Web-Based Tools for Mining the NCI Databases for Anticancer Drug Discovery

Xueliang Fang, Lei Shao, Hui Zhang, and Shaomeng Wang*

University of Michigan Comprehensive Cancer Center, Departments of Internal Medicine and Medicinal Chemistry, University of Michigan, 1500 E. Medical Center Drive, Ann Arbor, Michigan 48109-0934

Received September 18, 2003

In this paper, we describe the development of a set of integrated Web-based tools for mining the National Cancer Institute's (NCI) anticancer databases for anticancer drug discovery. For data mining, three different correlation algorithms were implemented, which included the commonly used Pearson's correlation algorithm available from the NCI's COMPARE program, the Spearman's and Kendall's correlation algorithms. In addition, we implemented the *p*-value test to evaluate the significance of the correlation results. These Web-based data mining tools allow robust analysis of the correlation between the in vitro anticancer activity of the drugs in the NCI anticancer database, the protein levels and mRNA levels of molecular targets (genes) in the NCI 60 human cancer cell lines for identification of potential lead compounds for a specific molecular target and for study of the molecular mechanism action of a drug. Examples were provided to identify PKC ligands using a lead compound and to identify potential ErbB-2 inhibitors using the mRNA levels of ErbB-2 in the NCI 60 tumor cell lines.

INTRODUCTION

Since 1990, the National Cancer Institute (NCI) has been carrying out in vitro screening of compounds to determine their in vitro inhibitory activity of cell growth in the NCI 60 human cancer cell lines for the purpose of anticancer drug discovery.[1,2] To date, approximately 70 000 "open" and proprietary compounds have been tested in the NCI 60 cell lines, which has generated a series of information-rich anticancer drug databases.[3] These databases include in vitro cell line screening data of small molecules, protein and mRNA levels of molecular targets (genes) in the NCI 60 human cancer cell lines, and two-dimensional and three-dimensional structures of small molecules in the NCI repository collected by the NCI over the last five decades. To better use these information-rich databases for anticancer drug discovery and for study of molecular mechanism of action of anticancer drugs, a number of database mining tools have been developed by the NCI scientists,[4−7] which included the COMPARE program.[4,5] The COMPARE program employed the Pearson's correlation algorithm to analyze the correlations between the anticancer activity patterns of compounds or the anticancer activity of compound and protein or mRNA levels of molecular targets (genes) to identify potential novel lead compounds or to study the possible molecular mechanism of action of a drug. Over the years, the COMPARE program has been successfully used for the discovery of promising lead compounds for several molecular targets.[8−11]

Despite the success of the COMPARE program, several improvements may be made. For example, the Pearson's correlation algorithm assumes normal data distributions. However, many data points in the NCI anticancer databases, including the in vitro anticancer activity of compounds and protein and mRNA levels of molecular targets, do not have normal distributions. Therefore, correlation algorithms which are more suitable for analyzing data with nonnormal distributions are needed. Furthermore, correlations between two data sets can be due to chance and assessment of the statistical significance of the correlation results between two variables is essential.

In this paper, we describe the development of a set of integrated Web-based tools for mining the NCI's anticancer databases. For data mining, we implemented three different correlation algorithms, which included the commonly used Pearson's correlation algorithm available from the COMPARE program as well as the Spearman's and Kendall's correlation algorithms. We included a *p*-value test to evaluate the statistical significance of the correlation results. We implemented a Web-based tool to analyze the two-dimensional structural similarity between compounds and linked the NCI compounds to several commercially available, large chemical databases. These Web-based data mining tools allow robust analysis of the correlation between the in vitro anticancer activity of the drugs in the NCI anticancer database, the protein levels and mRNA levels of molecular targets (genes) in the NCI 60 human cancer cell lines for identifying potential lead compounds for a specific molecular target and for studying the molecular mechanism action of a drug.

1. PROGRAM DESCRIPTION

The raw data were downloaded from the Developmental Therapeutics Program (DTP) Web site at the NCI (http://dtp.nci.nih.gov, March 2002) and were then decompressed to obtain the plain text files. The key information for each data file was summarized in Table 1.

A relational database management system was built to store these data in several different tables. Besides the

* Corresponding author phone: (734)615-0362; fax: (734)647-9647; e-mail: shaomeng@umich.edu.

10.1021/ci034209i CCC: $27.50
Published on Web 12/31/2003

Table 1. Information Derived from NCI Anticancer Screening Data Files[a]

data file	small molecules	molecular targets	gene profiles
entries	~37 000	~197	~15 000
records	~2 000 000	~11 000	~900 000

[a] Not all screening data for 60 cell lines are available.

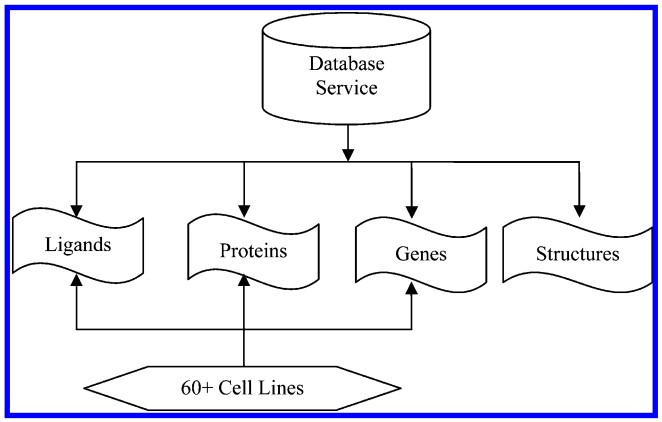

Figure 1. Schematic architecture of relational database management system for NCI anticancer screen data.

Table 2. Data Structure of a Typical Table in the Database

field name	note
ID[a]	primary identification
Data	experimental screening data
CELL	cell line
PANEL	panel name of the cell
MAX	maximal value
MIN	minimal value
POINTS	number of data
MEAN_VALUE	mean value
MEAN_BAR	code for plotting mean bar chart

[a] ID: NSC number, molecular target name, or gene description.

anticancer screen data, the 2-D and 3-D structures of small molecules and 3D structures of proteins were also included in our databases. The 3D coordinates of proteins were downloaded from the Protein Data Bank (PDB) Web site (http://www.rcsb.org/pdb/) and saved on a local hard disk in pdb file format.[12] The gene expressions in the NCI 60 cell lines were determined using DNA microarrays. The microarray data were generated by two groups: one set by Weinstein, Brown, and Botstein (about 9700 entries)[13,14] and the other set by the Millennium Pharmaceuticals (about 5300 entries). An in-house program was written to import these data into our databases, so that we can easily retrieve the information using the Structured Query Language (SQL). The mean bar of each entry that revealed the distribution of the cell line screening data was also generated by this program. The databases are managed by Oracle 9.0 for Windows 2000 Server. The schematic architecture of these databases is shown in Figure 1. The databases can be accessed by multiple remote users simultaneously. The information shown in Table 2 can be directly searched by entry identification, e.g. NSC number of a chemical compound, protein name of a molecular target, or a gene name.

Our program package was integrated on a Web server powered by Apache 1.2 for Windows 2000 Server. Most of the scripts were written in JAVA language. The system architecture of the program package is illustrated in Figure 2.

