Evaluating Chemical Structure Similarity as an Indicator of Cellular Growth Inhibition

Anders Wallqvist,*,‡ Ruili Huang,† Narmada Thanki,‡ and David G. Covell†

National Cancer Institute, Developmental Therapeutics Program, Screening Technologies Branch, and Science Applications International Corporation, P. O. Box B, Frederick, Maryland 21702

Received April 27, 2005

Chemical variations of small compounds are commonly used to probe biological systems and potentially discover lead-like compounds with selective target activity. Molecular probes are either generated by synthesis or acquired through directed searches of commercially available compound libraries. The data generated when testing the probes in various biological systems constitutes a structure/activity analysis. The ability to detect variations and classify biological responses requires the analysis of a compound in multiple assays. While the concept of a structure/activity relationship is straightforward, its implementation can vary considerably depending on the biological system under study and the probe library selected for testing. The analysis presented here will focus on the accumulated compound library used to screen for growth inhibition across the National Cancer Institute's panel of 60 tumor cells. The considerable chemical and biological diversity inherent in these data offers an opportunity to establish a quantifiable connection between chemical structure and biological activity. We find that the connection between structure and biological response is not symmetric, with biological response better at predicting chemical structure than vice versa. Structurally and functionally similar compounds can have distinguishable biological responses reflecting different mechanisms of action.

INTRODUCTION

A fundamental guiding principle in structure-based drug discovery is that of chemical similarity; that is, closely related molecules will elicit similar activity in a biological assay.[1−6] Optimization of drug properties through quantitative structure/activity relationship (QSAR) methods is based on this principle. The problem is, of course, that not all chemical similarities are equal, and the ability to abolish desired behavior by changing a single group or even atom is well-appreciated. Chemical similarity can also be used in an exclusive sense to construct diverse sets of probe molecules, for example, as in unfocused combinatorial chemistry approaches. In such efforts, the goal is to find as many different molecular structures/scaffolds as possible that generate a similar biological response.[7,8]

In this work, we will characterize the cellular response to chemical structure modifications using the publicly available data from the small molecule cancer cell screen at the NCI.[9−14] These cancer cells are from immortalized cell lines that have been selected on the basis of various independent quality-assurance criteria, notably their ability to grow in the developed media and to show reproducible results for growth and drug sensitivity. Chemicals that reduce the viability of the cell are tagged as potential leads for affecting particular pathways characteristic of each tumor cell's biology. The biological response of such a cell-based assay is rich in information as, in principle, the complete system's biology information is encoded in the assay. Even though such data is far more complicated and harder to interpret than a noncellular, direct molecular binding assay, the ability to monitor a complete biological system offers advantages when studying mechanisms of drug action.

In this assay, the response of all 60 cell lines to one drug as measured by 50% growth inhibition concentrations (GI_{50}) represents a pattern of biological activity. Differences in a compound's GI_{50} values across the tumor panel reflect the pathways and targets affected by the drug. The similarities of response patterns often relate to the mechanism of action, resistance, and structural properties within the screened compound set. GI_{50} patterns have been actively investigated since the inception of the NCI screen, beginning with the work of Paull and co-workers, using applications of statistical correlations and other computational techniques.[13,15−31] The experimental concept of using cell-line response patterns as a means to relate compounds to their activities has since been utilized by other groups.[4,32,33]

Studies of drug mechanisms based on a biological activity pattern have established that some drug classes, not necessarily with the same structure, cause a specific growth inhibition pattern across these tumor cell lines that can be related, by secondary testing, to a precise biological effect.[11−13,20,25,28,30,34−36] These "vetted" results confirm that these data can also be reverse-mined to suggest novel compounds that elicit a desired mechanism of action, based solely on analogy to characteristic growth inhibition patterns. This model further offers a strategy for deselecting newly screened compounds if their responses correspond to a previously well-exploited drug activity.

The capacity to examine a chemically diverse probe library across a large number of different types of cancer cells also reveals some general observations regarding connections between the similarity of GI_{50} patterns and the similarity of

* Corresponding author tel.: 301-846-5665; fax: 301-846-6798; e-mail: wallqvist@ncifcrf.gov.
† National Cancer Institute.
‡ Science Applications International Corporation.

10.1021/ci0501544 CCC: $33.50
Published on Web 11/16/2005

chemical structures. Most evident among these observations is that a highly correlated cellular growth inhibition pattern is a better predictor of chemical structure than the reverse case of high structural similarity at predicting growth inhibition. This does not necessarily mean that chemical structure is a poor predictor of biological outcome per se; quite the contrary, it highlights the fact that conclusions about target-selective chemistries may be improved by additional considerations of biological data similarities. Although our analysis has been developed using the NCI's tumor screening data, the results may be generalized to other screening systems. In particular, there are limits to the extent that structural similarity will translate into a similar biological response, and vice versa. Our analysis demonstrates that quantifiable measures can, however, be assigned to the degree of change that can be expected when extrapolating between these measures.

METHODS

The NCI-60 drug discovery panel was developed as a cell-based in vitro tool to assess the anticancer activity of compounds against a range of tumors, including lung, renal, colorectal, ovarian, breast, prostate, central nervous system, melanoma, and hematological malignancies.[11,12,14] The screening data contains concentration values (GI_{50}) determined from a dose response curve at which the tested drug resulted in a 50% reduction in the net protein increase compared to control cells during a 48 h drug incubation. The GI_{50} data for each of the 60 cell lines is used to construct a data vector for each compound tested. The final data vectors used in our analysis were log-transformed and selected to have a maximum of 20 missing data elements and a signal covariance of at least 0.02. Missing data elements were not included in any calculation. The GI_{50} values for typical compounds range from a high concentration of 10^{-4} M to very sensitive compounds that only require 10^{-8} M to evoke a GI_{50} response. The pattern of GI_{50} values across the tumor cell lines has proven effective for identifying mechanisms of action for some drug classes and aids in the classification of novel drugs submitted to the NCI's tumor screen.[13,37]

Using the NCI repository of compounds, we constructed a 2048-bit vector representation using DayLight tools for molecules that have been tested in the NCI tumor screen. The NCI structural database constitutes a relatively diverse and eclectic set of chemistries.[3] Approximately 16 000 compounds with both GI_{50} measurements and a bit-vector representation were used to calculate structural similarity using the Tanimoto coefficient between bit vectors and GI_{50} growth inhibition pattern similarity using the Pearson correlation coefficient.

