AD_____

Award Number: DAMD17-98-1-8323

TITLE: Deriving Structures for Lead Drug Discovery from Cell-Line Screens

PRINCIPAL INVESTIGATOR: Robert L. Jernigan, Ph.D.
Dr. David G. Covell

CONTRACTING ORGANIZATION: National Cancer Institute
Bethesda, Maryland 20892

REPORT DATE: October 2000

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

20010511 159

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	October 2000	Annual (1 Sep 99 - 1 Sep 00)

4. TITLE AND SUBTITLE
Deriving Structures for Lead Drug Discovery from Cell-Line Screens

5. FUNDING NUMBERS
DAMD17-98-1-8323

6. AUTHOR(S)
Robert L. Jernigan, Ph.D.
Dr. David G. Covell

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
National Cancer Institute
Bethesda, Maryland 20892

E-MAIL:
jernigan@structure.nci.nih.gov

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES
This report contains colored photos

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 Words)*

A suite of computational tools has been developed for detailed analysis of large-scale high-throughput screening data for the purpose of lead drug discovery and potential identification of novel molecular targets in the treatment of human cancers. The method has been developed and tested against the National Cancer Institute's 60 tumor cell panel. This suite of analytical and display tools is focused in the areas of data conditioning, pattern association, visualization and data presentation, with additional functionalities that address signal scaling issues, missing data elements, and locality/non-linearity features of the data-space. Careful considerations in these areas are found to significantly enhance the extraction of additional information from large, complicated, screening databases as well as provide a general tool well suited for drug discovery. These results find strong correlations between molecular structure and putative mechanism of action for large classes of anticancer agents; with a clear segregation of compounds according to their activities against specific molecular targets. More significantly, screening cells that are found within specific tumor cell panels are found to respond similarly to classes of molecular agents. This information can lead directly to the formulation of alternative chemical analogs and hypotheses about specific molecular targets and their affected biosynthetic pathways.

14. SUBJECT TERMS	15. NUMBER OF PAGES
Breast Cancer , Drug Screening, Anticancer	92
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. Z39-18
298-102

Table of Contents

1 INTRODUCTION

The goal of this proposed research is to apply modern, novel methods of data analysis useful for cataloging statistical correlations between response patterns of chemicals tested in the NCI's 60 tumor cell panel [1, 2] and deriving consensus structural details of active agents [3]. The tools developed in this research will directly improve the process of discovering novel lead compounds active in the treatment of breast, as well as other, cancers. Relating consensus structural details within classes of compounds to clear molecular target and putative mechanism of drug action [4] provides a powerful means to combine the results from these screens with the novel molecular design approach [5]. This report describes our research efforts to catalog biological response patterns within cellular screening data and tabulate consensus chemical structures. Our analysis finds that very careful consideration of factors related to data collection and statistical analyses are necessary to extract information from large datasets. The insight gained from this research can be directly applied to the analysis and mining of additional information gathered from large-scale screening efforts.

2 BODY

An integral component of our research has been the development of powerful computational tools for data analysis. A traditional 'cookbook' does not exist for the analysis of large and complex datasets [6]. Proper evaluation of information mined from such data requires a detailed understanding about decisions related to all areas of these experiments. The following section will briefly outline our efforts to address the effects of various decisions on data mining.

2.1 Data Conditioning

The National Cancer Institute's anticancer tumor cell screen measures growth retardation of selected immortalized tumor cell clones following exposure to a range of test concentrations for each agent [7]. Using as an endpoint the logarithm of the test concentration that leads to fifty percent growth inhibition, $\log(GI_{50})$, a biological response pattern for all tumor cell lines (typically 60 or more) is established for each tested compound. While the raw data generated in this screen determines cellular potency, much of the interest in this data lies in establishing the biological significance of these patterns; with the hope of identifying tumor specific reagents. A Z-score conditioning of each row of the raw data is used to provide a zero mean reference value, scaled according to its standard deviation, while preserving signal variation;

$$z_{ij} = \frac{(g_{ij} - \bar{g}_j)}{\sigma_j}, \tag{1}$$

where z_{ij} is the Z-score for element i, j, g_{ij} is the $\log(GI_{50})$ value, $\bar{g}_j$ is the mean and σ_j is the absolute deviation across row j (i.e. each compound). Using a metric related to data clustering (to be presented later) we find that the Z-score transformed data improves the quality of the clustering by $\sim 17\%$ when compared to the raw data. An additional consideration for data conditioning involves the intrinsic sensitivity of cell lines to chemical agents. Prior analysis found that, for example, the leukemia (leu) cell panel is most sensitive to chemical agents, whereas, the non-small cell lung (lns) panel is the least sensitive. Data normalization is necessary to assess growth inhibition across all cell lines, rather than detecting

agents active against only the most sensitive cell lines. Z-score normalization of each cell line's response to ALL tested compounds is used to establish a common reference for each cell line. Scaling the raw data across tumor cell types *and* across tested compounds provides a uniform means to assess pattern diversity within the complete set of tumor screening data generated within the NCI's publically available database(www.dtp.nci.nih.gov). Examples of the effects of data conditioning are shown in Figure 1, for pattern and response normalization.*

2.2 Self-Organizing Maps

Traditional methods for summarizing the results of large screening datasets seek to discover subsets of data where similarities in response are observed. The initial step in this process requires the selection of a pairwise measure of pattern similarity that assigns the highest score to the most similar datasets. Such pairwise measures include rank correlation and Euclidean, Mahalanobis or Minkowski measures of distance. These pairwise measures provide a simple and direct means to identify highly similar response subsets. Limitations in this procedure are known to occur, particularly when data is contaminated with large amounts of noise, resulting in a greater likelihood of random statistical correlations, and increased difficulty in determining 'real' relationships [8, 9]. Methods designed to treat noisy data include Principal Component Analysis and the related method of Singular Value Decomposition; where the data are reexpressed along directions that maximize the signal-to-noise ratio [10]. Recently the self-organizing map (SOM) method [11] has found great utility in studies of voice recognition and visual processing; datasets that often exhibit large amounts of random noise and missing data. Designed specifically to deal with extremely noisy and incomplete datasets, the algorithms associated with the SOM method are well-suited for mining the NCI anti-cancer tumor cell line screen for biochemical information useful in the analysis of drug screening data.

The SOM method can be divided into two regimes: clustering in high dimensional space,

*Data from all tumor cell lines was used in our analysis. This set consists of 80 cell-lines collected from leukemia, non-small cell lung, small cell lung, colon, central nervous system, melanoma, ovarian, renal, prostate and breast cancer tissues.

and projections into a lower dimensional display space (See Figure 2 and legend). Information 'linkages' between these two regimes provide a convenient means to validate clusters in low-dimensional space, by examinations of the original data, as well as assist in data interpretation and hypothesis generation. To accomplish this, the SOM method maps each compound's data vector (V^k) on to an optimally defined set of lower-dimensional response vectors, (R^k). This step is accomplished by minimizing the deviation between the data and response vectors:

$$\nabla R^k \propto \sum_j h(||V^j - R^k||)||V^j - R^k|| \tag{2}$$

where ∇R^k is the incremental change in position of the response vector R^k, V^j is the set of data vectors, and $||V^j - R^k||$ is the distance between data and response vector. The neighborhood kernel function $h(||V^j - R^k||)$, adjusts the position of the response vectors in order to collectively order the map clusters according to their underlying neighborhoods in data space. The form of the neighborhood kernal function exhibits a maximum when the data and response vectors coincide and goes to zero as these vectors become more distant. Often the neighborhood kernel is a Gaussian function, however, our analysis finds that Epanechicov function [max(0,1-$||V^j - R^k||^2$)] consistently yields a lower optimal value for ∇R^k, and will be used for our analysis.

The form of equation 2 determines how the response vectors orient themselves to mirror the data space, or alternatively, how the response vectors partition the data space into clusters (see Figure 3) Regions that are rich in data vectors attract many response vectors and as a result finely divide the most dense regions of data space. This process can be contrasted with the more conventional Principal Component Analyses [12], where data is oftentimes reoriented, in a linear fashion, on to the space of the top most principal components. The biochemically important regions of the cancer screening data are not uniformly distributed in the 80 dimensional tumor cell space, but rather are contained in densely populated subspaces. A key feature of the SOM method is its ability to transform the data space so as to make a two-dimensional projection map that is uniform in cluster neighbors. As a result the data rich regions are stretched such that the biochemically relevant cluster distinctions become apparent.

Two factors that make the SOM method particularly suited to drug exploration include the data-based clustering mentioned above, and the ability to display these results in an interpretable manner. The method for display is the uniform projection of the clustering in high-dimensional space to a low dimension display space (see Figures 2 and 3) This mapping is both simple and retains a great deal of the original high-dimensional information. An additional noteworthy feature of the SOM analysis is the treatment of missing data. No attempt is made estimate the missing data elements with, for example, the mean of the data vector; rather these data are regarded as elements with an unknown value and the SOM method simply skips over missing vector components. Attempts to analyze datasets where missing elements were replaced by the mean of existing data consistently resulted in poor cluster definition and increased error in measures of distance between data and response vectors. A simple example illustrates this point. In a dataset where equivalent measures of growth inhibition are obtained for only 64 of the 80 tumor cells, replacement of the missing values by the group mean would contribute 25% of the calculated distance $[16/(80-16) = 0.25]$. This large contribution from missing elements is unlikely to reflect the underlying biological response and degrades the ability to construct structure/function (referred to hereafter as S/F) correlations. The approach we have used is to keep missing elements as unknown values.

Much of the current excitement about drug discovery efforts is stimulated by the prospects of mining large biological screening databases [13]. Accompanying this interest is the realization that new tools will be needed to globally investigate and extract information from this data. Our computational method has been specifically developed to examine the complete drug discovery space contained in the biological screen under investigation. This method simultaneously examines the interrelation between all screened compounds and the biological response space that they probe. The ability to globally rearrange the response space follows from the form of the self-organizing map (SOM) mathematical algorithm, which has been designed to facilitate mapping the high-dimensional data space into a lower-dimensional cluster space that can be projected onto a two-dimensional map. Our results will demonstrate how our tools can be used in these data mining efforts.

2.3 Standard Anticancer Agents

We begin with an analysis of tumor growth inhibition by 122 standard anticancer agents compiled by Weinstein et al. annotated according to their putative mechanism of action (MOA) [14, 15]. Using this data as an example, the basic concepts of SOM clustering and its display capabilities can be illustrated. Figures 2 and 9 display the two-dimensional SOM map for this data. The final map could be projected on to a 9 by 17 hexagonal array, with the size of the hexagon at each node loci being proportional to the number of compounds assigned to this cluster and gray-scale color intensity between nodes reflecting the distance between neighboring clusters(See Figure 2 and Figure 9). Consistent with prior studies, these standard agents could be separated into those with MOA's involving inhibition of mitotic activity and those affecting nucleic acid biosynthesis; the former grouping includes the classes of taxanes, colchicines and vinca alkaloids, some of which are know to have selective activity against different breast cancer tumors. This division is quite sharp, and appears in Figure 9 at row six of the SOM map. The ability to segregate compounds by SOM clustering represents an important first step in the identification of novel compounds with specific activity within different cancer types. Within the anti-mitotic and nucleotide biosynthesis regions of the map, well defined sub-clusters exist that, upon inspection, consist of structurally similar compounds with stick-figure drawings of selected cluster members displayed at the map margins. This apparent consistency between molecular structure and function (putative MOA) was used to develop a metric for detailed sensitivity studies regarding the choice of parameters for our SOM optimization and their effect on quality of clustering.

2.4 Sensitivity Analysis

A detailed examination has been conducted to determine how the quality of the clustering is effected by choices in the experimental design parameters such as the number of cell lines in the screen, the size of the SOM clustering map, the treatment of noisy and incomplete data, and the importance of data conditioning. We assess the quality of clustering by correlating the SOM cluster memberships determined from the $\log(GI_{50})$ (i.e. functional) data with the SOM clustering based on chemical structure. This approach assumes that chemical

structure is a surrogate for the 'true' pharmacophore of the molecular target affecting cell growth. This is clearly a simplifying approximation for the true 'hidden' pharmacophore or molecular target [16]. It should be noted that implicit in this discussion is a primary goal of our research; the creation of a method to predict and understand the anticancer effect of chemotherapeutic agents and to identify their molecular targets.

To examine the correlation between cluster membership based on biological response and chemical structure, we have designed an extended mechanism of action (ExMOA) data set which consists of 362 compounds, based on the 122 standard anticancer agents discussed above, but expanded to include compounds with strong structural similarity [17] (Tanimoto Coeff .gt. 0.9) to these standard anticancer agents. SOM clustering of these compounds into structural classes is based on the E-state bit vectors available in the CACTVS suite of computational tools (www.cactvs.org). These bit vector assignments represent 431 chemical descriptors developed within CACTVS, with characteristics similar to assignments available within the MDL ISIS keys. SOM clustering treats the vectors of 431 structural descriptors for each agent in the same fashion as the vectors of $\log(GI_{50})$ values used for SOM clustering of the biological data.

We have clustered the set of 362 ExMOA compounds with the E-state structural bit vectors and have investigated compound clustering using the SOM method. The correlation between biological function clusters and structural clusters was accomplished with a heuristic matching algorithm that calculates the shared membership of clusters in both the function and structure sets. This results in a structure/function (S/F) plot where the linear correlation coefficient is the quantitative quality measure. It should be noted that what is chiefly of interest is the change in S/F correlation, not the absolute quantity. Therefore, any measure that accurately reflects relative correlation will serve as a surrogate marker for quality in the sensitivity analysis.

Table 1 lists the correlations between cluster memberships determined from biological response data ($\log(GI_{50})$) and chemical structure (bit vectors) for different data conditioning treatments. We have found a 15% [(0.9002-0.7820)/0.7820] improvement in the correlation coefficient with the Z-score normalization over an analysis based on raw data. This improvement is statistically significant, with an ANOV1 p value of 1.7e-15; a clear indication that

Z-score normalization enhances the quality of clustering. Based on these results, we believe it is important to routinely condition data to achieve a common reference and scale. In addition to Z-score normalization, the magnitude of any component of a data vector has been capped at a value of 3 absolute deviation units from the vector mean. Capping prevents the difference between two data vectors being dominated by a single or a few cell lines which have extreme values. Avoiding strong outliers by data capping improves the S/F correlation by 2.0% [(0.9185-0.9002)/0.9002]. This apparently small improvement is, however, statistically significant, with an ANOV1 p value of 4.4e-6, and has been adopted as a feature of data conditioning.

Another important design choice for data conditioning, that has been mentioned earlier, is the treatment of missing data. Oftentimes missing data are replaced by the mean value based on existing data. Our analysis indicates that this approach can substantially distort the information contained within the actual data. We find that retaining missing data elements as unknowns, rather that replacement by their vector mean, improves the S/F correlation coefficient by 6% [(0.9185-0.8654)/0.8654]. This improvement is statistically significant, with an ANOV1 p value of 7.6e-14, [18] and supports the earlier claim that missing data should be treated as unknown. See Figure 4 for representative structure/function correlation plots.

2.5 Map Dimensions

The possibility that map dimensions may affect the quality of clusters was investigated using S/F correlations. The SOM method contains a heuristic for the ratio of the relative sides of the two-dimensional SOM map based on the ratio of the two largest eigenvalues as the linear SVD solution to the dataset [12]. Using this heuristic and the ExMOA dataset, the SOM analysis recommends a map size of 17x9 to yield an eigenvalue ratio of 1.89. Figure 5 displays the dependence of the S/F correlation coefficient for a selection of map ratios (the cluster number $\sim$ 153 for all maps shown in the Figure 5.) The ratio that maximized the correlation coefficient matched the heuristic at 1.89. Ratios below this value generate thinner and narrower maps with a concomitant rapid decrease in the S/F correlation. This result suggests that the arrangement of ExMOA clusters cannot be easily ordered, as would occur for the clades in a single linkage hierarchical tree. More square maps, i.e. higher ratio, also

resulted in decreased S/F correlations, but more gradually than for lower eigenvalue ratios.

2.6 Number of Clusters

Perhaps the most controversial part of cluster analysis involves determination of cluster number. One popular approach repeatedly samples single linkage hierarchical cluster trees generated by removing one or more data elements. Cluster nodes that occur most frequently in the sample trees define the number of clusters. The approach we have used calculates the dependence of the S/F correlation on cluster size, and the uses the percent of maximal clustering to determine cluster number. Cluster size in SOM clustering is equivalent to map dimensions. Using the ExMOA dataset, SOM clusters were generated for a range of map dimensions, and the results are displayed in Figure 6. For 99% of maximal clustering the ExMOA is calculated to have $\sim$ 110 clusters. Since there is exponential improvement in the percent maximal clusters with cluster number, the S/F correlation decreases with increasing number of clusters. Based on the heuristic that a cluster number above 110 is sufficient to achieve at least 99% coverage of S/F correlations our selection of 153 (17x9) clusters exceeds this criterion.

2.7 Number of Cell Lines

Our analysis explored the role of tumor cell number in our SOM analysis using the structure/function correlation. The correlation with number of cell lines has two or three basic regimes (see Figure 7). Below $\sim$ 20 cell lines the S/F correlation drops off dramatically, between 20-50 cell lines the correlation rapidly increases, while for greater than 70 cell lines the correlation achieves a maximum. Although further analysis of this result will not be presented here, there is a clear indication that a near optimal clustering result can be achieved with fewer than the 80 tumor cell lines analyzed herein.

2.8 Robustness

We have investigated the behavior of our method of analysis towards noisy and degraded input data. Figure 8 (upper panel) shows the sigmoidal decrease in S/F correlation with

decreasing completeness of the input data. The data set was degraded by systematically removing data elements with the most extreme Z-score values. The results show that from 100% thru 70% completeness of data the S/F correlation is resistant to this degradation. Below 70% the correlation coefficient rapidly decreases approaching a minima at 0.45.

This analysis illustrates the importance of diversity within a data vector. The behavior of the S/F correlation with degraded the data is relatively stable against datasets which exhibit above a 10% coefficient of variation. Below Z-score of $\sim$1.1 absolute deviation units the S/F correlation is drastically decreased (see Figure 8 lower panel). Consistent with intuition, a data vector with a large amount of diversity can be more easily assigned to a cluster when compared to data vectors with a small absolute deviation. Based on this result, our analysis excludes data vectors with a mean absolute deviation below 8%.