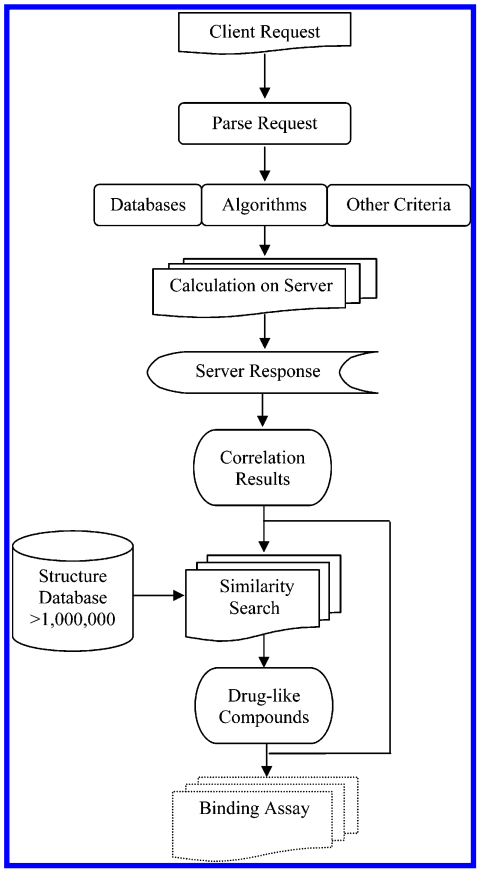

Figure 2. Systematic architecture of program package.

2. MINING TOOLS

The COMPARE program was initially developed by the NCI to facilitate the drug discovery efforts at the NCI.[4,5] The COMPARE program evaluates the correlation between compounds in terms of their activities in the NCI 60 cell lines, and between the activities of compounds in the NCI 60 cell lines and the mRNA or the protein levels of a specific molecular target, and between the mRNA or protein levels of different molecular targets in the NCI 60 cell lines. The correlations were expressed as the Pearson's correlation coefficient (PCC). In the COMPARE program, a procedure (PROC CORR) integrated in a commercial statistical package (the Statistical Analysis System) has been used to obtain PCC values.

In our Web-based database mining tools, three independent correlation algorithms were implemented and integrated into the databases, which included the Pearson's correlation coefficient (PCC), Spearman's correlation coefficient (SCC), and Kendall's correlation coefficient (KCC), respectively. The core programming source codes were written in C++.

2.1. Pearson's Correlation Coefficient. Pearson's correlation coefficient (PCC) analysis is widely used in statistical studies. PCC measures the strength of the linear relationship between two variables. It is assumed that both variables are interval/ratio and approximately normally distributed, and their joint distribution is bivariate normal. These two variables are often called $\vec{X}$ and $\vec{Y}$, $\vec{X} = (x_1,x_2,...,x_N)^T$ and $\vec{Y} = (y_1,y_2,...,y_N)^T$. In our database they are experimentally

determined activities of compounds in the NCI 60 cell lines and expression levels of molecular targets. The formula for calculating PCC is given in many forms. Equations 1 and 2 are two commonly used forms

$$r_p = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2} \quad (1)$$

$$r_p = \frac{\sum_{i=1}^{N} x_i y_i - \frac{1}{N}\sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{\sqrt{\left(\sum_{i=1}^{N} x_i^2 - \frac{1}{N}(\sum_{i=1}^{N} x_i)^2\right)\left(\sum_{i=1}^{N} y_i^2 - \frac{1}{N}(\sum_{i=1}^{N} y_i)^2\right)}} \quad (2)$$

where r_p denotes the PCC value, N is the size of the data pairs of these two variables, x_i and y_i are the relative measurements of the ith experiment (e.g. $-\log GI_{50}$ value) of $\vec{X}$ and $\vec{Y}$, and $\bar{x}$ and $\bar{y}$ are mean values of data $\vec{X}$ and $\vec{Y}$, respectively.

PCC can take on values from -1.0 to 1.0, where -1.0 represents a perfect negative (inverse) correlation, 0.0 represents no correlation, and 1.0 represents a perfect positive correlation. Since the PCC method is a widely used algorithm, it has been integrated in well-known commercial statistical analysis software.

2.2. Spearman's Rank Order Correlation Coefficient. PCC makes an implicit assumption that the two variables are jointly normally distributed. When this assumption is not valid, a nonparametric measure, such as the Spearman's correlation coefficient (SCC),[15] may be more suitable. Therefore, the second algorithm implemented in our program is based upon the discriminant derived from the SCC conception. It is usually calculated when it is not convenient, economic, or even possible to assign actual values to variables, but only to give a rank order to instances of each variable. Consequently, the SCC is also referred to as the Spearman rank order correlation coefficient.[15]

Since Spearman's method works by assigning a rank to each observation in each group separately, it may also be a better indicator that a relationship exists between two variables $\vec{X} = (x_1,x_2,...,x_N)^T$ and $\vec{Y} = (y_1,y_2,...,y_N)^T$ when the relationship is nonlinear. Let R_i be the rank of x_i among the other x's, and S_i be the rank of y_i among the other y's. SCC is given by eq 3

$$r_s = \frac{\sum_{i=1}^{N}(R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^{N}(R_i - \bar{R})^2}\sqrt{\sum_{i=1}^{N}(S_i - \bar{S})^2}} \quad (3)$$

where $\bar{R}$ and $\bar{S}$ are the mean values of R and S, respectively. When there are no ties in the measurement, r_s can be expressed as a conventional measure:

$$r_s = 1 - \frac{6\sum_{i=1}^{N}(R_i - S_i)^2}{N(N^2 - 1)} \quad (4)$$

However, when there are ties in the measurement, the exact relationship is slightly more complicated, as follows

$$r_s = \frac{1 - \frac{6}{N(N^2-1)}\left[\sum_{i=1}^{N}(R_i - S_i)^2 + \frac{1}{12}\sum_{i=1}^{N}(f_i^3 - f_i) + \frac{1}{12}\sum_{i=1}^{N}(g_i^3 - g_i)\right]}{\sqrt{1 - \frac{\sum_{i=1}^{N}(f_i^3 - f_i)}{N(N^2-1)}}\sqrt{1 - \frac{\sum_{i=1}^{N}(g_i^3 - g_i)}{N(N^2-1)}}} \quad (5)$$

where f_i denotes the number of ties in the ith group of ties among the R_i and g_i denotes the number of ties in the ith group of ties among the S_i. If all the f_i and all the g_i are equal to one, meaning there are no ties, then eq 5 reduces to eq 4.

Spearman's rank correlation is a distribution-free analogue of the correlation analysis. Like regression, it can be applied to compare two independent random variables, each at several levels (which may be discrete or continuous). Unlike regression, Spearman's rank correlation works on ranked (relative) data, rather than directly on the original data itself. Like the R^2 value produced by the PCC regression, the r_s value indicates the agreement between two variables.

A value of r_s near one indicates a good agreement, and a value near zero indicates a poor agreement. As a distribution-free method, the Spearman rank correlation does not make any assumption about the distribution of the data. Therefore, the SCC method generally produces more satisfactory results than the PCC method in our studies. The SCC method has also been used in consensus scoring for ligand and protein interactions by Clark.[16]

2.3. Kendall's τ Correlation Coefficient. Compared to Spearman's rank order correlation coefficient r_s, Kendall's τ is even more nonparametric.[15] Instead of using the numerical difference of ranks, Kendall's algorithm only concerns the relative ordering of ranks: higher in rank, lower in rank, or the same in rank. Nevertheless, ranking the data is not necessary.

