Quantitative Structure-Antitumor Activity Relationships of Camptothecin Analogues: Cluster Analysis and Genetic Algorithm-Based Studies

Yi Fan,† Leming M. Shi,‡ Kurt W. Kohn, Yves Pommier, and John N. Weinstein*

Laboratory of Molecular Pharmacology, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892

Received December 4, 2000

Topoisomerase 1 (top1) inhibitors are proving useful against a range of refractory tumors, and there is considerable interest in the development of additional top1 agents. Despite crystallographic studies, the binding site and ligand properties that lead to activity are poorly understood. Here we report a unique approach to quantitative structure−activity relationship (QSAR) analysis based on the National Cancer Institute's (NCI) drug databases. In 1990, the NCI established a drug discovery program in which compounds are tested for their ability to inhibit the growth of 60 different human cancer cell lines in culture. More than 70 000 compounds have been screened, and patterns of activity against the 60 cell lines have been found to encode rich information on mechanisms of drug action and drug resistance. Here, we use hierarchical clustering to define antitumor activity patterns in a data set of 167 tested camptothecins (CPTs) in the NCI drug database. The average pairwise Pearson correlation coefficient between activity patterns for the CPT set was 0.70. Coherence between chemical structures and their activity patterns was observed. QSAR studies were carried out using the mean 50% growth inhibitory concentrations (GI_{50}) for 60 cell lines as the dependent variables. Different statistical methods, including stepwise linear regression, principal component regression (PCR), partial least-squares regression (PLS), and fully cross-validated genetic function approximation (GFA) were applied to construct quantitative structure-antitumor relationship models. For our data set, the GFA method performed better in terms of correlation coefficients and cross-validation analysis. A number of molecular descriptors were identified as being correlated with antitumor activity. Included were partial atomic charges and three interatomic distances that define the relative spatial dispositions of three significant atoms (the hydroxyl hydrogen of the E-ring, the lactone carbonyl oxygen of the E-ring, and the carbonyl oxygen of the D-ring). The cross-validated r^2 for the final GFA model was 0.783, indicating a predictive QSAR model.

Introduction

Camptothecin (CPT) topoisomerase 1 (top1) inhibitors are proving useful against a range of refractory tumors, most prominently against some colon and ovarian cancers.[1−3] Two of the CPTs, topotecan and CPT-11, have received Food and Drug Administration approval, and several others are in clinical trials. The continuing interest in development of better top1 inhibitors prompted us to analyze structure−activity relationships involving the binding site of top1 and the presumed ternary cleavable complex of top1 with its inhibitors and DNA. Despite recent crystallographic structures for top1, its complex binding site is poorly understood, and the structural characteristics of a ligand that promote potency have been only partially determined.[4,5] Here we report an unusual approach to quantitative structure−activity relationship (QSAR) analysis: the large drug activity databases generated over the last 11 years by the National Cancer Institute (NCI) are used in conjunction with cluster analysis and a genetic algorithm-based method for nonlinear analysis to predict functionally important molecular features.

Since 1990, the NCI has screened >70 000 chemical compounds against a panel of 60 human cancer cell lines.[6−9] The 50% growth inhibitory concentration (GI_{50}) for any particular cell line is an index of cytotoxicity or cytostasis. Similarity in GI_{50} activity patterns across the 60 cell lines very often indicates similarity in mechanism of action, mode of drug resistance, and molecular structure of tested compounds.[10,11] A number of different algorithms have been used to study the GI_{50} activity patterns.[11,12] The COMPARE program[6,11] developed by K. D. Paull uses statistical correlation to find agents with activity patterns across the 60 cell lines similar to that of a "seed" compound. Back-propagation neural networks,[13] Kohonen self-organizing maps,[14] principal component regression,[12,15] multidimensional scaling,[12] hierarchical cluster analysis,[12,16] and clustered image maps (CIMs)[11,16−19] have been used to predict mechanism of drug action or to cluster compounds or cell lines based on activity patterns. The CIM has proved a particularly useful tool for visualization of patterns in high-dimensional data sets such as these. This overall "information-intensive" approach to molecular pharmacology has demonstrated that the patterns of GI_{50} values are useful for identifying subgroups of compounds

* To whom correspondence should be addressed: LMP/CCR, NCI, NIH, Bethesda, MD 20892. Phone: 301-496-9571. Fax: 301-402-0752. E-mail: weinstein@dtpax2.ncifcrf.gov.

† Present address: Wyeth-Ayerst Research, CN8000, Princeton, NJ 08543-8000. E-mail: fank@war.wyeth.com.

‡ Present address: ChipScreen BioSciences, Ltd., Shenzhen, China.

10.1021/jm0005151
Published on Web 08/31/2001

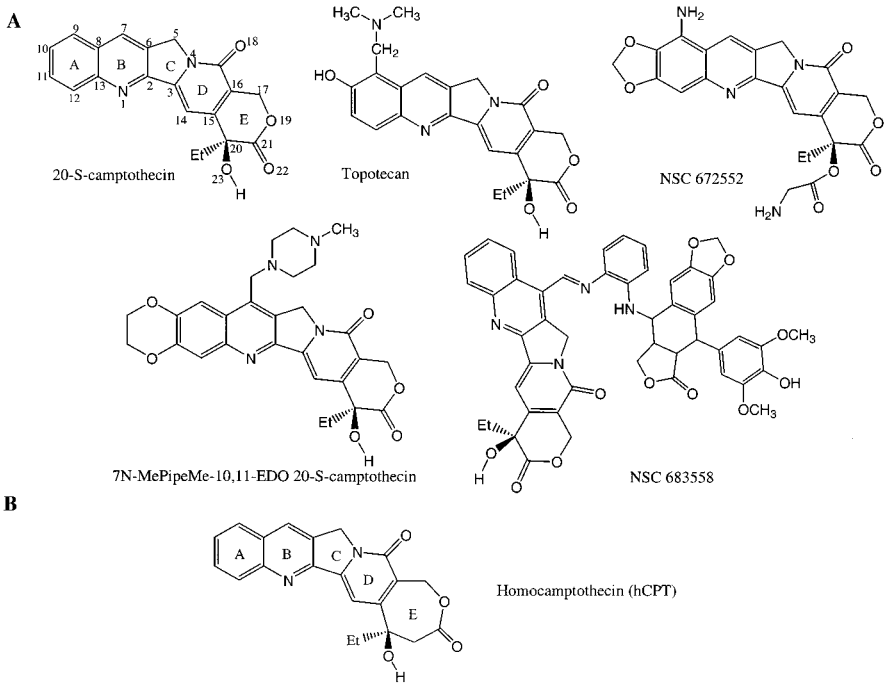

Figure 1. (A) The numbering system for the camptothecin core structure and examples of the training data set used for clustering and QSAR analysis. (B) The structure of homocamptothecin (hCPT).

related to particular biological targets and for investigating the mechanisms of action of screened compounds.

