On the Physical Interpretation of QSAR Models†

David T. Stanton*

Corporate Research, Chemical Technology Division, Procter & Gamble, Miami Valley Laboratories, 11810 East Miami River Road, Cincinnati, Ohio 45252

Received April 8, 2003

Multidimensional quantitative structure−activity models (QSAR) developed using molecular structure descriptors and regression analysis techniques have found wide utility and acceptance. However, it is often difficult to extract a physical interpretation of such models because of the types of descriptors involved and the multidimensional nature of the model. The work described here illustrates a method of model interpretation that employs partial least squares (PLS) analysis. Structure−activity relationship information is derived from the positions of specific sets of structures in the PLS score plots and the weights for each variable in the PLS components. Using these data, information regarding major structure−activity trends, trend exceptions, and unique or outlying observations is easily obtained. Examples of this methodology are illustrated using QSAR equations developed for the inhibition of quinolone-resistant bacterial DNA gyrase and human topoisomerase-II inhibition by a series of quinolone antibacterial agents.

INTRODUCTION

One of the most important aspects of QSAR model development in the drug design process is the extraction and communication of structure−activity relationship information that is encoded in the model. Simply put, in order for a model to be perceived as credible, it is critical to be able to explain the physical significance of the model. Unless it is a very simple equation, a routine examination of a model generated using conventional QSAR techniques is insufficient for a number of reasons. The coefficients of the equation are not directly interpretable because they represent a composite picture of two or more structural trends typically buried in the model. The structural descriptors involved may include those that are derived using a mathematical construct that may not have a direct physical interpretation without providing a structural context (e.g. molecular connectivity indices[1,2]). Also, a descriptor selected for inclusion in the model using conventional regression methods, or more advanced methods such as a genetic algorithm, may be acting as a surrogate measure for a structural characteristic of the molecule that is not as accurately characterized by another descriptor with a more intuitive physical interpretation. The purpose of conventional QSAR descriptors is to provide a quantitative measurement of some specific feature of molecular structure. The descriptors one chooses to compute are usually selected because it is hypothesized that a meaningful correlation can be found that relates changes in specific structural features to changes in a specific physical or chemical property of a molecule. For example, measurement of topological changes in a molecule can be related to changes in geometry and thus act as a measure of the shape the molecule presents to other interacting molecules. However, changes in branching can affect the distribution of electronic charge, which can be reflected as changes in reactivity or polarity. Similarly, a measurement of the distribution of electronic charge can, in some cases, act as a better measure of branching, and therefore of shape, than do the typical topological descriptors.[3] Thus, simply knowing the type of features measured by a descriptor is insufficient by itself.

The act of developing a multivariate linear model for some observed experimental property of a set of structures involves selecting a set of one or more descriptors that provide a statistical correlation with the experimental property values. A preconceived notion of the physical interpretation of the descriptors selected can result in a misunderstanding of the underlying structure−activity relationship (SAR) in the model. There are two important pieces of information one needs in order to generate a clear SAR interpretation. The first is the knowledge of what features of structure are measured by a given descriptor. The second is knowledge of where the changes in structure are taking place that are correlated to changes in the experimentally observed property. In the case of whole-molecule descriptors, such as the molecular connectivity indices, changes in the features of structures measured by the descriptor may be occurring in several places. However, the important changes, those affecting the observed property, may be localized in a particular region of the molecule. Thus one must know both the *features* and the *context* to make a clear SAR interpretation.

The method of partial least squares (PLS) regression analysis[4,5] can be used to help extract this information. We have found that PLS can be used quite effectively as a tool for interpreting QSAR models and that the information extracted is much more detailed than that obtained by simply considering the overall model equation. However, PLS is not necessarily the best method for developing the model. Since the latent variables (scores) obtained using PLS represent a linear combination of all the descriptors provided

* Corresponding author e-mail: stanton.dt@pg.com.

† The report of this work is dedicated to the memory of Glen E. Mieling, Ph.D., a respected scientist, a valued colleague, and a good friend, who inspired me to employ PLS analyses for this purpose.

10.1021/ci0340658 CCC: $25.00
Published on Web 08/01/2003

at the start of the analysis, identification of the key descriptors is hampered because of the existence of noise (extra, unnecessary descriptors). The resulting model generally performs poorly in cross-validation compared to a simpler optimized ordinary least squares (OLS) equation.[6,7] Because of these problems, a process was sought that would provide the ability to develop optimized models that yield a clear physical description of the underlying structure activity relationship. The work described here illustrates the process that we have found to provide detailed structure−activity information useful for the design of new molecules.

The example data set comprises a collection of quinolones that were synthesized for evaluation as antibacterial agents.[8] The desired activity is inhibition of bacterial DNA gyrase derived from a quinolone-resistant strain of *Staphlococcus aureus* (QR-gyrase). An undesirable effect is the inhibition of a related enzyme, human topoisomerase-II (TOPO-II). The overall goal of the model development work was to identify how the structure of the molecules in question could be modified to produce good inhibition of gyrase and simultaneously reduce inhibition of TOPO-II.[9]

EXPERIMENTAL SECTION

Data Set. The data set used in this study comprised 32 structures. These structures are shown in Figure 1. The structures were sketched into the computer and stored in a database using the SYBYL package.[10] The initial conformation chosen for each structure was fixed according to the following convention (see Figure 2). The structure was drawn with the carboxylic acid to the right. In this orientation, the cyclopropyl group attached to the nitrogen at the 1-position was drawn such that it extended out of the plane of the page. Any substituent at the 8-position larger than a single nonhydrogen atom (e.g., methoxy) was drawn such that it extended into the plane of the page. Last, the torsion angle about the bond connecting the ring attached at the 7-position was set to −30 degrees when the atoms are identified as shown in the example provided in Figure 2. Specific stereochemistry (where applicable) was included as shown in Figure 1. Once the structures were entered in this fashion, the strain energy was minimized using the Tripos force field,[11] with electrostatic terms included. Atomic partial charges used in the energy-minimization step were calculated in SYBYL using the Gasteiger−Huckel method,[12] and the final structures and partial atomic charges were then transferred to the ADAPT package[13,14] for descriptor calculation and model development. The biological data obtained for each structure for both QR-gyrase and TOPO-II are provided in Table 1.

Descriptor Calculation. A set of 144 descriptors was calculated for each of the 32 structures in the data set. The descriptor set was chosen to capture important topological, geometric, and electronic structural features. The topological descriptors are derived using graph-theoretical approaches to defining chemical structures (chemical graph theory).[15] Such descriptors capture detailed information concerning molecular shape and complexity and have the added advantage of being independent of conformation. Additional conformation-independent information is captured as counts of specific structural fragments (i.e., counts of carbon and heteroatoms, counts of single, double, triple, and aromatic bonds, etc.). Geometric descriptors capture conformation-dependent shape characteristics of structure, such as surface area and volume,[16] width, length, and thickness,[17,18] whereas electronic descriptors provide information concerning the distribution of charge in the molecule.[19] Additionally, some descriptors employ structural representations that capture two or more of these structural feature types (e.g., surface area and partial atomic charge). This class of descriptor is represented by the CPSA descriptors[20,21] and the related hydrogen bonding-specific descriptors[22] that have been shown to be useful in past studies. The partial charges used in the calculation of the CPSA and related descriptors were those obtained using the Gasteiger−Huckel method during the strain-energy step in SYBYL.