In our implementation, we started with the N pairs of measurements in the NCI cancer cell lines $\vec{X} = (x_1,x_2,...,x_N)^T$ and $\vec{Y} = (y_1,y_2,...,y_N)^T$. We considered all $C_N^2 = N(N-1)/2$ pairs of data points for each variable, where a data point cannot be paired with itself, and where the points in either order count as one pair. If the relative ordering of the ranks of the two x's is the same as the relative ordering of the ranks of two y's, a pair "*concordant*" is counted. If the relative ordering of the ranks of the two x's is opposite from the relative ordering of the ranks of two y's, a pair "*discordant*" will be counted. If there is a tie in either the ranks of the two x's or the ranks of the two y's, then we do not call the pair either concordant or discordant. If the tie is in the x's, an "*extra_y*" pair is counted. If the tie is in the

Table 3. Tools Integrated in the Package

no.	module (seed vs targets)
1	compound vs compounds
2	compound vs molecular targets
3	compound vs gene expressions
4	molecular target vs molecular targets
5	molecular target vs compounds
6	molecular target vs gene expressions
7	gene expression vs gene expression
8	gene expression vs compounds
9	gene expression vs molecular targets

Table 4. Pairwise and Casewise Deletion Example[a]

experiment ID	cell 1	cell 2	cell 3	cell 4	cell 5
1	√	○	√	√	√
2	√	√	√	√	√
3	√	√	√	√	○

[a] √: denotes data present. ○: denotes data missing. For pairwise deletion: 1. Data of cell 1, 3, 4, and 5 will be applied to analyze the correlation between case 1 and case 2. 2. Data of cell 1, 3, and 4 will be applied to analyze the correlation between case 1 and case 3. 3. Data of cell 1, 2, 3, and 4 will be applied to analyze the correlation between case 2 and case 3. For casewise deletion: only data of cell 1, 3, and 4 will be applied for correlation analysis.

y's, an "*extra_x*" pair is counted. If the tie is in both *x*'s and *y*'s, we do not call the pair anything at all. Kendall's τ is defined as

$$\tau = \frac{C - D}{\sqrt{C + D + E_y}\sqrt{C + D + E_x}} \tag{6}$$

where C is the number of "*concordant*" pairs, D is the number of "*discordant*" pairs, and E_y and E_x denote "*extra_y*" and "*extra_x*" pairs, respectively.

As shown, $-1 \le \tau \le 1$. A larger τ value shows a good rank agreement.

2.4. Handling Missing Data. In the NCI databases, not all compounds, molecular targets, or microarrays were analyzed against all the 60 cell lines; in other words, some data were missing. There are two ways of handling missing values: pairwise deletion and casewise deletion. A simple example for pairwise deletion and casewise deletion is provided in Table 4. The results for pairwise deletion of the example can be described as follows: (1) for 1−2 pair, cells 1, 3, 4, and 5 are used; (2) for 1−3 pair, cells 1, 3, and 4 are be used; (3) for 2−3 pair, cells 1, 2, 3, and 4 are used. While for casewise deletion, the results of the example depend on the total system, both cells 2 and 5 are deleted for all pairs of correlations, so only cells 1, 3, and 4 are used for all the correlation evaluations.

We selected the pairwise deletion method for correlation calculation so that we can keep as much useful information as possible.

2.5. *p*-Value Test. Once we calculate the correlation coefficient for two variables, we need to determine what is the likelihood that the correlation occurs by chance. When a correlation is known to be significant, correlation coefficient (r) value is one conventional way of summarizing its strength. In fact, the value of r can be translated into a statement about the root-mean-square deviation (RMSD) being expected if the data are fitted to a straight line by the least-squares method.

Unfortunately, r is a rather poor statistic for deciding whether an observed correlation is statistically significant and/or whether an observed correlation is significantly stronger than another. The reason is that r is ignorant of the individual distributions of $\vec{X}$ and $\vec{Y}$, so there is no universal way to compute its distribution in the case of the null hypothesis. About the only general statement that can be made is this: If the null hypothesis is that $\vec{X}$ and $\vec{Y}$ are uncorrelated, if the distributions for $\vec{X}$ and $\vec{Y}$ each have enough convergent moments ("tails" die off sufficiently rapidly), and if N is large enough (typically >20), then r is distributed approximately normally, with a mean of zero and a standard deviation of $1/\sqrt{N}$. In that case, the significance of the correlation, that is, the possibility that $|r|$ should be larger than its observed value in the null hypothesis, is

$$p = erfc\left(\frac{|r|\sqrt{N}}{\sqrt{2}}\right) \tag{7}$$

where $erfc(x)$ is the complementary error function.

$$\begin{aligned} erfc(x) &= 1 - erf(x) \\ &= 1 - \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt \\ &= \frac{2}{\sqrt{\pi}}\int_x^\infty e^{-t^2}dt \end{aligned} \tag{8}$$

A smaller p-value indicates that the two distributions are significantly correlated.

3. STRUCTURAL SIMILARITY SEARCH

In the last five decades, approximately 500 000 small molecules (primarily synthetic organic compounds and natural products) have been collected by the NCI, and approximately half of the compounds ("open" compounds) are publicly accessible.[6] However, only 70 000 compounds out of the 500 000 have been tested for their anticancer cell activity in the NCI 60 cell lines and 37 000 are "open" compounds.[3] To further facilitate the use of the NCI databases for drug discovery, we have developed a Web-based tool for a structural similarity search and linked the NCI structural database to approximately 1 000 000 compounds (listed in Table 5) from different commercial chemical suppliers.

Table 5. 2D Structure Searchable Databases

database	entries	notes
Aldrich	129 000	catalog of Sigma-Aldrich, Inc
ChemDiv	318 000	catalog of Chemical Diversity Labs, Inc.
ComGenex	213 000	catalog of ComGenex, Inc.
Merck Index	6085	build in-house, Merck Index XIII
CHM	9367	build in-house, Chinese Herb Medicine
NCI	249 000	NCI small organic molecules
Ryan	153 000	catalog of Ryan Scientific, Inc.
total	>1 000 000	

For the structural similarity search, we used the Tanimoto coefficient (S),[17] which is expressed in eq 9