Table 1. Data Conditioning Structure/Function Correlation

Missing Data	Normalization	Capping	S/F Corr. Mean	S/F Corr. Std.	Samples
NaN	Raw	No Cap	0.7820	0.0648	20
NaN	Z-score	No Cap	0.9002	0.0182	40
NaN	**Z-score**	**Cap ± 3**	**0.9185**	**0.0147**	**40**
Mean	Z-score	Cap ± 3	0.8654	0.0339	40

Multiple SOM maps were generated from random starting conditions for different combinations of data conditioning with respect to normalization(Raw vs. Z-score), capping (none vs ± 3) and treatment of missing data (replace with mean vs NaN:unknown). The correlation coefficient is determined between each of these SOM maps and the SOM clustering based on the structural descriptors. The basis of this comparison is that maps with the highest structure/function correlations are most desirable. Values represent averages for total number of samples.

4 Figure Captions:

Figure 1:

Pattern Normalization - The growth inhibition (GI_{50}) biological response pattern of a test compound is dependent on a variety of experimental features. This results in a wide variance in signal strength masking the underlying biological response of the cell lines to the given test compound. The $log(GI_{50})$ data for a given compound is normalized by transforming with the Z-score function across the cell lines [$z_{ij} = (g_{ij} - \bar{g}_j)/\sigma_j$ for each element across row j; where z_{ij} is the Z-score for element i, j, g_{ij} is the $log(GI_{50})$ value, $\bar{g}_j$ is the mean and σ_j is the standard deviation across row j]. The magnitude of each element is capped at +/- 3 standard deviation units.

Response Normalization - Different cell lines have varying sensitivities in response to the introduction of test compounds. For example, multi-drug resistant cell lines would be expected to give low signal strength because these cells can efficiently transport small molecules out of the cell. The response pattern is normalized across the cell lines by transforming each column of the data matrix with the Z-score function. The magnitude of each element is capped at +/- 3 standard deviation units. Data conditioning by normalizing pattern and response improves clustering and makes the patterns in the data visually accessible. The ordering of cell lines is as follows: leukemia (leu) 1-8, lung not small cell (lns) 9-20, small cell lung (scl) 21-23, colon (col) 24-32, central nervous system (cns) 33-41, melanoma (mel) 42-51, ovarian (ova) 52-57, renal (ren) 58-67, prostate (pro) 68-69, breast (bre) 70-80. The data set shown consists of 533 data vectors which comprise an extended mechanism of action (ExMOA) data set.

Figure 2: *Clustering(Top Panel)* - The conditioned growth inhibition ($\log(GI_{50})$) data consists of an $M \times N$ matrix of data elements. In the example above there are 533 data vectors for the compounds ($M = 533$), and there are 80 components for each data vector measuring the response across the different cell lines ($N = 80$). Two of the 80 dimensions for the 533 data vectors are shown in the figure (blue dots.) A set of P cluster vectors are chosen to represent the data space. The number of cluster vectors and the map dimensions are chosen to reflect the information contained in the data space as measured by the number

of data vector samples and the extent of the first two principle components found with single value decomposition (SVD) method. The cluster vectors are shown as open red circles (in the example above $P = 153$.) The plot on the left shows the initial coordinates for the cluster vectors (2 of the 80 dimensions are shown.) These initial coordinates where calculated by gridding the vectors along the principle component vectors found by solving the linear equations via the SVD method. The right plot displays the cluster vectors found with the self-organizing map (SOM) method. The SOM method minimizes the sum square error distance between the data and the cluster vectors which represent the local biological response space.

Mapping(Bottom Panel) - To make the information contained in the high dimensional clustering space accessible for drug discovery, the P clusters in N dimensions are projected on to a two dimensional map which represents the biological response space. The mapping is a to non-linearly function which transforms the data space such that each cluster vector is uniformly represented in the two dimensional map. The SOM clustering and uniform projection stretches the data space such that the map has a finer discrimination where more data is present. This ability to faithfully represent the biologically important information makes drug exploration and hypothesis generation possible with the large and high-dimensional data sets. The map shown above and enlarged in Figure 3, makes the mechanism of action and compound class readily accessible.

Figure 3: Sample of SOM clustering. Insert region on complete SOM map is expanded and partitioned according to response vectors (open red circles) and data vectors (closed blue dots). Partitions in the insert display relationship between information-rich and information-poor regions on the map and the representative SOM data vectors. Regions that are rich in data attract many response vectors which finely divide that region of data space and enhance discrimination between response patterns for the cell screening data.

Figure 4: Sample of structure/function correlation for different forms of data conditioning. Top panel: z-score, capping and no consideration for unknown data. Middle panel: z-score, no capping and no consideration for unknown data. Bottom panel: z-score, no capping and replacement of unknowns with group average. Best cases occur for z-score, capping adn NaN data conditioning.

Figure 5: Structure/function correlation versus ratio of map dimensions. Maximum average structure/function correlations occur for a ratio of map dimension of 1.89. This corresponds to the SOM map dimensions of 17x9. Points represent averages of correlation coefficient and standard deviation of correlation coefficient.

Figure 6: Structure/function correlation versus number of clusters. Repeat SOM maps were generated for different cluster numbers. The 99 percent asymptote occurs at 110 clusters and suggests that, on average, the highest structure/function correlations will arise for more that 99 percent of the SOM maps when cluster size exceeds 110.

Figure 7: Structure/function correlation versus number of cell lines. Highest correlations occur for greatest number of cell lines. Plateaus are observed in the average structure/function correlation for 50-60 cell lines and 20-30 cell lines. Fewer that 20 cell lines drastically reduce the structure/function correlation coefficient.

Figure 8: Structure/function correlation versus data completeness (upper panel) and versus z-score threshold (lower panel). Incomplete datasets are reasonably well tolerated above 60 percent. When greater that 40 percent of the data is removed, the structure/function correlation declines continuously. Open circles represent maximum response. Lower panel displays correspondence between structure/function correlation coefficient and z-score. Low z-scores indicate a relatively flat cellular response pattern. A uniform cellular response is accompanied by poor placement in SOM map clusters, and can be used to locate random regions of SOM maps.

Figure 9: Complete SOM map for the ExMOA dataset (see text). Map consists of 17x9 clusters, with the number of compounds in each cluster indicated by the size of hexagon at each loci (also highlighted with actual numerical count). Som analysis display a clear separation between anti-mitotic agents (top) and nucleic acid affecting agents (bottom). Compounds within selected clusters are shown at map borders. Consistent with the structure/function correlation analysis, a strong correspondence is seen between structurally similar compounds and their appearance in a functional cluster.

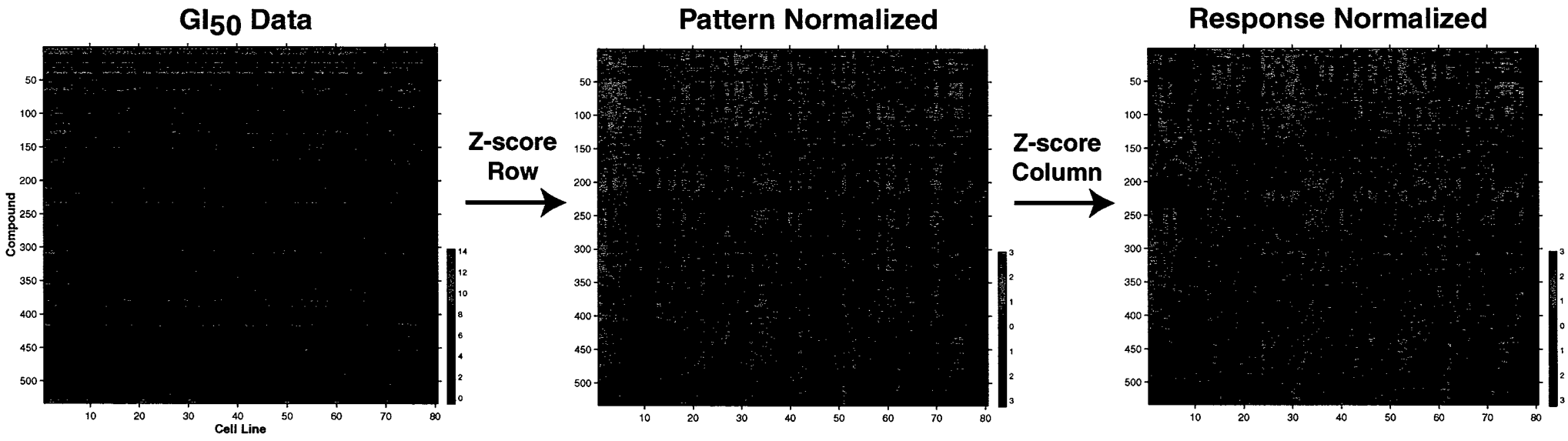
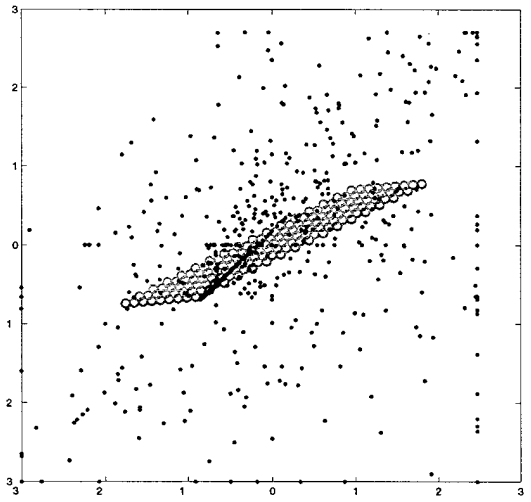
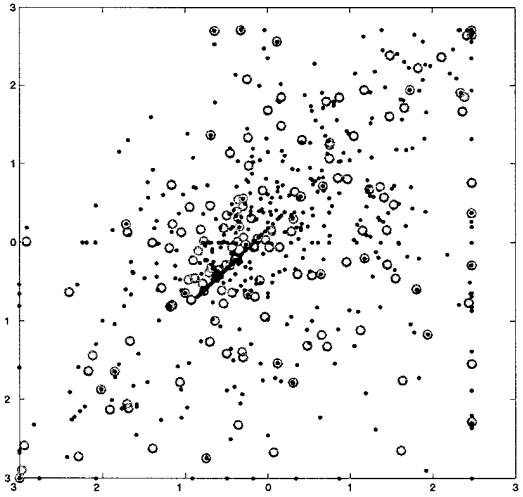


Figure 1 - Covell



SOM Clustering

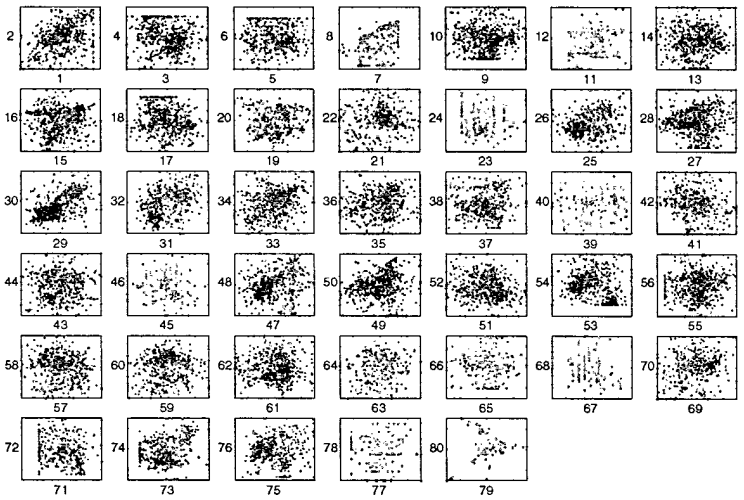
(2 of the 80 dimension shown)



: - Cluster Vector (2 of the 80 dimensions shown)

• - Data Vector (2 of the 80 dimensions shown)

Optimized Placement of 153 Cluster Vectors



153 Cluster Vectors in 80 dimensions

Mapping: 2-d Uniform Projection

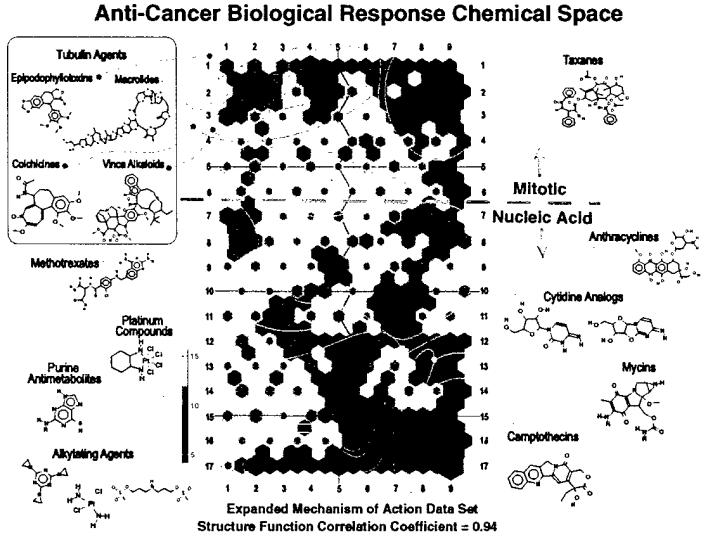
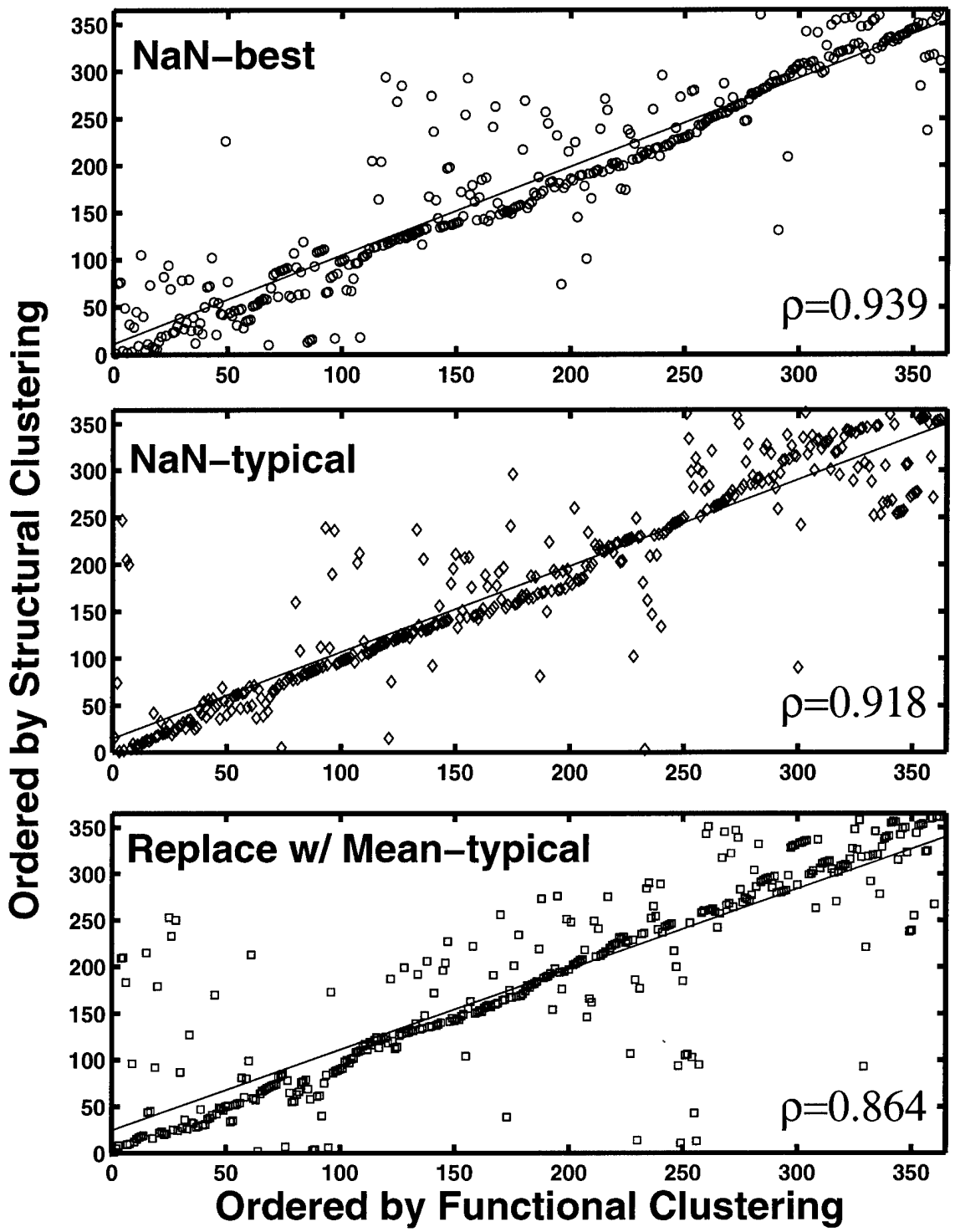