Model Development and Validation. Model development was carried out in ADAPT using a generalized simulated annealing approach.[23] Additional statistical evaluation of the model was performed using linear regression methods in the Minitab package.[24] Prior to the acceptance of a final model, PLS analysis was performed using the SCAN program[25] to ensure that the model was not overfit. A model was considered to be overfit if the PLS analysis showed the number of validated components to be less than the number of original descriptors in the model (e.g., a seven-variable model yielded six or fewer validated PLS components). Robust regression analysis[26] was also performed to ensure that unusual or outlying observations were not having a detrimental effect on the model. Once a final model was obtained, PLS analysis was repeated to obtain the score plots and X-variable weights for the components that explained the majority of the variance in the observed property values (Y-variable). This information is used to develop the final physical interpretation of the model.

Model Interpretation. Interpretation of the final models required two pieces of information. The first is the PLS score plot for each validated component in the PLS version of the model. The validated components are those for which the predicted sum of squared errors (PRESS) does not increase from the previous component. Alternatively, one can also use the cross-validated R^2 value, also called Q^2 (typically calculated using the leave-one-out method) criteria. A valid component is defined as one where the Q^2 value does not decrease from the preceding component's value (i.e., Q^2 increases with each component).[27] The position of each of the training set structures in the score plot is one of the key pieces of information needed for model interpretation. Statistics packages that provide the ability to perform PLS analysis and that have the ability to identify the individual points on a graph, sometimes referred to as "brushing", help in this part of the analysis. The second important set of information are the weights of each of the original descriptors in each PLS component. Since each component is a linear combination of all of the descriptors presented for analysis, the weight assigned to each descriptor provides a way of determining which structural features are being emphasized in a given component. In practice, we have found that PLS analysis of an optimized model yields a number of valid components equal to the number of descriptors used in the original model. It has also been observed that only a few of the original descriptors were highly weighted in each component. Using both the sign and magnitude of the weights, it is possible to determine what types of structural

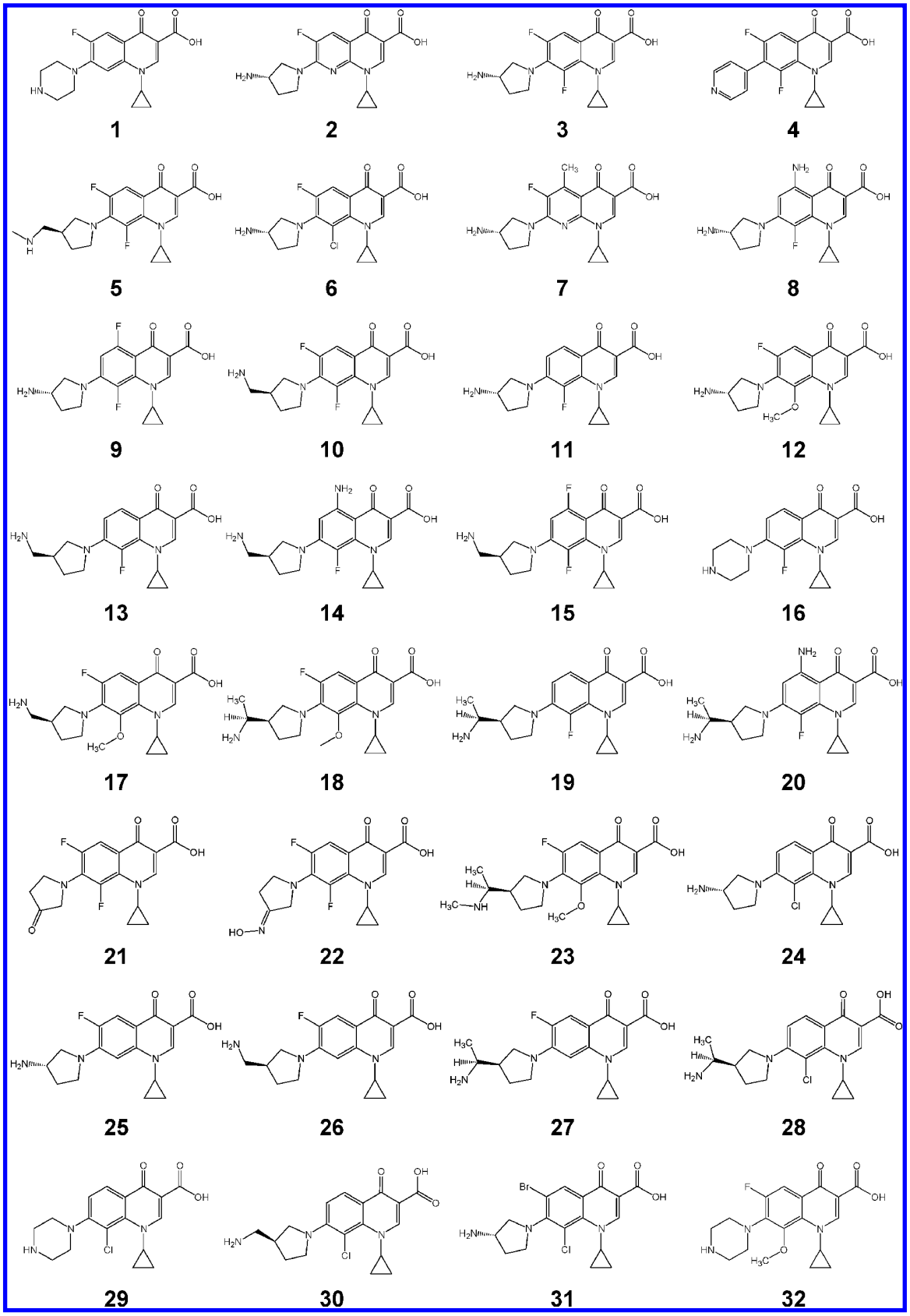

Figure 1. Structural diagrams for the compounds that form the training sets for this study. The configuration of the stereocenters used for each structure is shown.

features are being highlighted in each component and how changes in these features affect the property being modeled. In this paper, this will be referred to as a *structural trend*. By observing the structures for individual compounds and their position in the score plots, one can determine where these structural features are changing on the molecules. Using this combined information (both weights and score plots), a very detailed picture of the underlying structure−activity

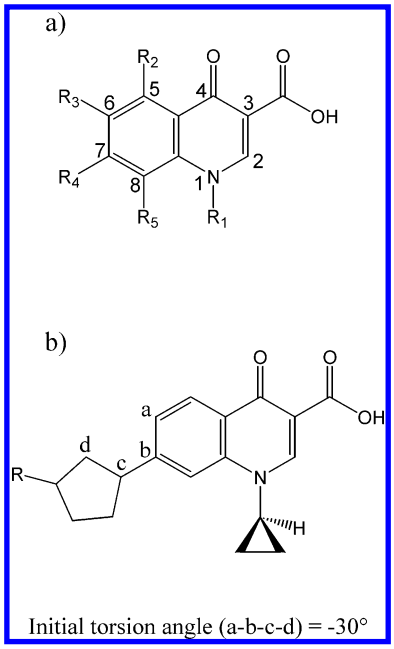

Figure 2. (a) The atom numbering for the quinolone core used throughout this study is shown. (b) Structures were entered into the computer system using the conventions shown. In all cases, the molecule was oriented as shown with the cyclopropyl ring at position 1 extending out of the plane toward the viewer, and the torsion angle for the bond at the 7-position was set to 30° prior to starting molecular mechanics calculations.