The Tanimoto coefficient or Jaccard similarity measure is defined as

$$t(\vec{x},\vec{y}) \equiv \frac{\vec{x} \times \vec{y}}{|\vec{x}| + |\vec{y}| - \vec{x} \times \vec{y}}$$

where $\vec{x}$ and $\vec{y}$ are the extracted bit vectors of zeros and ones for each compound. Typically 200−400 bit features per molecule are turned on in our representation. $t(\vec{x},\vec{y})$ measures the similarity of two sets of bit-vector descriptors using the ratio of the size of their intersection divided by the size of their union and can range between 0 and 1 but is limited by the largest fraction of bit-vector features:

$$\mathrm{wt} \equiv \max\,[t(\vec{x},\vec{y})] = \frac{\min\,(|\vec{x}|,|\vec{y}|)}{\max\,(|\vec{x}|,|\vec{y}|)}$$

This effectively weighs larger Tanimoto measure to molecules that have roughly the same number of bit features. A consideration of wt may be more suitable to substructure searching. An artifact of coding molecular bits is that two chemically different molecules can have the same bit-vector representation. In the present analysis, we are restricting the included structural pairs to have a wt $\geq$ 0.70 and a min $(|\vec{x}|,|\vec{y}|)$ of 100 bit features.

The Pearson or sample correlation coefficient of the GI_{50} response is defined as

$$r(\vec{u},\vec{v}) = \frac{\sum_i (u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\sum_i (u_i - \bar{u}) \sum_i (v_i - \bar{v})}}$$

where $\vec{u}$ and $\vec{v}$ are the log-transformed vectors of GI_{50} measurements of different cancer cells and $\bar{u}$ denotes the average of all elements in $\vec{u}$. A typical data vector has between 50 and 60 valid entries for cell line measurements, where each cell line entry is an average of GI_{50} values for repeated tests of the compound. The correlation coefficient measures the fidelity of a linear fit of $v(u)$ and takes on values between −1 and +1. A correlation coefficient of 1 indicates that each vector is linearly dependent on the other. It does not mean that the vectors are the same.

The ability of either structural or growth inhibition similarities to correctly recover the corresponding growth inhibition or bit-vector pattern was measured as a function of similarity. We define $N(t \geq a; r \geq b)$ as the number of molecule pairs in the data set that have a Tanimoto coefficient of at least a between the structural bit vectors and a Pearson correlation coefficient of at least b between the growth inhibition data vectors. We can then construct

$$F(t = a|r = b) \equiv \frac{N(t \geq a; r \geq b)}{N(t \geq a; r \geq -1.0)}$$

and

$$F(r = a|t = b) \equiv \frac{N(t \geq b; r \geq a)}{N(t \geq 0.0; r \geq a)}$$

$F(t = a|r = b)$ measures the fraction of all pairs given a Tanimoto cutoff of a that have a corresponding GI_{50} Pearson correlation coefficient of b or larger. Similarly, $F(r = a|t = b)$ measures the fraction of all pairs given a GI_{50} Pearson correlation coefficient cutoff of a that have a corresponding Tanimoto coefficient of b or larger. As an example, $F(t = 0.60|r = 0.80)$ gives the fraction of pairs that were selected to have a Tanimoto coefficient of 0.60 or larger that also have a GI_{50} correlation of 0.80 or larger. Trivially, $F(t = 0.00|r = -1.00)$ and $F(r = -1.00|t = 0.00)$ are 1. $F(r = 1.00|t = 0.90)$ is the fraction of pairs that have a perfect GI_{50} correlation and whose structural similarity, gauged by the Tanimoto coefficient, is larger than or equal to 0.90. The connection between the constructs $F(t = t_c|r = r_c)$ and $F(r$

$= r_c|t = t_c$) is that whereas the former answers the question of what fraction of all molecule pairs that share structural components also share biological responses (positive predictive value), the latter gives the fraction of all molecule pairs that share a biological response that also share structural similarity (sensitivity or coverage). These functions were calculated for all pairs of compounds in the NCI-60 screening data.

RESULTS

Since the 3D configuration of a molecule in solution is not known, nor can it's configuration in a biologically relevant bound state be predicted from the chemical structure alone,[38] a general approach is to rely on 2D descriptors. If 3D descriptors are available, they can complement the information derived only from the chemical connectivity.[39] Here, we use a 2D description of the compound, intended to capture the primary connectivity of the molecule, which is often employed to search and characterize large databases of chemical structures. The best possible representation is context-dependent and continues to be an active field of research.[40] The most recent and commonly used measures of chemical similarity are based on bit vectors,[2,41] with the Tanimoto coefficient as an index of bit-vector similarity. Here, we use a two-dimensional DayLight bit-vector representation of chemical structure, ignoring stereo effects, and measure chemical similarity using Tanimoto coefficients (t).

The biological response to a chemical probe will vary considerably depending on the system being studied and the ability to quantify the data. Here, we use the growth inhibition response of immortalized cancer cells measured in a fluorescent assay 48 h after drug exposure. The drug concentration at which 50% growth inhibition has been achieved, compared to nonexposed cells, is estimated from a dose−response curve and recorded for each of the NCI's cell lines.[11,14] The combined GI_{50}'s for all cell types encode the biological response to a particular compound, and here, we use these patterns as an effective definition of a cellular mechanism of action. Compounds with similar biological response profiles are, thus, assumed to share a similar cellular mechanism of action. When GI_{50} profiles were used, broad ranges of compounds have been classified and recognized to function as antimitotic, DNA-interfering, or stress-response active agents. Testing the biological response similarity using GI_{50} correlation coefficients (r), we can evaluate and contrast these measures with a bit-vector-based Tanimoto coefficient of structural similarity (t). Likewise, testing the structural similarity between compounds, we can evaluate their biological response similarity. The goal, here, is to assess the strength of the structure/activity relations in the general sense and to provide some specific examples illustrating this duality.