Figure 2 - Covell

MSsomPartitionDataFig0.2

Figure 3 - Cowell

Structure–Function Correlation



Expanded Mechanism of Action Data Set

Figure 5 - Covell

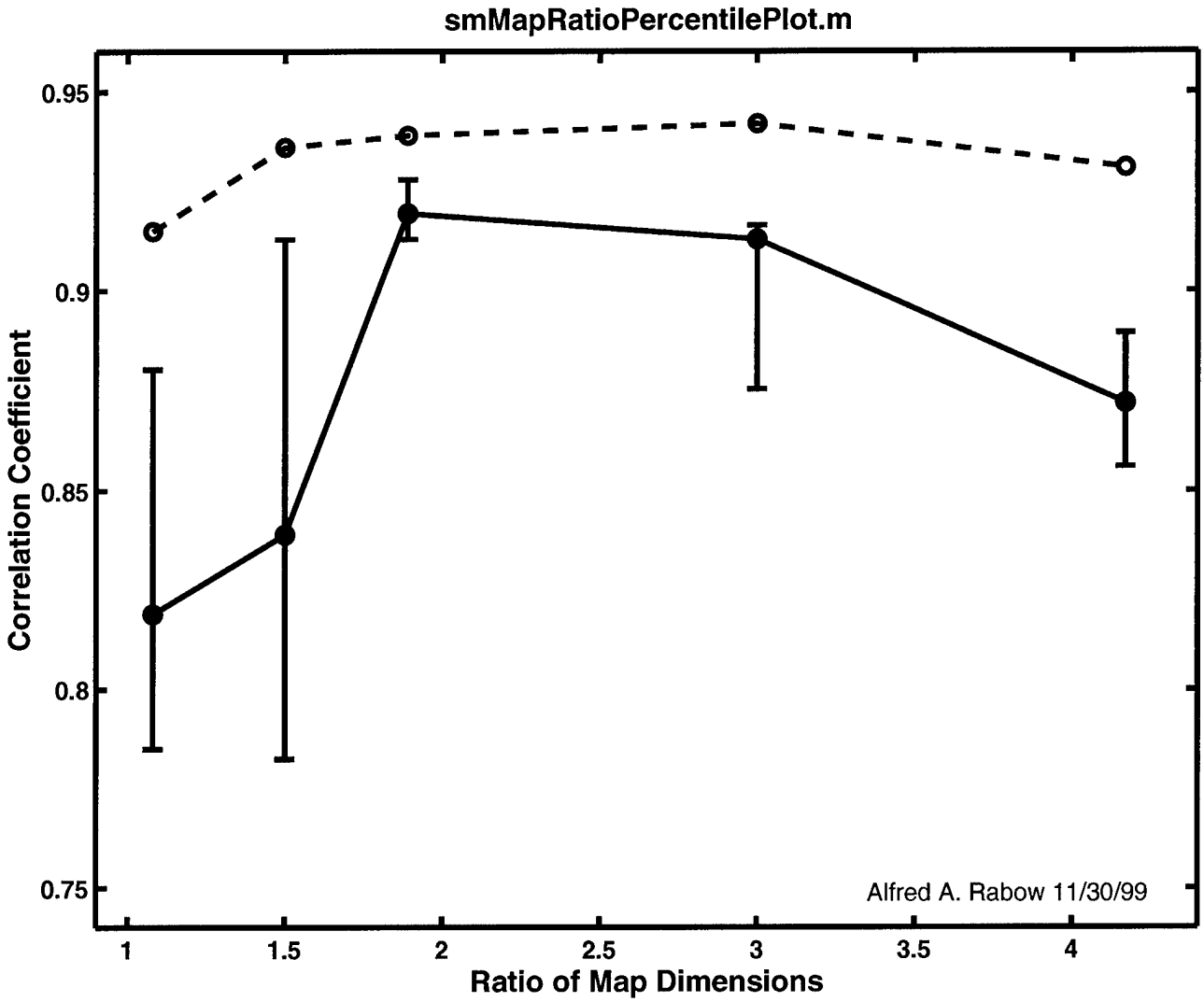
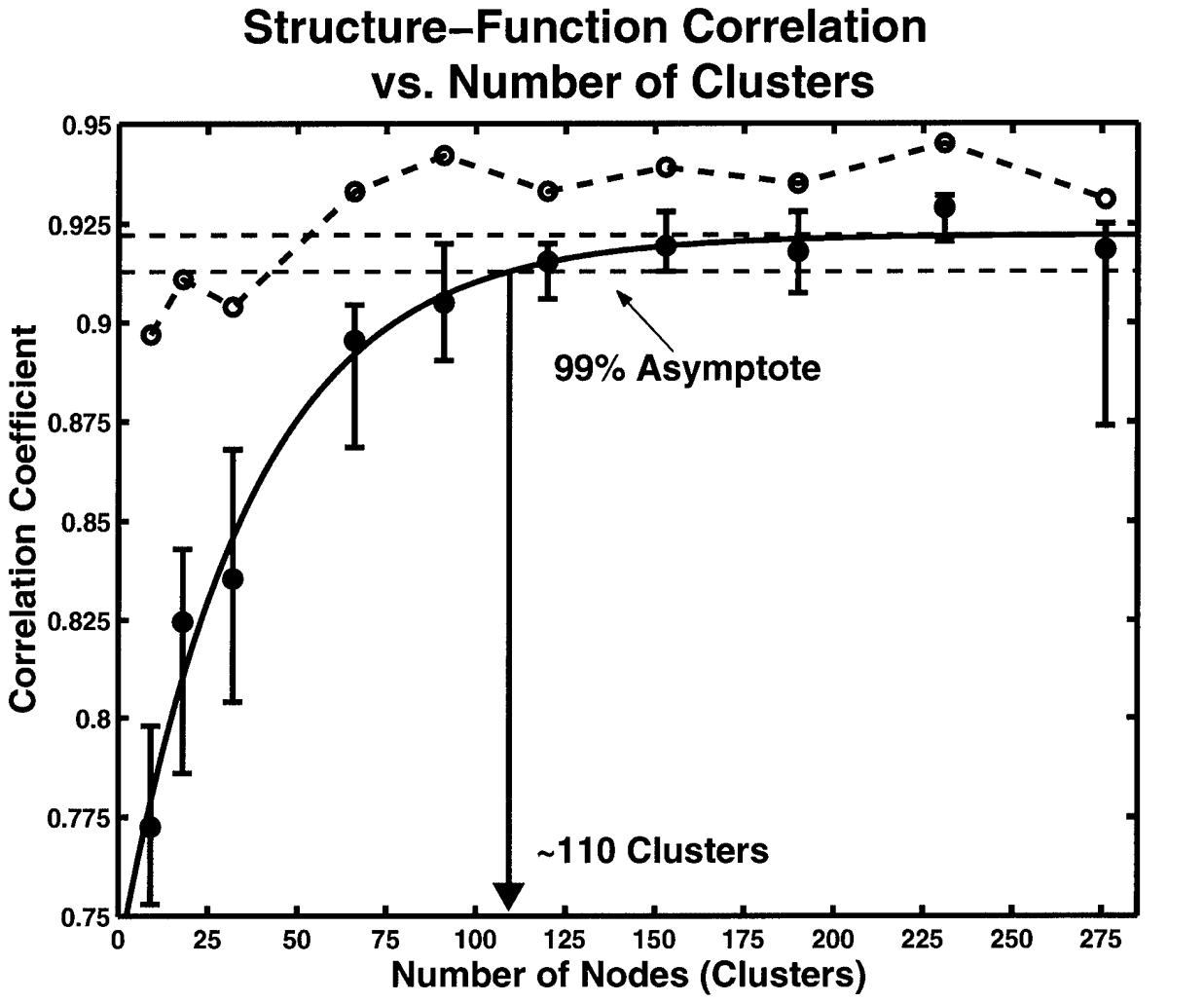


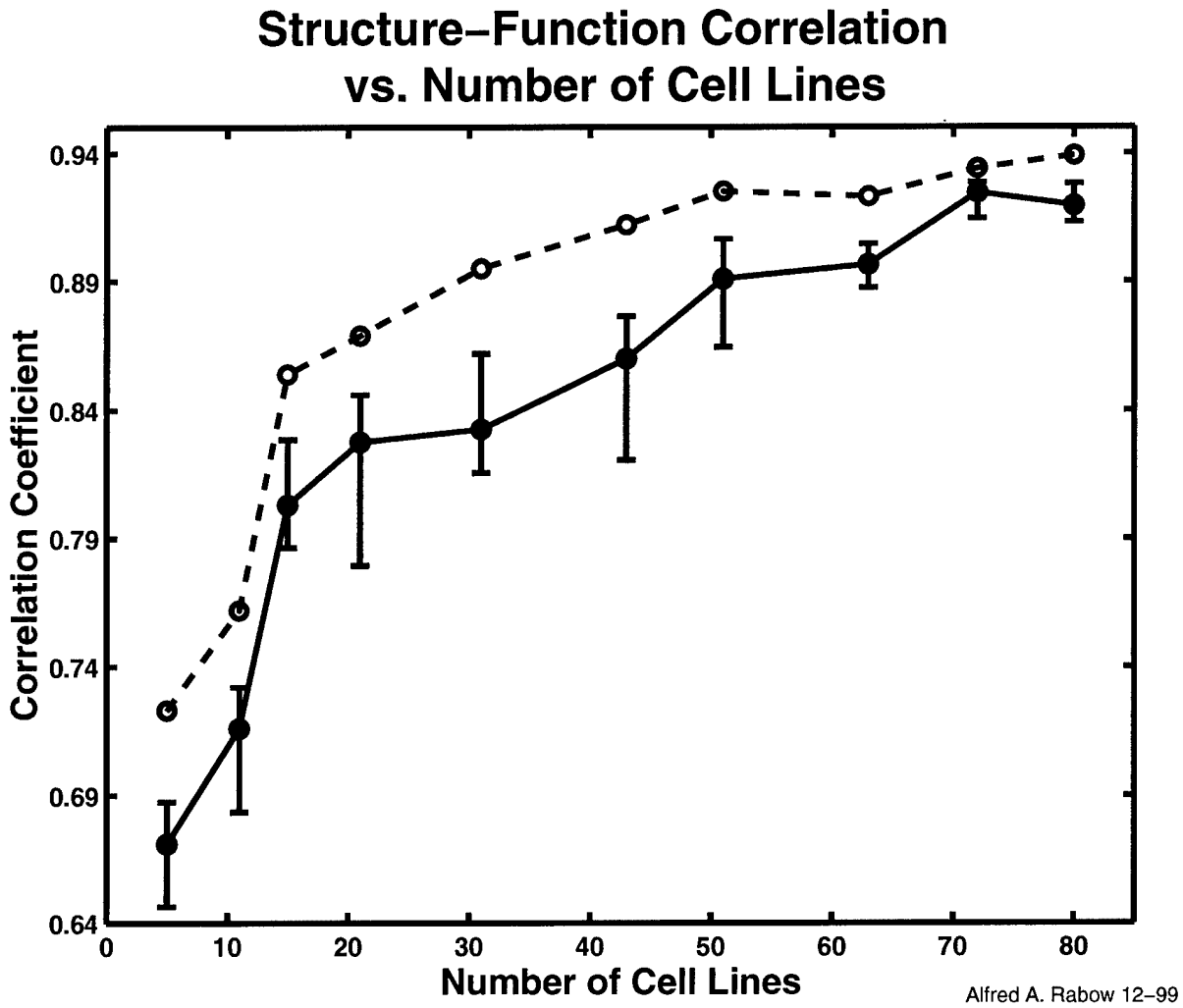
Figure 6 - Covell



Alfred A. Rabow 12–99

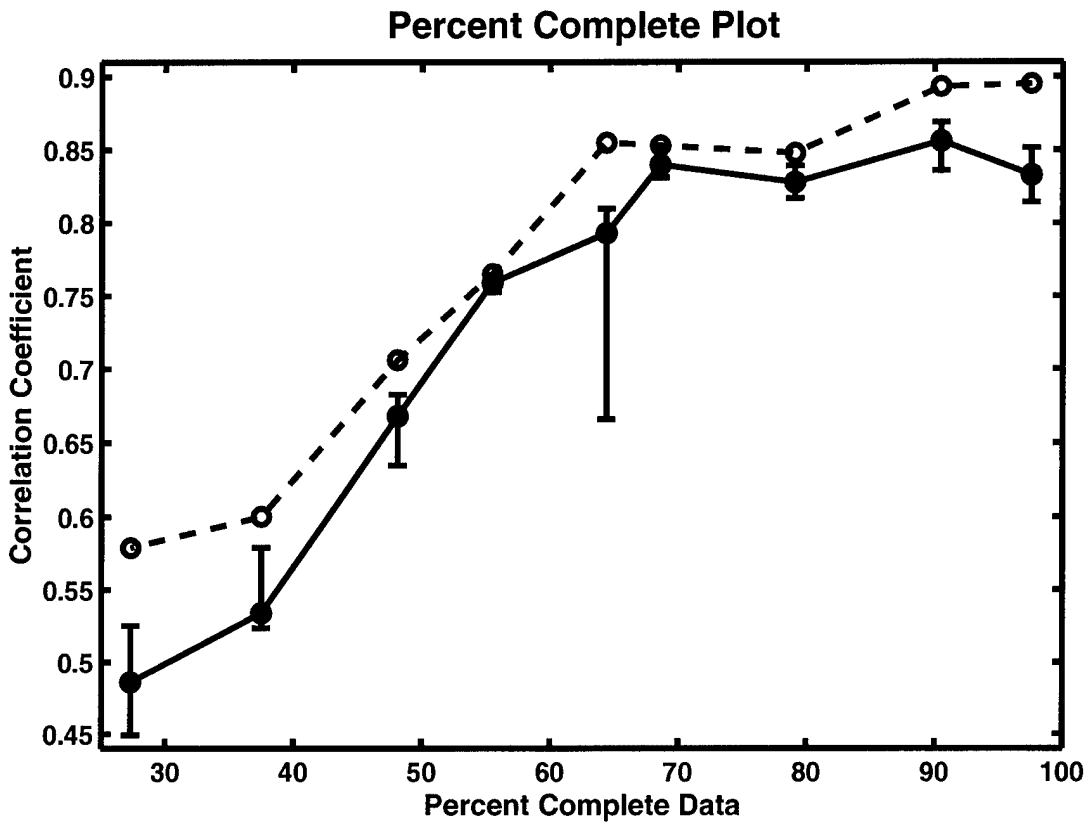
Expanded Mechanism of Action Data Set

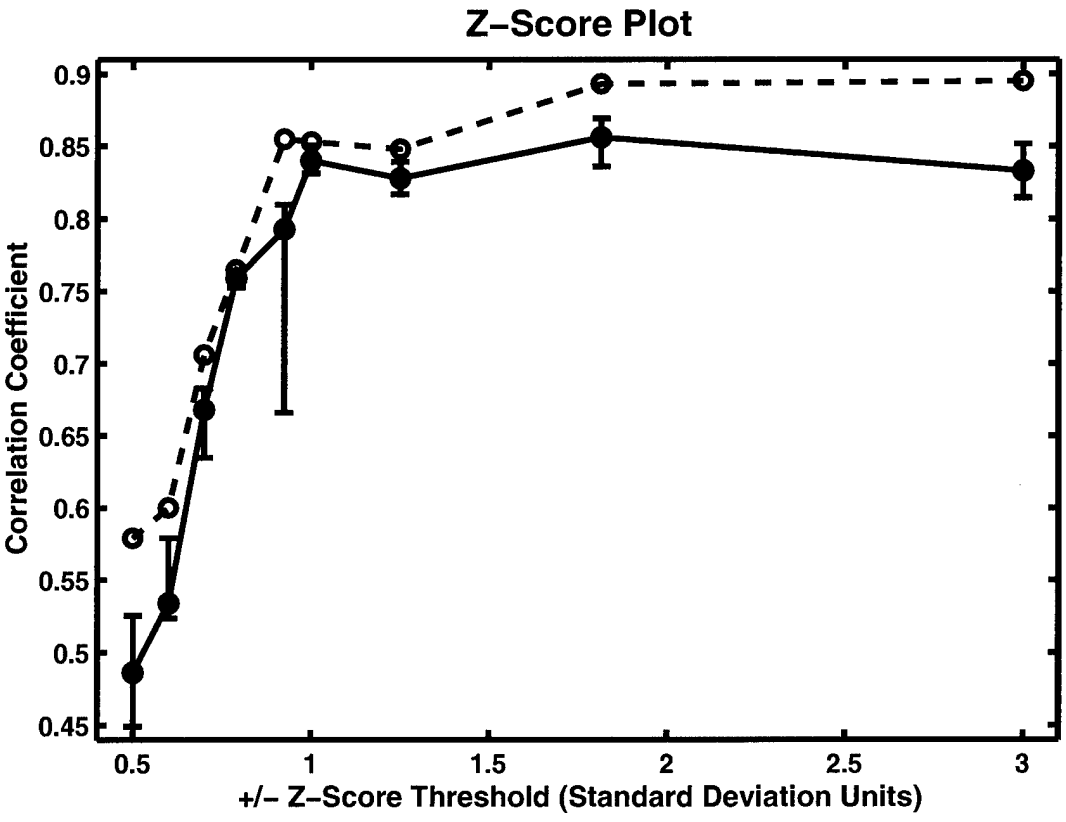
Figure 7 - Covell



Expanded Mechanism of Action Data Set

Structure–Function Correlation vs. Degradation of Data Set

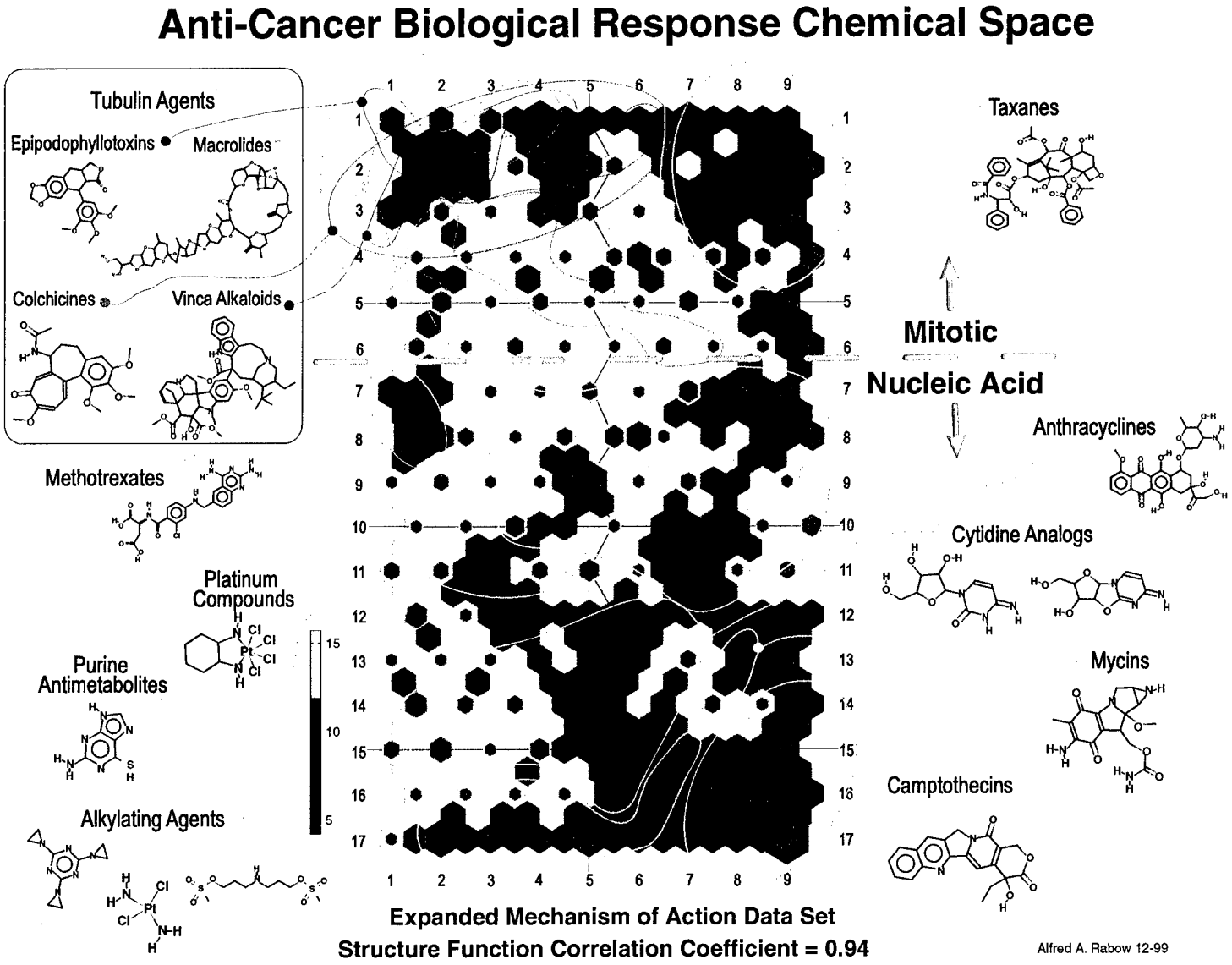




Alfred A. Rabow 12–99

Expanded Mechanism of Action Data Set

Figure 9 - Covell



5 KEY RESEARCH ACCOMPLISHMENTS:

The primary accomplishments in this research project are to clearly establish that screening patterns can be used effectively to organise large biological datasets into a meaningful form useful for hypothesis generation about mechanism of action for anticancer agents. This foundation provides the basis for a complete analysis of ALL the screening data publically available at the National Cancer Institute's Drug Screening effort.