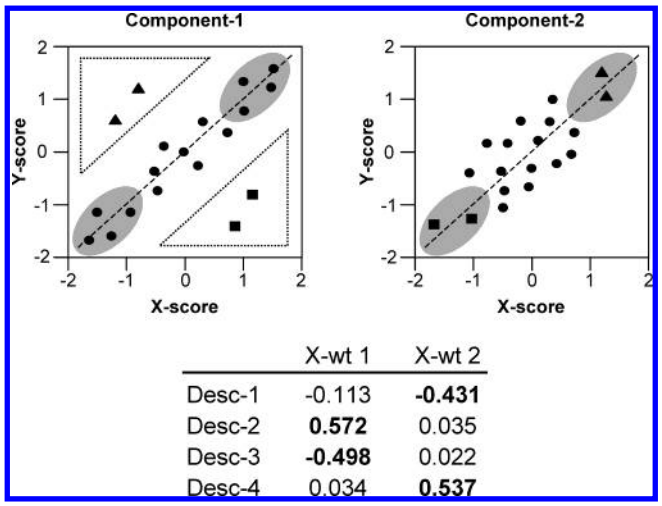

Figure 3. Example illustrating the use of the PLS score plots for model interpretation. Compounds represented by the points in the shaded regions (upper right and lower left quadrants) of component 1 are the focus of the first step. The descriptors, Desc-2 and Desc-3, have the largest weights for this component indicating that these are the structural measures responsible for explaining the positions of points representing the compounds in this plot. Points in the off-diagonal regions (upper left and lower right) of component 1 represent the compounds that are the focus of component 2. These compounds are incompletely explained by the SAR trend in component 1 and Desc-1 and Desc-4, the highly weighted descriptors in component 2, measure structural features that correct for this.

relationship can be obtained.

The process of model interpretation is illustrated in Figure 3. The plot shows score plots for the first two components of a hypothetical four-descriptor model. The weights for each of the four molecular descriptors in each component are also provided. The structural trend in the first component is explained primarily by the structural descriptors, Desc-2 and Desc-3. Using the typical QSAR convention, the compounds in the upper right quadrant of the first component are active, and the inactive compounds are in the lower left quadrant. By focusing on the changes in the structures of the active and inactive molecules with respect to the structural features measured by Desc-2 and Desc-3, it is possible to develop a clear understanding of the structural trend encoded in the first component. Typically, there are often compounds that are poorly explained by the first component, and these usually lie on the off-diagonal axis of the plot. Compounds whose activities are overestimated in the first trend are in the lower right quadrant of the plot, and those that are underestimated by the first trend lie in the upper left quadrant. The model uses other descriptors to correct these errors in subsequent components. In the example shown in Figure 3, these errors are corrected using the structural features encoded by descriptors Desc-1 and Desc-4. Now the compounds that were previously poorly estimated by the first trend are now correctly accounted for by the second structural trend. Thus, by using this type of examination of the score plots and component X-variable weights, the overall model can be broken down into its underlying trends, and these can be used to generate a detailed understanding of the structure−activity relationships for compounds in the training set. This methodology was applied to evaluate the structure−activity relationships for the quinolones.

Table 1. Observed and Calculated Biological Responses for the 32-Observation Data Set

	QR-gyrase activity log(1/IC_{50}, mM)		TOPO-II activity log(1/IC_{50}, mM)	
compd no.	obsd	calcd	obsd	calcd
1	1.122	1.211	−0.179	0.005
2	1.521	1.589	0.345	0.320
3	2.066	1.958	0.066	0.318
4	1.126	1.098	ND[a]	
5	2.099	2.086	0.673	0.961
6	2.563	1.942	−0.039	0.126
7	1.062	0.822	0.937	0.565
8	0.664	0.685	0.238	0.461
9	1.543	1.911	0.543	0.460
10	2.560	2.214	0.958	1.230
11	1.520	1.423	0.219	0.397
12	1.859	2.014	−0.045	−0.384
13	0.936	1.175	1.538	1.427
14	ND[a]		0.954	1.343
15	1.453	1.942	1.560	0.959
16	0.918	0.787	−0.082	0.098
17	2.097	2.116	0.097	0.460
18	2.289	2.409	−0.188	0.180
19	ND[a]		1.379	1.066
20	1.874	1.999	0.971	0.824
21	1.754	1.757	ND[a]	
22	1.241	1.242	0.939	0.824
23	2.128	2.046	0.003	−0.107
24	ND[a]		−0.237	0.029
25	1.520	1.644	0.520	0.273
26	1.538	1.616	1.061	0.741
27	2.555	2.340	0.777	0.773
28	1.876	1.859	0.730	0.759
29	0.542	0.872	0.064	−0.064
30	1.604	1.684	0.957	0.899
31	1.523	1.088	ND[a]	
32	1.751	1.776	−0.141	−0.327

[a] ND = not determined.

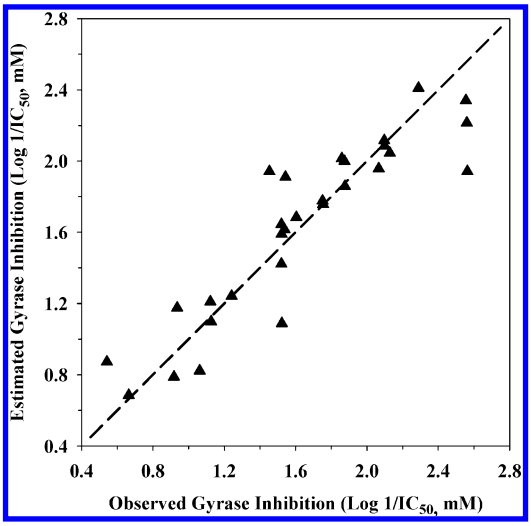

Figure 4. Plot showing the comparison of the estimated and observed bacterial gyrase inhibition values for the 29-observation training set model.

Table 2. Summary of the Regression Analysis Equation for the Quinolone-Resistant Bacterial Gyrase Inhibition Model[a]

descriptor	regression coeff	SD of coeff	t-value	variance inflation factor
PNSA-3	−0.149	0.0184	−8.14	3.4
ACGD	−25.8	3.70	−6.98	2.9
RSAA	−0.155	0.0254	−6.13	3.5
NN	−0.754	0.158	−4.78	2.3
MOMH-5	0.378	0.0793	4.76	1.5
Y-intercept	11.9	1.72	6.95	

[a] $R^2 = 82.2\%$, $s = 0.254$, $N = 29$, F-value = 21.2.

RESULTS AND DISCUSSION

Quinolone-Resistant Gyrase Inhibition. The details of model obtained for the inhibition of quinolone-resistant gyrase are shown in Table 2, and the fit plot for the model is shown in Figure 4. This model yielded a good fit to the observed inhibition values ($R^2 = 0.822$) using five variables. The five descriptors are as follows: PNSA-3 (type-3 partial negative surface area),[20] ACGD (the average difference in charge between a heteroatom and its donatable hydrogens),[22] RSAA (relative solvent-accessible surface area of a hydrogen-bond acceptor atom,[22] NN (simple count of nitrogen atoms), and MOMH-5 (ratio of the first and third principal moments of inertia).[18] The model is statistically sound, with an overall F-value[28] of 21.2 (compared to a critical F-value of 2.64 with 5 and 23 degrees of freedom and an alpha of 0.05), and the lowest partial F-values[29] of 22.7 (compared to a critical F-value of 4.28 with 1 and 23 degrees of freedom and an alpha of 0.05). There is little collinearity between the model descriptors as is evidenced by a maximum variance inflation factor, VIF,[30] of 3.5 and an average VIF of 2.7. Ideally, VIF values should be less than 10 and have an average close to 1. Last, the model performed well in PLS cross-validation ($Q^2 = 0.746$).