Structure to Biological Response. We will first investigate the implications of structural similarity. The reliability of structure as a predictor of biological response similarity is given in Figure 1. From pairs of molecules in the data set (excluding identities), we can calculate the Tanimoto coefficient for each compound pair. If we introduce a Tanimoto threshold and only look at molecular pairs above this cutoff, we can determine the fraction of these pairs satisfying a minimum desired biological response correlation, r. This is

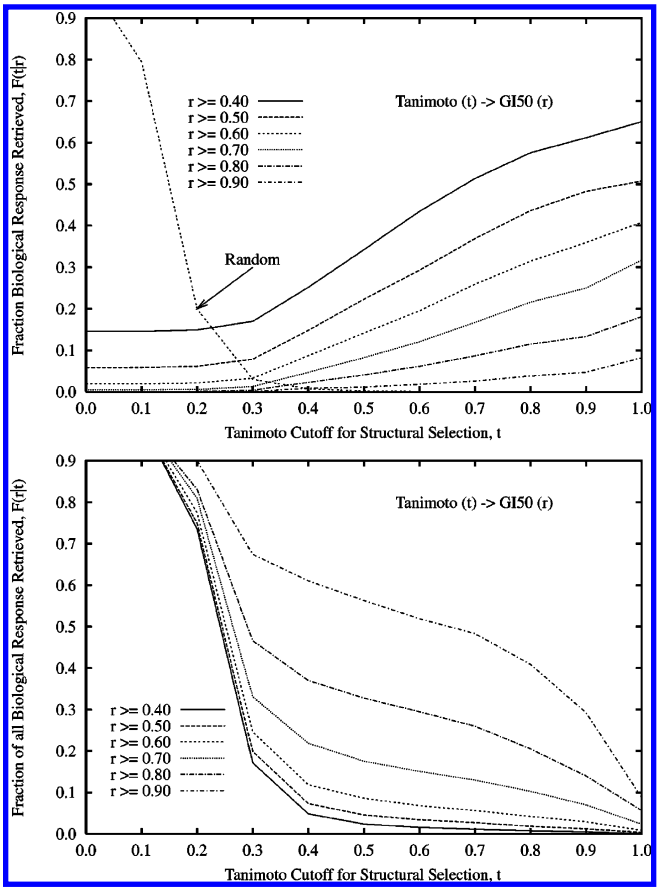

Figure 1. Structure to biology. (A) The graph shows the fraction of correctly identified molecular pairs $F(t|r)$ that evoke the same biological response as that measured by the growth inhibition correlations (r) given as a function of the structural similarity Tanimoto coefficient (t) cutoff. This function is given for a selection of six growth inhibition correlation thresholds, $r \geq 0.40$, $r \geq 0.50$, $r \geq 0.60$, $r \geq 0.70$, $r \geq 0.80$, and $r \geq 0.90$. The interpretation of, e.g., $F(t = 0.70|r = 0.60)$ is that, when using a Tanimoto cutoff of 0.70 for selecting similar molecules, with a measure of biological response similarity measured via the GI_{50} correlation at $r \geq 0.60$, 25% of all molecules retrieved have the same biological response. Despite the structural similarity, 75% of all molecules retrieved have a different GI_{50} response profile, indicative of an unrelated mechanism of action for these compounds. Also included in the graph is the probability of finding a pair of molecules with the same or higher structural similarity regardless of biological response. (B) The fraction of all molecules with a particular biological response that is retrieved using structural similarity to identify compounds $F(r|t)$, given as a function of the structural similarity Tanimoto coefficient (t) cutoff.

a test of the concept that structurally similar compounds yield similar biological responses, and here, we quantify this feature as the fraction of compounds that jointly satisfy structural similarity ($t \geq t_{\text{threshold}}$) and response similarity ($r \geq r_{\text{threshold}}$). This fraction is denoted as $F(t|r)$, see Methods section. Since the threshold for either structural or biological response similarity can vary depending on the study and level of tolerance for inclusion, it is appropriate to study these variations on the basis of thresholds. In Figure 1A the x axis gives the threshold of structural similarity for selecting pairs and the y axis gives the fraction of the selected pairs that have the desired biological response property, which is indicated by the curves for different r-value limits. Each curve in the figure represents the variation in response depending on the reliability of structural similarity (t) as a function of the strength of their biological response correla-

tion (r). If we are interested in the case of using a structural similarity of $t \geq 0.50$ to retrieve molecules with a biological response correlation of $r \geq 0.40$, the data indicate that, using such a structural filter, 35% of the molecule pairs retrieved will share the desired biological response, [$F(t = 0.50|r = 0.40) = 0.35$]. As we increase the threshold of structural similarity, the fraction of correctly identified biological response correlations increases but approaches a ceiling of 65% at the highest structural similarity. The predictive capacity of structural similarity for biological response is, thus, limited. A salient extrapolation of these results is that small variations in structure can cause large changes in biological response. Higher correlation thresholds for biological response result in a diminished fraction of acceptable pairs, with the decrease occurring nearly linearly with increasing r. Commonly used Tanimoto thresholds ranging from 0.70 to 0.80 retrieve structure sets where 50% of the GI_{50} correlations at $r \geq 0.40$ are the same as the parent compound, and this changes to 30% at $r \geq 0.60$, with a low positive predictive value of 5% at $r \geq 0.90$. In Figure 1B, we can gauge the total fraction of molecules with the desired biological response we retrieve from the data on the basis of a structural comparison. In the same Tanimoto range of 0.70−0.80, roughly 20% of all molecules with the biological response defined as $0.70 \leq r \leq 0.80$ are retrieved. Using a higher threshold for biological response similarity increases the number of false positives retrieved but also increase the overall number of structures identified to cause that response.