- Construction of an *in silico* toolkit for the comprehensive analysis of large screening databases.
- Application of a suite of graphical displays as aids in data analysis.
- The development of a rigorous method for correlating biological function, measured in terms of cell killing, and structural classifications of selected compounds.
- The application of structure/function maps for pharmacophore searching and receptor mapping.
- Applications of this information for exploring efficacy of molecular isotypes.
- The development of a means for systematic investigations of cellular response patterns and their application towards an improved molecular taxonomy of chronic diseases such as cancer.
- The application of this information for the purpose of identifying and modifying representative elements of combinatorial libraries as potential lead compounds.

6 Reportable Outcomes:

- Keskin, O., Bahar, I, Jernigan, R.L., Beutler, J.A., Shoemaker, R.H., Sausville, E.A. and Covell, D.G.: Characterization of anticancer agents by their growth inhibitory activity asnd relationships to mechanism if action and structure. Anticancer Drug Design 15, 79-98, 2000.

- The development of tools for anticancer drug discovery. Rabow, A.R., Keskin, O., Shoemaker, R.H., Sausville, E.A., Jernigan, R.L. and Covell, D.G. Presentation at Protein Society Meetings, 2000.

- In silico drug discovery. A.R. Rabow, Shoemaker, R.H., Sausville, E.A., Jernigan, R.L. and Covell, D.G. Presentation at annual National Cancer Institute Retreat, 2000.

- Deriving Structures for Lead Drug Discovery. A.R. Rabow, Shoemaker, R.H., Sausville, E.A. Jernigan, R.L. and Covell, D.G., Presented at the Era of Hope Meeting in Atlanta, GA, 2000.

- Bai, R., Covell, D. G., Pei, X-F, Ewell, J. B., Nguyen, N. Y., Brossi, A. and Hamel, E.: Mapping the binding site of colchicinoids on beta-tubulin: 2-chloroacetyl-2-demethylthiocolchicine covalently reacts predominately with cysteine 239 and secondly with cysteine 354, *J Biol. Chem*, to appear, 2000.

- Rabow, A.R., Shoemaker, R.A., Sausville, E.A. and Covell, D. G.: Analysis of the NCI's tumor screening panel; Assessment of relationships between chemical structure and mechanism of action. submitted, 2000.

3 CONCLUSION

A suite of computational tools has been developed for detailed analysis of large-scale high-throughput screening data for the purpose of lead drug discovery and potential identification of novel molecular targets in the treatment of human cancers. The method has been developed and tested against the National Cancer Institute's 60 tumor cell panel. This suite of analytical and display tools is focused in the areas of data conditioning, pattern association, visualization and data presentation, with additional functionalities that address signal scaling issues, missing data elements, and locality/non-linearity features of the data-space. Careful considerations in these areas are found to significantly enhance the extraction of additional information from large, complicated, screening databases as well as provide a general tool well suited for drug discovery. These results find strong correlations between molecular structure and putative mechanism of action for large classes of anticancer agents; with a clear segregation of compounds according to their activities against specific molecular targets. More significantly, screening cells that are found within specific tumor cell panels are found to respond similarly to classes of molecular agents. This information can lead directly to the formulation of alternative chemical analogs and hypotheses about specific molecular targets and their affected biosynthetic pathways [19].

References

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

[2] A. Monks, D. Scudiero, P. Skehan, R. Shoemaker, K. Paull, D. Vistica abd C. Hose, C. Langely, P. Cronise, A. Vaigro-Wolff, M. Grey-Goodrich, L. Cambell, J. Mayo, and M. R. Boyd. Feasibility of a high flux anticancer drug screen using a diverse panel of cultured human tumor cell lines. *J Natl. Canc. Inst*, **83**, 757-766, (1991).

[3] K.D. Paull, R. H. Shoemaker, L. Hodes, A. Monks, D.A. Scudiero, L. Rubenstein, J. Plowman, and M. R. Boyd. 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.*, **81**, 1088-1092, (1989).

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

[5] S.M. Swanson E.F. Meyer and J.A. Williams. Molecular modeling and drug design. *Pharmacology and Therapeutics*, **85**, 113-121, (2000).

[6] T. K. Attwood. The quest to deduce protein function from sequence: the role of pattern databases. *International Journal of Biochemistry and Cell Biology*, **32**, 139-155, (2000).

[7] M. R. Boyd. The NCI in vitro anticancer drug discovery screen. In B. Teicher, editor, *Anticancer Drug Development Guide: Preclinical Screening, Clinical Trials and Approval*, pages 23–41. Humana Press, Totowa, New Jersey, (1995).

[8] P. H. A. Sneath and R. R. Sokal. *Numerical Taxonomy.* W. H. Freeman and Company, San Francisco, (1973).

[9] A. Giuliani, A. Colosimo, R. Benigni, and J. Zbilut. On the constructive role of noise in spatial systems. *Physics Letters A*, **247**, 47-52, (1998).

[10] O. Keskin, I. Bahar, R.L. Jernigan, J.A. Beutler, R.H. Shoemaker, E.A. Sausville, and D.G. Covell. Characterization of anticancer agents by their growth-inhibitory activity and relationships to mechanism of action and structure. *Anticancer Drug Discovery*, **15**, 79-98, (2000).

[11] W.W. vanOsdol, T. G. Myers, K. D. Paull, K. W. Kohn, and J. N. Weinstein. Use of kohonen self-organizing map to study the mechanism of action of chemotherapeutic agents. *Journal of the National Cancer Institute*, **86**, 1853-1859, (1994).

[12] K. Liu. Application of SVD in optimization of structural modal test. *Computers and Structures*, **63**, 51-59, (1997).

[13] J. Bellenson. Integrating information technology and drug discovery processes. *Nature Biotechnology*, **16(7)**, 597-598, (1998).

[14] O'Connor P.M., Jackman J., Bae I., Myers TgG., Fan S., Mutoh M., Scudiero D.A., Monks A., Sausville E.A., Weinstein J.N., Friend S., Fornace A.J. Jr., and Kohn K.W. Characterization of the p53 tumor suppressor pathway in cell lines of the national cancer institute anticancer drug screen and correlations with the growth-inhibitory potency of 123 anticancer agents. *Cancer Research*, **57(19)**, 4285-4300, (1997).

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

[16] Y. C. Martin and P. Willet. *Designing Bioactive Molecules.* American Chemical Society, Washington D.C., (1998).

[17] G. Maggiora and M.A. Johnson. *Concepts and Applications of Molecular Similarity.* John Wiley, New York, NY, (1990).

[18] K.Hrach. Comparison of survival between two groups using software SAS, S-PLUS and STATISTICA. *Journal of Medical Informatics*, **45(1-2)**, 31-33, (1997).

[19] Benton D. Integrated access to genomic and other bioinformation: an essential ingredient of the drug discovery process. *Sar and Qsar in Environmental Research*, **8(3-4)**, 121-155, (1998).

7 APPENDIX:

A description of a significant portion of this effort is in the final stages of preparation. When this manuscript has been completed, a copy can be forwarded to the ARMY for inclusion in this report.

The two following references will be provided herein:

- Keskin, O., Bahar, I, Jernigan, R.L., Beutler, J.A., Shoemaker, R.H., Sausville, E.A. and Covell, D.G.: Characterization of anticancer agents by their growth inhibitory activity asnd relationships to mechanism if action and structure. Anticancer Drug Design 15, 79-98, 2000.

- Bai, R., Covell, D. G., Pei, X-F, Ewell, J. B., Nguyen, N. Y., Brossi, A. and Hamel, E.: Mapping the binding site of colchicinoids on beta-tubulin: 2-chloroacetyl-2-demethylthiocolchicine covalently reacts predominately with cysteine 239 and secondly with cysteine 354, *J Biol. Chem*, to appear, 2000.

8 FINAL REPORT:

Dr. Alfred R. Rabow has been completely funded by the IDEAS award. His efforts have been instrumental and vital for establishing a larger program involved in data analysis of screening measurements. This effort would not have been possible without the support of the ARMY Breast Cancer Project. Duing the award period, one SGI workstation was purchased and used as a dedicated computer for this project.

Anti-Cancer Drug Design (2000), 15, 79–98

Characterization of anticancer agents by their growth inhibitory activity and relationships to mechanism of action and structure

Ozlem Keskin[1,2], Ivet Bahar[1], Robert L. Jernigan[2], John A. Beutler[3], Robert H. Shoemaker[4], Edward A. Sausville[4] and David G. Covell[2]

[1]Chemical Engineering Department and Polymer Research Center, Bogazici University, TUBITAK Advanced Polymeric Materials Research Center, Bebek 80815, Istanbul, Turkey, [2]Molecular Structure Section, Laboratory of Experimental and Computational Biology, NCI, NIH, SAIC, Frederick, MD 21702 and Bethesda, MD 20892 USA, [3]Laboratory of Drug Discovery Research and Development, DTP, DCTDC, NCI, SAIC, Frederick, MD 21702 and [4]Developmental Therapeutics Program, Divison of Cancer Treatment and Diagnosis NCI, NIH, Frederick, MD 21702 and Bethesda, MD 20892, USA

Summary

An analysis of the growth inhibitory potency of 122 anticancer agents available from the National Cancer Institute anticancer drug screen is presented. Methods of singular value decomposition (SVD) were applied to determine the matrix of distances between all compounds. These SVD-derived dissimilarity distances were used to cluster compounds that exhibit similar tumor growth inhibitory activity patterns against 60 human cancer cell lines. Cluster analysis divides the 122 standard agents into 25 statistically distinct groups. The first eight groups include structurally diverse compounds with reactive functionalities that act as DNA-damaging agents while the remaining 17 groups include compounds that inhibit nucleic acid biosynthesis and mitosis. Examination of the average activity patterns across the 60 tumor cell lines reveals unique 'fingerprints' associated with each group. A diverse set of structural features are observed for compounds within these groups, with frequent occurrences of strong within-group structural similarities. Clustering of cell types by their response to the 122 anticancer agents divides the 60 cell types into 21 groups. The strongest within-panel groupings were found for the renal, leukemia and ovarian cell panels. These results contribute to the basis for comparisons between log(GI_{50}) screening patterns of the 122 anticancer agents and additional tested compounds.

Key words

clustering behavior/SVD/tumor cell-line screen

Introduction

Development of high-throughput screening technologies in drug discovery has led to dramatic increases in the diversity of compounds that can be tested (Gordon *et al.*, 1994; Ganesan, 1998; Gray *et al.*, 1998) and in the types of targets available for testing (Monks *et al.*, 1991; Grever *et al.*, 1992; Boyd and Paull, 1995; Kauver *et al.*, 1995; Chee *et al.*, 1996; Botstein and Cherry, 1997; Castell and Gomes-Lechon, 1997; Zhang *et al.*, 1997). Accompanying these advances has been the development of a diverse collection of general approaches for mining the large quantity of data generated by these systems (Marchington, 1995; O'Connor *et al.*, 1997; Ajay *et al.*, 1998; Bellenson, 1998; Benton, 1998; Gillet *et al.*, 1998; Sadowski *et al.*, 1998; Shi *et al.*, 1998b,c). Database-related, information-intensive drug discovery efforts (Myers *et al.*, 1997) are showing promise in revealing relationships between drug screening profiles and potential therapeutic targets. Extending these efforts by further exploration of relationships between screening profiles and chemical structures may enhance the discovery of novel chemotherapeutic agents.

In this paper we re-examine the publicly available data from the cancer drug discovery program at the National Cancer Institute (NCI). Our goal is to systematically analyze the relationship between (i) the growth inhibitory activities for a set of anticancer agents from the panel of 60 tumor cell lines; (ii) the structural features of the tested agents; and (iii) their apparent mechanism of growth inhibitory action (MOA). Based on the hypothesis that selective *in vitro* activity of a compound against cancer cell lines might be predictive of its activity against the corresponding specific type of human tumor, the NCI has developed and made available results of primary drug screens against 60 different human cancer cell lines (http://dtp.nci.nih.gov). Among other endpoints available in the NCI's database, the growth inhibitory activity of each compound, expressed as the drug concentration (GI_{50})

Correspondence to: D.G. Cavell. Received 23 December 1999; accepted 10 March 2000

required to inhibit tumor cell growth by 50% compared with an untreated cell was selected for analysis. Log(GI_{50}) values for a given compound across all tumor cell lines provide its activity pattern for comparison with patterns from other tested compounds. Similarities in patterns of *in vitro* inhibitory activity have been shown to be related to MOAs, modes of resistance and molecular structure (Boyd, 1995; Boyd and Paull, 1995; Paull *et al.*, 1995; Hrach, 1997; Myers *et al.*, 1997; O'Connor *et al.*, 1997; Shi *et al.*, 1998b,c). To date, the NCI has screened >70 000 chemical compounds and a similar number of natural product extracts against a panel of 60 different tumor cell lines.

Several algorithms have previously been applied to analyze activity patterns. These algorithms utilize, in various ways, the tools of multivariate statistical clustering (Hrach *et al.*, 1997). As an example, the internet-accessible program COMPARE (Paull *et al.*, 1989, 1995) uses Pearson correlation coefficients (PCCs) to extract compounds with screening patterns similar to a 'seed' compound. Applications of back-propagation neural networks (Weinstein *et al.*, 1992) and Kohonen self-organizing maps (Koutsoukos *et al.*, 1994) have demonstrated varying success when predicting MOAs and grouping compounds based on similar activity patterns. These methods also complement the COMPARE program by identifying clusters of 'seed' compounds, thus addressing the important question of whether a 'seed' compound appears on the lists of highly correlated activity patterns for all other 'seeds' in the data set. Statistical and artificial intelligence techniques, including principal component analysis, hierarchical cluster analysis, stepwise linear regression and multidimensional scaling, have begun to be applied to the NCI's screening data (van Osdol *et al.*, 1994; Shi *et al.*, 1998a).

Structurally similar compounds can have similar physicochemical properties and thus are thought to have similar biological activities, consistent with the similarity property principle (Johnson and Maggiora, 1990). For example, a dramatic coherence between molecular structures and activity patterns was observed for 112 ellipticine analogs (Shi *et al.*, 1998c). Detailed crystallographic and NMR studies further support the similarity property concept by demon- strating that ligand–receptor interactions are characterized by complementary shapes and chemical characteristics (Janin and Chothia, 1990; Clackson and Wells, 1995; Schreiber and Fersht, 1995; D.G. Covell *et al.*, manuscript in preparation). Cell-based screening assays represent a complex array of interactions that is monitored as cell growth or killing [e.g. log(GI_{50})]. Differential activity patterns in these measurements can result from the activity of compounds that interact well, poorly or not at all with one or many targets within the panel of cell types. Earlier attempts to establish correspondences between activity patterns, MOAs and chemical structure found general clustering (i) for compounds of similar chemical structure, and (ii) for compounds classified as having a similar mechanism of action (MOA), yet having diverse chemical structures (Shi *et al.*, 1998a). Distant clustering was also found for compounds similar in chemical structure but having different MOAs (Shi *et al.*, 1998a). Earlier studies by Paull *et al.* (Paull *et al.*, 1995; O'Connor *et al.*, 1997) demonstrated that anticancer agents having similar functional groups (e.g. chloroethylating agents, platinum analogs and nitrosoureas) produce similar activity patterns in cell-based screens. However, there are some compounds that display a relatively strong structural similarity, and yet exhibit drastically different activity patterns. Alternatively, compounds with similar activity patterns can have little structural correspondence to one another.