Once satisfied with the statistical validity of the model, the physical interpretation can be pursued as illustrated above. The first step is calculation of the PLS model. The summary of the PLS analysis is shown in Table 3a. While all five components are validated using either the PRESS or Q^2 criteria, the majority of the variance in the data set is accounted for in the first three components. Table 3b shows the weights for each of the descriptors in the first three components. The score-plot for the first of the three components of interest is shown in Figure 5. Extraction of the structure−activity relationship for this model was begun by examining component 1. Two examples each of more active and less active compounds from this component are identified in Figure 6. The highly weighted descriptors are measuring features that are related to the solvent accessible surface area of negatively charged atoms (PNSA-3), hydrogen-bond donor characteristics (ACGD), and molecular shape (MOMI-5). The PNSA-3 descriptor takes a negative sign due to the way it is computed. When taken with the negatively signed weight in this component, increases in negative surface area are correlated with increasing activity (a negative value multiplied by a negative coefficient). Increases in the value of ACGD are correlated negatively with increases in activity, and activity increases with increasing the ratio of the first and third moments of inertia. An examination of

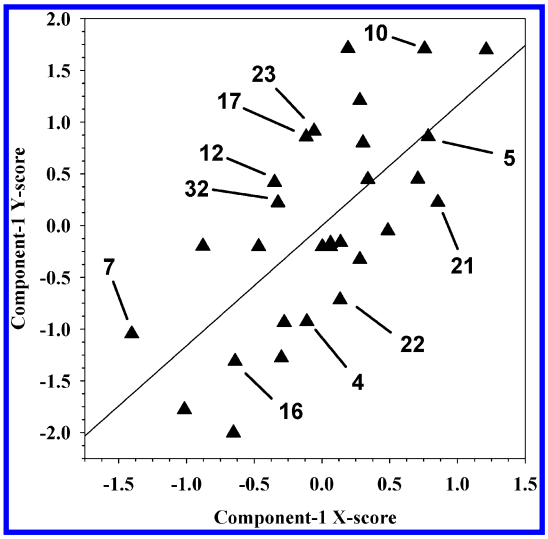

Figure 5. Score plot for component 1 from the PLS analysis of the gyrase inhibition model. Points representing compounds considered explicitly as part of the model interpretation are labeled with the compounds numbers.

Table 3. (a) Summary of the Results of the PLS Analysis of the Quinolone-Resistant Bacterial Gyrase Inhibition Model and (b) X-Weights for Each Descriptor in the First Three Components of the PLS Analysis of the Quinolone-Resistant Bacterial Gyrase Inhibition Model

Part a

component	residual sum of squared error	R^2 (cumulative)	PRESS	Q^2 (cumulative)
1	4.37	0.473	7.59	0.0843
2	2.44	0.706	3.73	0.550
3	1.58	0.810	2.24	0.730
4	1.48	0.821	2.12	0.745
5	1.48	0.822	2.10	0.746

Part b

descriptor	component 1 X-weights	component 2 X-weights	component 3 X-weights
PNSA-3	−0.597	0.206	−0.531
ACGD	−0.507	−0.476	0.405
RSAA	−0.234	−0.698	−0.171
NN	−0.360	0.241	−0.363
MOMH-5	0.451	−0.431	−0.627

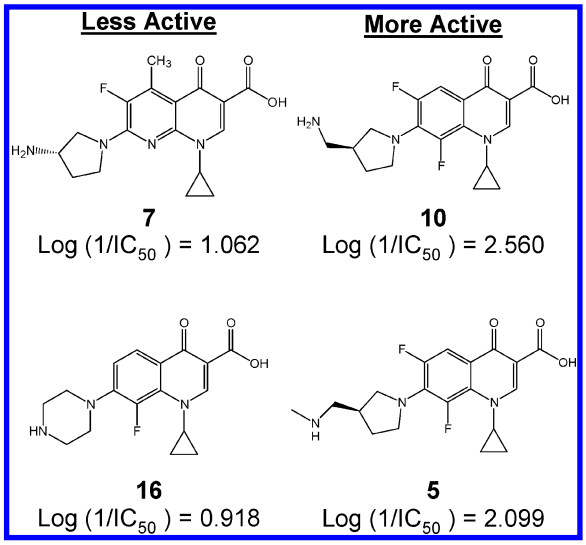

Figure 6. Structures selected as part of the interpretation process for component 1 of the gyrase inhibition model. These represent structures at the extreme ends of the on-diagonal portion of the score plot. Active compounds are in the upper right quadrant by convention.

the structures shows that the regions of the molecule changing most with regard to negative surface area are positions 6 and 8 on the quinolone core. Compounds **5** and **10** both have halogens at both these positions, while only one position is substituted with a halogen on compounds **7** and **16**. The values of ACGD change as the electronic environment and the type of terminal amine changes at the end of the group at the 7-position. For example, the magnitude of the atomic partial charge on the amine nitrogen decreases if the nitrogen is attached directly to the pyrrolidine ring instead of being attached to an intervening methylene. Also, charge differences and hydrogen-bonding characteristics are different between primary and secondary amines (amino pyrrolidine vs piperazine rings). Last, the length of the group attached at the 7-position of the quinolone core is important. In active compounds, the terminal amine is moved farther away from the core than in less active compounds. All three of these bits of structural information combine to explain over 47% of the variance in the data set. The physical impact of these changes can be seen by examining what these combined changes do to the molecule. In addition to having a relatively large negative partial atomic charge, halogens also are quite hydrophobic. While the charge on a chlorine atom is smaller, its solvent-accessible surface area is large. Thus the combination of having two hydrophobic groups at the 6- and 8-positions of the quinolone core appear to increase activity. If one also considers that the model is bringing attention to the nature and position of the terminal amine of the group at position 7, one sees that moving the hydrophilic group away from the core increases the hydrophobic nature of the compound in the region close to the core. This can be seen more clearly in Figure 7. Here, the solvent accessible surface area of the structure is color coded in terms of hydrophobicity using MOLCAD.[31] Thus, the physical interpretation of component 1 (the first structural trend) is that increasing the hydrophobic nature of the quinolone core in the region of positions 6, 7, and 8 increases activity.

Component 2 accounts for an additional 23.3% of the variance in the data (70.6% cumulative) and is focused

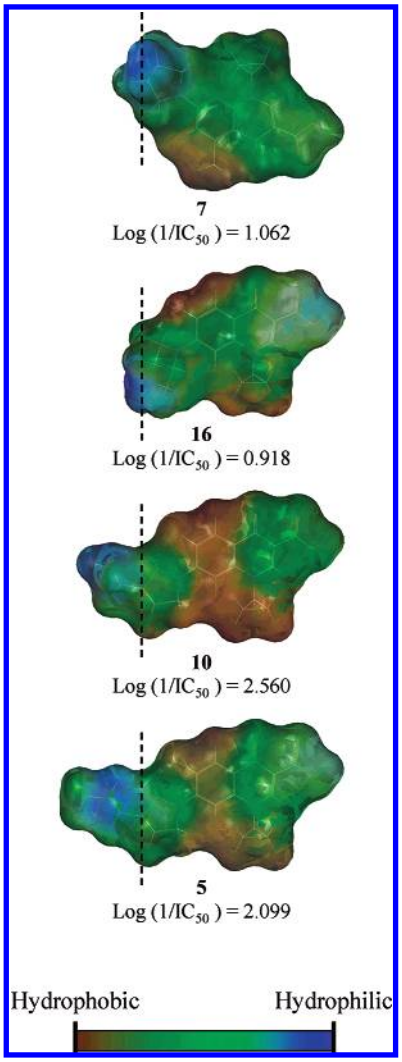

Figure 7. A comparison of the hydrophobicity of the solvent accessible surface for the four example quinolones. The dashed lines indicate the position of the hydrophilic group on the end of the tether at the 7-position (note that the structure for compound **7** is inverted in order to be able to see the hydrophilic portion).

almost entirely on three compounds, **4**, **21**, and **22**, as is illustrated in the component 2 score plot shown in Figure 8. The descriptors that are highly weighted are RSAA, and to a lesser extent, ACGD. Both of these descriptors have negative weights and capture information about hydrogen bonding features. In this case the compounds **4**, **21**, and **22** are moved to the extreme lower left (lower-activity) corner of the score plot. In component 1 these compounds had much higher X-scores. However, they were overestimated by the first component, and the model corrects for this in component 2 using their RSAA and ACGD values. These descriptors were designed to measure aspects of hydrogen bonding characteristics of molecular structure. An examination of these structures with regard to their hydrogen bonding characteristics shows that while these compounds meet the criteria for activity based on component 1, they lack the basic amine that is usually placed at the end of the group at position 7 on the quinolone core. Thus the second structural trend suggests that a basic amine is needed in this portion of the molecule.