Table 6. Top 20 NCI Compounds ($GI_{50} \leq 1\ \mu M$) Correlated with NSC 239072 Using Pearson's Correlation Method

rank	NSC#	γ_p	best activity[a] (GI_{50}^{min}:nM)	mean value[b] ($-\log GI_{50}$)	selectivity (fold)[c]	*p*-value
1	239072[d,e]	1.000	0.500	7.295	3020	<1E−4
2	266186[d,e]	0.808	10	5.880	2042	<1E−4
3	654239[d,e]	0.802	612	4.332	162	<1E−4
4	623310	0.776	205	3.699	4898	<1E−4
5	703749[d,e]	0.662	109	5.361	162	<1E−4
6	703750[d,e]	0.650	81	5.355	214	<1E−4
7	688228	0.627	977	4.574	102	<1E−4
8	252940[e]	0.615	100	4.634	10000	<1E−4
9	627960[e]	0.578	63	3.346	10000	0.0002
10	688220[d,e]	0.571	463	5.446	43	<1E−4
11	631939[d,e]	0.565	10	5.435	2512	0.0001
12	688222	0.558	136	5.874	60	<1E−4
13	688235[e]	0.555	13	6.180	339	<1E−4
14	703751[d,e]	0.551	902	5.059	23	<1E−4
15	645597[d,e]	0.543	838	4.348	120	0.0001
16	629156[d,e]	0.524	700	4.863	145	0.0001
17	329507	0.520	44	5.311	1202	<1E−4
18	688239[d,e]	0.518	912	4.833	28	<1E−4
19	686038[e]	0.505	321	4.178	309	0.0001
20	631941[d,e]	0.505	10	5.166	3890	0.0010

[a] Best GI_{50} value, means the compound will be against at least 1 cell line at the specified GI_{50}. [b] Mean value of $-\log GI_{50}$ (concentration unit of GI_{50}: M). [c] Selectivity: $GI_{50}^{min}/GI_{50}^{max}$ (fold). [d] Also discovered at top 20 by Spearman's correlation method, overlapped 60%. [e] Also discovered at top 20 by Kendall's τcorrelation method, overlapped 80%.

$$S = \frac{N_{A\&B}}{N_A + N_B - N_{A\&B}} \tag{9}$$

where S is the Tanimoto similarity coefficient; N_A and N_B are the number of bits set in compounds A and B, respectively; and $N_{A\&B}$ is the number of bits that are set in both.

The similarity threshold specifies a lower limit for the Tanimoto coefficient. If a Tanimoto coefficient is greater than the threshold, then the query structure and the given structure in our database are considered similar. The search engine is powered by ChemAxon.[18]

4. APPLICATIONS OF THE WEB-BASED TOOLS FOR LEAD DISCOVERY

4.1. Identification of PKC Ligands Using a "Seed" Compound. Previously, we screened the NCI anticancer database for discovery of novel protein kinase C (PKC) liagnds using the standard Pearson's correlation analysis, which led to the identification of two novel PKC ligands (NSC 631939 and NSC 631941).[11]

Herein, we performed a similar analysis using three different algorithms to identify PKC ligands using a "seed" compound, mezerein (NSC 239072). Mezerein binds to PKC with a subnanomolar affinity and potently inhibits cell growth in the NCI 60 cell lines.[11] We analyzed the correlations of the anticancer activities of 7269 compounds which were found to have a GI_{50} value (concentration needed to inhibit 50% of cancer cell growth in the screening) less than 1 μM against at least one cancer cell line with that of mezerein. The results are summarized in Tables 6−8. The *p*-value test reveals that the results are significant (generally less than 0.001). Among the top 20 hits identified by three different algorithms, 12 of them are overlapped. Among these compounds, the following compounds have been confirmed as potent PKC ligands: NSC 266186 (huratoxin) is a potent PKC ligand. NSC 654239 (Cytoblastin) is an analogue of teleocidin and indolactam (ILV), a structurally distinct class of potent PKC ligands. NSC 631939 and NSC 631941 are iridal analogues, which were confirmed in our previous study as potent PKC ligands.[11] Figure 3 illustrates the representative structures and cytotoxicity profiles of these compounds discovered by these correlation analyses.

Table 7. Top 20 NCI Compounds ($GI_{50} \leq 1\ \mu M$) Correlated with NSC 239072 Using Spearman's Correlation Method

rank	NSC#	γ_s	best activity[a] (GI_{50}^{min}:nM)	mean value[b] ($-\log GI_{50}$)	selectivity (fold)[c]	*p*-value
1	239072[d,e]	1.000	0.50	7.295	3020	<1E−4
2	654239[d,e]	0.732	612	4.332	162	<1E−4
3	688220[d,e]	0.691	463	5.446	43	<1E−4
4	631939[d,e]	0.649	10	5.435	2512	0.0001
5	13484[e]	0.609	800	5.325	22	0.0003
6	688239[d,e]	0.607	912	4.833	28	<1E−4
7	266186[d,e]	0.596	10	5.880	2042	<1E−4
8	645597[d,e]	0.585	838	4.348	120	0.0001
9	703749[d,e]	0.565	109	5.361	162	<1E−4
10	703751[d,e]	0.562	902	5.059	23	<1E−4
11	3090[e]	0.538	24	6.856	27	0.0004
12	703750[d,e]	0.538	81	5.355	214	0.0001
13	674066[e]	0.525	191	6.013	10	0.0001
14	631941[d,e]	0.524	10	5.166	3890	0.0012
15	629156[d,e]	0.518	700	4.863	145	0.0004
16	683831	0.506	426	5.102	57	0.0002
17	631578	0.503	10	4.825	10000	0.0010
18	85442	0.503	791	5.303	14	0.0007
19	688230[e]	0.502	21	5.918	204	0.0002
20	688542	0.500	270	5.033	170	0.0002

[a−c] Same as Table 6. [d] Also discovered at top 20 by Pearson's correlation method, overlapped 60%. [e] Also discovered at top 20 by Kendall's τ correlation method, overlapped 80%.

Table 8. Top 20 NCI Compounds ($GI_{50} \leq 1\ \mu M$) Correlated with NSC 239072 Using Kendall's Correlation Method

rank	NSC#	τ	best activity[a] (GI_{50}^{min}:nM)	mean value[b] ($-\log GI_{50}$)	selectivity (fold)[c]	*p*-value
1	239072[d,e]	1.000	0.50	7.292	3020	<1E−4
2	654239[d,e]	0.586	612	4.332	162	<1E−4
3	688220[d,e]	0.504	463	5.446	43	<1E−4
4	688239[d,e]	0.488	912	4.833	28	<1E−4
5	631939[d,e]	0.471	10	5.435	2512	<1E−4
6	266186[d,e]	0.459	10	5.880	2042	<1E−4
7	645597[d,e]	0.430	838	4.348	120	<1E−4
8	13484[e]	0.421	800	5.325	22	0.0003
9	703751[d,e]	0.420	902	5.059	23	<1E−4
10	703749[d,e]	0.417	109	5.361	162	<1E−4
11	703750[d,e]	0.405	81	5.355	214	<1E−4
12	686038[d]	0.398	321	4.178	309	<1E−4
13	631941[d,e]	0.387	10	5.166	3890	0.0005
14	674066[e]	0.374	191	6.013	10	<1E−4
15	629156[d,e]	0.373	700	4.863	145	0.0002
16	3090[e]	0.370	24	6.856	27	0.0003
17	627960[d]	0.370	63	3.346	10000	0.0013
18	688230[e]	0.370	21	5.918	204	0.0001
19	252940[d]	0.364	100	4.634	10000	0.0001
20	688235[d]	0.355	13	6.180	339	0.0001

[a−c] Same as Table 6. [d] Also discovered at top 20 by Pearson's correlation method, overlapped 80%. [e] Also discovered at top 20 by Spearman's τcorrelation method, overlapped 80%.