For this study, we identified 167 camptothecin analogues among the compounds in the NCI Drug Information System (DIS) database. There has been renewed interest in this class of compounds for anticancer drug development since the demonstration that camptothecin acts by selectively inhibiting eukaryotic topoisomerase I (top I) and since the identification of several clinically active derivatives,[20,21] including 9-amino CPT (NSC-603071), topotecan (NSC-609699), and CPT-11 (NSC-616348). Camptothecin analogues have been characterized by numerous research groups,[22−28] notably by Wall and Wani. These studies have provided the basis for our present understanding of structure−activity relationships among camptothecin analogues. Significant factors include the stereospecificity at the 20-position (20*S* CPT being active, whereas 20*R* CPT is not), the activating effects of substituents at the 9- and 10- positions of the A-ring, the inactivating effects of substituents at the 11- and 12- positions of the A-ring, substituent effects at the 7-position of the B-ring, and the role of the E-ring lactone in antitumor activity[23,25] (see Figure 1A). Despite enormous efforts in this area, however, many aspects of the cytotoxicity and antitumor activity of camptothecins remain unclear. We recently proposed a hypothetical computer model for the formation of a ternary cleavable complex of top1, DNA, and camptothecin.[29] Simultaneously, a complementary model based on the crystallographic resolution of the top1−DNA complex was reported.[4] These studies provided a plausible explanation for the observed stabilization of the DNA−top1 cleavable complex by CPT and its derivatives.

More generally, QSAR[30,31] began with the pioneering work of Hansch, who used multiple linear regression (MLR) to build predictive models of the biological activity of a series of compounds. However, MLR cannot be used when there are more descriptors than compounds (i.e., when the problem is overdetermined). More recently, PLS (partial least-squares regression) has been invoked to reduce the number of variables and optimize them, for example, in comparative molecular field analysis (COMFA).[32] Here we have used an alternative approach, genetic function approximation (GFA), developed by Rogers and Hopfinger,[33,34] and compared its results with those obtained by stepwise regression, principal component regression (PCR), and PLS.

Method

The Data Set. We searched the NCI DIS database of ∼460 000 compounds for CPT analogues and identified 167 that had been screened against the 60 cell lines. For cluster analysis, we added a number of compounds found previously to have top1 activity. Included were four saintopin analogues and one nitidine analogue. Two VP-16 analogues that are topoisomerase 2 (top2) inhibitors were also included because we were interested in the activity patterns of four so-called "bridge compounds", in which the 7-position of CPT was substituted by VP-16. Activity data for these 174 compounds included 4% missing values, each of which was replaced by the mean value over all remaining cell lines for the compound in question. These compounds are listed by NSC number at http://discover.nci.nih.gov.

Fifty-eight of the 167 CPT analogues (see Figure 1A) were selected for QSAR analysis on the basis of the clear characterization of stereochemistry at the 20-position. Since orientation at the 20-position is known to be especially important, the compounds were removed if their stereochemistry at the 20-position of CPT was not specified or if they were submitted as racemic mixtures. Several compounds, including 9-amino-20*R*-CPT (NSC 639173) and 12-nitro-17-hydroxy-camptothecin (NSC 684918) were excluded, because their activity was too

weak to produce potency values above threshold in the screening cell lines.

Structures for all camptothecin molecules were built using the Cerius[2] molecular modeling package (Molecular Simulations, Inc., San Diego, CA).[35] Each structure was energy-minimized with a convergence criterion of 0.01 kcal/mol, using the universal force field developed by Rappé and co-workers.[36] Partial atomic charges were computed by an equilibration approach.[37] The MCSG (maximum common subgroup) method in Cerius[2] was used to superimpose the molecules in the series.

Cluster Analysis. The patterns of activity across 60 cell lines were analyzed using the "hclust" (hierarchical clustering) function implemented in the S-Plus statistical package (StatSci Division, MathSoft, Inc., Seattle, WA).[38] Compounds were clustered on the basis of their patterns of GI_{50} values. For this study, we used the "average linkage" clustering algorithm and distance metric $(1 - r)$, where r is the Pearson correlation coefficient.

Genetic Function Approximation (GFA). The genetic function approximation (GFA) method developed by Rogers and Hopfinger[33,34] was used to derive QSAR models. This method combines Holland's genetic algorithm (GA)[39] with Friedman's multivariate adaptive regression splines (MARS).[40,41]

In addition to linear and quadratic terms for each descriptor variable, the MARS algorithm provides "truncated power spline" terms for construction of regression models. A spline term can be of 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. The spline term introduces nonlinearity into the regression model. It can provide high levels of accuracy, and MARS often competes well with neural network approaches, given moderate numbers of descriptors. However, the algorithm becomes computationally intensive with a large number of descriptors, e.g. more than 20.

GFA uses a GA to search the MARS descriptor space to evolve multiple models that best fit the training data. Two advantages can be expected: (i) the GA searches the MARS descriptor space efficiently, and (ii) it can find models containing combinations of descriptors or features that predict well as a group but poorly individually. The GFA procedure as used in this study was as follows: (i) An initial population of 100 equations was generated by a random choice of descriptors and basis functions (linear, quadratic, or spline). For each spline term, the initial knot was randomly initialized and later optimized during GFA evolution. (ii) Pairs of "parent" equations were chosen randomly from the set of 100, and "crossover" operations were performed at randomly chosen points within the equations to produce "progeny" models that contained characteristics of both parents. Because the crossover points in the two equations were allowed to differ, progeny equations could have more or fewer terms than the parents. (iii) The "goodness" of each progeny model was assessed by a fitness function using Friedman's lack of fit (LOF) measure, which assigns each equation of the population a score

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

where LSE is the least-squares error, c is the number of basis functions in the model, d is a smoothing parameter, p is the number of descriptors used in the model, and m is the number of observations in the training set. The LOF score resists overfitting by penalizing for the addition of terms (i.e., descriptors and basis functions). The smoothing parameter in the equation allows user control over the amount of penalty imposed. In our calculations, the smoothing parameter d was set to the default value of unity. (iv) If the new equation's fitness score (LOF) was among the top 100, it was kept, and equation number 100 was dropped; otherwise the progeny equation was discarded. Crossover steps ii−iv were repeated a preset number of times. The process selects models with improved performance by recombination of terms. The evolution from a population of randomly constructed models can thus lead to the discovery of highly predictive QSARs.

Results and Discussion

Camptothecin Antitumor Activity and Topoisomerase I Inhibition. The activity values from the NCI drug discovery program were used in both cluster analysis and QSAR studies. Activity is expressed as the quantity $-\log(GI_{50})$, where GI_{50} is the 50% growth inhibitory concentration compared with untreated controls. For each compound, 60 activity values (one for each cell line) constitute the activity pattern or "fingerprint".