The same analysis can be done for component 3 (see Figure 9). This component accounts for an additional 10.4% of the data set variance (81.0% cumulative). There are two

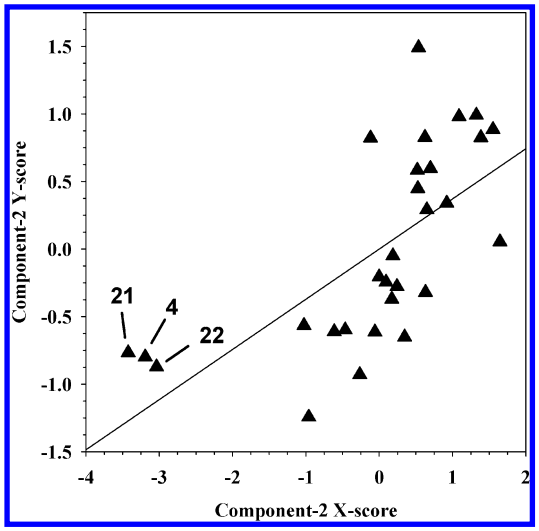

Figure 8. Score plot for component 2 from the PLS analysis of the gyrase inhibition model. Points representing structures of interest are labeled by their compound numbers.

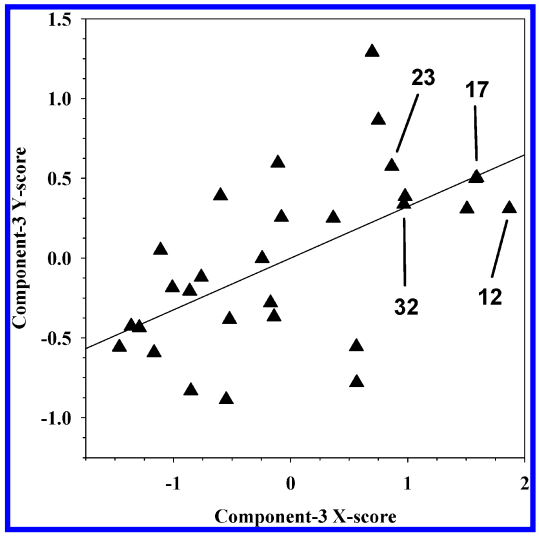

Figure 9. Score plot for component 3 from the PLS analysis of the gyrase inhibition model. Points representing structures of interest are labeled using their compound numbers.

highly weighted descriptors in this component, MOMH-5 and PNSA-3. In this component, MOMH-5 takes a negative weight. This suggests a change in the structure that reduces the ratio of length of the molecule to its thickness can increase activity. Since PNSA-3 is also highly weighted, attention must be paid to changes that include negatively charged atoms. In this case, compounds **12**, **17**, **23**, and **32** are considered. These compounds were underestimated in component 1 because the oxygen of the methoxy group in position 8 makes it less hydrophobic than other substituents (e.g., chloro and trifluoromethyl). However, the model uses MOMH-5 and PNSA-3 together here to correct for that underestimation. This suggests that the methyl group on the oxygen provides a hydrophobic shield for this otherwise hydrophilic atom, reducing its impact on the overall hydrophobicity in that region of the quinolone core.

The next component only accounts for an additional 1.1% of the variance for the data set, so not much additional structure−activity information can be gathered from its analysis. However, over 80% of the variance has been accounted for in the analysis by this point, and a detailed

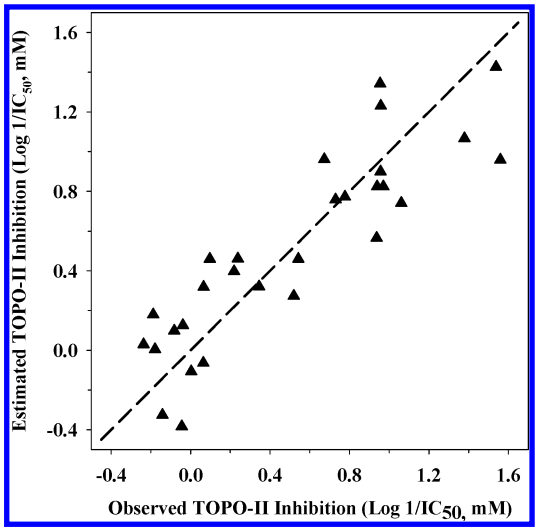

Figure 10. Plot showing the comparison of the estimated and observed topoisomerase-II (Topo-II) inhibition values for the 29-observation training set.model.

Table 4. Summary of the Regression Analysis Equation for the Mammalian Topoisomerase-II (TOPO-II) Inhibition Model[a]

descriptor	regression coeff	SD of coeff	t-value	variance inflation factor
S3C	−2.30	0.451	−5.10	2.3
S7CH	10.2	1.52	6.69	3.6
GEOH-1	0.156	0.558	2.80	3.0
MOMH-4	12.1	2.13	5.65	2.2
Y-intercept	−19.8	3.01	−6.59	

[a] $R^2 = 77.8\%$, $s = 0.280$, $N = 29$, F-value $= 21.0$.

picture of the structure−activity relationships encoded in the model has emerged. The overall theme of the SAR seems to focus on the requirement for a hydrophobic environment in the region around the positions 6, 7, and 8 on the quinolone core.

Mammalian Topoisomerase Inhibition. The final model for topoisomerase inhibition is summarized in Table 4. The fit-plot is shown in Figure 10, which indicates that the 4-variable model yields a good fit to the experimental results ($R^2 = 0.778$, $s = 0.280$). The descriptors included in this model are very different from those selected for the gyrase model. The descriptors are as follows: S3C (simple third-order cluster molecular connectivity),[1] S7CH (simple seventh-order chain molecular connectivity),[1] GEOH-1 (length of molecule along primary geometric axis),[17] and MOMH-4 (ratio of the first and second moments of inertia).[18] The statistical validation of this model showed the relationship to be significant, with an overall F-value of 21.0 (compared to a critical F-value of 2.78 with 4 and 24 degrees of freedom and an alpha of 0.05), and the lowest partial F-value is 7.84 (compared to a critical F-value of 4.26 with 1 and 24 degrees of freedom and an alpha of 0.05). There is also little collinearity between the model descriptors based on the VIFs, with the maximum VIF for the model of 3.6, and an average VIF of 2.8. The model also performed well in PLS cross-validation, yielding a Q^2 of 0.675.