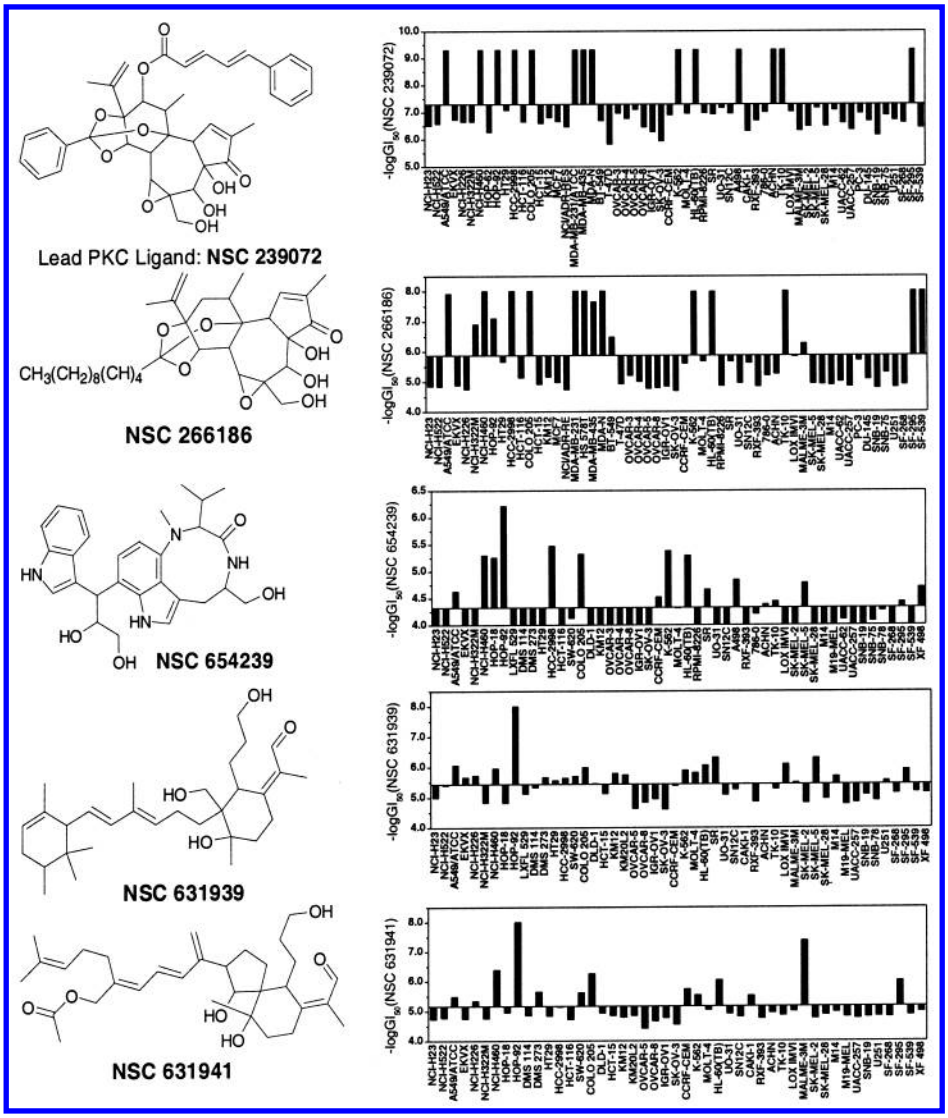

Figure 3. Chemical structures and mean bar charts of antiproliferation activities against NCI 60 human cancer cell lines.

4.2. Identification of Potential ErbB-2 Inhibitors. The protein levels for hundreds and the mRNA levels for thousands of molecular targets have been measured in the NCI 60 human cancer cell lines. The interaction between an anticancer agent and its specific molecular target may be suggested by a strong correlation between the anticancer activity pattern of the agent and the expression pattern of the molecular target in the NCI 60 cell lines.[19,20] Therefore, we can readily use correlation analysis to identify potential inhibitors (ligands) for a specific molecular target. Herein, we present our analysis to identify potential ErbB-2 inhibitors using three correlation algorithms we implemented.

Her-2/neu/C-erbB2 (ErbB-2) is a member of the EGFR receptor family proteins and has been pursued as a promising molecular target for the design of new anticancer drugs.[21−24] Herceptin, a humanized ErbB-2 monoclonal antibody is an effective therapy for the treatment of ErbB-2 overexpressing human breast cancer.[25] Currently small molecule inhibitors targeting ErbB-2 and EGFR are in advanced clinical trials.[26]

The mRNA expression levels of ErbB-2 in the NCI panel of 60 human tumor cell lines are shown in Figure 4.[27] As can be seen, ErbB-2 is remarkably overexpressed in SK-OV-3 ovarian and MDA-N breast cancer cell lines. The correlation results based upon these three algorithms are summarized in Tables 9−11, respectively. Spearman's algorithm and Kendall's algorithm produce consistent results with 75% overlaps among the top 20 compounds. Several interesting compounds were identified using these two methods. NSC 382584 (Figure 4) was identified by the Spearman's algorithm as a potential ErbB-2 inhibitor (rank #17, $r = 0.3476$, $p = 0.0122$) and also by the Kendall's algorithm (rank #16, $r = 0.2680$, $p = 0.0046$). It has a very similar chemical structure to a known EGFR inhibitor (shown in Figure 6).[28] The mean bar chart of the antiproliferation activities of this compound against NCI cell lines is illustrated in Figure 5. NSC 669364 is a known potent EGFR and ErbB-2 inhibitor.[29,30] In our analyses, NSC 669364 was identified by the Spearman's algorithm as a potential ErbB-2 inhibitor (rank #18, $r = 0.3407$, $p = 0.0095$) and also by the Kendall's algorithm (rank #41, $r = 0.2270$, $p = 0.0111$). The structure and mean bar chart of antiproliferation activities of NSC 669364 is shown in Figure 7. Using the Pearson's correlation method, we found that NSC 382584 did not have a correlation with the mRNA level of ErbB-2 (rank #1383, $r = -0.004$ and $p = 0.9800$), but NSC 669364 has some

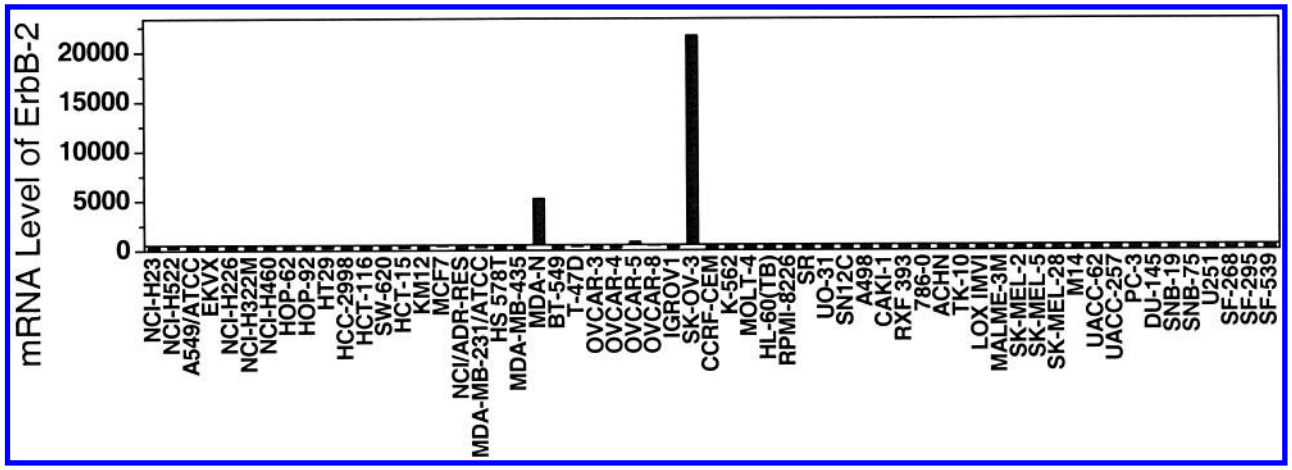

Figure 4. The ErbB-2 mRNA expression levels (relative phosphorimager signal) in the NCI 60 human cancer cell lines. (Data from DTP Web site, February 2003 release, http://dtp.nci.nih.gov/webdata.html).