We have investigated the activity profiles of all 167 CPT analogues in the database. These include 20*R* CPT derivatives, compounds with substituents on the A-ring and B-ring, and some 14- and 17-substituted analogues. In general, the activity for any single cell line is simply an index of cytotoxicity or cytostasis. It reflects an in vitro summation of effects that might arise from multiple mechanisms of action under cell culture conditions. Interestingly, recent studies show that the activity profiles of cell lines revealed the same trends as were found in various structure−activity relationship studies of top1.[4,22,23,25,26,42−44] For example, (i) 20*R* CPT is essentially inactive, whereas the 20*S* CPT is highly active; (ii) 21-lactam *S*-camptothecin is inactive; (iii) substitution at the 7-, 9-, or 10-position of most CPT analogues enhances antitumor activity, and small substituents at position 11 are allowed, whereas addition at position 12 is inactivating; (iv) activity is retained when a methoxy group is added at position 10 or 11, and addition of a methylenedioxy group to form a five-membered ring across positions 10 and 11 (10,11-MDO CPT) enhances potency; however the simultaneous addition of methoxy groups to positions 10 and 11 (10,11-DMO CPT) is inactivating.

Cluster Analysis. Our previous studies[11−19,45] have demonstrated coherent mapping between chemical structure and in vitro cell screen activity patterns. Here, we also find that the in vitro cell screen activity patterns reflect the biological behavior of tested compounds. Four saintopin analogues, two etoposide analogues, and one nitidine analogue were included in the data set. Two of the saintopins (UCE6 and UCE1022) have been shown in biochemical assays to be top1 agents; the other two saintopins show both top1 and top2 inhibition. The nitidine analogue has been reported to be a top1 agent,[24]

and the two etoposide analogues are top2 agents.[46] We wished to test whether cluster analysis of the in vitro cell screen activity patterns could distinguish these compounds from one another on the basis of mechanism of action. The cluster tree (distance metric, $1 - r$; clustering method, average linkage) for a total of 174 compounds is shown in Figure 2.

In general, we found that compounds similar in chemistry and presumed mechanism of action tended to group together. The camptothecin analogues clustered side by side in the cluster tree, as shown in Figure 2. However, compounds 1 and 2 in the tree were very different from the rest of the CPT analogues in terms of their activity patterns. These two compounds, 12-nitro-17-hydroxy-CPT (**1**) and 9-amino-20R−CPT (**2**), were essentially inactive in the screen; only one or two of the 60 cell lines were sensitive enough for 50% growth inhibition at the highest concentration tested (hiconc = 10^{-4} mol/L). In other words, there was not enough information encoded in the patterns to characterize the biological behavior of these two compounds. We treated them as outliers in the QSAR analysis that will be discussed later. With the exclusion of compounds **1** and **2**, as well as **3** and **4** (etoposide analogues), the average Pearson correlation coefficient (r) for all pairwise relationships between activity patterns for the rest of the CPT data set was 0.703 (SD = 0.181). This observation indicates the unique pattern of antitumor activity for CPT analogues and perhaps reflects the current view that CPTs act by a single and specific primary mechanism, top1 inhibition.

There were two major branches in the cluster tree shown in Figure 2: **1**−**167** and **168**−**174**. The first group consisted almost entirely of CPTs. A group of very potent CPTs (**34**−**52**) and a set of saintopin derivatives (**16**−**18**, **20**) were among the large group. The activity patterns of the four saintopin derivatives were similar to those of the CPTs. It was known on the basis of previous biochemical assay data that UCE6 (**17**) and UCE1022 (**18**) are top1 agents; the other two saintopins have been found to have both top 1 and top 2 activities.[46]

In the four "bridge compounds" (NSC 683555−683558), the 7-position of CPT was substituted by etoposide. These compounds (**76**−**79**), which formed one small subgroup within the middle branch of the cluster tree, showed activity patterns quite similar to that of topotecan (NSC 609699, **61**) ($r > 0.82$) but very different from those of the etoposide derivatives (**3** and **4**). It is possible that the "bridge compounds" were hydrolyzed under cell culture conditions and that the top1 activity dominated because the potency of CPT on a molecular basis is greater than that of etoposide—or that the top1 dominated without hydrolysis. Another possible explanation is that the etoposide moiety was released in an inactive form. The small subgroup (**168**−**174**) consisted entirely of relatively inactive 5-substituted CPTs. In summary, the cluster analysis demonstrated, even at a "micro" level, that the activity patterns can encode incisive information about the selective cytotoxicity of compounds and their mechanisms of action.

Biochemical assays of top1 inhibition have shown inactivation of CPT when the 20-OH is substituted with $OCOCH_2NH_3^+$,[17] but most such compounds in our data set were highly potent in the cell screen. This apparent discrepancy suggests that these compounds are converted to the normal CPT analogues by the hydrolytic reaction shown in Figure 3A. Therefore, we will refer to these CPTs hence forth as prodrugs. Most of the prodrugs in the data set have a 20-OH substituted by $OCOCH_2R$. CPT-11 is also a prodrug by virtue of substitution at the 10-position (See Figure 3B).[47,48] The activity patterns of these prodrugs appeared similar to those of the corresponding normal CPTs, suggesting that they can be converted efficiently under tissue culture condition.

Table 1. The Choice of Molecular Descriptors

model	no. of descriptors	LOF	r^2	CV r^2	description[a]
1	49	1.057	0.415	0.344	default
2	73	0.632	0.738	0.642	default + Q
3	77	0.496	0.795	0.706	default + Q + D
4	132	0.451	0.800	0.702	all descriptors
5	43	0.406[b]	0.812[b]	0.735[b]	descriptor subset

[a] "Default" refers to the 49-default descriptor database suggested by the Cerius[2] program. "Q" refers to 24 atomic charges generated for core atoms of CPT. "D" refers to four interatomic distances between atoms of the particular functional groups of CPT. [b] There is no direct comparison between these numbers and the numbers in other tables.

QSAR Model for CPT Analogues. In this study, we used the genetic function approximation (GFA) method to construct QSAR models based on 58 CPT analogues (see Figure 1A). The compound selection criteria were described in the Methods section. The mean activity value across 60 cancer cell lines was taken as the dependent variable to be predicted. For each calculation, we did 100 000 crossover steps, after which LOF scores for the 100 final models remained almost unchanged, indicating convergence of the calculation. Both randomization tests and full cross-validation procedures showed the QSAR model to be predictive. Homocamptothecin, a novel CPT analogue that differs from CPT by the presence of an additional methylene group in the E-ring (see Figure 1B), was predicted to be among the more active compounds by our GFA models. The QSAR results can be summarized as follows.