The present analysis identifies clusters of anticancer compounds based on their log(GI_{50}) activity patterns in NCI's data for 60 tumor cell lines. The analysis is performed on the set of 122 standard anticancer agents available in the NCI's Developmental Therapeutic Program's database. Here we adopt singular value decomposition (SVD) (Harary, 1971; Golub and Loan, 1989; Berry *et al.*, 1995; Liu, 1997; Bahar *et al.*, 1998) and hierarchical clustering methods (Sneath and Sokal, 1973) to cluster the chemotherapeutic agents. Compounds clustered with these methods are to be compared by their assigned MOAs and their structural similarities.

Methods

Variance-based measures of similarity rely on the spread in a data set to determine membership within a cluster. Principal component analysis (PCA), SVD, *D*-optimal design and *k*-nearest neighbor clustering are commonly used as variance-based methods. These have as their overall goal the minimization of the noise-to-signal ratio (Giuliani *et al.*, 1998). The SVD approach has been shown to be a powerful method to filter noise and enhance the information content of the original data (Harary, 1971; Golub and Loan, 1989; Berry *et al.*, 1995; Liu, 1997). Similar to PCA, SVD defines rotation of axes (principal components) so that columns in the data matrix maximize their standard deviation with respect to other columns in the data set. This transformation yields a new space where the columns of data exhibit maximum variance (i.e. minimum correlation) with respect to one other. The original data can be re-expressed approximately as a linear combination of a few dominant principal components. This new space, referred to as the SVD space, has previously been effectively used, for example, to classify words within texts (Berry *et al.*, 1995) and protein structures with respect to their amino acid composition (Bahar *et al.*, 1998).

SVD analysis is used here to classify anticancer agents by examining their log(GI_{50}) values in the 60-dimensional space of the cancer cell lines. This space is transformed into an SVD space, where the anticancer agents are represented by activity

arrays emphasizing their differences. The compounds are clustered on the basis of their pairwise distances in SVD space, by using hierarchical clustering algorithms (Sneath and Sokal, 1973). The calculations discussed below have been coded into a Fortran program, which is available upon request. Many of these calculations can also be completed using the SAS library of utilities.

In general, the SVD of a given matrix **A** yields three matrices Λ, **U** and **V** which comprise (i) the singular eigenvalues λ_i of **A**, organized in ascending order in the diagonal matrix Λ; (ii) the orthonormal transformation matrix **U** that defines the relationship between the original coordinate frame and the SVD frame; and (iii) the normalized representation, $\mathbf{V}^T$, of the original matrix in the SVD space. **A** can thus be decomposed, hence the term 'singular value decomposition', into the product of these three matrices

$$\mathbf{A}_{mxn} = \mathbf{U}_{mxm}\Lambda_{mxm}\mathbf{V}^T{}_{mxn} \quad (1)$$

where the subscripts denote the dimensions of the two-dimensional matrices and the superscript T indicates the transpose. In general, the columns of **A** each represent a given quantity (here anticancer agents) characterized by m properties (activity patterns for 60 cell lines), whereas those of the product $\Lambda\mathbf{V}^T$ are the same quantities expressed in the SVD frame which best describes the similarities/differences between these quantities on the basis of their n properties. In the present application of the SVD method to anticancer compound screening data, each column of **A**, conveniently denoted as $\mathbf{a}_i$, is a 60-dimensional vector describing the activity pattern of a given drug i ($1 \leq i \leq 122$), expressed in terms of the $\log(GI_{50})$ values observed against the 60 tumor cell lines. Therefore the SVD of a 60 × 122 **A** matrix is performed, using the data set of $n = 122$ anticancer agents screened against $m = 60$ cell lines. The $\mathbf{a}_{ij}$ element of the **A** matrix is then row and column normalized by first subtracting the column average [i.e. the average $\log(GI_{50})$ value for each compound] and then subtracting the row average (i.e. the average for each cell line). The resulting relative cytotoxic potencies are thought to eliminate the differences arising from the generic characteristics of the particular cell lines and permits us to emphasize more clearly the differences among activity patterns of the anticancer agents. The activity pattern of the ith agent in the SVD space is used to define its distances from the activity patterns for the remaining ($n = 121$) compounds. The activity pattern of the ith agent in SVD space is represented by the ith column $\mathbf{v}^T{}_i$ of $\mathbf{V}^T$ pre-multiplied by Λ, and designated as $\mathbf{a}_i^* = \Lambda\mathbf{v}^T{}_i$ such that the SVD distance between agents i and j is

$$d_{ij} = [(\mathbf{a}_i^* - \mathbf{a}_j^*)\,(\mathbf{a}_i^* - \mathbf{a}_j^*)]^{1/2} = [(\Lambda\mathbf{v}^T{}_i - \Lambda\mathbf{v}^T{}_j)\,(\Lambda\mathbf{v}^T{}_i - \Lambda\mathbf{v}^T{}_j)]^{1/2}$$

These SVD distances constitute the basic measure for clustering the anticancer agents into groups in the present analysis. The analyzed set includes 122 compounds with six putative MOAs: 35 alkylating agents, 24 antimitotic agents, 16 topoisomerase I inhibitors, 19 topoisomerase II inhibitors, 16 RNA–DNA antimetabolites and 13 DNA antimetabolites.

Results

The results of clustering compounds according to their pairwise SVD distances are listed in Table I. Clusters obtained from pairwise distances place compounds with the most similar activity patterns adjacent to one another. Using this approach, clusters are ordered such that compounds with the greatest and least similarities in their SVD distances are presented first and last, respectively, in Table I. Figure 1 displays the 2-D structures of the compounds within each cluster.

Statistical clustering of these patterns was obtained using the SAS/STAT clustering algorithms. The cubic clustering criterion (CCC) was selected to determine cluster membership. This criterion estimates the number of clusters based on minimizing the within cluster sum of squares. The CCC calculation generates a rough approximation to a 'goodness of fit' measure under the null hypothesis that the data are sampled from a uniform distribution on a hyperbox (P-dimensional right parallelpiped). A t-test statistic with one degree of freedom ($t = 3.078$, $P < 0.05$, $n = 1$) is generated for testing the null hypothesis that a compound's SVD distance pattern is not different from a given cluster (i.e. cannot be excluded from the cluster). This method has been shown to help determine cluster number for both univariate and multivariate data with small sample sizes ($n \approx 20$). See SAS Technical Report A-108 for additional details (SAS, 1992).

The results of this analysis find that the 122 standard agents can be clustered into 25 groups, labeled Groups 1–25, and listed in Table I. Fifteen of these groups have at least two members, while the final 10 groups consist of a single agent. Figure 1 displays the molecular structures of these compounds, ordered according to the Groups 1–25 in Table I. The list of compounds in each group in Table I includes their putative MOAs and characteristic structural/functional groups. Multiple compounds within each group cannot be further subdivided on the basis of their $\log(GI_{50})$ patterns. However, structural similarities within clusters can be easily found by inspection of Figure 1.

Group 1 is composed of 38 compounds consisting predominantly of alkylating agents (23 compounds), topoisomerase II inhibitors (nine compounds), DNA antimetabolites (five compounds) and a single RNA–DNA antimetabolite. Alkylating agents are antitumor drugs that act through covalent binding of their alkyl groups to cellular molecules (Pratt *et al.*, 1994; Chabner and Longo, 1996). Many of these are proposed to attack the N-7 or O-6 atoms on guanine in

Table I

Compounds ordered according to pattern similarity

Cluster	*Name*	*NSC*	*MOA*	*Structural group*
1	teroxirone	296934	1	epoxide
1	AZQ	182986	1	aziridine
1	CHIP	256927	1	platinum
1	cis-platinum	119875	1	platinum
1	carboplatin	241240	1	platinum
1	hepsulfam	329680	1	alkane sulfonate
1	Yoshi-864	102627	1	alkane sulfonate
1	Busulfan	750	1	alkane sulfonate
1	cyclodisone	348948	1	alkane sulfonate
1	clomesone	338947	1	alkane sulfonate
1	guanazole	1895	6	
1	pyrazoloimidazole	51143	6	
1	ftorafur (pro-drug)	148958	5	
1	hydroxyurea	32065	6	hydroxyurea
1	melphalan	8806	1	nitrogen mustard
1	chlorambucil	3088	1	nitrogen mustard
1	br-propionyl piperazine	25154	1	nitrogen mustard
1	fluorodopan	73754	1	nitrogen mustard
1	mitozolamide	353451	1	nitrogen mustard
1	BCNU (carmustine)	409962	1	nitrosourea–nitrogen mustard
1	spirohydantoin mustard	172112	1	nitrogen mustard
1	methyl CCNU	95441	1	nitrosourea–nitrogen mustard
1	chlorozotocin	178248	1	nitrosourea–nitrogen mustard
1	PCNU	95466	1	nitrosourea–nitrogen mustard
1	CCNU	79037	1	nitrosourea–nitrogen mustard
1	3-HP	95678	6	hydrazinecarbonthio-amide
1	5-HP	107392	6	hydrazinecarbonthio-amide
1	asaley	167780	1	nitrogen mustard
1	amonafide	308847	4	–
1	hycanthone	142982	1	–
1	pyrazoloacridine (PZA)	366140	4	acridine
1	oxanthrazole	349174	4	anthracene
1	anthrapyrazole derivative	355644	4	anthracene
1	rubidazone	164011	4	anthracene dione
1	doxorubicin (Adriamycin)	123127	4	anthracene-daunorubicin
1	daunorubicin	82151	4	anthracene-daunorubicin
1	deoxydoxorubicin	267469	4	anthracene-daunorubicin
1	VP-16	141540	4	podophyllotoxin
2	thio-tepa	6396	1	aziridine
2	triethylenemelamine	9706	1	aziridine
2	dianhydrogalactitol	132313	1	epoxide
2	nitrogen mustard	762	1	nitrogen mustard
2	uracil nitrogen mustard	34462	1	nitrogen mustard
2	piperazine analog	344007	1	nitrogen mustard
2	piperazinedione	135758	1	piperazine
2	camptothecin derivative	643833	3	camptothecin
2	camptothecin, Na salt	100880	3	camptothecin
2	menogaril	269148	4	anthracene–daunorubicin
3	mitomycin C	26980	1	mitomycin
3	porfiromycin	56410	1	mytomycin
3	camptothecin	94600	3	camptothecin
3	camptothecin derivative	95382	3	camptothecin
3	camptothecin derivative	107124	3	camptothecin
3	*m*-AMSA (amsacrine)	249992	4	anthracene
3	camptothecin derivative	295501	3	camptothecin
3	camptothecin derivative	606173	3	camptothecin
3	camptothecin derivative	364830	3	camptothecin
3	camptothecin derivative	374028	3	camptothecin
3	aminocamptothecin	603071	3	camptothecin
3	camptothecin derivative	606172	3	camptothecin
3	camptothecin derivative	606985	3	camptothecin
3	camptothecin derivative	610457	3	camptothecin
3	camptothecin derivative	610458	3	camptothecin
3	camptothecin derivative	618939	3	camptothecin
4	camptothecin derivative	249910	3	camptothecin
4	camptothecin derivative	606947	3	camptothecin
4	camptothecin derivative	606499	3	camptothecin
4	camptothecin derivative	610456	3	camptothecin
4	camptothecin derivative	610459	3	camptothecin
4	camptothecin derivative	629971	3	camptothecin
5	camptothecin derivative	176323	3	camptothecin
5	camptothecin derivative	295500	3	camptothecin
6	VM-26 (teniposide)	122819	4	podophyllotoxin
6	mitoxantrone	301739	4	anthracene
7	aphidicolin glycinate	303812	6	aphidicolin
8	tetraplatin	363812	1	platinum
8	carboxyphthalato-platinum	271674	1	platinum
8	acivicin	163501	5	amino acid analog
8	dichlorallyl lawsone	126771	5	napthoquinone
8	thioguanine	752	6	guanine
8	alpha-TGDR	71851	6	guanine
8	beta-TGDR	71261	6	guanine
8	inosine glycodialdehyde	118994	6	guanine
8	5-azacytidine	102816	5	cytidine

Table I (continued)

Cluster	*Name*	*NSC*	*MOA*	*Structural group*
8	cyanomorpholino-doxorubicin	357704	1	anthracene-daunorubicin
8	morpholinodoxo-rubicin	354646	3	anthracene–daunorubicin
8	*N,N*-dibenzyl daunomycin	268242	4	anthracene–daunorubicin
9	macbecin II	330500	6	lactone
9	rhizoxin	332598	2	macrolide
9	maytansine	153858	2	macrolactam
9	vinblastine sulfate	49842	2	vinca alkaloid
9	halichondrin B	609395	2	polyether macrolide
9	trityl cysteine	83265	2	triphenyl
9	bisantrene HCL	337766	4	anthracene
9	dolastatin 10	376128	2	modified peptide
10	L-alanosine	153353	5	aspartate analog
10	*N*-(phosphonoacetyl)-L-aspartate	224131	5	aspartate analog
10	5-fluorouracil	19893	5	uracil analog
10	brequinar	368390	5	folate analog
11	taxol	125973	2	taxane
11	taxol derivative	608832	2	taxane
12	colchicine derivative	33410	2	colchicine
12	allocolchicine	406042	2	colchicine
12	thiocolchicine	361792	2	colchicine
13	colchicine	757	2	colchicine
13	vincristine sulfate	67574	2	vinca alkaloid
14	methotrexate	740	5	folate analog
14	methotrexate derivative	174121	5	folate analog
15	L-ornithine	633713	5	folate analog
15	trimetrexate	352122	5	folate analog
16	thiopurine	755	6	purine
17	5-aza-2′-deoxycytidine	127716	6	cytidine
18	2′-deoxy-5-fluorouridine	27640	6	uridine
19	ara-C	63878	6	uridine
20	5,6-dihydro-5-azacytidine	264880	5	cytidine
21	pyrazofurin	143095	5	pyrazofurin
22	cyclocytidine	145668	6	cytidine
23	Baker's antifol soluble	139105	5	folate
24	an antifol	623017	5	folate analog
25	aminopterin derivative	184692	5	folate analog
25	aminopterin derivative	134033	5	folate analog
25	aminopterin derivative	132483	5	folate analog

the DNA major groove, and to cross-link DNA strands (Pratt *et al.*, 1994; Chabner and Longo, 1996). Cross-linked products are removed by an alkyltransferase DNA repair enzyme, via a repair mechanism known to be deficient in certain tumors. The first two members of this group are compounds bearing two or more aziridine or oxirane groups (296934 and 182986). These are analogs of the putative closed-ring intermediates of the nitrogen mustards, but are believed to be less reactive (Chabner and Longo, 1996). Three of the five platinum containing compounds are found next within this group (119875, 256927 and 241240). The next set of compounds in this group is composed of alkyl alkane sulfonates (329680, 102627, 750, 348948 and 338947). Busulfan (750) has been shown to attack the N-7 atom of guanine, but its ability to cross-link DNA is not certain. Pyrazoloimidazole (51143) and guanazole (1895) appear next, and are highly reactive DNA antimetabolites with nitrogen containing ring structures. The prodrug ftorafur (148958) appears next. The remaining members of Group 1 fall into two structural classes: the first composed of nitrosoureas, either alone or in combination with nitrogen mustards or guanidine groups (32065, 8806, 3088, 25154, 73754, 353451, 409962, 171112, 95441, 178248, 95466, 79037, 95678, 107392 and 167780), and the second composed of anthracyclines, anthracenediones and epipodophyllotoxins (308847, 142892, 366140, 349174, 355644, 164011, 123127, 82151, 267469 and 141540). The nitrosourea compounds bearing both chloroalkylating and carbamoylating (carbamoyl: –R-N–C=O) groups can produce interstrand cross-links in DNA by preferentially attacking the O-6 position on guanine. The greater antitumor activity of the compounds in the modified nitrosourea class, when compared with the parent nitrosourea, has been attributed partly to their greater lipophilic character (Chabner and Longo, 1996). The latter subclass of compounds in this group are doxorubicin analogs, thought to inhibit DNA topoisomerase II and protein kinase C mediated signal transduction pathways (Chabner and Longo, 1996). The structural similarity of these latter compounds originates in their anthracene scaffold. The various congeners in this group do not appear to effectively affect growth inhibitory behavior, since they all exhibit similar activity patterns in the SVD space when compared with the complete set of 122 compounds. Three of the compounds within the group of anthracyclines share a dimethyl or diethyl amine group (308847, 142892 and 366140). Amonifide (308847) is a topoisomerase II inhibitor that acts as a DNA intercalator or binder (Chabner and Longo, 1996), while pyrazoloacridine (366140) and hycanthone (142982) share an acridine moiety which may contribute to their similar activities.