The structure/activity similarity principle gradually breaks down as structural similarity is diminished. This effect is mostly due to the fact that less similar molecules will affect different biological pathways in different manners, though still, not even at the highest similarity can all compounds causing the same biological response be identified using a parent compound. The unidentified compounds remain hidden and can only be retrieved by testing molecules that are not similar to the parent compound. The graphs in Figure 1 can also be used to establish minimum Tanimoto thresholds for maximizing the variation of the GI_{50} response. So, if we turn our query around and assume that a GI_{50} correlation of 0.40, in this case, constitutes an upper limit for what we can accept as a biological response correlation, a diverse library should consist of structures having Tanimoto coefficients no higher than 0.30 between any pair of molecules to ensure that 80% of the biological response is not correlated.

The question of whether structure is a good predictor for biological activity is, thus, one of degrees. Strong structural similarity in the bit-vector representation of molecules does not necessarily imply similarity in activity. The cellular systems studied here are, by nature, complex compared to biomolecular assays, as these latter assays are designed to answer different questions. In the case of molecular assays designed to assess activity against a single molecule or biomolecular complex, structural similarity may be a more-reliable estimator of specific activity against a single target. However, the impact of these changes on the biological response could be larger than indicated in the biomolecular assay. Alternatively, the capacity of a small structural variant having a broader range of biological targets would not be revealed in a biomolecular assay.

Biological Response to Structure. The connection between structure and biological response is not symmetric. Whereas it is generally correct that highly similar molecules are affecting the same target, the converse claim of similar biological effect does not generally imply that the structures of the molecules causing the effect are the same. Such examples may include different classes of molecules binding to the same binding pocket, for example, HIV-protease inhibitors,[42] or binding/affecting different sites of the same macromolecule, for example, different alkylators or topoisomerase poisons/suppressors.[34]

In the present analysis, we can quantify these questions using the GI_{50} responses across cell lines as our biological readout. This pattern of growth inhibition response has proven effective for identifying specific mechanisms of action, although these patterns have no general interpretation by themselves. The fact that we can observe a pattern in the first place is due to differences in growth of each cancer cell upon drug insult. The primary interest in this response is as a biological marker characteristic of a mechanistic class of drug action. Any postulated connection between target response and drug action must be validated through further experiments.

The correlation of GI_{50} patterns, as shown in Figure 2, can be used to gauge the chemical similarity of compounds causing such a response. If we arbitrarily use a minimum GI_{50} correlation coefficient of, for example, 0.70 to retrieve similar growth inhibition patterns, the fraction of retrieved molecule pairs that have a Tanimoto coefficient of 0.40 or larger is 25%, $F(r = 0.70|t = 0.40) = 0.25$. To boost the retrieval of chemically similar compounds to 50%, at this level of structural similarity, we must only consider those GI_{50} patterns that at least have a correlation coefficient of $r = 0.85$. The difference in the number of retrieved pairs at $r = 0.85$ is not strongly dependent on the desired Tanimoto similarity; that is, at a t value of 0.60, 40% [$F(r = 0.85|t = 0.60) = 0.40$] of the retrieved pairs have this property, and 25% have this property at a minimum similarity level of $t = 0.90$ [$F(r = 0.85|t = 0.90) = 0.25$]. On the other hand, the fraction of structurally similar molecules retrieved from all compounds with that structural similarity based on biological response, shown in Figure 2B, is typically below 10% at $0.70 \leq r \leq 0.90$ with $0.40 \leq t \leq 0.90$.

The behaviors of these two sets of curves in Figures 1 and 2 are fundamentally different, depending on whether the goal is to retrieve a structure from the response or vice versa. Increasing the GI_{50} similarity above an r value of 0.60 rather dramatically increases the likelihood of finding similar structures in the retrieved set. Roughly, the average change in the fraction of structural pairs with a high similarity per unit change in r ($\Delta F/\Delta r$) is about 1.0−2.0. Thus, an increase in r of 0.1 gains 10−20% in the positive prediction of structurally similar compounds. However, with a close to perfect GI_{50} pattern correlation of 1.00, still only 50% of the compounds retrieved have a minimum structural similarity threshold of 0.90. Increasing the response similarity by selecting a higher structural similarity thus depends on the desired minimum level of GI_{50} correlation and, hence, biological response coherence. The employment of higher r-value thresholds does not necessarily lead to an improvement in defining a structurally similar set of compounds. At a Tanimoto coefficient of 0.50, this change is roughly linear in r, ranging from a high $\Delta F(t = 0.5|r)/\Delta t$ of 0.9 at $r \geq 0.40$ to 0.1 at $r \geq 0.90$.