The physical interpretation of this model was extracted using PLS analysis as described above. The results of the PLS analysis are shown in Table 5a. While all four PLS components were validated, the majority of the variance

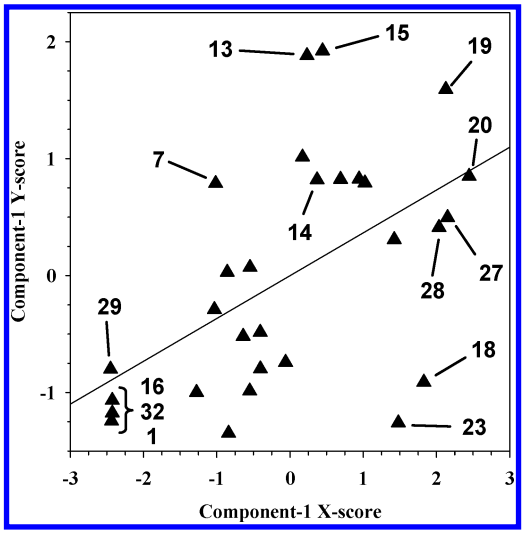

Figure 11. Score plot for component 1 from the PLS analysis of the Topo-II inhibition model. Points representing structures of interest are labeled using their compound numbers.

Table 5. (a) Summary of the Results of the PLS Analysis of the Mammalian Topoisomerase-II (TOPO-II) Inhibition Model and (b) X-Weights for Each Descriptor in the First Three Components of the PLS Analysis of the Mammalian Topoisomerase-II (TOPO-II) Inhibition Model

Part a

component	residual sum of squared error	R^2 (cumulative)	PRESS	Q^2 (cumulative)
1	6.03	0.287	7.09	0.162
2	2.58	0.695	3.95	0.533
3	1.92	0.774	2.76	0.674
4	1.88	0.778	2.75	0.675

Part b

descriptor	component 1 X-weights	component 2 X-weights	component 3 X-weights
S3C	0.256	−0.528	−0.809
S7CH	0.778	0.267	0.044
GEOH-1	0.572	−0.060	0.257
MOMH-4	−0.049	0.804	−0.527

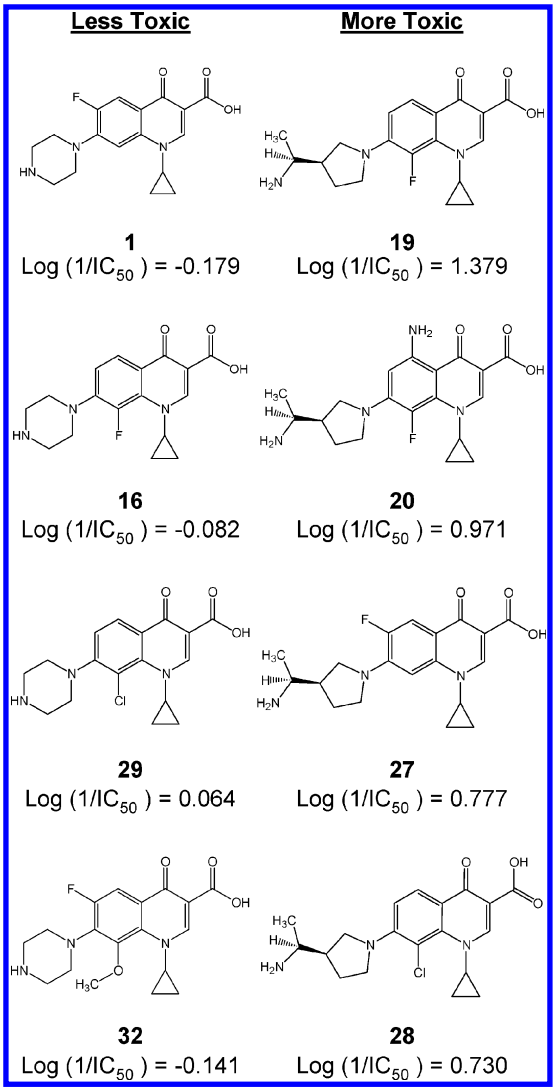

Figure 12. Structures selected as part of the interpretation process for component 1 of the Topo-II inhibition model. These represent structures at the extreme ends of on-diagonal portion of the score plot. Active compounds are in the upper right quadrant by convention.

accounted for by the model (77.4% of the total 77.8%) is treated in the first three components. The variable weights for these three components are given in Table 5b. The score plot for component 1 is shown in Figure 11. Examination of the results of the PLS analysis shows that a very different structure−activity relationship emerges for topoisomerase inhibition than was observed for the inhibition of gyrase. The attention of this model is focused on the nature and position of the terminal amine on the group attached at the 7-position of the quinolone core. The first component of the PLS analysis accounts for 28.7% of the variance in the data set and shows the descriptors S7CH and GEOH-1 to be highly and positively weighted. Example structures are shown in Figure 12. The length of the linkage between the quinolone core and the terminal amine is the major difference between the structures shown. Both descriptors capture information regarding this region of the molecule. The topological descriptor, S7CH, helps to discriminate the pyrrolidine from the piperazines, and the geometric descriptor, GEOH-1 captures the length of the molecule explicitly. The structure−activity relationship in this first trend explains that compounds with short linkages between the terminal amine and the quinolone core will be less toxic (poorer inhibitors of topoisomerase) than those with long linkages.

The second structural trend accounts for an additional 40.8% of the variance (cumulative 69.5%) and shows that not all structures with long linkers will be toxic. The descriptors that are highly weighted in component 2 are MOMH-4 (positively weighted) and S3C (negatively weighted). The score plot for component 2 is shown in Figure 13. Example structures for this trend are shown in Figure 14. Both MOMH-4 and S3C capture information regarding the substituents at positions 6 and 8 on the quinolone core. The descriptor S3C codes for a structural feature that has one heavy atom attached to three other heavy (non-hydrogen) atoms (e.g., a substituent on an aromatic ring). The values of this descriptor increase when a non-hydrogen atom is attached to either or both the 6- and 8-positions of the quinolone. Since S3C takes a negative weight, structures substituted at both positions would be predicted to be less toxic than structures with either one or no substituents at those positions. The MOMH-4 descriptor captures the effect of the size of the substituents, especially those at the 8-position. This descriptor is the ratio of the length and the

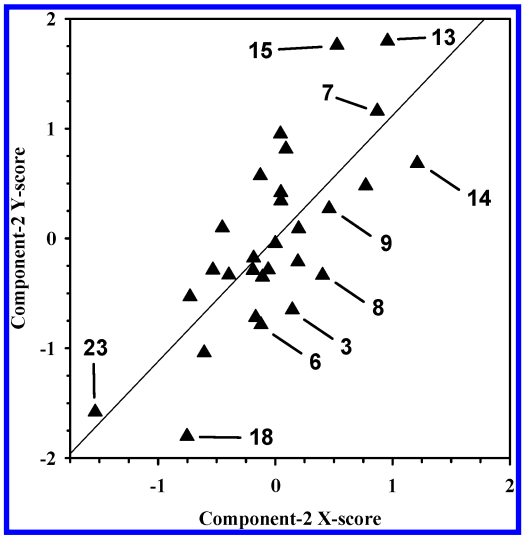

Figure 13. Score plot for component 1 of the Topo-II inhibition model. Points representing structures of interest for the interpretation are labeled using their compound numbers.

width of the structure. Since MOMH-4 is positively weighted in this component, structures with large substituents favor reduced toxicity. The effect of large substituents at the 8-position is a change in the position of the terminal amine on the group at the 7-position relative to the plane of the quinolone core. There is a great deal of steric crowding when a large group exists at the 8-position, especially with a cyclopropyl group at the 1-position. The crowding causes significant conformational changes in the molecule. The S3C descriptor also captures the presence of the methyl group adjacent to the terminal amine (see compounds **18** and **23**). Such branching also affects the geometry of the linker. The overall result is that the terminal amine cannot take the same position in space relative to the plane of the quinolone core, as would be the case where fewer steric interactions affect the geometry of the system. Thus, the second component shows that not all compounds with long linkages at the 7-position will be toxic.