(1) Significance of Molecular Descriptors. The Cerius[2] QSAR+ module provides more than 160 descriptors divided into seven categories: conformational, electronic, receptor, quantum mechanical, shape, spatial, thermodynamic, information, and topological. Among these, 49 molecular descriptors constitute a "default" set. Using this default set, we did not obtain any good QSAR models. The average cross-validated (CV) r^2 was only 0.344 (see Table 1). Therefore, the descriptor set was extended to 132 descriptors, including (i) all applicable 2D and 3D descriptors in the Cerius[2] QSAR+ package, (ii) 24 partial atomic charges on CPT core atoms, and (iii) four "pharmacophoric" distances (the interatomic distances between significant functional atoms present in molecules of the data set). With these additions, the models were greatly improved. The results are shown in Table 1.

To investigate how well GFA performed, we also evaluated descriptors in the extended descriptor set by examining how they correlated with the mean activity values. Descriptors that did not correlate with mean activity values were excluded from the set. We then used

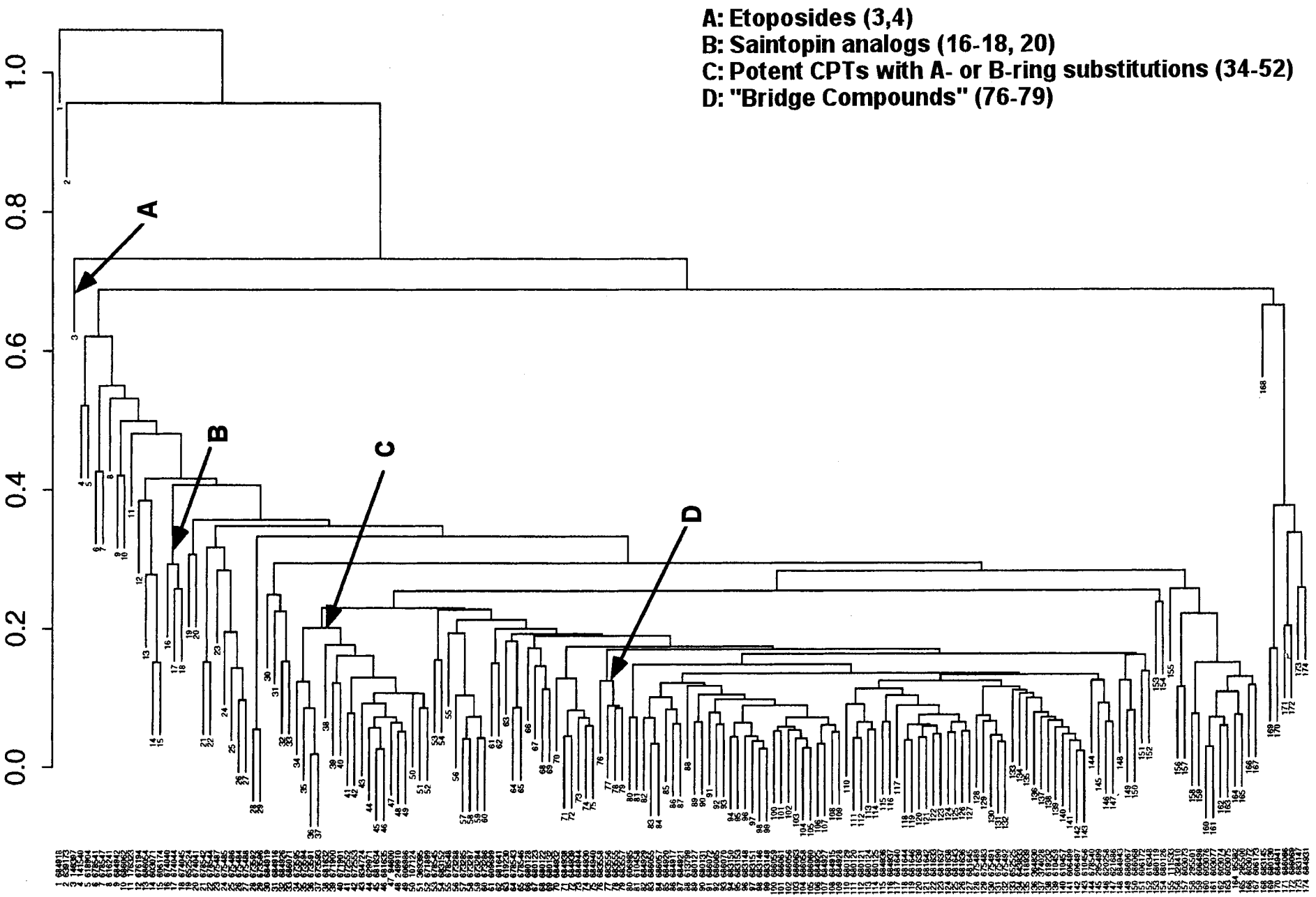

Figure 2. Hierarchical cluster analysis of 167 camptothecin analogues and seven related compounds based on their activity patterns across 60 human cancer cell lines. The compounds are numbered in cluster order, and the NSC numbers are shown at the bottom. The average linkage algorithm with correlation metric was used.

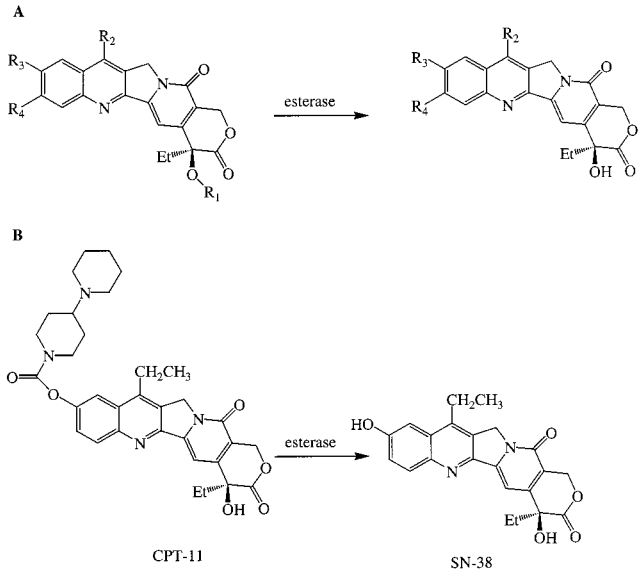

Figure 3. (A) Conversion of 20-ester prodrugs to camptothecin. (B) Conversion of CPT-11 to SN-38.

Table 2. Performance of GFA Models and the Models Generated by Other Statistical Methods

method	r^2	CV r^2	description
GFA	0.805	0.783	
forward stepwise regression	0.695	0.609	$F = 4.0$[a]
backward stepwise regression	0.865	0.635	$F = 4.0$[a]
backward stepwise regression	0.783	0.671	$F = 6.0$[a]
PLS	0.622	0.486	no. of components 2
PCR	0.597	0.489	no. of components 6

[a] The number of variables in the QSAR equation from forward stepwise regression analysis was 4, whereas the numbers of variables maintained in backward stepwise regression with $F = 4.0$ and $F = 6.0$ were 16 and 9, respectively.

the remaining 43 descriptors. The value of CV r^2 for the best QSAR model generated using the set of 43 was 0.735, slightly better than that 0.702 obtained using 132 descriptors. Overall, the QSAR model based on the 132-descriptor set and the one based on the 43-descriptor set had similar qualities. As long as the six or seven descriptors that showed up in the final model were included in the descriptor set, it made little difference what other descriptors were also included. GFA efficiently optimized the model and gave reasonably good values of CV r^2 with low LOF. The 43-descriptor set was used in all further GFA calculations.