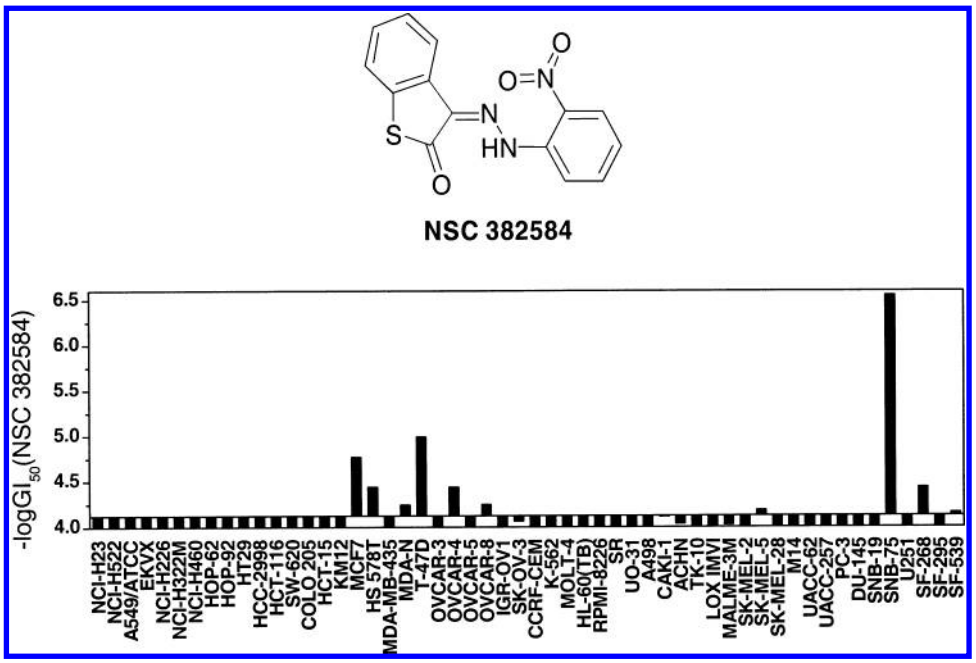

Figure 5. Chemical structure of NSC 382584 and mean bar chart for its inhibitory activity (GI_{50} value from the DTP Web site) against the NCI 60 human cancer cell lines.

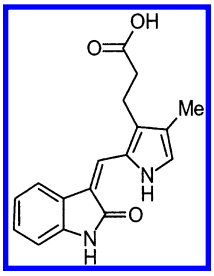

Figure 6. Chemical structure of a known EGFR/ErbB-2 inhibitor developed by SUGEN, Inc.

correlation with the mRNA level of ErbB-2 (rank #84, $r = 0.2571$, $p = 0.0494$). Taken together, our analysis showed that correlation analysis using the Spearman's algorithm and the Kendall's algorithm can effectively identify known potent ErbB-2 small molecule inhibitors and may be superior to the Pearson's correlation analysis in cases the data is not normally distributed.

SUMMARY

In this paper, we present our development of a set of integrated Web-based tools for mining the NCI anticancer databases for the purpose of drug discovery and molecular mechanism studies. We implemented three correlation algorithms based upon Pearson's correlation coefficient, Spearman's rank order coefficient, and Kendall's τ coefficient for data mining. Although Pearson's algorithm has been widely used for correlation analysis, Spearman's and Kendall's algorithms may offer certain advantages when the data do not have normal distributions. We also implemented the *p*-value test tool to evaluate the significance of the correlation results. Furthermore, we included a structural similarity analysis tool and linked the NCI structural database to several

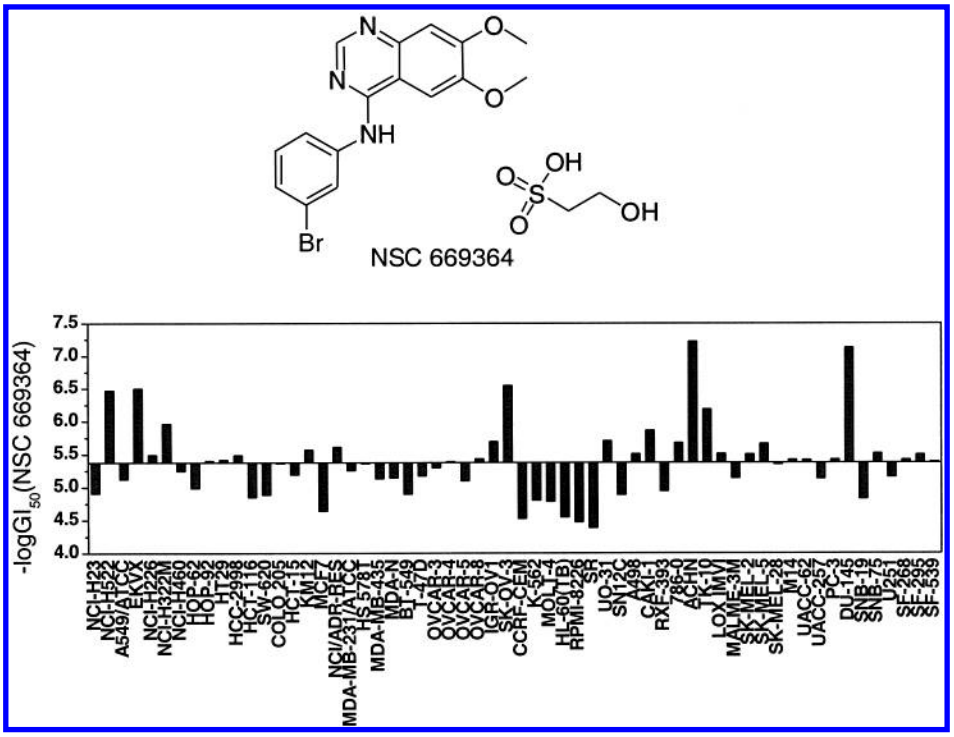

Figure 7. Chemical structure of NSC 669364 and mean bar chart for its inhibitory activity (GI_{50} value from the DTP Web site) against the NCI 60 human cancer cell lines.