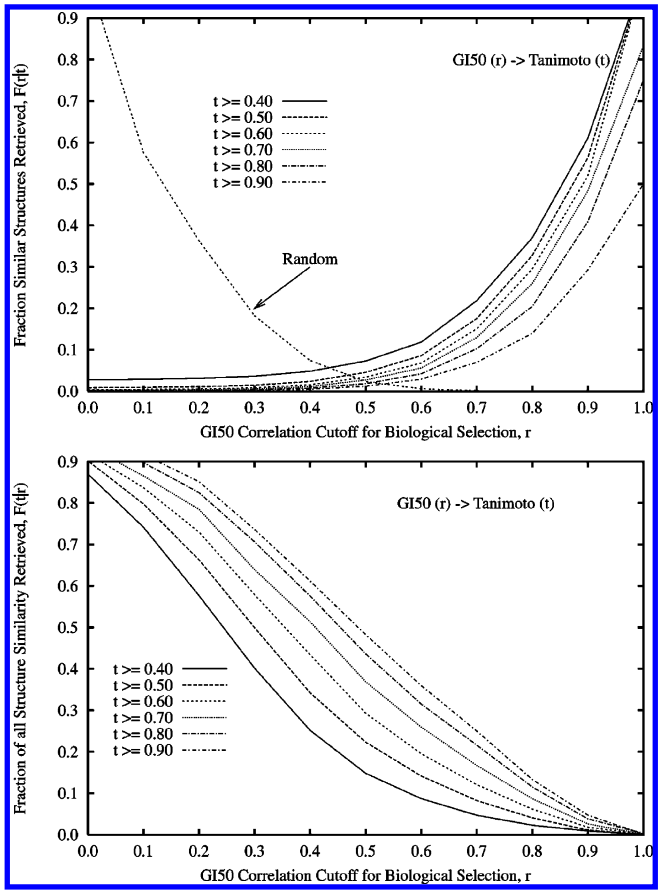

Figure 2. Biology to structure. (A) This figure is the counterpart of Figure 1 but with structure and response reversed; thus, the figure displays the fraction of correctly identified molecular pairs $F(r|t)$ that have a similar structure as identified by a Tanimoto coefficient (t) given as a function of biological response similarity as measured by the growth inhibition correlations (r). This function is given for a selection of six structural similarity thresholds, $t \geq 0.40$, $t \geq 0.50$, $t \geq 0.60$, $t \geq 0.70$, $t \geq 0.80$, and $t \geq 0.90$. The interpretation of, e.g., $F(r = 0.90|t = 0.90)$ is that, when using a GI_{50} correlation cutoff of 0.90 for selecting similar growth inhibition responses, with a measure of structural similarity measured via a Tanimoto coefficient at $r \geq 0.90$, 30% of all molecules retrieved have a similar structure. Also included in the graph is the probability of finding a pair of molecules with the same or higher GI_{50} correlation regardless of structural similarity; e.g., finding a pair of molecules with a GI_{50} correlation of 0.30 or larger is roughly 20%, whereas finding a correlation of 0.90 or larger is virtually zero. (B) The fraction of all molecules with a particular structure similarity that is retrieved using biological response to identify compounds $F(t|r)$, given as a function of biological response similarity as measured by the growth inhibition correlations (r).

The probability that a molecule pair in the data set either has a Tanimoto coefficient of $t_{threshold}$ regardless of biological response or a GI_{50} correlation of $r_{threshold}$ regardless of structural similarity is also shown in Figures 1A and 2A, respectively. The random chance of picking a pair of molecules in this data set as having a Tanimoto coefficient above 0.20 is roughly 20%; at $t \geq 0.50$, the chance is 0.20%, and at $t \geq 0.80$, it is 0.02%. Conversely, the random chance of selecting a pair of molecules, regardless of structure, that have a GI_{50} correlation coefficient above 0.50 is 2%, and at $r \geq 0.80$, it is less than 0.01%. Thus, the enrichment in selecting structures based on chemical similarity greatly enhances the probability that they will have a similar biological response.

Mechanism of Action Classifications. Drugs are generically grouped into categories reflective of the type of interaction the compound, or similar compounds, have with a target. These groups can encompass many diverse molecules, for example, alkylators that have DNA as their primary target, or more selective ones, for example, boronic acid analogues that target specific residues in the 26S proteasome.[43] The distinction between functionality, molecular structure, and the mechanism of action is not always made. Functionality reflects the primary interaction between the drug and targets, molecular structure defines the 2D chemistry of the tested unmetabolized drug, and the mechanism of action reflects a drug interacting with targets within the cellular environment and causing a phenotypic change of the cell population, for example, growth inhibition, necrosis, or apoptosis. It is the latter two characterizations of structure and the mechanism of action that are intended to be captured by the Tanimoto and GI_{50} correlation coefficients.

Table 1. Evaluation of Structural Similarity and GI_{50} Correlations for a Set of Drug Classes with Varying Targets and Mechanisms of Action[a]

drug class	N	$\langle t \rangle_{intra}$ (σ_t)	$\langle r \rangle_{intra}$ (σ_t)
acetogeneins	26	0.81 (0.19)	0.52 (0.16)
alkylators	251	0.20 (0.13)	0.29 (0.23)
antibiotics	84	0.52 (0.27)	0.45 (0.24)
antifolates	48	0.53 (0.22)	0.56 (0.19)
boronic acids	13	0.64 (0.23)	0.71 (0.11)
CDK inhibitors	21	0.52 (0.35)	0.42 (0.32)
channel agents	46	0.47 (0.32)	0.26 (0.27)
chelating agents	5	0.65 (0.20)	0.55 (0.10)
DNA polymerase inhibitors	11	0.80 (0.10)	0.60 (0.14)
direct membrane	130	0.45 (0.18)	0.36 (0.24)
geldanamycin analogues	21	0.87 (0.08)	0.43 (0.20)
golgi disruptive agents	23	0.49 (0.20)	0.20 (0.34)
intercalating agents	251	0.45 (0.15)	0.31 (0.20)
mitotic	66	0.50 (0.31)	0.48 (0.22)
phosphatase/kinase	32	0.42 (0.20)	0.41 (0.30)
purine antimetabolites	14	0.40 (0.41)	0.52 (0.23)
pyrimidine antimetabolites	69	0.51 (0.20)	0.32 (0.26)
steroids	66	0.35 (0.20)	0.23 (0.28)
topoisomerase I	62	0.75 (0.21)	0.78 (0.16)
topoisomerase II	54	0.32 (0.25)	0.44 (0.21)

[a] The average Tanimoto and correlation coefficient are evaluated for all pairs of molecules within each group. The standard deviation is also given for each calculation. Both the structural similarity and biological coherence vary considerably for these drug classes, though higher structural similarity, in general, implies higher biological coherence.

We investigated the structural similarity from a bit-vector representation and cellular response similarity as recorded by the GI_{50} for a group of 20 different and diverse drug classifications,[34] as given in Table 1. These small molecule drugs include both noncovalent- and covalent-type agents capable of targeting proteins, DNA, and other macromolecular assemblies in the cell. The Tanimoto coefficient from all pairwise structural comparisons within each classification group as well as between groups and the corresponding average GI_{50} correlation coefficients are shown in Figure 3. Encouragingly, the structural measure is correlated with the biological measure ($r = 0.61$, $p = 0.004$) for the defined groups, and correspondingly, the between-group Tanimoto coefficient and GI_{50} correlation are not correlated with each other ($r = 0.10$, $p = 0.18$). For these classifications, structure and effect are correlated, though not all of them are immediately useful for analyzing structure/activity relationships.