(2) Performance of GFA Models. The GFA algorithm offers a new nonlinear approach to the construction of QSAR models. For comparison, stepwise regression, principal component regression (PCR), and PLS were also performed on the same data and the same 43-descriptor set. The results are compiled in Table 2, along with the results obtained by GFA. Reasonable performance was obtained by a cross-validated forward stepwise regression procedure (with $F = 4.0$ as the threshold value for adding variables). The r^2 and CV r^2 were 0.695 and 0.609, respectively, as compared with 0.805 and 0.783 for the GFA method. The r^2 was 0.865 for backward stepwise regression analysis with $F = 4.0$. However, the CV r^2 was only 0.635. Only four variables were used in the model generated by forward stepwise regression, whereas 16 variables were used in that generated by backward stepwise regression with $F = 4.0$. For backward stepwise regression with $F = 6.0$, r^2 and CV r^2 were 0.783 and 0.671, respectively, but nine variables still remained in the QSAR model. There might be an overfitting problem for the backward stepwise regression analysis performed here.

The QSAR models from forward and backward stepwise regression are shown in Table 4. The molecular descriptors included in the equation by forward stepwise regression were also ones frequently obtained by GFA. PLS yielded an r^2 of 0.621, comparable with that of stepwise regression. However, CV r^2 was only 0.486 with two components used; hence, the PLS model was not very predictive. As to the PCR analysis with six components, both r^2 (0.597) and CV r^2 (0.489) were poor. We believe that the superior performance of GFA was due largely to the inclusion of spline terms in building QSAR models, because the spline terms permit modeling of nonlinearities.[33,34,40,41]

The top eight models for prediction of mean activities generated using GFA are listed in Table 3. The most frequently used descriptors in the population of 100 best QSAR models were partial atomic charges at the 11- and 12-positions of the A-ring and three interatomic distances that reflect pharmacophoric patterns involving the D- and E-rings (see Figure 4). These results are consistent with those of our earlier molecular modeling studies,[29] indicating that three functional groups (oxygen of 20-OH, oxygen of 21-C=O in the E-ring, and 18-O in the D-ring) are important for inhibitory activity. The

Table 3. Summary of the Eight Best GFA Models for the Optimized 43-Descriptor Set

model no.	QSAR equation[a]
1	$Y = 4.49 + 3.45Q_{11} - 25.85\langle D_{18-22} - 6.11\rangle - 4.95Q_{12} - 17.66\langle 5.86 - D_{18-23}\rangle - 23.15Q_5 + 226.68\langle 0.20 - Q_{20}\rangle$ $r^2 = 0.805$, CV $r^2 = 0.783$, LOF = 0.461
2	$Y = 3.55 + 2.90D_{22-23} + 3.33Q_{11} - 4.96Q_{12} - 21.26Q_5 - 27.08\langle D_{18-22} - 6.11\rangle - 17.57\langle 5.86 - D_{18-23}\rangle$ $r^2 = 0.790$, LOF = 0.474
3	$Y = 3.48 + 2.89D_{22-23} + 3.32Q_{11} - 4.96Q_{12} - 17.65\langle 5.86 - D_{18-23}\rangle - 21.07Q_5 - 27.02\langle D_{18-22} - 6.11\rangle$ $r^2 = 0.790$, LOF = 0.474
4	$Y = 0.94 + 0.46D_{22-23}^2 + 3.34Q_{11} - 4.96Q_{12} - 21.39Q_5 - 27.00\langle D_{18-22} - 6.11\rangle - 17.56\langle 5.86 - D_{18-23}\rangle$ $r^2 = 0.790$, LOF = 0.474
5	$Y = -3.47 + 3.32Q_{11} - 4.95Q_{12} - 20.99Q_5 + 2.88D_{22-23} - 27.14\langle D_{18-22} - 6.11\rangle - 17.65\langle 5.86 - D_{18-23}\rangle$ $r^2 = 0.790$, LOF = 0.474
6	$Y = 0.98 + 0.46D_{22-23}^2 + 3.33Q_{11} - 4.95Q_{12} - 17.64\langle 5.86 - D_{18-23}\rangle - 21.20Q_5 - 26.94\langle D_{18-22} - 6.11\rangle$ $r^2 = 0.790$, LOF = 0.475
7	$Y = 0.99 - 21.12Q_5 + 0.46D_{22-23}^2 + 3.33Q_{11} - 27.06\langle D_{18-22} - 6.11\rangle - 4.95Q_{12} - 17.64\langle 5.86 - D_{18-23}\rangle$ $r^2 = 0.790$, LOF = 0.475
8	$Y = -3.37 + 2.86D_{22-23} + 3.30Q_{11} - 4.96Q_{12} - 26.93\langle D_{18-22} - 6.11\rangle - 20.75Q_5 - 17.78\langle 5.86 - D_{18-23}\rangle$ $r^2 = 0.790$, LOF = 0.475

[a] Q refers to atomic charge and D refers to the interatomic distance between two atoms.

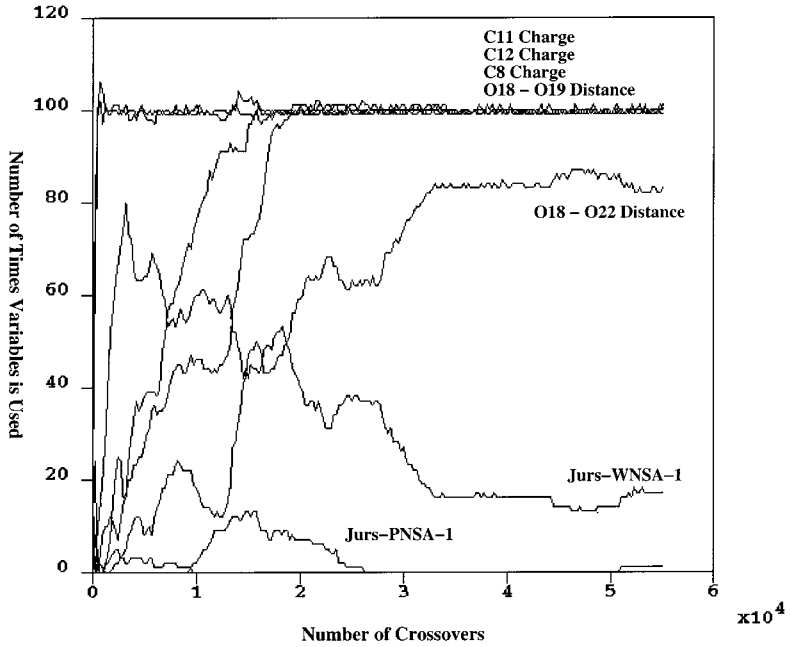

Figure 4. Frequency of descriptor use for the GFA model. The six descriptors used in the best model shown in Table 3. These descriptors are highly represented in the final 100 models, whereas other descriptors are rarely used. These observations indicate convergence of the GFA process.