The second group of compounds shares structural similarity with members of Group 1, but has SVD distance patterns different from the first group. Three of these compounds have aziridine or oxirane groups (6396, 9706 and

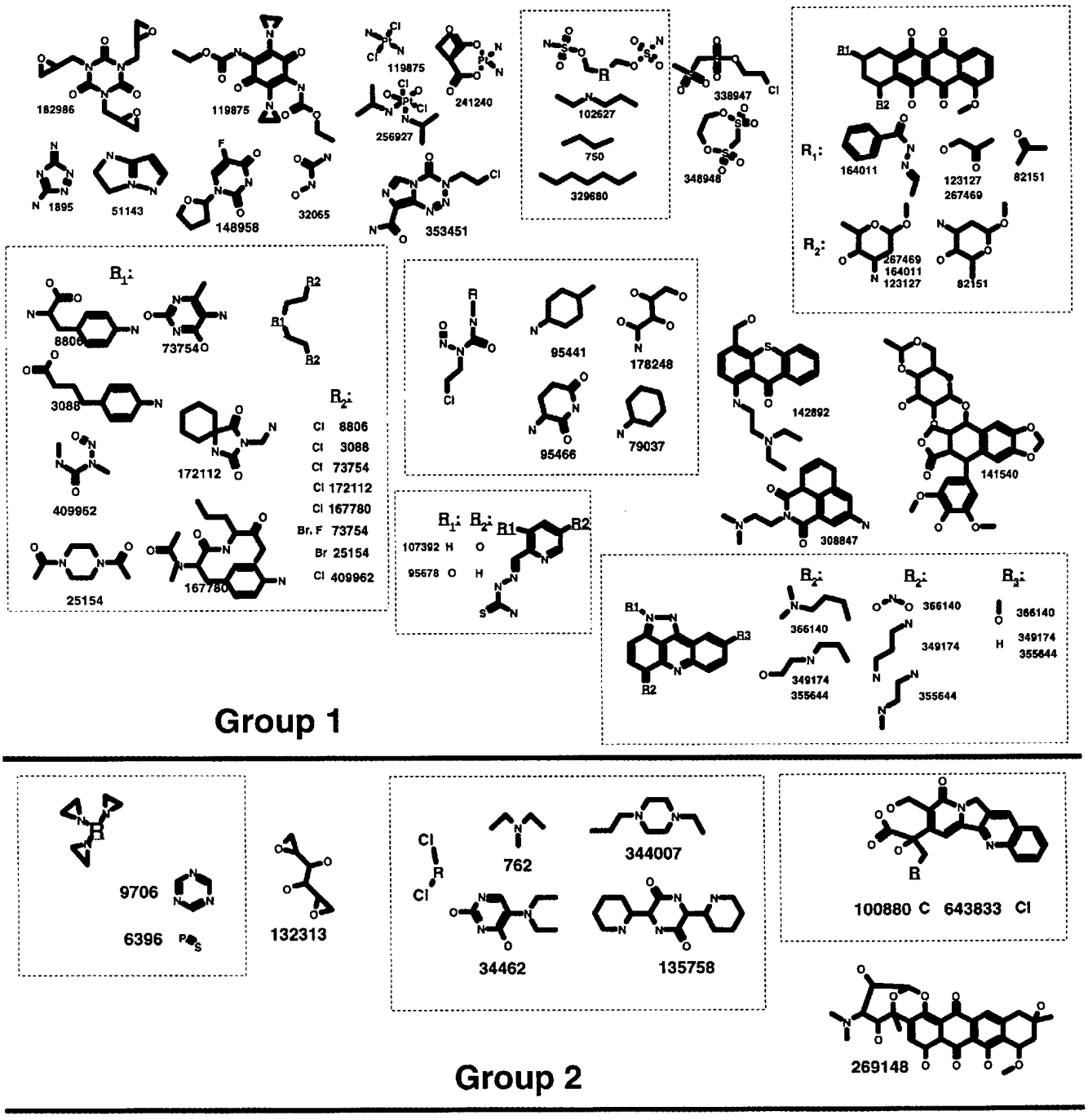


Figure 1

Two-dimensional representations of the chemical structures of the 122 compounds analyzed in this study. Compounds are ordered into 25 groups as described in the text. Structurally similar compounds are displayed together within each group. This figure has been prepared using the ISIS/DRAW software package.

132313), four compounds are nitrogen mustards (762, 34462 and 344007) and one is a doxorubicin analog (269148). The diepoxides in the oxirane, dianhydrogalactitol (132313), are presumably responsible for its antitumor activity. Also within this group are two camptothecin analogs (643833 and 100880) and piperazinedione (135758), two of these compounds exhibiting an alkylation capacity probably because of their chloride groups.

The third group (Group 3) comprises 16 compounds, including two mitomycins (26980 and 56410), the only known

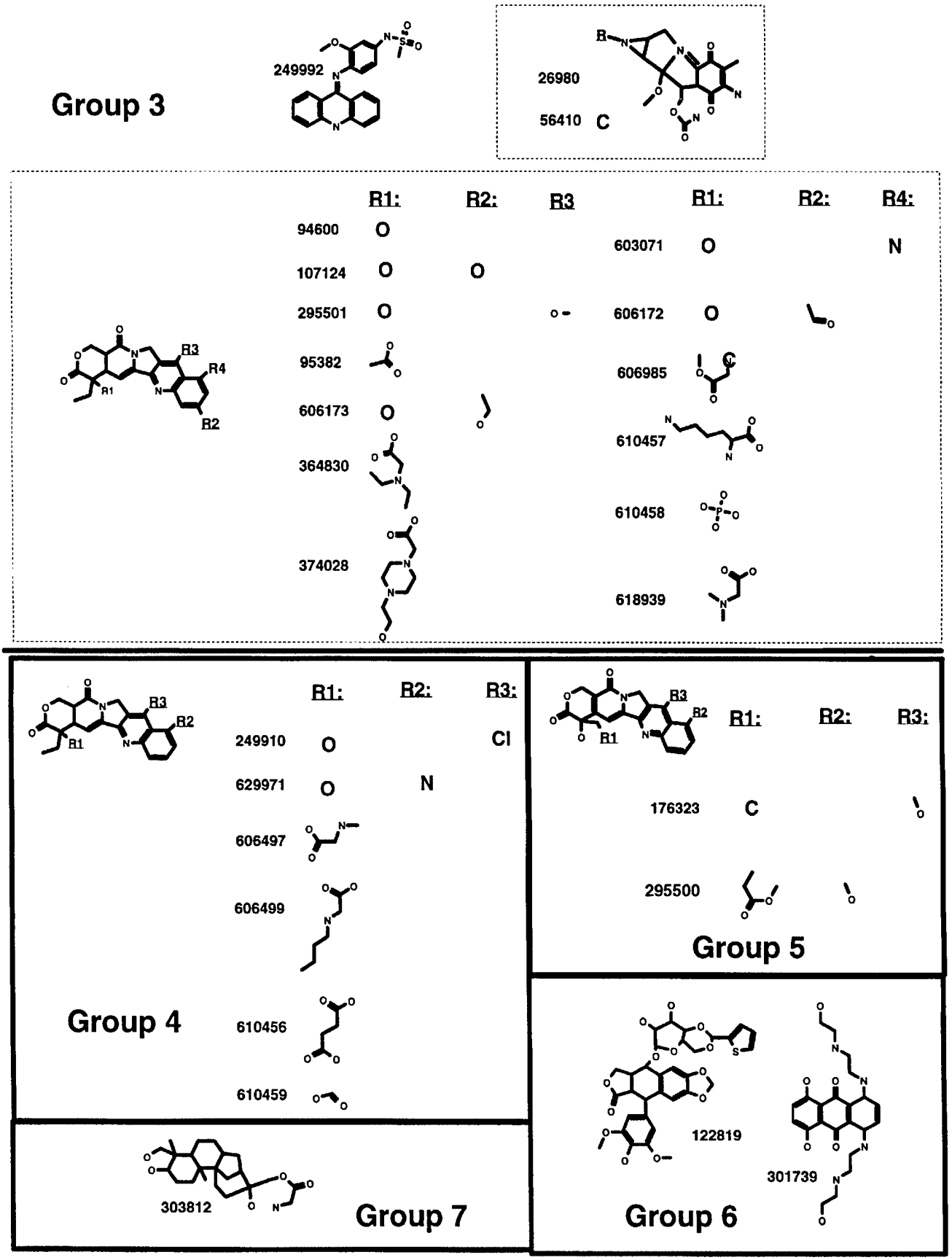


Figure 1(continued)

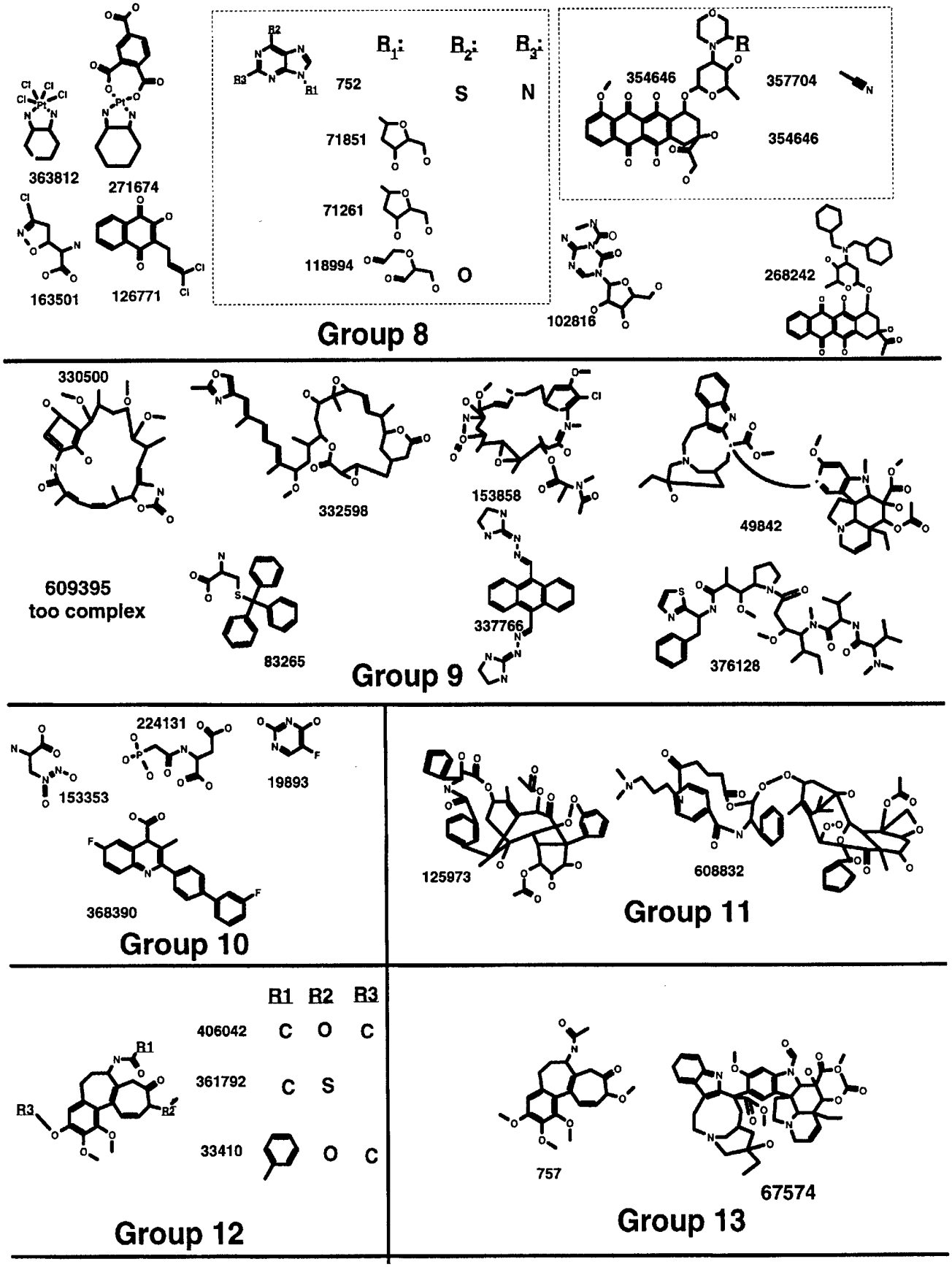


Figure 1 (continued)

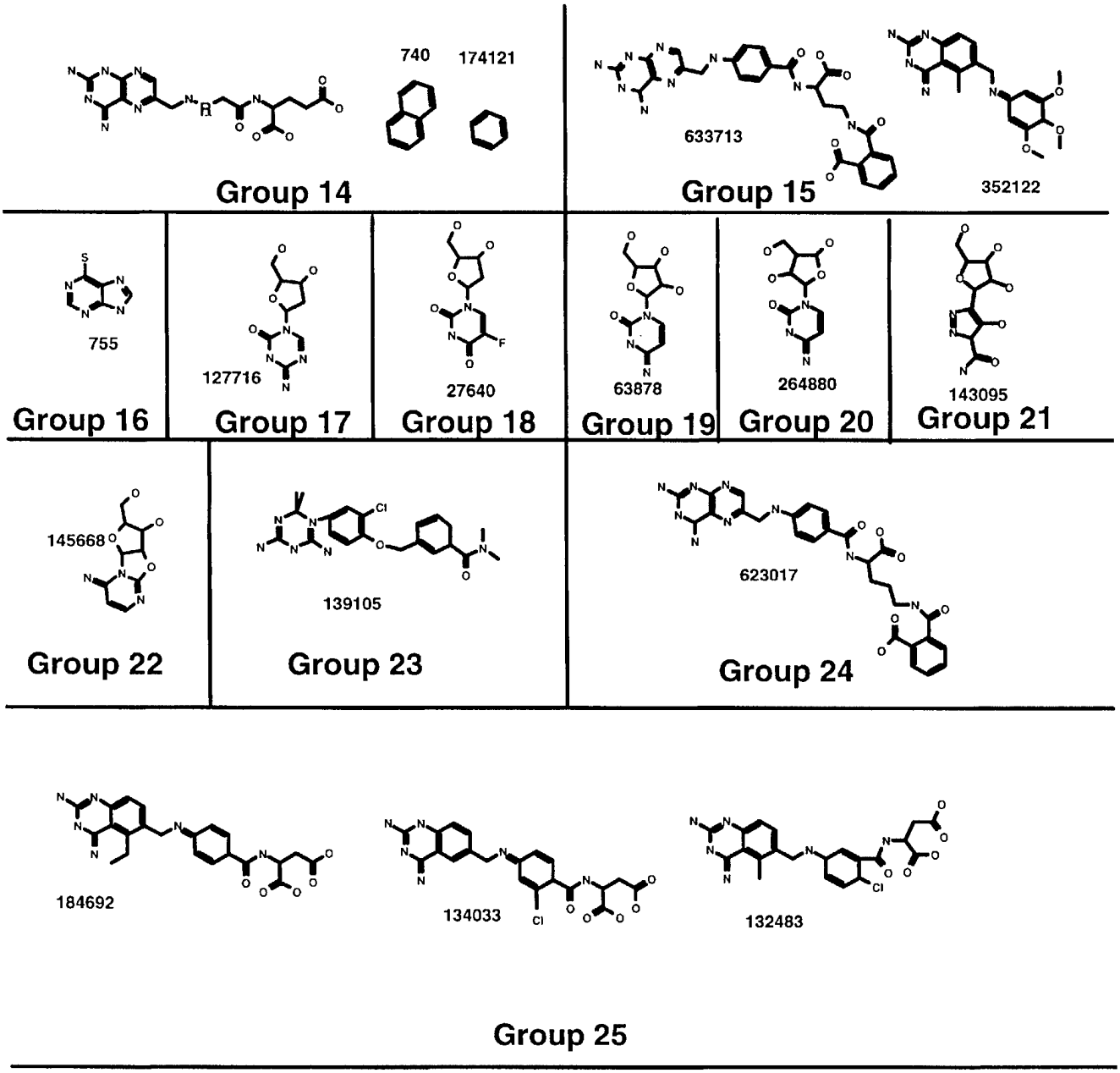


Figure 1(continued)

natural compounds containing an aziridine ring (Chabner and Longo, 1996). These compounds alkylate guanine at the N-2 position in the DNA minor groove (Chabner and Longo, 1996) and differ from one another only by a methyl group. With the exception of the topoisomerase II inhibitor 249992, the remaining compounds in this group are camptothecin analogs that are thought to inhibit the DNA gyrase enzyme topoisomerase I. The strong structural similarity within the camptothecin derivatives is thus also exhibited in their SVD distance patterns. Groups 4 and 5 consist of six and two camptothecin analogs, respectively. The cellular activities of the compounds in these two groups are sufficiently different from the larger set in Group 3 to include them as separate groups. The structural features responsible for this different activity are not clearly apparent. These compounds may exhibit similar activity patterns on the basis of solubility, or cell permeability.

Group 6 consists of only two compounds, the podophyllotoxin Teniposide (122819) and the topoisomerase II inhibitor 301379. Although both of these compounds share structural similarity and activity patterns with the alkylating compounds in Group 1, their location adjacent to the group of topoisomerase I agents suggests that their structural differences produce a distinctly different activity pattern.

Cluster 7 is a singlet, composed of aphidicolin glycinate (303812). Although this compound is thought to be a DNA polymerase inhibitor, it shares structural similarity with the camptothecin family, and its placement in a cluster near the camptothecin analogs in Groups 3, 4 and 5 suggests that its cellular activity may also mimic that of topoisomerase I inhibitors.

Twelve compounds are found in Group 8. Included in this set are the platinum containing, DNA intercalating compounds tetraplatin (363812) and carboxyphthalatoplatinum (271674). These compounds contain a stabilizing cyclohexane group that may contribute to their distinctive activity patterns when compared with the three platinum containing compounds in Group 1. Seven nucleoside analogs appear within this group (163501, 126771, 752, 71851, 71261, 118994 and 102816), most of which share a guanine or uracil moiety linked to a pentose. These compounds are thought to be directly incorporated into DNA (Myers *et al.*, 1997). The antibiotic acivicin (163501) and dichloroallyl-lawsone (126771) are thought to act as an inhibitor of pyrimidine biosynthesis, and their location within the family of nucleoside analogs is reasonable. The three doxorubicins that complete this group, morpholinodoxorubicin (354646), cyanomorpholino-doxorubicin (357704) and *N,N*-dibenzyl duanomycin (268242), share a unique hexopyranosyl moiety. The two platinum containing alkylating agents and the three doxorubicin analogs act by directly damaging DNA, while the remaining compounds in this group are inhibitors of nucleotide synthesis, acting as DNA/RNA antimetabolites.

The antitubulin agents are found to cluster into five groups. The first group (Group 9) is composed of six antitubulin agents (330500, 332598, 153858, 49842, 609395 and 376128), one topoisomerase II inhibitor (337766) and trityl cysteine (83265). The second group (Group 11) includes taxol (125973) and a taxol derivative (608832). The third and fourth groups (Groups 12 and 13, respectively) include the colchicines (757, 67574, 406042, 361792) and 33410. These compounds show weak pattern similarity to other anticancer agents, which suggests that these antitubulin agents share similar growth inhibitory mechanisms in the cell screen.

Group 10, which has an activity pattern that places it between the antitubulin Groups 9 and 11, consists of a nucleoside analog (19893), two amino acid analogs (153353 and 224131) and a folate analog (368390). Group 10 is the first cluster of compounds that lack close SVD distances to members of Groups 1–8. Thus its activity pattern lacks near SVD distances to groups containing alkylating agents and topoisomerase I and II inhibitors, with close SVD distances restricted mostly to members within its group. As will be shown later, this type of activity pattern may reflect agents that primarily act as inhibitors of nucleotide biosynthesis, rather than as DNA damaging agents.