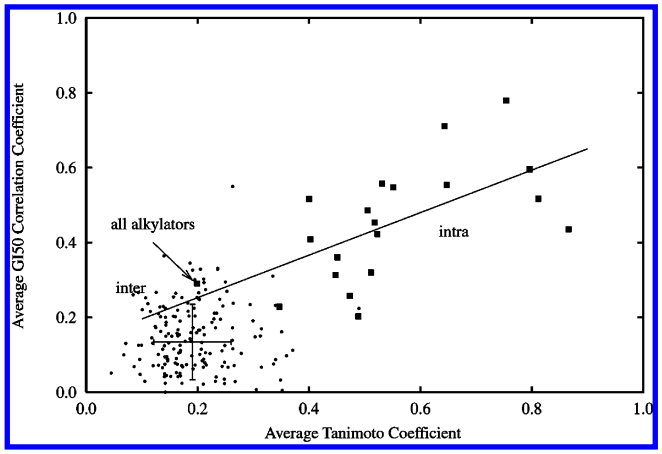

Figure 3. Drug MOA classification relations to structure and biological response. Correspondence between structural similarity and biological response coherence as measured via averaged Tanimoto coefficients and GI_{50} correlations for groups of compounds with characterized modes of action from literature sources.[34] The graph shows very little intergroup coherence, whereas intragroup values exhibit correlation between structural and biological response measures.

The alkylator group contains 251 compounds, and this classification does not make much sense in terms of cause and effect. The within-group correlation cannot be distinguished from between-group values. From Figure 1A, we can immediately gauge the significance of this correlation ($t = 0.20$, $r = 0.29$) as not being different from a random one. The gross structural classification of alkylators from this grouping is, thus, not meaningful, even though the functionality of these molecules is shared. Since there is no coherence in this group of compounds, it is appropriate to further dissect this group of compounds to establish whether alkylation can generate distinguishable biological responses.

We uniquely partitioned the alkylator compound set, on the basis of either pairwise Tanimoto or correlation coefficients, into clusters on the basis of expanding our compound lists to include all pairs above a given similarity threshold. Singlet clusters were not included in the subsequent analysis. This clustering scheme identifies compound groups that are highly correlated within the group and lack correlations between groups. This structural classification of alkylators yields compound groups that have, on the average, an intragroup Tanimoto coefficient of 0.84. The concomitant biological response correlation within these groups has an average value of 0.55. Using Figure 1A, we can now ascertain how well the structural classification of alkylators perform; at $t = 0.84$ and $r = 0.55$, we can interpolate to find the fraction of compounds with the same biological response that is retrieved, that is, roughly 40%. Thus, despite the high Tanimoto coefficient, we are missing about 60% of all compounds that evoke this biological response. The intergroup Tanimoto and correlation values for compound groups assembled via Tanimoto-score clustering still show low values comparable to the average for the entire class of alkylators, at $\langle t \rangle_{\text{inter}} = 0.14$ and $\langle r \rangle_{\text{inter}} = 0.26$. This implies that the structural classification can differentiate between biological responses that are not solely due to indiscriminate alkylation.

One example of two such classes is compounds that contain the alkylating moiety bis(2-chloroethyl)amino group. These compounds can be separated on the basis of structure alone into two groups, which, in addition to the bis(2-chloroethyl)amino group, contain acridine, for example, quinacrine mustard dihydrochloride, or a quinolin ring system. These groups have an average Tanimoto coefficient of 0.42 and a GI_{50} correlation of 0.18 between them, whereas among themselves, they are characterized by Tanimoto coefficients of 0.80 and 0.85 and GI_{50} correlation coefficients of 0.59 and 0.55, respectively. Although the apparent mechanism for this compound group appears to be roughly similar, alkylation versus intercalation, modification of the chemical structure not related to this function per se results in a separable and characteristic biological response.

The acetogenins[34] are a group of structurally similar compounds composed of an unbranched C_{32} or C_{34} fatty acid ending in a γ-lactone and are thought to affect the mitochondrial respiratory chain complex I; however, their diversity in observed effect points to additional targets or mechanisms involved in other cellular functions. The acetogeneins are structurally homogeneous, with an average Tanimoto coefficient of 0.81 among all molecules. Their average biological response similarity is 0.51 among this set. Sorting this group of compounds on the basis of pairwise Tanimoto scores allows us to form three groups of acetogenins compounds where the average intragroup Tanimoto and GI_{50} correlation coefficients are 0.98 and 0.55, respectively. Thus, even though we have increased the structural similarity almost to the limit in this set, the biological response remains about the same within these clusters as compared to the average between all acetogenins. Again, we can estimate the success of this group classification by finding the ($t = 0.98$, $r = 0.55$) value in Figure 1A. This indicates that, for each group, approximately 45% of the identified compounds share a similar ($r \geq 0.55$) response profile.

In general, the biological response is not known for the compound set being tested, but in our case, we have the associated GI_{50} profiles for each molecule. If we sort the acetogenins on the basis of their GI_{50} response similarity, they cluster into four groups with average intragroup Tanimoto and correlation coefficients of 0.78 and 0.81, respectively. In this case, the higher response similarity did not select compounds with an average Tanimoto coefficient higher than 0.78.

Topoisomerase II inhibitors show both a low structural and a low biological coherence; the averaged intra-Tanimoto and GI_{50} correlations are 0.32 and 0.44, respectively. In comparison to topoisomerase I inhibitors, the different ways in which topoisomerase II can be inhibited via complex stabilizing and catalytic inhibitors are reflected in these values. Sorting topoisomerase II inhibitors on the basis of structural similarities raises the average intracluster coherence to a Tanimoto score of 0.88 and a GI_{50} correlation of 0.70. From these values, we can use Figure 1A to find the fraction of correct biological response associated with this structural classification, which is roughly 25%. From Figure 1B, we can see that the relatively high structural similarity given by the Tanimoto score of 0.88 misses roughly 90% of all other structures that evoke a similar biological response. On the other hand, sorting topoisomerase II inhibitors on the basis of biological responses raises the average GI_{50} correlation to 0.76, while lowering the Tanimoto score to 0.58. A stronger similarity in biological response is, thus, achievable when putting less emphasis on structural similarities.