Table 4. Summary of QSAR Models Generated by Forward and Backward Stepwise Regression

method	QSAR equation[a]
forward stepwise $F = 4.0$	$Y = 58.07 - 0.86$"RadOfGyration" $+ 3.55Q_{11} - 5.91Q_{12} - 11.61D_{18-19}$ $r^2 = 0.695$, CV $r^2 = 0.635$, no. of variables = 4
backward stepwise $F = 4.0$	$Y = 17.39 + 8.41$"IAC-Mean" $+ 4.01$"Kappa-3" $- 4.04$"Kappa-3-AM" $- 12.12$"Density" $+ 36.86$"Jurs-RASA" $- 0.07$"Jurs-TASA" $+ 7.53Q_{11} - 5.17Q_{12} - 13.97Q_4 - 65.45Q_{17} + 22.69Q_8 + 18.60Q_{24} + 25.95D_{22-23} - 28.28D_{18-22} - 6.24D_{23-24} + 28.16D_{25-23}$ $r^2 = 0.865$, CV $r^2 = 0.635$, no. of variables = 16
backward stepwise $F = 6.0$	$Y = 15.14 + 8.41$"Jurs-RASA" $- 0.02$"Jurs-TASA" $+ 4.98Q_{11} - 16.62Q_4 + 15.08Q_8 - 15.87D_{18-19} + 16.05D_{22-23} - 8.45D_{24-22} + 16.31D_{25-23}$ $r^2 = 0.783$, CV $r^2 = 0.671$, no. of variables = 9

[a] Q refers to atomic charge and D refers to the interatomic distance between two atoms in the QSAR equation.

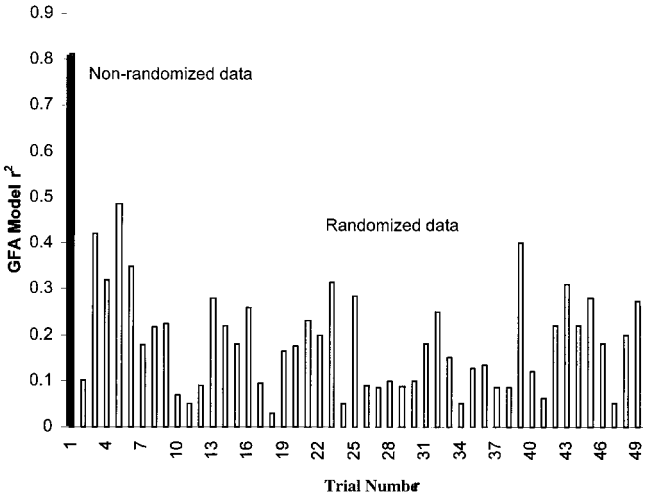

Figure 5. GFA Randomization test. The first bar (solid) shows the r^2 value for the model based on the actual dataset; the other 48 bars (open) show r^2 for 48 models based on permuted data.

electrostatic interaction of the A-ring with the binding site could also be important.

(3) Randomization Tests and Full Cross-Validation Test. To be useful, a QSAR model must be predictive so that it can provide estimates of the activity of untested compounds similar to those in the data set used to construct the model. To determine the model's reliability and significance, both randomization and full cross-validation procedures were performed.

The randomization was done by repeatedly permuting the dependent variable set (i.e., the mean activity data). If the score of the original QSAR model proved better than those from the permuted data sets, the model would be considered statistically significant. The results of 49 randomization tests are presented in Figure 5. The correlation coefficient, r^2, for the nonrandom QSAR model was 0.805, significantly better than those obtained from randomized data (mean $r^2 = 0.266$, SD = 0.128). None of the 49 permuted sets produced an r^2 comparable with 0.805; hence, the value obtained for the original GFA model for the 58 compounds could be considered significantly different from zero with $p < 0.05$.

A full cross-validation has also been done for one of the best GFA models. Standard cross-validation in GFA encompasses only the optimization of regression coefficients; it does not encompass optimization of the choice of descriptors. That is, the regression model is validated only for the specific subset of descriptors obtained from GFA. In contrast, full cross-validation encompasses the entire algorithm, including both the choice of descriptors and the optimization of regression coefficients. For the jackknife "leave-1-out" rule, each full cross-validation step finds the best subset of descriptors for a training set of $N - 1$ compounds. Here, the full CV r^2 was computed using the predicted values of the missing molecules. The results based on the rules of "leave-1-

Table 5. Results of Full Cross-Validation

rule	PRESS	ΣSD	CV r^2
leave-1-out	20.29	76.74	0.783
leave-2-out	20.17	79.74	0.747
leave-5-out	21.82	79.74	0.726
leave-7-out	20.45	79.74	0.744
leave-10-out	29.28	79.76	0.633

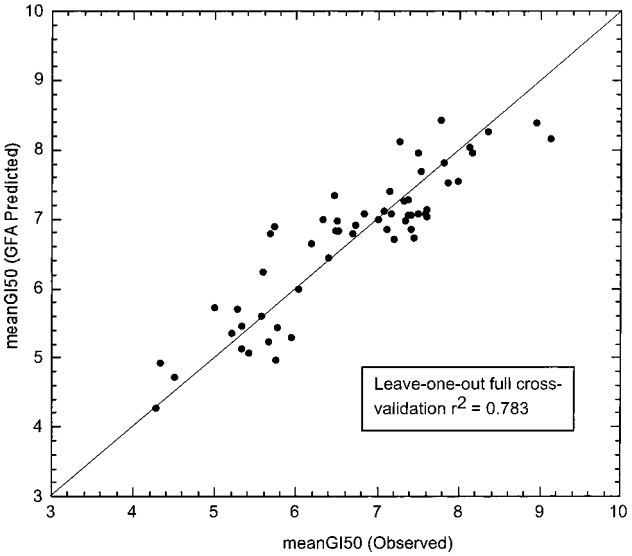

Figure 6. Full cross-validation test. The entire algorithm (including both the variable selection and fitting steps) was repeated 57 times, leaving out each compound in turn and then predicting its activity.

out", "leave-2-out", "leave-5-out", "leave-7-out", and "leave-10-out" are shown in Table 5. In other words, at each step one, two, five, seven, or 10 of the 58 compounds were left out in the GFA training process. The process was repeated until every compound had been left out and predicted once. CV r^2 was then calculated on the basis of predictions by the models obtained from the remaining compounds in the data set. The GFA models proved to be very predictive, with good full CV r^2 values obtained when up to seven molecules were left out at a time (i.e., $p \ll 0.01$ with respect to the null hypothesis that $r^2 = 0$). The observed mean activity values and those predicted by full cross-validation based on the "leave-1-out" rule are shown in Figure 6.