Table 9. Top 20 NCI Compounds ($GI_{50} \leq 1\ \mu M$) Correlated with ErbB-2 Using Pearson's Correlation Method

rank	NSC#	γ_p	best activity[a] (GI_{50}^{min}:nM)	mean value[b] ($-\log GI_{50}$)	selectivity (fold)[c]	*p*-value
1	655396	0.9989	634	5.022	16	<1E−4
2	653851	0.9972	630	4.045	158	<1E−4
3	655400	0.9953	85	4.072	1175	<1E−4
4	624392	0.9935	1000	4.037	100	<1E−4
5	655402	0.9852	325	4.056	309	<1E−4
6	638462	0.9845	428	4.056	234	<1E−4
7	653855	0.9824	693	4.056	145	<1E−4
8	630450	0.9725	885	4.045	112	<1E−4
9	672085	0.9714	495	4.041	204	<1E−4
10	672083	0.9714	513	4.041	195	<1E−4
11	35790	0.9713	237	4.049	427	<1E−4
12	672081	0.9710	778	4.041	129	<1E−4
13	653857	0.9674	394	4.067	251	<1E−4
14	672429	0.9646	578	4.041	174	<1E−4
15	692423	0.9572	292	4.082	347	<1E−4
16	630705	0.9496	180	4.740	331	<1E−4
17	93033	0.9444	10	4.087	10000	<1E−4
18	63549	0.9178	13	4.143	7586	<1E−4
19	630258	0.9167	53	4.090	1862	<1E−4
20	655737	0.9061	41	4.233	2455	<1E−4

[a−c] Same as Table 6.

Table 10. Top 20 NCI Compounds ($GI_{50} \leq 1\ \mu M$) Correlated with ErbB-2 Using Spearman's Correlation Method

rank	NSC#	γ_s	best activity[a] (GI_{50}^{min}:nM)	mean value[b] ($-\log GI_{50}$)	selectivity (fold)[c]	*p*-value
1	676497[d]	0.5195	0.257 pM	11.251	10000	0.0001
2	676495[d]	0.4989	0.257 pM	11.115	10000	0.0002
3	682335[d]	0.4796	173	5.475	195	0.0003
4	178249[d]	0.4313	23	5.442	5370	0.0038
5	671456[d]	0.4141	125	5.661	126	0.0016
6	684480[d]	0.4111	977	5.492	40	0.0017
7	672883[d]	0.4032	975	4.874	102	0.0057
8	682767[d]	0.3972	44	4.481	2291	0.0030
9	659331[d]	0.3876	740	5.039	13	0.0048
10	668814[d]	0.3816	136	5.930	288	0.0070
11	98892[d]	0.3694	10	4.131	10000	0.0062
12	703099	0.3609	10	4.481	10000	0.0069
13	650396[d]	0.3608	2.8	8.026	7	0.0065
14	9223	0.3597	320	5.761	12	0.0071
15	657149	0.3574	479	5.216	209	0.0093
16	622481[d]	0.3555	10	4.322	10000	0.0212
17	382584[d]	0.3476	280	4.125	355	0.0122
18	669364	0.3407	60	5.377	692	0.0095
19	686420[d]	0.3406	15.6	4.525	6457	0.0123
20	696991	0.3370	102	4.721	46	0.0238

[a−c] Same as Table 6. [d] Also discovered at top 20 by Kendall's correlation method, 75% overlapped.

large chemical databases from commercial chemical suppliers.

These Web-based data mining tools allow robust analysis of the correlation between the in vitro anticancer activity of the drugs in the NCI anticancer database, the protein levels and mRNA levels of molecular targets (genes) in the NCI 60 human cancer cell lines for identification of potential lead compounds for specific molecular targets and for study of the molecular mechanism of actions for a drug molecule. Two examples were provided to identify PKC ligands using a lead PKC ligand and to identify potential ErbB-2 inhibitors using the mRNA levels of ErbB-2 in the NCI 60 cell lines. Our results show that these integrated Web-based informatics tools may be used to aid the discovery of promising novel lead compounds for specific molecular targets of interest.

Of note, since our Web-based tools are integrated with an extensible database management system, it is very convenient to add new experimental data and to perform correlation analyses without changing any source code.

Table 11. Top 20 NCI Compounds ($GI_{50} \leq 1\ \mu M$) Correlated with ErbB-2 Using Kendall's Correlation Method

rank	NSC#	τ	best activity[a] (GI_{50}^{min}:nM)	mean value[b] ($-\log GI_{50}$)	selectivity (fold)[c]	*p*-value
1	676497[d]	0.3825	0.257 pM	11.251	10000	<1E−4
2	676495[d]	0.3768	0.257 pM	11.115	10000	<1E−4
3	659331[d]	0.3253	740	5.039	13	0.0005
4	682335[d]	0.3190	173	5.475	195	0.0004
5	178249[d]	0.3093	23.5	5.442	5370	0.0024
6	98892[d]	0.3005	10	4.131	10000	0.0011
7	650396[d]	0.2993	2.8	8.032	7	9.0E−4
8	684480[d]	0.2915	977	5.492	40	0.0011
9	682767[d]	0.2905	44	4.481	2291	0.0014
10	671456[d]	0.2897	125	5.661	126	0.0012
11	672883[d]	0.2884	975	4.874	102	0.0038
12	655501	0.2831	5.0	5.087	1995	0.0041
13	662579	0.2800	1.0	5.175	10000	0.0090
14	645804	0.2744	0.187	8.279	145	0.0021
15	622481[d]	0.2705	10.0	4.323	10000	0.0106
16	382584[d]	0.2680	280	4.125	355	0.0046
17	615554	0.2643	10	4.251	10000	0.0136
18	695057	0.2635	131	4.148	759	0.0045
19	668814[d]	0.2631	135	5.931	288	0.0064
20	686420[d]	0.2616	15.6	4.524	6547	0.0048

[a−c] Same as Table 6. [d] Also discovered at top 20 by Spearman's correlation method, 75% overlapped.

ACKNOWLEDGMENT

We would like to thank Karen Kreutzer for her excellent assistance on the manuscript. We thank the Developmental Therapeutics Program NCI/NIH for providing Web-accessible anticancer screening data (http://dtp.nci.nih.gov/webdata.html, February 2003 Release). A request to access these Web-based tools may be sent to Dr. Shaomeng Wang, shaomeng@umich.edu (e-mail).

REFERENCES AND NOTES

(1) Boyd, M. R.; Paull, K. D.; Rubinstein, L. R. In *Cytotoxic Anti-cancer Drugs: Models and Concepts for Drug Discovery and Development*; Vleriote, F. A., Corbett, T. H., Baker, L. H., Eds.; Kluwer Academic: Hingham, MA, 1992; pp 11−34.

(2) Boyd, M. R. In *Cancer: Principles and Practice of Oncology*; DeVita, V. T., Jr.; Hellman, S., Rosenberg, S. A., Eds.; Lippincott: Philadelphia, PA, 1989; Vol. 3, pp 1−12.

(3) Weinstein, J. N.; Myers, T. G.; O'Connor, P. M.; Friend, S. H.; Fornace, A. L., 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.

(4) Zaharevitz, D. W.; Holbeck, S. L.; Bowerman, C.; Svetlik, P. A.; COMPARE: a Web accessible tool for investigating mechanisms of cell growth inhibition. *J. Mol. Graph Model* **2002**, *20*, 297−303.

(5) Paull, K. D.; Hamel, E.; Malspeis, L. COMPARE http://dtp.nci.nih.gov/docs/compare/compare.html, 2002.