CONCLUSIONS

Variations in probe structure will almost always result in a change of specific cellular activity. In the traditional viewpoint, this may be an increase or decrease of the binding constant or activity to a particular set of targets. When a large panel of cancer cell lines is used, these changes can, in turn, be observed as changes in the pattern of GI_{50} response for each cell line. A change in correlation coefficient between GI_{50} patterns for different agents reflects a change in the specific cellular response. For pairs of molecules that have a Tanimoto coefficient of 0.80 or higher, only 12% of these pairs will have a GI_{50} correlation coefficient of 0.80 or higher, yet still, 57% of them will have an r value of 0.40 or higher. The specificity of the cellular responses is, thus, to a large extent, encoded in the molecular structure.

Although we do not have direct control over the growth inhibition response, we can still a posteriori investigate the effect of changes in GI_{50} patterns. Practically, the GI_{50} patterns are an indication of cellular sensitivity for a particular compound, through direct or indirect mechanisms. Small changes of r are compatible with relatively large changes in the chemistry of the compounds. At a GI_{50} pair correlation of 0.80, we can expect 20% of the selected pairs to have a Tanimoto coefficient of 0.80 or larger; 37% of the pairs will, on the other hand, have a Tanimoto coefficient of 0.40 or larger. Thus, a majority of the selected pairs will not have strongly similar structures.

The variations in biological response are characteristic of broad mechanistic processes that underlie the health and growth of the cell. In evaluating the biological effects of chemical changes to a molecule, activity to all possible targets and pathways has to be taken into account. Using multiple biological assays to construct a biological response pattern for a test compound provides a basis for hypothesizing modes of action and how changes in the structure affect cellular phenotypes. It is the biological consequences of changing the structure that are outlined and quantified in this work. Our analysis links structural aspects of molecules to the biological response in a quantifiable manner and can be used to gauge either's effect on the other.

ACKNOWLEDGMENT

This project has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under Contract NO1-CO-12400. The content of this publication does not necessarily reflect the view or policies of the Department of Health and Human Services, nor does the mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

REFERENCES AND NOTES

(1) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activity? *J. Med. Chem.* **2002**, *45*, 4350−4358.

(2) Willett, P. Similarity-based approaches to virtual screening. *Biochem. Soc. Trans.* **2003**, *31*, 603−606.

(3) Voigt, J. H.; Bienfait, B.; Wang, S.; Nicklaus, M. C. Comparison of the NCI Open Database with Seven Large Chemical Structural Databases. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 702−712.

(4) Fliri, A. F.; Loging, W. T.; Thadeio, P. F.; Volkmann, R. A. Biological spectra analysis: Linking biological activity profiles to molecular structure. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 261−266.

(5) Diller, D. J.; Hobbs, D. W. Deriving knowledge through data mining high-throughput screening data. *J. Med. Chem.* **2004**, *47*, 6373−6383.

(6) Oprea, T. I. Chemoinformatics in Drug Discovery. In *Methods and Principles in Medicinal Chemistry*; Mannhold, R., Folkers, H. K. G., Eds.; Wiley-VCH: Weinheim, Germany, 2005.

(7) Matter, H.; Rarey, M. Design and diversity analysis of compound libraries for lead discovery. *Combinatorial Chemistry. Synthesis, Analysis, Screening*; Wiley-VCH: Weinheim, Germany, 2000; pp 409−439.

(8) Agrafiotis, D. K.; Lobanov, V. S.; Salemme, F. R. Combinatorial informatics in the post-genomics ERA. *Nat. Rev. Drug. Discovery* **2002**, *1*, 337−346.

(9) Boyd, M. R. The NCI in vitro anticancer drug discovery screen. *Anticancer Drug Development Guide: Preclinical Screening, Clinical Trials and Approval*; Humana Press: Totowa, New Jersey, 1995; pp 23−41.

(10) Boyd, M.; 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.

(11) Shoemaker, R. H.; Monks, A.; Alley, M. C.; Scudiero, D. A.; Fine, D. L.; McLemore, T. L.; Abbott, B. J.; Paull, K. D.; Mayo, J. G.; Boyd, M. R. Development of human tumor cell line panels for use in disease-oriented drug screening. *Prog. Clin. Biol. Res.* **1988**, *276*, 265−286.

(12) Scudiero, D. A.; Shoemaker, R. H.; Paull, K. D.; Monks, A.; Tierney, S.; Nofziger, T. H.; Currens, M. J.; Seniff, D.; Boyd, M. R. Evaluation of a soluble tetrazolium/formazan assay for cell growth and drug sensitivity in culture using human and other tumor cell lines. *Cancer Res.* **1988**, *48*, 4827−4833.

(13) 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.

(14) Monks, A.; Scudiero, D.; Skehan, P.; Shoemaker, R.; Paull, K.; Vistica, D.; Hose, C.; Langley, J.; Cronise, P.; Vaigro-Wolff, A. Feasibility of a high-flux anticancer drug screen using a diverse panel of cultured human tumor cell lines. *J. Natl. Cancer Inst.* **1991**, *83*, 757−766.

(15) Lee, J. S.; Paull, K.; Alvarez, M.; Hose, C.; Monks, A.; Grever, M.; Fojo, A. T.; Bates, S. E. Rhodamine efflux patterns predict P-glycoprotein substrates in the National Cancer Institute drug screen. *Mol. Pharmacol.* **1994**, *46*, 627−638.

(16) Osdol, W. W. v.; 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.

(17) Alvarez, M.; Paull, K.; Monks, A.; Hose, C.; Lee, J.-S.; Weinstein, J.; Grever, M.; Bates, S.; Fojo, T. Generation of a drug resistant profile by quantitation of mdr-1/P-Glycoprotein in the cell lines of the National Cancer Institute anticancer drug screen. *J. Clin. Invest.* **1995**, *95*, 2205−2214.

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

(19) Freije, J. M.; Lawrence, J. A.; Hollingshead, M. G.; de la Rosa, A.; Narayanan, V.; Grever, M.; Sausville, E. A.; Paull, K.; Steeg, P. S. Identification of compounds with preferential inhibitory activity against low-Nm23-expressing human breast carcinoma and melanoma cell lines. *Nat. Med.* **1997**, *3*, 395−401.

(20) 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.

(21) Scala, S.; Akhmed, N.; Rao, U. S.; Paull, K.; Lan, L. B.; Dickstein, B.; Lee, J. S.; Elgemeie, G. H.; Stein, W. D.; Bates, S. E. P-glycoprotein substrates and antagonists cluster into two distinct groups. *Mol. Pharmacol.* **1997**, *51*, 1024−1033.