An equally distinct activity pattern is also found for the antifolate compounds composing Groups 14 and 15. Group 14 consists of methotrexate (740) and the folate analog (174121), while Group 15 includes the antimetabolites 633713 and 352122. It should be noted that in general, clustering of compounds in this subgroup is based largely on their SVD distance dissimilarities, rather than similarities, to the other members in the set of 122 compounds.

Groups 16–22 all comprise single compounds, all of which are nucleosides that act as antimetabolites of nucleotide biosynthesis. As with the folate analogs discussed above, their activity patterns are sufficiently unique for these compounds to share no pattern similarities with any of the standard 122 agents.

Folate analogs complete the final three groups. Groups 23 and 24 consist of single compounds (139105 and 623017, respectively), while Group 25 consists of three folates (184692, 134033 and 132483). These latter RNA–DNA antimetabolites have alcohols or ethers substituted at positions C-7 or C-11 of the parent compound that may contribute to their increased water solubility and unique activity pattern.

The results described here are consistent with earlier classifications by Koutsoukos *et al.* (1994) and van Osdol *et al.* (1994) that divided these compounds into two large clusters. Our analysis finds a similar division of compounds, while providing further subclustering of compounds within these two major divisions. The largest division consists of compounds with the most similar activity patterns, compounds which appear at the top of Table I, comprised primarily of DNA-damaging agents (Groups 1–8). Compounds in the lower portion of Table I comprise the second major division and act by targeting a biosynthetic pathway or part of the mitotic machinery.

Each of the groups described above can be further examined for their average activity patterns across the 60 tumor cell lines. Figure 2 displays the mean activity for the 25 different groups across all 60 tumor cell lines. These results provide an indication of the diversity of activity patterns associated with the 25 clusters identified above, and can be used to identify which groups of compounds are more or less active against individual cell lines or within panels of cells. The results in Figure 2 are displayed according to the cluster order in Table I, from Group 1 to Group 25. The average sensitivity of the 60 tumor cells against the compounds within each cluster is indicated by color. Tumor cells with progressively more sensitive activity patterns when compared with their group averages are shown in yellow to orange to red. Cells with progressively less sensitivity are shown from pale blue to dark blue. Cells with activity patterns near their group averages are shown in light green.

Examination of the mean activity patterns for the 25 clusters obtained from the cubic clustering algorithm in the SAS Technical Report (SAS, 1992) can be used to qualitatively assess differences between each group. The agents within

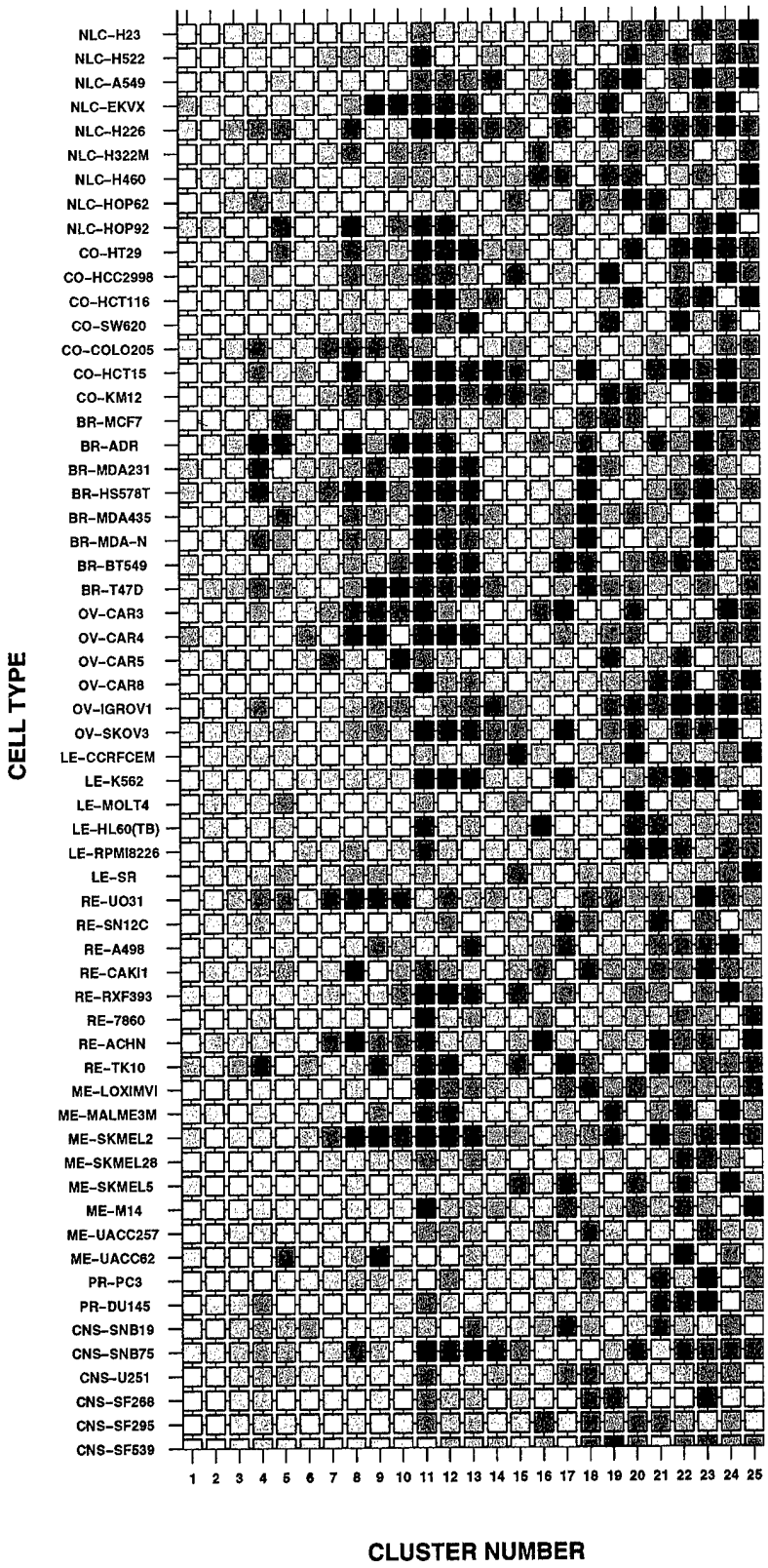


Figure 2

Average activity across the 60 cell lines for compounds in each of the 25 groups. Panels of cells are ordered from bottom to top as follows: CNS, PROSTATE, MELANOMA RENAL, LEUKEMIA, OVARIAN, BREAST, COLON and NLC. Groups with a positive mean activity pattern are displayed from least, to intermediate, to greatest, in orange, red and brown, respectively. Groups with negative mean activity patterns are shown, from least to greatest, in light blue, blue and dark blue, respectively. Groups with mean activity patterns near zero are shown in green.

Groups 1–3 exhibit a uniformly weak mean activity pattern across all 60 cell types, as indicated by the near-baseline light green color for all cell types. Groups 4 and 5 begin to exhibit a more diverse activity pattern, with a greater sensitivity (orange color) to the panel of CNS cells, as well as selected RENAL, LEUKEMIA and BREAST cells. Group 4 is composed of five camptothecin analogs that have an apparent, albeit weak, selectivity for the CNS panel of cells, with a strong activity against the single BREAST-ADR cell line. Groups 6 and 7 also have a relatively uniform activity pattern with the exception of an insensitivity to RENAL-ACHN, RENAL-UO-31 and OVARIAN-OVCAR5.

Group 8 has a diverse activity pattern with high sensitivity to MEL-SKMEL2, BREAST-HS578T, COLON-COLO205, COLON-HCC2998, COLON-HT29 and with low sensitivity to RENAL-ACHN, RENAL-CAKI-1, RENAL-UO-31, OVARIAN-OVCAR4 and BREAST-ADR. Groups 9–13, the antitubulin active agents, display high sensitivity to most of the COLON tumor cells, and a variable sensitivity to BREAST and MELANOMA tumor cells. Groups 14–16 showed a low sensitivity within the BREAST panel and variable sensitivity to cells within the COLON panel. Group 17 displays a consistent sensitivity against most of the cells within the BREAST and COLON panels. The single compound in Group 18 is uniformly sensitive to the BREAST panel, while Groups 19–25 exhibited a widely diverse range of activity patterns, with both sensitive and insensitive cellular activity patterns. Cells with the least sensitivity to the 122 standard agents are: NLC-EKVX, NLC-H226, BREAST-T47D, -HS578T and -MDA231, OVARIAN-OVCAR4, RENAL-RXF393 and CNS-SNB75.

Our analysis can be used to cluster members of the 60 cell panel according to their response to the 122 standard anticancer agents. In contrast to the previous analysis, where 122 agents were examined for their activity pattern across the 60 cell lines, a similar analysis can be performed whereby the 60 cell lines are examined for their activities against the 122 standard agents. Clustering of the cell types on this basis can be used to identify each cell type's differential response to these standard anticancer agents. Fifteen clusters are obtained using the cubic clustering analysis (CCC) within the SAS Technical report. Figure 3 displays a cladogram for clusters obtained in this analysis, with each branch labeled and color coded according to cell type. Cells are initially separated into two major branches, with one branch consisting of 15 cell types, the remaining 45 cell types appearing in the other major branch.

The smaller of the two major branches appears at the rightmost portion of Figure 3, and is subdivided into four clusters. The largest of these four clusters consist of RENAL cell types, with UO-31, 786-0, ACHN, CAKI-1 and RXF-393 along with two MELANOMA cell lines, LOX-IMVI and M14. Four of the five RENAL cells in this panel are know to exhibit multidrug resistance (MDR). MDR is a resistance modulator for many chemotherapeutic agents associated with either an increased expression of the P-170 membrane glycoprotein MDR1 or the presence of the multidrug resistance protein (Lee *et al.*, 1994; Alvarez *et al.*, 1995). Both of these mechanisms act by lowering the effective drug concentration, enhancing drug efflux (Chabner and Longo, 1996) and reducing drug efficacy. The remaining three sub-branches within this major branch are composed of four LEUKEMIA, two NLC, one CNS and one MELANOMA cell type. The LEUKEMIA cell line has the greatest average sensitivity in mean deviation ($\Delta x = [\log GI_{50}] - \langle \log GI_{50} \rangle$) for the 122 standard agents. The LEUKEMIA cell type SR appears as a singlet, thus having no comparable cell type with a similar response to the 122 standard agents.

The larger of the two major branches found in this analysis is clustered into four sub-branches, which are further divided into 17 branches. The leftmost sub-branch (as viewed in Figure 3) is divided into seven clusters. The largest cluster in this group consists of seven cell types, appearing as the leftmost branch of the cladogram. This cluster includes three OVARIAN, two NLC and one MELANOMA cell type. Adjacent to this cluster are four branches comprising only a single cell type: (RE)SN12C, (CNS)SF-268, (BR)BT-549 and (ME)MALME-3M. Two BREAST cell types (T-47D and MCF7) along with the LEUKEMIA cell line RPMI-8226 appear in the next cluster. Membership in this leftmost sub-branch is completed by a cluster comprising only two OVARIAN cell types (SK-OV-3 and OVCAR-8) and the singlet (NLC)HOP-92. The remaining clusters in this major sub-branch consist primarily of NLC, COLON, BREAST and MELANOMA cell types. Within the clusters formed by these cell types, a clear separation according to these panels is not apparent based on their response to the 122 standard agents. An apparent coherence between the COLON, BREAST and LEUKEMIA panels is clearly indicated; however, the basis for this clustering is not evident. These results indicate that many tumor cell types, both within and between different panels, exhibit similar sensitivities to the set of 122 compounds studied here. Additional studies with a larger set of test compounds will be needed to more thoroughly determine which cell types share the most similar response patterns.

Prediction of MOAs

Mechanism of action classifications can be based on applications of a wide range of statistical tools (Harary, 1971; Golub and Loan, 1989; Berry *et al.*, 1995). The results in Table I show that there is a substantial similarity between the clusters of compounds based on GI_{50} activity patterns and their classification based on their previously assigned MOAs. Yet, subclusters interspersed between clusters of a given MOA are observable, which call for a more systematic

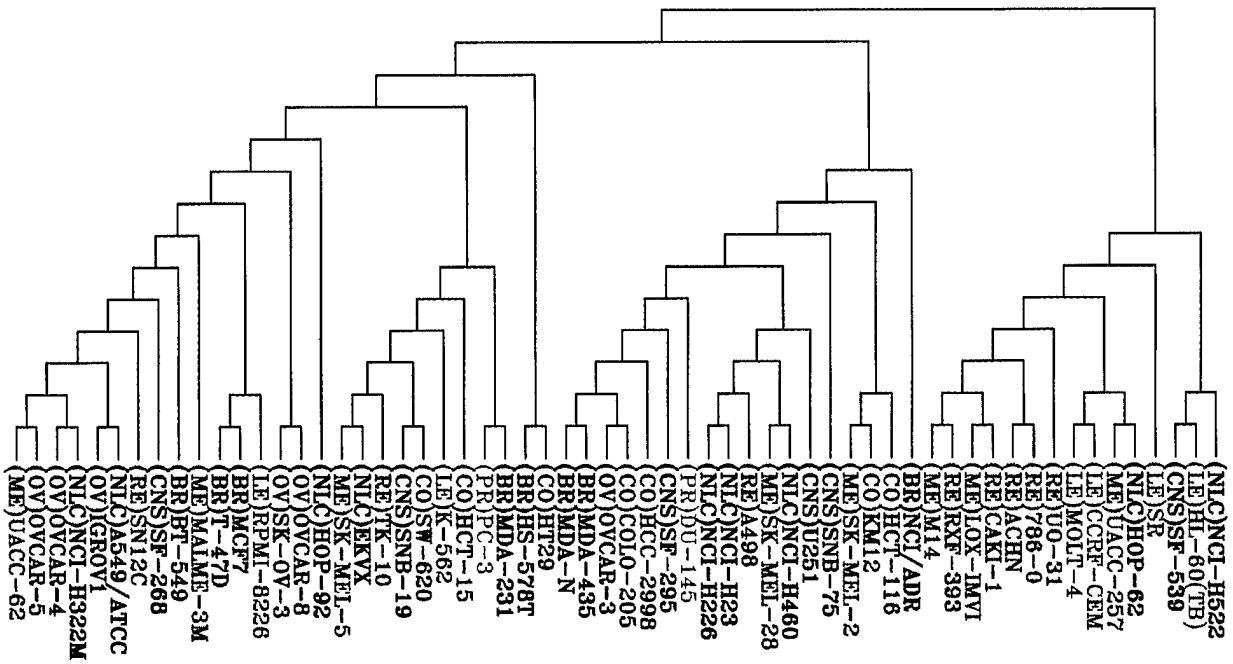


Figure 3

Cladogram of SVD distances for the 60 cell types determined from the activity data for the standard 122 anticancer agents. Branch labels are colored according to cell panels: black, non-small cell lung carcinoma (NLC); light green, COLON; magenta, LEUKEMIA; red, OVARIAN; dark green, RENAL; brown, MELANOMA; light blue, PROSTATE; black, CNS. (Note that the color black has been used for both NLC and CNS.) The abbreviations for each panel also appear in the label for each branch. The GROWTREE utility from the GCG software package has been used to generate this figure. Cluster assignments, from left to right, are as follows: Cluster 1: (ME)UACC-62, (OV)OVCAR-5, (OV)OVCAR-4, (NLC)NCI-H322M, (OV)IVGROV1, (NLC)A549/ATCC, (RE)SN12C. Cluster 2: (CNS)SF-268. Cluster 3: (BR)BT-549. Cluster 4: (ME)MALME-3M. Cluster 5: (BR)T-47D, (BR)MCF7, (LE)RPMI-8226. Cluster 6: (OV)SK-OV-3, (OV)OVCAR-8. Cluster 7: (NLC)HOP-92. Cluster 8: (ME)SK-MEL-5, (NLC)EKVX, (RE)TK-10, (CNS)SNB-19, (CO)SW-620, (LE)K-562. Cluster 9: (CO)HCT-15. Cluster 10: (PR)PC-3, (BR)MDA-231. Cluster 11: (BR)HS-578T, (CO)HT29. Cluster 12: (BR)MDA-N, (BR)MDA-435, (OV)OVCAR-3, (CO)COLO-205, (CO)HCC-2998, (CNS)SF-295. Cluster 13: (PR)DU-145. Cluster 14: (NLC)NCI-H226, (NLC)NCI-H23, (RE)A498, (ME)SK-MEL-28, (NLC)NCI-H460, (CNS)U251. Cluster 15: (CNS)SNB-75. Cluster 16: (ME)SK-MEL-2, (CO)KM12, (CO)HCT-116. Cluster 17: (BR)NCI/ADR. Cluster 18: (ME)M14, (RE)RXF-393, (MEL)LOX-IMVI, (RE)CAKI-1, (RE)ACHN, (RE)786-0, (RE)UO-31. Cluster 19: (LE)-MOLT-4, (LE)CCRF-CEM, (ME)UACC-257, (NLC)HOP-62. Cluster 20: (LE)SR. Cluster 21: (CNS)SF-539, (LE)HL-60(TB), (NLC)NCI-H522.

analysis of the degree of correlation between the GI_{50} data and MOAs. To this aim we performed the following analysis: mean activity fluctuation vectors in the SVD space were found for each of the six MOAs using

$$<\mathbf{a}^*>_{\mathrm{MOA}} = \Sigma_i \mathbf{a}_i^* / N_{\mathrm{MOA}} \qquad (2)$$

Here N_{MOA} is the number of agents exhibiting a given MOA, and the summation is performed over this particular subset of agents. The average activity patterns are thus obtained for each MOA. The departure of the behavior $\mathbf{a}_i^*$ of individual agents from these averages are examined for an assessment of the accuracy of the MOAs assigned to the different agents. The deviation of each drug from the mean activity fluctuation vector for the six MOA classes is thus

$$\Delta\mathbf{a}_i^*{}_{\mathrm{MOA}} = \mathbf{a}_i^* - <\mathbf{a}^*>_{\mathrm{MOA}} \qquad (3)$$

The smallest of the six distances obtained for each drug is used to identify its most likely MOA. Application of this test to all compounds in the training set of 122 standard agents shows that the correct MOAs are assigned with an average accuracy level of 96.7%. Column 2 in Table II summarizes the results for the six different MAOs. Weinstein *et al.* (1992) obtained an accuracy level of 91.5% by using a neural network model and 85.8% by linear discriminant analysis.

