996**, *6*, 1331−1334.
(45) Vistica, D. T.; Kenney, S.; Hursey, M. L.; Boyd, M. R. Cellular uptake as a determinant of cytotoxicity of quaternized ellipticines to human brain tumor cells. *Biochem. Biophys. Res. Commun.* **1994**, *200*, 1762−1768.
(46) Kenney, S.; Vistica, D. T.; Linden, H.; Boyd, M. R. Uptake and cytotoxicity of 9-methoxy-N2-methylellipticinium acetate in human brain and non-brain tumor cell lines. *Biochem. Pharmacol.* **1995**, *49*, 23−32.
(47) Kohn, K. W.; Waring, M. J.; Glaubiger, D.; Friedman, C. A. Intercalative binding of ellipticine to DNA. *Cancer Res.* **1975**, *35*, 71−76.
(48) Sainsbury, M. In *The Chemistry of Antitumor Agents*; Wilman, D., Ed.; Chapman and Hall: New York, 1990.
(49) Froelich, A. S.; Patchan, M. W.; Osheroff, N.; Thompson, R. B. Topoisomerase II binds to ellipticine in the absence or presence of DNA. Characterization of enzyme-drug interactions by fluorescence spectroscopy. *J. Biol. Chem.* **1995**, *270*, 14998−15004.
(50) Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A.; Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.* **1992**, *114*, 10024−10035.
(51) Rappe, A. K.; Goddard, W. A. *J. Phys. Chem.* **1991**, *95*, 3358.
(52) Hansch, C.; Leo, A.; Hoekman, D. *Exploring QSAR, vol. 1. Fundamentals and Applications in Chemistry and Biology; vol. 2. Hydrophobic, Electronic, and Steric Constants*; American Chemical Society: Washington, D.C., 1995.
(53) *Chemometric Methods in Molecular Design*; van de Waterbeemd, H., Ed.; VCH: Weinheim, 1995; Vol. 2.
(54) *Advanced Computer-Assisted Techniques in Drug Discovery*; van de Waterbeemd, H., Ed.; VCH: Weinheim, 1995; Vol. 3.
(55) *Structure−Property Correlations in Drug Research*; van de Waterbeemd, H., Ed.; Academic and R. G. Landes: San Diego, 1996.
(56) 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.
(57) Rogers, D. In *Genetic Algorithms in Molecular Modeling*; Devillers, J., Ed.; Academic: London, 1996.
(58) Friedman, J. H. *Multivariate Adaptive Regression Splines*, Technical Report No. 102, Nov/Rev.1990 Aug; Laboratory of Computational Statistics, Department of Statistics, Stanford University: Standford, CA, 1988.
(59) Friedman, J. H. Multivariate adaptive regression splines (with discussion). *Ann. Stat.* **1991**, *19*, 1−141.
(60) Holland, J. *Adaptation in Artificial and Natural Systems*; University of Michigan: Ann Arbor, MI, 1975.
(61) Shi, L. M.; Myers, T. G.; Fan, Y.; Weinstein, J. N. Fifth Conference on Current Trends in Computational Chemistry (CCTCC), Vicksburg, Mississippi, 1996; pp 131−135.
(62) Shi, L. M.; Fan, Y.; Myers, T. G.; Weinstein, J. N. *Proceedings of the 1997 International Conference on Neural Networks (ICNN'97)*, Houston, Texas, 1997; pp 2490−2493.
(63) *StatSci S.-PLUS Reference Manual*; MathSoft: Seattle, WA, 1993.

CI970085W