(22) Wosikowski, K.; Schuurhuis, D.; Johnson, K.; Paull, K. D.; Myers, T. G.; Weinstein, J. N.; Bates, S. E. Identification of epidermal growth factor receptor and c-erbB2 pathway inhibitors by correlation with gene expression patterns. *J. Natl. Cancer Inst.* **1997**, *89*, 1505−1515.

(23) Shi, L. M.; Fan, Y.; Myers, T. G.; Paull, K. D.; Weinstein, J. N. Mining the anticancer activity database generated by the U. S. National Cancer Institute's drug discovery program using statistical and artificial intelligence techniques. *Model. Sci. Comput.* **1998**, *38*, 189−196.

(24) Shi, L. M.; Myers, T. G.; Fan, Y.; O'Connor, P. M.; Paull, K. D.; Friend, S. H.; Weinstein, J. N. Mining the NCI anticancer drug

discovery database: genetic function approximation for the QSAR study of anticancer ellipticine analogues. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 189−199.
(25) 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. Pharmacol.* **1998**, *53*, 241−251.
(26) Bradshaw, T. D.; Shi, D. F.; Schultz, R. J.; Paull, K. D.; Kelland, L.; Wilson, A.; Garner, C.; Fiebig, H. H.; Wrigley, S.; Stevens, M. F. Influence of 2-(4-aminophenyl)benzothiazoles on growth of human ovarian carcinoma cells in vitro and in vivo. *Br. J. Cancer* **1998**, *78*, 421−429.
(27) Koo, H.-M.; Gray-Goodrich, M.; Kohlhagen, G.; McWilliams, M. J.; Jeffers, M.; Vaigro-Wolff, A.; Alvord, W. G.; Monks, A.; Paull, K. D.; Pommier, Y.; Vande Woude, G. F. The ras oncogene-mediated sensitization of human cells to topoisomerase II inhibitor-induced apoptosis. *J. Natl. Cancer Inst.* **1999**, *91*, 236−244.
(28) 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.
(29) Staunton, J. E.; Slonim, D. K.; Coller, H. A.; Tamayo, P.; Angelo, M. J.; Park, J.; Scherf, U.; Lee, J. K.; Reinhold, W. O.; Weinstein, J. N.; Mesirov, J. P.; Lander, E. S.; Golub, T. R. Chemosensitivity prediction by transcriptional profiling. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 10787−10792.
(30) Fan, Y.; Shi, L. M.; Kohn, K. W.; Pommier, Y.; Weinstein, J. N. Quantitative structure−antitumor activity relationships of camptothecin analogues: cluster analysis and genetic algorithm-based studies. *J. Med. Chem.* **2001**, *44*, 3254−3263.
(31) Blower, P. E.; Yang, C.; Fligner, M. A.; Verducci, J. S.; Yu, L.; Richman, S.; Weinstein, J. N. Pharmacogenomic analysis: correlating molecular substructure classes with microarray gene expression data. *Pharmacogenomics J.* **2002**, *2*, 259−271.
(32) Dan, S.; Tsunoda, T.; Kitahara, O.; Yanagawa, R.; Zembutsu, H.; Katagiri, T.; Yamazaki, K.; Nakamura, Y.; Yamori, T. An integrated database of chemosensitivity to 55 anticancer drugs and gene expression profiles of 39 human cancer cell lines. *Cancer Res.* **2002**, *62*, 1139−1147.
(33) Nakatsu, N.; Yoshida, Y.; Yamazaki, K.; Nakamura, T.; Dan, S.; Fukui, Y.; Yamori, T. Chemosensitivity profile of cancer cell lines and identification of genes determining chemosensitivity by an integrated bioinformatical approach using cDNA arrays. *Mol. Cancer Ther.* **2005**, *4*, 399−412.
(34) Rabow, A. A.; Shoemaker, R. H.; Sausville, E. A.; Covell, D. G. Mining the National Cancer Institute's tumor-screening database: Identification of compounds with similar cellular activities. *J. Med. Chem.* **2002**, *45*, 818−840.
(35) Covell, D. G.; Wallqvist, A.; Huang, R. L.; Thanki, N.; Rabow, A. A.; Lu, X. J. Linking tumor cell cytotoxicity to mechanism of drug action: An integrated analysis of gene expression, small-molecule screening and structural databases. *Proteins: Struct., Funct., Bioinf.* **2005**, *59*, 403−433.
(36) Huang, R. L.; Wallqvist, A.; Covell, D. G. Anticancer metal compounds in NCI's tumor-screening database: putative mode of action. *Biochem. Pharmacol.* **2005**, *69*, 1009−1039.
(37) Sausville, E. A.; Feigal, E. Evolving approaches to cancer drug discovery and development at the National Cancer Institute, USA. *Ann. Oncol.* **1999**, *10*, 1287−1291.
(38) Nicklaus, M. C.; Wang, S.; Driscoll, J. S.; Milne, G. W. A. Conformational changes of small molecules binding to proteins. *Bioorg. Med. Chem.* **1995**, *3*, 411−428.
(39) Oprea, T. I. On the information content of 2D and 3D descriptors for QSAR. *J. Braz. Chem. Soc.* **2002**, *13*, 811−815.
(40) Whittle, M.; Gillet, V. J.; Willett, P.; Alex, A.; Loesel, J. Enhancing the effectiveness of virtual screening by fusing nearest neighbor lists: a comparison of similarity coefficients. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1840−1848.
(41) Flower, D. R. On the properties of bit string-based measures of chemical similarity. *J. Chem. Inf. Comput. Sci* **1998**, *38*, 379−386.
(42) Covell, D. G.; Jernigan, R. L.; Wallqvist, A. Structural analysis of inhibitor binding to HIV-1 protease: identification of a common binding motif. *THEOCHEM* **1998**, 423, 93−100.
(43) Adams, J.; Kauffman, M. Development of the proteasome inhibitor Velcade (Bortezomib). *Cancer Invest.* **2004**, *22*, 304−311.

CI0501544