The third PLS component accounts for 7.9% of the variance (cumulative 77.4%). The score plot for component 3 is shown in Figure 15. It is interesting to note that the two most highly weighted descriptors in component 3 (S3C and MOMH-4) are the same ones that are the focus of component 2. However, in component 3, the sign of the weight for MOMH-4 is inverted. The reason for this is that the model incorrectly assigns the activity of several structures in component 2, and these misassignments are corrected in component 3. For example, structures **3**, **6**, **8**, and **9** are overpredicted by the trend in component 2 (see Figure 13), and the third component corrects for this. In component 2, the flatness of the structure contributed most strongly to the inhibition of TOPO-II. These structures are relatively flat (larger values of MOMH-4). However, they also possess two substituents each, either both the 5- and 8-positions or the 6- and 8-positions. In component 3 the flatness of the structure is de-emphasized (the weight of MOMH-4 takes a negative sign), and the weight of S3C is large and negative. This moves the overestimated structures to the proper position relative to their observed TOPO-II activity.

The overall SAR for TOPO-II derived in this fashion suggests that the location of the terminal amine relative to

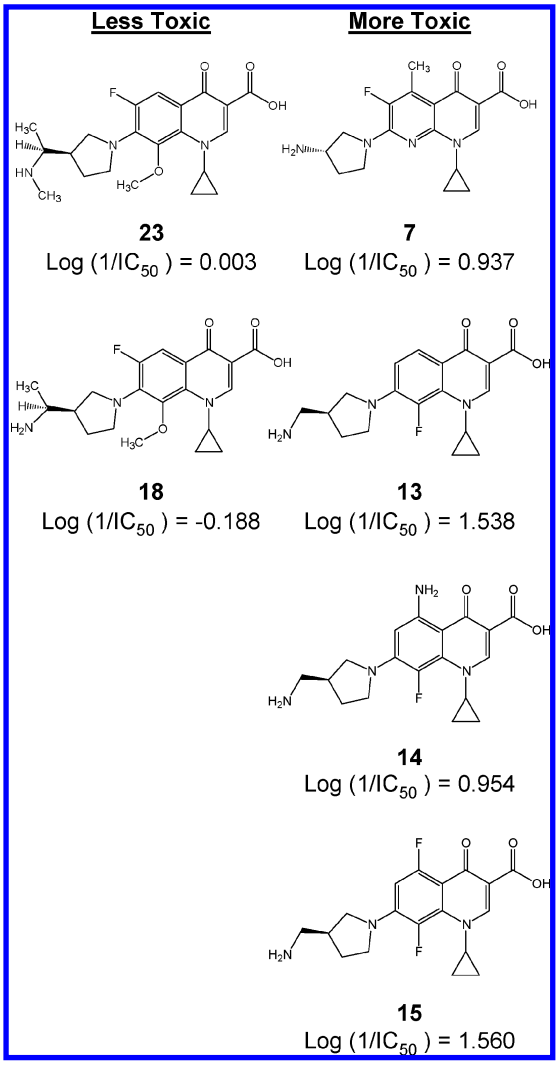

Figure 14. Structures selected as part of the interpretation process for component 2 of the Topo-II inhibition model.

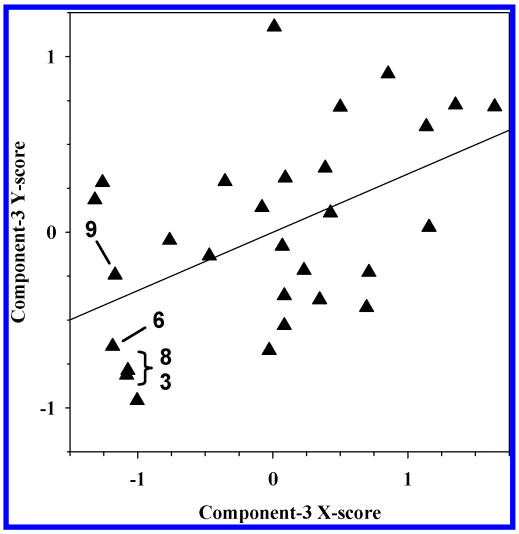

Figure 15. Score plot for component 3 from the Topo-II inhibition model.

the quinolone core is the primary driving factor for the activity of these compounds. Shorter tethers are preferred over longer. However, longer tethers are tolerated if the structure is not very flat, putting the terminal amine in the plane of the quinolone core. One way to change the position of the amine is to put large substituents at positions 6 and 8,

with position 8 being most preferable, especially if a cyclopropyl group occupies position 1.

One hypothesis that can explain the observed TOPO-II SAR is that there is a point of strong interaction of the quinolone terminal amine with the TOPO-II protein at the site where the quinolone binds. If we assume that the quinolone core binds strongly in a relatively fixed position, then the modifications that decrease TOPO-II activity prevent the terminal amine from reaching a potential protein-binding site.

The original reason for developing the models for both QR-gyrase and TOPO-II was to determine if a differential SAR could be identified. A comparison of the SARs derived for these two related endpoints yields important information in this regard. While shorter structures will be less toxic (decreased TOPO-II activity), the shorter linkers bring the terminal amine closer to the region near the quinolone core, which prefers to contain hydrophobic groups, thus reducing the gyrase activity. Moving the amine away from the core helps improve QR-gyrase inhibition with the potential to also increasing TOPO-II inhibition. Addition of large hydrophobic substituents at both the 6- and 8-positions of the quinolone alters the position of the terminal amine on compounds with long linkers and reduces TOPO-II inhibition. Thus, compounds with long tethers and large, hydrophobic substituents at the 8-position satisfy the SAR requirements for good gyrase inhibition and reduced TOPO-II activity.

CONCLUSIONS

A great deal of detailed structure−activity relationship information can be derived from conventional QSAR models by using PLS to extract the underlying SAR trends. Similar information cannot be obtained from an examination of the overall regression equation because it represents a composite picture of the combined trends. Also, since weights for the descriptors in these trends (PLS components) can take different signs, an interpretation of the sign of the coefficient from the overall regression equation will not provide the true SAR information available from a particular descriptor. The detail of the SAR extracted in this way rivals that typically obtained with more sophisticated 3D-QSAR methodologies. In the present case, very detailed structure−activity information was obtained by examining relatively simple conventional QSAR models. The ability to extract detailed SAR information from these types of QSAR equations yields new value for their use in a compound design process.

It is clear from this analysis that all the descriptors involved encode very specific information about what molecular features are changing, where they are changing on the molecules, and how those changes affect the properties of these compounds. Even though most of these descriptors are whole-molecule descriptors, they highlight specific changes on specific portions of the structure. It is also clear that all the descriptors, including the more theory-based molecular connectivity descriptors, yield physically meaningful information. This is contrary to the general criticism of the whole class of topological descriptors.[32] Last, the results of this work warn of the potential for errors in interpretation of SAR if one employs a preconceived notion of what the descriptors mean. In this work, topological descriptors provided information about important differences in geometry and hydrophobicity. Descriptors typically thought to only measure changes in geometry highlighted potentially important hydrogen-bonding or electrostatic interactions and descriptors designed to describe polar and hydrogen-bonding features actually provide information concerning potential hydrophobic interactions. This work supports the notion that one should employ a broad diversity of descriptors and allow the statistical process to identify those that are most useful. The subsequent analysis using PLS will provide the physical interpretation that is needed as part of the overall model validation process.