Conclusion

The patterns of GI_{50} values across 60 cancer cell lines can provide rich information on chemical structure classification and mechanism of action, even within quite homogeneous data sets such as those for the CPTs. In the present study, we find that the activity profiles of CPTs for the 60 cell lines reflect those found in various structure−activity relationship studies of top1 at the biochemical level. The only apparent discrepancy observed is for compounds with the 20-$OCOCH_2R$ substituents (instead of 20-OH), which are much more active in the NCI cell screen than in biochemical assays. This finding can, however, be explained by a hydrolytic reaction mechanism that probably converts these "prodrugs" to their normal CPT analogues.

GFA has several possible advantages over traditional statistical methods of multivariate analysis. Unlike traditional multiple regression methods, it offers a nonlinear approach to the construction of QSAR models, using a variety of basis functions including spline terms. Spline terms make the models relatively unstable, but that problem can be ameliorated by reducing the size of the descriptor set and increasing the number of crossover operations. Like PLS, GFA is able to produce robust equations when the number of independent variables vastly exceeds the number of observations. However, PLS reduces the dimensionality of the independent variable set by extracting correlated components using PCA, whereas GFA efficiently selects correlated independent variables using GA. The algorithm tests full-size models rather than incrementally building them as most other techniques do. It is better at discovering combinations of correlated variables, although there remains uncertainty as to how many degrees of freedom should be considered lost per spline term. Finally, one of the important differences between GFA and other method is the construction and use of multiple models. All models in the finally selected population have roughly the same high productivity, but each model may provide different insights into the problem. The utility of the modeling process can sometimes be increased by averaging the results of multiple models with the aid of scientific intuition, rather than relying on an individual model.

The QSAR model relates molecular descriptors to mean activity values. The frequently used descriptors in the best QSAR models were partial atomic charges at the 11- and 12-positions of the A-ring. The three pharmacophoric distance descriptors (the interatomic distances between significant functional atoms of O18, O19, O22, and O23) also appeared in the QSAR models. These functional groups have been found in our previous modeling studies to be important.[29] On the basis of the current QSAR results and earlier molecular modeling studies, a four-center pharmacophore model has been constructed for use in searching the NCI Drug Information System Database (Fan et al., unpublished studies).

Acknowledgment. We thank Dr. David Rogers of MSI for helpful discussions on GFA. We also thank members of the NCI Developmental Therapeutics Program (DTP), particularly Dr. Timothy G. Myers and Dr. Daniel Zaharevitz, for providing the anticancer activity and 2D chemical structure data used in this study. The extraordinary efforts of DTP members in developing and maintaining the cancer cell screen have made these theoretical analyses possible. We wish, in particular, to cite the exceptional contributions of Dr. Kenneth D. Paull, who died in 1998. His seminal work initiated analysis of patterns in the NCI screen.

References

(1) Burris, H. A.; Fields, S. M. Topoisomerase I inhibitors. An overview of the camptothecin analogues. *Hematol. Oncol. Clin. North. Am.* **1994**, *8*, 333−355.
(2) Pourquier, P.; Pommier, Y. Topoisomerases I: new targets for the treatment of cancer and mechanisms of resistance. *Bull Cancer* **1998**, Spec, 5−10.
(3) Takimoto, C. H.; Wright, J.; Arbuck, S. G. Clinical applications of the camptothecins. *Biochim. Biophys. Acta* **1998**, *1400*, 107−119.
(4) Stewart, L.; Redinbo, M. R.; Qiu, X.; Hol, W. G.; Champoux, J. J. A model for the mechanism of human topoisomerase I. *Science* **1998**, *279*, 1534−1541.