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

(7) 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.

(8) Bai, R.; Paull, K. D.; Herald, C. L.; Pettit, G. R.; Hamel, E. Halichondrin B and Homohalichondrin B, Marine Natural Products Binding in the Vinca Domain of Tubulin: Discovery of Tubulin-based Mechanism of Action by Analysis of Differential Cytotoxicity Data. *J. Biol. Chem.* **1991**, *266*, 15882−15889.

(9) Paull, K. D.; Lin, C. M.; Malspeis, L.; Hamel, E. Identification of Novel Antimitotic Agents Acting at the Tubulin Level by Computer-Assisted Evaluation of Differential Cytotoxicity Data. *Cancer Res.* **1992**, *52*, 3892−3900.

(10) Zaharevitz, D. W.; Gussio, R.; Leost, M.; Senderowicz, A. M.; Lahusen, T.; Kunick, C.; Meijer, L.; Sausville, E. A. Discovery and Initial Characterization of the Paullones, a Novel Class of Small-Molecule Inhibitors of Cyclin-Dependent Kinases. *Cancer Res.* **1999**, *59*, 2566−2569.

(11) Shao, L.; Lewin, N. E.; Lorenzo, P. S.; Hu, Z.; Enyedy, I. J.; Garfield, S. H.; Stone, J. C.; Marner, F.; Blumberg, P. M.; Wang, S. Iridals Are a Novel Class of Ligands for Phorbol Ester Receptors with Modest Selectivity for the RasGRP Receptor Subfamily. *J. Med. Chem.* **2001**, *44*, 3872−3880.

(12) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.

(13) Scherf, U.; Ross, D. T.; Waltham, M.; Smith, L. H.; Lee, J. K.; Tanabe, L.; Kohn, K. W.; Reinhold: W. C.; Myers, T. G.; Andrews, D. T.; Scudiero, D. A.; Eisen, M. B.; Sausville, E. A.; Pommier, Y.; Botstein, D.; Brown, P. O.; Weinstein, J. N. A Gene Expression Database for the Molecular Pharmacology of Cancer. *Nature Genetics* **2000**, *24*, 236−244.

(14) Ross, D. T.; Scherf, U.; Eisen, M. B.; Perou, C. M.; Rees, C. Spellman, P.; Iyer, V. Jeffrey, S. S.; Van de Rijn, M.; Waltham, M.; Pergamenschikov, A.; Lee, J. C. E.; Lashkari D.; Shalon, D.; Myers, T. G.; Weinstein, J. N.; Botstein, D.; Brown, P. O. Systematic Variation in Gene Expression Patterns in Human Cancer Cell Lines. *Nature Genetics* **2000**, *24*, 227−235.

(15) Siegal, S.; Castellan, N. J., Jr. *Nonparametric Statistics for the Behavioral Sciences*, 2nd ed.; McGraw-Hill Book Company: New York, 1988; pp 235−244.

(16) Clark, R. D.; Strizhev, A. Leonard, J. M.; Blake, J. F.; Mattew, J. B. Consensus Scoring for Ligand/Protein Interactions. *J. Mol. Graph. Model*. **2002**, *20*, 281−295.

(17) Godden, J. W.; Xue, L.; Bajorath, J. Combinatorial Preferences Affect Molecular Similarity/Diversity Calculations Using Binary Fingerprints and Tanimoto Coefficients. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 163−166.

(18) Csizmadia, F. J Chem: Java Applets and Modules Supporting Chemical Database Handling from Web Browsers. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 323−324.

(19) Amundson, A. S.; Myers, T. G.; Scudiero, D.; Kitada, S.; Reed, J. C.; Fornace, A. J., Jr. An Informatics Approach Identifying Markers of Chemosensitivity in Human Cancer Cell Lines. *Cancer Res.* **2000**, *60*, 6101−6110.

(20) Tamm, I.; Kornblau, S. M.; Segall, H.; Krajewski, S.; Welsh, K.; Kitada, S.; Scudiero, D. A.; Tudor, G.; Qui, Y. H.; Monks, A.; Andreeff, M.; Reed, J. C. Expression and Prognostic Significance of IAP-Family Genes in Human Cancers and Myeloid Leukemias. *Clinical Cancer Res.* **2000**, *6*, 1796−1803.

(21) Yarden, Y.; Sliwkowski, M. X. Untangling the ErbB Signaling Network. *Nat. Rev. Mol. Cell Biol.* **2001**, *2*, 127−137.

(22) Hynes, N. E.; Horsch, K.; Olayioye, M. A.; Badache, A. The ErbB Receptor Tyrosine Family as Signal Integrators. *Endocrine-Related Cancer* **2001**, *8*, 151−159.

(23) Mendelsohn, J.; Baselga, J. The EGF Receptor Family as Targets for Cancer Therapy.*Oncogene* **2001**, *19*, 6550−6565.

(24) Arteaga, C. L. The Epidermal Growth Factor Receptor: From Mutant Oncogene in Nonhuman Cancers to Therapeutic Target in Human Neoplasia. *J. Clin. Oncol.* **2001**, *19*, 32−40.

(25) Slamon, D. J.; Leyland-Jones, B.; Shak, S.; Fuchs, H.; Paton, V.; Bajamonde, A.; Fleming, T.; Eiermann, W.; Wolter, J.; Pegram, M.; Baselga, J.; Norton, L. Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *N. Engl. J. Med.* **2001**, *344*, 783−92.

(26) Grünwald, V.; Hidalgo, M. Developing Inhibitors of the Epidermal Growth Factor Receptor for Cancer Treatment. *J. Natl. Cancer Inst.* **2003**, *95*, 851−867.

(27) 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.

(28) Sun, L.; Tran, N.; Liang, C. Hubbard, S.; Tang, F.; Lipson, K.; Schreck, R.; Zhou, Y.; McMahon, G.; Tang, C. Identification of Substituted 3-[(4,5,6,7-Tetrahydro-1H-indol-2-yl)methylene]-1,3-dihydroindol-2-ones as Growth Factor Receptor Inhibitors for VEGF-R2 (Flk-1/KDR), FGF-R1, and PDGF-R Tyrosine Kinases. *J. Med. Chem.* **2000**, *43*, 2655−2663.

(29) Fry, D. W.; Kraker, A. J.; McMichael, A,; Ambroso, L. A.; Nelson, J. M.; Leopold, W. R.; Connors, R. W.; Bridges, A. J. A Specific Inhibitor of the Epidermal Growth Factor Receptor Tyrosine Kinase. *Science* **1994**, *265*, 1093−1095.

(30) Tsou, H.; Mamuya, N.; Johnson, B. D.; Reich, M. F.; Gruber, B. C.; Ye, F.; Nilakantan, R.; Shen, R.; Discafani, C.; DeBlanc, R.; Davis, R.; Koehn, F. E.; Greenberger, L. M.; Wang, Y.; Wissner, A. 6-Substitue-4-(3-bromophenylamino) quinazolines as Putative Irreversible Inhibitors of the Epidermal Growth Factor Receptor (EGFR) and Human Epidermal Growth Factor Receptor (HER-2) Tyrosine Kinases with Enhanced Antitumor Activity. *J. Med. Chem.* **2001**, *44*, 2719−2734.

CI034209I