The utility of this methodology is not limited to models for biological activity alone. Models involving a variety of physical and biological properties have been evaluated in this way, and clear and detailed structure−property relationships have been obtained with each. Nor does there seem to be any limitation regarding the types of descriptors that can be evaluated. What is required is that the details of the calculation of the descriptor be understood so that the proper structural features can be identified. Indeed, application of this methodology often leads to a clearer understanding of precisely how particular descriptors work and helps to suggest how improvements in existing descriptors can be made.

ACKNOWLEDGMENT

The author gratefully acknowledges the work of B. Ledoussal, J. L. Gray, J. Almstead, and X. E. Hu of Procter & Gamble Pharmaceuticals who designed and synthesized the majority of the compounds considered in this study as well as the work of T. L. Twinem, also of Procter & Gamble Pharmaceuticals, who provided all the biological screening data. I am also grateful for the comments and suggestions of Prof. W. D. Dunn III and Prof. A. J. Hopfinger, both of University of Illinois at Chicago, regarding the PLS analysis.

REFERENCES AND NOTES

(1) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic: New York, 1976.
(2) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; Wiley & Sons: New York, 1986.
(3) Stanton, D. T.; Jurs, P. C. Computer-Assisted Study of the Relationship Between Molecular Structure and Surface Tension of Organic Compounds. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 109−115.
(4) Glen, W. G.; Dunn III, W. J.; Scott, D. R. Principal Components Analysis and Partial Least Squares Regression. *Tetrahedron Comput. Methodol.* **1989**, *2*, 349−376.
(5) Geladi, P.; Kowalski, B. R. Partial Least-Squares Regression: A Tutorial. *Anal. Chim. Acta* **1986**, *185*, 1−17.
(6) Spiegelman, C. H., McShane, M. J.; Goetz, M. J.; Motamedi, M.; Yue, Q. L.; Coté, G. L. Theoretical Justification of Wavelength Selection in PLS Calibration: Development of a New Algorithm. *Anal. Chem.* **1998**, *70*, 35−44.
(7) Baroni, M.; Clementi, S.; Crucuani, G.; Costantino, G.; Riganelli, D.; Oberrauch, E. Predictive Ability of Regression Models. Part II. Selection of the Best Predictive PLS Model. *J. Chemom.* **1992**, *6*, 347−356.
(8) Ledoussal, B.; Almstead, J. K.; Flaim, S. M.; Gallagher, C. P.; Gray, J. L.; Hu, X. E.; Kim, N. K., McKeever, H. D.; Miley, C. J.; Twinem, T. L.; Zheng, S. X. Novel Non-Fluoroquinolones (NFQs), Structure−Activity, and Design of New, Potent and Safe Agents. Book of Abstracts, 39th Interscience Conference on Antimicrobial Agents and Chemotherapy, San Francisco, CA, Sept. 1999; poster F-0544.
(9) Twinem, T. L.; Stanton, D. T. Determination of the Structural Features Responsible for Selectivity of Quinolones for Bacterial Gyrase over Mammalian Topoisomerase II Using Computer-Assisted Pattern Recognition and QSAR Techniques. Book of Abstracts, 39th Interscience Conference on Antimicrobial Agents and Chemotherapy, San Francisco, CA, Sept. 1999; poster F-0545.

(10) SYBYL molecular modeling software, Version 6.3; Tripos, Inc.: St. Louis, MO, 1996.
(11) SYBYL Version 6.3, Force Field Manual; Tripos, Inc.: St. Louis, MO, 1996; p 196.
(12) Gasteiger−Huckel partial atomic charges are calculated using the Gasteiger−Marsili method to calculate the σ-electron contributions and the Huckel method for calculating the π-electron contributions. Sybyl Version 6.3 Force Field Manual; Tripos: St. Louis, MO, 1996; p 290.
(13) Stuper, A. J.; Jurs, P. C. ADAPT: A Computer System for Automating Data Analysis using Pattern-Recognition Techniques. *J. Chem. Inf. Comput. Sci*. **1976**, *2*, 99−105.
(14) Jurs, P. C.; Chou, J. T.; Yuan, M. In *Computer-Assisted Drug Design*; Olson, R. C., Christoffersen, R. E., Eds.; American Chemical Society: Washington, DC, pp 103−129.
(15) Ivanciuc, O.; Balaban, A. T. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach: The Netherlands, 1999; pp 59−167.
(16) Pearlman, R. S. In *Physical Chemical Properties of Drugs*; Yalkowsky, S. H., Sinkula, A. A., Valvani, S. C., Eds.; Marcel Dekker: New York, 1980.
(17) Brugger, W. E.; Stuper, A. J.; Jurs, P. C. Generation of Descriptors from Molecular Structures. *J. Chem. Inf. Comput. Sci*. **1976**, *16*, 105−110.
(18) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Mannhold, R., Kubinyi, H., Timmerman, H., Eds.; Wiley-VCH: Weinheim, Federal Republic of Germany, 2000; Methods and Principles in Medicinal Chemistry, Vol. 11, p 352.
(19) Dixon, S. L.; Jurs, P. C. Atomic Charge Calculations for Quantitative Structure−Property Relationships. *J. Comput. Chem*. **1992**, *13*, 492−504.
(20) Stanton, D. T.; Jurs, P. C. Development and Use of Charged Partial Surface Area Descriptors in Computer-Assisted Quantitative Structure−Property Relationship Studies. *Anal. Chem.* **1990**, *62*, 2323−2329.
(21) Stanton, D. T.; Dimitrov, S.; Grancharov, V.; Mekenyan, O. G. Charged Partial Surface Area (CPSA) Descriptors. QSAR Applications. *SAR QSAR Environ. Res.* **2002**, *13*, 341−351.
(22) Stanton, D. T.; Egolf, L. M.; Jurs, P. C.; Hicks, M. G. Computer-Assisted Prediction of Normal Boiling Points of Pyrans and Pyrroles. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 306−316.
(23) Sutter, J. M.; Jurs, P. C. Selection of Molecular Descriptors for Quantitative Structure−Activity Relationships. *Data Handl. Sci. Technol.* **1995**, *15*, 111−132.
(24) Minitab for Windows, Release 12.23; Minitab, Inc.: State College, PA, U.S.A.
(25) SCAN for Windows, Release 1.1; Minitab, Inc.: State College, PA, U.S.A.
(26) Rousseeuw, P. J.; Leroy, A. M. *Robust Regression and Outlier Detection*; Wiley and Sons: New York, 1987; p 226.
(27) Baroni, M.; Constantino, G.; Cruciani, G.; Riganelli, D.; Valigi, R.; Clementi, S. Generating Optimal Linear PLS Estimations (GOLPE): An Advanced Chemometric Tool for Handling 3D-QSAR Problems. *Quant. Struct.-Act. Relat.* **1993**, *12*, 9−20.
(28) Neter, J.; Wasserman, W.; Kutner, M. H. *Applied Linear Statistical Models*, 2nd ed.; Irwin: Homewood, IL, 1985; p 240.
(29) Neter, J.; Wasserman, W.; Kutner, M. H. *Applied Linear Statistical Models*, 2nd ed.; Irwin: Homewood, IL, 1985; p 281.
(30) Neter, J.; Wasserman, W.; Kutner, M. H. *Applied Linear Statistical Models*, 2nd ed.; Irwin: Homewood, IL, 1985; p 391.
(31) SYBYL Graphics Manual, MOLCAD, Version 6.8.1; Tripos, Inc.: St. Louis, MO, 2002.
(32) Kubinyi, H. *QSAR: Hansch Analysis and Related Approaches*; Mannhold, R., Krogsgaard-Larsen, P., Timmerman, H., Eds.; VCH: New York, NY, 1993; Methods and Principles in Medicinal Chemistry, Vol. 1, pp 50−53.

CI0340658