(5) Redinbo, M. R.; Stewart, L.; Kuhn, P.; Champoux, J. J.; Hol, W. G. Crystal structures of human topoisomerase I in covalent and noncovalent complexes with DNA. *Science* **1998**, *279*, 1504−1513.
(6) Boyd, M. R.; Paull, K. D. Some practical considerations and applications of the National Cancer Institute in vitro anticancer drug discovery screen. *Drug Dev. Res.* **1995**, *34*, 91−109.
(7) 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.
(8) 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.
(9) Boyd, M. R. The NCI in vitro anticancer drug discovery screen: concept, implementation, and operation, 1985−1995. In *Anticancer Drug Development Guide: Preclinical Screening, Clinical Trials, and Approval*; Teicher, B. A., Ed.; Humana Press: Totowa, NJ, 1997; pp 23−42.
(10) 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.
(11) 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.
(12) Shi, L. M.; Fan, Y.; Lee, J. K.; Waltham, M.; Andrews, D. T.; Scherf, U.; Paull, K. D.; Weinstein, J. N. Mining and visualizing large anticancer drug discovery databases. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 367−379.
(13) 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.; et al. Neural computing in cancer drug development: predicting mechanism of action. *Science* **1992**, *258*, 447−451.
(14) 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.
(15) 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 antitumour drug screen: predicting biochemical mechanism of action. *Stat. Med.* **1994**, *13*, 719−730.
(16) Shi, L. M.; Myers, T. G.; Fan, Y.; O'Connor, P. M.; Paull, K. D.; Friend, S. H.; Weinstein, J. N. Mining the National Cancer Institute Anticancer Drug Discovery Database: cluster analysis of ellipticine analogues with p53-inverse and central nervous system-selective patterns of activity. *Mol. Pharmac.* **1998**, *53*, 241−251.
(17) 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.; et al. 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.
(18) 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.
(19) Scherf, U.; Ross, D. T.; Waltham, M.; Smith, L. H.; Lee, J. K.; Tanabe, L.; Kohn, K. W.; Reinhold: W. C.; Myers, T. G.; Andrews, D. T.; Scudiero, D. A.; Eisen, M. B.; Sausville, E. A.; Pommier, Y.; Botstein, D.; Brown, P. O.; Weinstein, J. N. A gene expression database for the molecular pharmacology of cancer. *Nat. Genet.* **2000**, *24*, 236−244.
(20) Bleiberg, H.; Rothenberg, M. L. CPT-11: From DNA topology to clinical activity. 1996; Vol. 23(1), Suppl.3.
(21) Slichenmyer, W. J.; Rowinsky, E. K.; Donehower, R. C.; Kaufmann, S. H. The current status of camptothecin analogues as antitumor agents. *J. Natl. Cancer Inst.* **1993**, *85*, 271−291.
(22) Wall, M. E.; Wani, M. C. Camptothecin and taxol: discovery to clinic— thirteenth Bruce F. Cain Memorial Award Lecture. *Cancer Res.* **1995**, *55*, 753−760.
(23) Jaxel, C.; Kohn, K. W.; Wani, M. C.; Wall, M. E.; Pommier, Y. Structure−activity study of the actions of camptothecin derivatives on mammalian topoisomerase I: evidence for a specific receptor site and a relation to antitumor activity. *Cancer Res* **1989**, *49*, 1465−1469.
(24) Pommier, Y.; Pourquier, P.; Fan, Y.; Strumberg, D. Mechanism of action of eukaryotic DNA topoisomerase I and drugs targeted to the enzyme. *Biochim. Biophys. Acta* **1998**, *1400*, 83−105.
(25) Hsiang, Y. H.; Liu, L. F.; Wall, M. E.; Wani, M. C.; Nicholas, A. W.; Manikumar, G.; Kirschenbaum, S.; Silber, R.; Potmesil, M. DNA topoisomerase I-mediated DNA cleavage and cytotoxicity of camptothecin analogues [published erratum appears in *Cancer Res.* **1989**, *49* (23), 6868]. *Cancer Res.* **1989**, *49*, 4385−4389.
(26) Kingsbury, W. D.; Boehm, J. C.; Jakas, D. R.; Holden, K. G.; Hecht, S. M.; Gallagher, G.; Caranfa, M. J.; McCabe, F. L.; Faucette, L. F.; Johnson, R. K.; et al. Synthesis of water-soluble (aminoalkyl)camptothecin analogues: inhibition of topoisomerase I and antitumor activity. *J. Med. Chem.* **1991**, *34*, 98−107.
(27) Wang, X.; Zhou, X.; Hecht, S. M. Role of the 20-hydroxyl group in camptothecin binding by the topoisomerase I-DNA binary complex. *Biochemistry* **1999**, *38*, 4374−4381.
(28) Luzzio, M. J.; Besterman, J. M.; Emerson, D. L.; Evans, M. G.; Lackey, K.; Leitner, P. L.; McIntyre, G.; Morton, B.; Myers, P. L.; Peel, M.; et al. Synthesis and antitumor activity of novel water soluble derivatives of camptothecin as specific inhibitors of topoisomerase I. *J. Med. Chem.* **1995**, *38*, 395−401.
(29) Fan, Y.; Weinstein, J. N.; Kohn, K. W.; Shi, L. M.; Pommier, Y. Molecular modeling studies of the DNA − topoisomerase I ternary cleavable complexe with camptothecin. *J. Med. Chem.* **1998**, *41*, 2216−2226.
(30) Hansch, C.; Muir, R. M.; Fujita, T.; Maloney, P. P.; Geiger, F.; Streich, M. The correlation of biological activity of plant growth regulators and chloromycetin derivatives with Hammett constants and partition coefficients. *J. Am. Chem. Soc.* **1963**, *85*, 2817−2824.
(31) Hansch, C.; Leo, A.; Hoekman, D. Exploring QSAR, v.1 Fundamentals and applications in chemistry and biology; v.2 Hydrophobic, electronic, and steric constants; American Chemical Society, Washington, DC, 1995.
(32) Cramer, R. D., III.; Patterson, D. E.; Bunce, J. D. Comparative molecular filed analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.
(33) 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.
(34) Rogers, D. Some theory and examples of genetic fuction approximation with comparison to evolutionary techniques. In *Genetic Algorithms in Molecular Modeling*; Devillers, J., Ed.; Academic Press: London, 1996; pp 87−107.
(35) MSI Cerius2 Version 2.1, Molecular Simulations, http://www.msi.com/.
(36) 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.
(37) Rappe, A. K.; Goddard, W. A. *J. Phys. Chem.* **1991**, *95*, 3358.
(38) StatSci S−PLUS Reference Manual. **1993**; MathSoft, Inc., Seattle,
(39) Holland, J. *Adaptation in Artificial and Natural Systems*; University of Michigan Press: Ann Arbor, 1975.
(40) Friedman, J. H. Multivariate Adaptive Regression Splines, Technical Report No. 102, Nov/Rev.1990 Aug. 1988; Laboratory of Computational Statistics, Department of Statistics, Stanford University, Standford.
(41) Friedman, J. H. Multivariate adaptive regression splines (with discussion). *Ann. Statistics* **1991**, *19*, 1−141.
(42) Pommier, Y.; Jaxel, C.; Heise, C. R.; Kerrigan, D.; Kohn, K. W. *Structure−activity relationship by camptothecin derivatives: evidence for the existence of a ternary cleavable complex*; Oxford University Press: New York, 1991.
(43) Hertzberg, R. P.; Caranfa, M. J.; Holden, K. G.; Jakas, D. R.; Gallagher, G.; Mattern, M. R.; Mong, S. M.; Bartus, J. O.; Johnson, R. K.; Kingsbury, W. D. Modification of the hydroxy lactone ring of camptothecin: inhibition of mammalian topoisomerase I and biological activity. *J. Med. Chem.* **1989**, *32*, 715−720.
(44) Emerson, D. L.; Besterman, J. M.; Brown, H. R.; Evans, M. G.; Leitner, P. P.; Luzzio, M. J.; Shaffer, J. E.; Sternbach, D. D.; Uehling, D.; Vuong, A. In vivo antitumor activity of two new seven-substituted water-soluble camptothecin analogues. *Cancer Res.* **1995**, *55*, 603−609.
(45) Shi, L. M.; Fan, Y.; Myers, T. G.; O'Conner, P. M.; Paull, K. D.; Friend, S. H.; Weinstein, J. N. Mining the NCI anticancer drug discovery databases: genetic function approximation for the

QSAR study of anticancer ellipticine analogues. *J. Chem. Inf. and Comput. Sci.* **1998**, *38*, 189−199.
(46) Leteurtre, F.; Fujimori, A.; Tanizawa, A.; Chhabra, A.; Mazumder, A.; Kohlhagen, G.; Nakano, H.; Pommier, Y. Saintopin, a dual inhibitor of DNA topoisomerases I and II, as a probe for drug−enzyme interactions. *J. Biol. Chem.* **1994**, *269*, 28702−28707.
(47) Tanizawa, A.; Fujimori, A.; Fujimori, Y.; Pommier, Y. Comparison of topoisomerase I inhibition, DNA damage, and cytotoxicity of camptothecin derivatives presently in clinical trials. *J. Natl. Cancer Inst.* **1994**, *86*, 836−842.
(48) Tanizawa, A.; Kohn, K. W.; Kohlhagen, G.; Leteurtre, F.; Pommier, Y. Differential stabilization of eukaryotic DNA topoisomerase I cleavable complexes by camptothecin derivatives. *Biochemistry* **1995**, *34*, 7200−7206.

JM0005151