The accuracy of the MOA assignments for anticancer agents has additionally been examined by jack-knife tests. The jack-knife test, also called the leave-one-out test (Mardia *et al.*, 1979), is a method often utilized for small samples which cannot be divided into training and testing sets without

Table II

Performance of SVD analysis for determining MOA*

MOA[a]	*Success %*	
	Training set	*Prediction set*
1 (35)	97	97
2 (13)	92	85
3 (24)	96	96
4 (15)	100	87
5 (19)	100	63
6 (16)	94	63
Mean (122)	96.7	84.4

Each % success represents the correctly predicted compounds for each MOA [e.g. all 15 of the topoisomerase II inhibitors were predicted correctly in MOA class 4 for the training set, while 87% (n = 13 of 15) of these agents were correctly predicted in the jack-knife procedure].
[a]1, alkylating; 2, antimitotic; 3, topoisomerase I inhibitors; 4, topoisomerase II inhibitors; 5, RNA–DNA antimetabolites; 6, DNA antimetabolites.

Table III

MOA classification for incorrectly predicted MOAs

NSC no.	*Name*	*Assigned MOA*	*Predicted MOA*
357704	cyanomorpholinodoxorubicin	1	3
153858	maytansine	2	6
67574	vincristine sulfate	2	6
354646	morpholinodoxorubicin	3	4
268242	*N,N*-dibenzyl daunomycin	4	1
366140	pyrazoloacridine	4	1
148958	Ftorafur	5	6
102816	5-azacytidine	5	4
264880	5,6-dihydro-5-azacytidine	5	1
174121	methotrexate derivative	5	6
139105	Baker's soluble antifol	5	2
132483	aminopterin derivative	5	3
623017	an antifol	5	6
63878	ara-C	6	1
27640	2′-deoxy-5-fluorouridine	6	1
127716	5-aza-2′-deoxycytidine	6	4
330500	Macbecin II	6	1
95678	3-HP	6	1
32065	hydroxyurea	6	1

loss of information. In this procedure each compound to be tested is removed from the training data set and the identification of the activity fluctuation $\Delta\mathbf{a}_i{}^*{}_{MOA}$ for each MOA is carried out using the GI_{50} data of the remaining 121 drugs. The most probable MOA of the test compound is then predicted using the same distance criteria (equation 3), with the basic difference that the mean fluctuation vectors $<\mathbf{a}^*>_{MOA}$ are now extracted from a set of data excluding the test compound. The average accuracy level reached by this method was 84.4%. A summary of these results is presented in the third column of Table II. The mispredicted compounds and their predicted MOAs are listed in Table III. Most of the 19 mispredicted compounds were classified as topoisomerase II agents or DNA–RNA antimetabolites, with the majority of these agents predicted to behave as alkylators. Since topoisomerases act to create covalent damage in DNA, their functional activity may be similar to alkylating agents.

Discussion

NCI's 60 cell line screening assay provides a measure of growth inhibition for human cancer cells exposed to candidate anticancer compounds. Activity data accumulated in these screens can be used to group agents that exhibit similar activity patterns across a broad variety of tumor cell lines. Compounds grouped according to pattern similarities can be further examined for possible relationships between their activities, their chemical substructures and/or their MOAs. The results presented here apply the standard statistical method of SVD to the $\log(GI_{50})$ data to define measures of distances between compounds in a space that best distinguishes their similarities and dissimilarities. Hierarchical clustering of these SVD-derived distances divides these 122 compounds into 25 groups. The first eight groups are predominantly formed by DNA-damaging agents, while the latter 17 groups (9–25) mostly consist of agents that inhibit nucleic acid biosynthesis or mitosis. Compounds in the first class comprise MOAs assigned as alkylators, and inhibitors of topoisomerases I and II, along with a few DNA antimetabolites, while the latter class is dominated by antimitotic agents and antimetabolites.

DNA damaging agents (Groups 1–8), when observed together, exhibit strongly similar activity patterns. Agents such as DNA alkylators and DNA metalators (platinum agents) are equally effective against slowly dividing or non-dividing cells (termed G_0 cells). Since strong pattern similarities are observed among alkylators and platinum analogs, it is reasonable to conclude that these compounds have comparable activities against all cell types, as evidenced by the uniform activity pattern for these groups. Thus compounds that act directly on DNA, either by cross-linking or less directly by inhibiting enzymes responsible for processing DNA (i.e. unwinding), fall into this first group. While alkylating agents would be expected to be included in the class of DNA-damaging agents, the present finding that topoisomerase inhibitors behave similarly to alkylating agents is unexpected. However, inhibition of topoisomerases result in DNA damage, with repair modulated by the impact of the damage. Earlier studies have found that some topoisomerases are constitutively expressed at relatively constant levels throughout the cell cycle, even in cells that are not actively dividing (Hwang *et al.*, 1989). Thus inhibitors of topoisomerases may potentially be active in tumors that have low growth fractions (Chabner and Longo, 1996) and as a result exhibit cytotoxic behavior similar to alkylating agents.

The second major class of compounds identified in our analysis acts against the enzymatic machinery required for cell division. Most of these compounds inhibit purine or

pyrimidine biosynthesis or act as antitubulin agents. Evidence to support this claim can be found in the crystallographic complexes between biosynthetic enzymes and ligands that are either identical to those included in the set of 122 compounds or close structural analogs. Although it is not our intention here to present a systematic analysis of structural data in support of this claim, the Appendix summarizes our survey of the crystallographic database of proteins complexed with ligands that bear strong structural similarity to many of the antimetabolite agents in the set of 122 compounds.

A strong correspondence was not observed between specific MOAs of compounds assigned to each cluster. For example, alkylating agents and topoisomerase I and II inhibitors appear in most of the first eight clusters. The results of this analysis are, however, sufficiently meaningful to yield an MOA prediction accuracy of >84%. Inspection of the subclusters obtained from this analysis finds compounds that both share and lack structural similarity.

Many approaches are available for classification of compounds by chemical structure (Johnson and Maggiora, 1990; Martin and Willet, 1998). Some approaches are based on one-dimensional (1-D) global features such as polarizability, molecular weight and number of hydrogen bond donors/acceptors (Shemetulskis *et al.*, 1995; Cummins *et al.*, 1996). Alternative approaches attempt to maximize a selection of 2-D and 3-D indices assigned to each compound (Good and Lewis, 1997; Lewis *et al.*, 1997; Weininger *et al.*, 1997). Some of the more commonly used descriptors are based on chemical formula (Weininger *et al.*, 1997), 2-D topological similarity (Burden, 1989; Brown and Martin, 1996; Randic, 1997; Pearlman *et al.*, 1998) and 3-D superposition (Miller, 1995). Using sets of indices representative of these descriptors, compounds can be assigned a 'fingerprint' which can be used for assessing similarities within groups of compounds (Gillet and Smith, 1998). Clusters of the 122 compounds examined here, based on a set of 54 1-D descriptors available in the Cerius package and based on 2-D SMILES descriptors, found no statistically significant correlation with the activity patterns from the screening assay. Taken separately or together, no combination of these 1-D or 2-D descriptors could be found to produce a statistically significant correlation with the activity patterns observed for the 122 agents examined here. Although examination of Figure 1 provides clear evidence that many compounds within each group have common substructural features, a systematic means of assigning the compounds to these groups, on the basis of 1-D and 2-D descriptors alone, was not apparent. These results are consistent with widespread observations such as those of Brown and Martin (1996), where small chemical modifications can result in quite different biological responses. The family of camptothecins offers a clear example of such behavior, i.e. small differences in the parent structure resulted in quite different activity patterns. Our results emphasize the importance of assessing structural information together with screening data to assess biological activity.

One important question arises about studies such as that presented here: what is the effect of data errors on the results? Single compounds, such as those clustered in Groups 16–24 above, are easily distinguished in this type of analysis. Hierarchical clustering of SVD distances alone identifies these singlets on the basis of their position in a separate branch of the tree. The additional classification based on pairwise differences in SVD distances with respect to the whole set of compounds can be further used to determine whether compounds isolated in a single branch of the tree have an important different activity pattern or lack any such feature.

Measurement errors that appear in the reported $\log(GI_{50})$ values represent another type of error. These errors result from experimental conditions as well as errors in data reporting. In an attempt to address the importance of these types of errors on our results, the current data set was perturbed with random noise and the SVD distances were recalculated. Figure 4 displays the results of perturbing the current set of $\log(GI_{50})$ values by an error that ranges from zero to 40%. The ordinate in Figure 4 represents the correlation coefficient (Snedecor and Cochran, 1980) between the matrix of SVD distances calculated for the unperturbed and perturbed data sets. There we see that perturbing the existing data with 20% error yields an SVD distance matrix whose entries are still correlated with the original data with a correlation coefficient of 0.9. By contrast, a 40% error produces a correlation coefficient near 0.7. From this analysis we believe that data error in the range of 10–20% should yield results extremely similar to those reported here. The actual error in these data is difficult to establish. An estimate of the maximum error can be obtained by calculating the coefficient of variation [C.V. = $\sigma/\ \overline{\log(GI_{50})}$] for the $\log(GI_{50})$ values obtained for each compound. The variance (σ) is estimated therein as the squared sum of x_{ij} calculated in equation (3). This method yields a coefficient of variation of 0.87 (or a percentage error of 13%), which according to Figure 4 corresponds to a correlation coefficient of 0.95. We conclude that the results of our analysis are robust enough to sustain errors lower than 15% without significant degradation. The experimental data used in our study include results from multiple replicate analysis performed between two to 50 replicates, which would reduce the measurement noise.

Based on the above observation that selected cell types could be clustered according to their response to the 122 standard agents, we explored whether differences in SVD distance clusters would occur from analyses based on subsets of selected cell types that are known to exhibit MDR. Based on the relative expression of MDR1 mRNA and the immunocytochemical characterization of P-glycoprotein expression (Wu *et al.*, 1992) eight MDR1 expressing cell types are

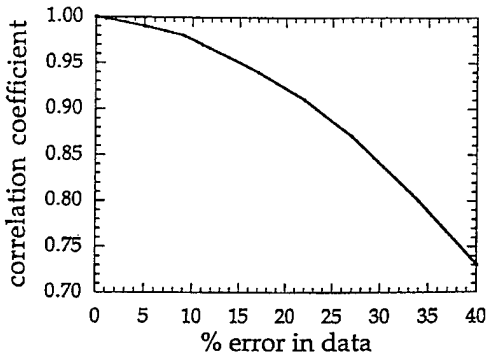


Figure 4

Sensitivity analysis of present SVD results. Correlation coefficients between the results found from SVD derived distances based on original log(GI_{50}) data, and those based on the randomly perturbed log(GI_{50}) data. The ordinate represents the percentage error introduced upon perturbation of the original data set.

identified: HCT-15(CO), SF-295(CNS), HOP-62(NLC), UO-31(RE), A498(RE), ACHN(RE), CAKI-1(RE) and RXF-393(RE). This selection conforms most closely to those cells exhibiting the highest rhodamine efflux measurements as posted on the Developmental Therapeutics' web page (http://dtp.nci.nih.gov). Clustering analysis was performed using (*i*) the log(GI_{50}) values from the eight MDR1 expressing cell lines and (ii) the log(GI_{50}) values from the 52 non-MDR1 expressing cell lines in the screen. The latter analysis clustered compounds in a qualitatively similar way to that obtained for the complete set of 60 cell lines. The analysis performed on the eight MDR1 expressing cells found that the activity patterns within this group had similar SVD distances, and their activity pattern with respect to their response to the 122 standard agents was quite similar to that found for the previously classified DNA-damaging agents. In particular, the antitubulin agents found in Groups 9, 11, 12 and 13 exhibit SVD distances that are similar to the members of the DNA damaging agents in Groups 1–8. In addition to this subset of antimitotic agents, the antimetabolites found in Groups 14–25 also display SVD distance patterns that reflect patterns closely resembling that of the DNA damaging agents. This result is consistent with the view that MDR is associated with the increased efflux of etoposides, anthracyclines (topoisomerase II inhibitors), colchicines and vinca alkaloids (antimitotic agents) (Pratt *et al.*, 1994; Chabner and Longo, 1996), and also demonstrates that agents that inhibit nucleotide biosynthesis are also affected. The result of multi drug resistance is a more uniform activity pattern across all cell panels, a feature characteristic of DNA damaging agents.

The results presented herein can be contrasted with those available from the web-accessible program COMPARE. The SVD distances, used in our procedure, and the PCCs, used in COMPARE, both represent measures of similarity between activity patterns in the tumor cell screen. A calculation of the correlation coefficient between these two measures is statistically significant ($r = 0.51$, $P < 0.001$). A scatter plot of PCC versus SVD distances finds the correlation to be strongest for the high values of PCC (PCC > 0.75) and low SVD distances. Consistent with this observation, compounds with high PCC values also appear in our SVD-derived cluster sets. As the PCC values become lower and SVD distances become greater, their correlation becomes weaker, albeit statistically significant. The major difference between the two methods involves identification of cluster membership. The CCC clustering criterion used in our analysis grouped these standard agents into 25 distinct clusters. The COMPARE program generates a PCC for a selected 'seed' compound. Since a PCC above 0.38 is statistically significant ($P < 0.05$, $n = 59$), compounds with higher PCCs would be included as neighbors of this 'seed'. Constructing clusters according to this procedure often yields many compounds. As an example, a COMPARE analysis based on a 'seed' selected from compounds in Groups 1–6 from our analysis found statistically significant 'hits' for over half of the 122 standard agents, many of which were found to have large SVD distances. Instances where statistically significant PCC values corresponded to near SVD distances were observed for compounds in Groups 8, 10, 11 and 12 and the single compounds in Groups 14–24. The agreement between cluster membership for the two approaches becomes increasingly better when selection is based on higher PCC values. In support of this observation, the correlation between PCC values above 0.75 and their SVD distances is 0.72 ($P < 0.0001$). Our application of the SVD approach is based on its documented performance in the analysis of systems with data corrupted by noise. While it is not our intention here to produce a detailed comparison of these two methods, it is clear that compounds with the highest pattern similarities will be found by both methods. However, in circumstances where these patterns are less similar, each approach can be expected to yield varying degrees of agreement.

In summary, statistical clustering tools have been used to analyze the growth inhibitory potency data available from the NCI's 60 tumor cell line screen. Analysis of the results for 122 standard anticancer agents finds that this set of compounds can be clustered according to screening patterns into 25 groups, with eight of these groups consisting of DNA damaging agents and the remaining groups consisting of agents that act to inhibit either nucleotide biosynthesis or mitosis. Structural similarities are found between compounds assigned to these two broad categories. Clustering of the cell types based on their response to the 122 standard agents divided the cells into two major branches which were further subdivided into 21 groups. Strongest within-panel responses were found for the RENAL, OVARIAN and LEUKEMIA panels. The current analysis provides a reference for evaluating larger data sets of compounds for similarities in their

screening patterns with respect to the standard 122 anticancer agents. Analyses of these larger data sets may be able to relate more precisely chemical substructure to activity.

Acknowledgements

The authors are grateful for discussions with Drs Anne Monks, Dominic Scuderio and Timothy G. Myers about the cell screening data. I.B. gratefully acknowledges partial support from NATO-CRG 951240. We would also like to acknowledge the TUBITAK fellowship of O. Keskin, the IRSP program at SAIC Frederick and Grant DAMD17-98-1-8323 from the US Army. The contents of this publication does not neces-sarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products or organizations imply endorsement by the US Government.

References

Ajay A., Walters W.P., Murcko M.A. (1998) Can we learn to distinguish between drug-like and nondrug-like molecules? *Journal of Medicinal Chemistry*, **41**, 3314.

Alvarez M., Paull K., Monks A., Hose C., Lee J.S., Weinstein J., Grever M., Bates S., Fojo T. (1995) Generation of a drug resistance profile by quantitation of MDR-1/P-glycoprotein in the cell lines of the National Cancer Institute Anticancer Drug Screen. *Journal of Chemical Investigation*, **95**, 2205.

Bahar I., Wallqvist A., Covell D.G., Jernigan R.L. (1998) Correlation between native state hydrogen exchange data and cooperative residue fluctuations from a simple model. *Biochemistry*, **37**, 1067.

Bellenson J. (1998) Integrating information technology and drug discovery processes. *Nature Biotechnology*, **16**, 597.

Benton D. (1998) Integrated access to genomic and other bioinformation: an essential ingredient of the drug discovery process. *SAR and QSAR in Environmental Research*, **8**, 121.

Bernstein F.C., Koetzle T.F., Williams G.J., Meyer E.E. Jr, Brice M.D., Rogers J.R., Kennard O., Shimanouchi T., Tasumi M. (1977) The Protein Data Bank: a computer based archival file for macromolecular structures. *Journal of Molecular Biology*, **112**, 535.

Berry M.W., Dumais S.T., O'Brien G.W. (1995) Using linear algebra for intelligent information retrieval. *SIAM Review*, **37**, 573.

Botstein D., Cherry J.M. (1997) Molecular linguistics: extracting information from gene and protein sequences. *Proceedings of the National Academy of Sciences, USA*, **94**, 5506.

Boyd M.R. (1995) *The NCI* In Vitro *Anticancer Drug Discovery Screen*. Humana Press: Totwa, NY.

Boyd M., Paull K.D. (1995) Some practical considerations and applications of the National Cancer Institute *In Vitro* Anticancer Drug Discovery Screen. *Drug Discovery Research*, **34**, 91.

Brown R.D., Martin Y.C. (1996) The information content of 2D and 3D structural descriptors relevant to ligand–receptor binding. *Journal of Chemical Information and Computer Sciences*, **37**, 1.

Burden F.R. (1989) Molecular identification number for substructure searches. *Journal of Chemical Information and Computer Sciences*, **29**, 225.

Castell J.V., Gomes-Lechon M.J. (eds) (1997) In Vitro *Methods in Pharmaceutical Research*. Academic Press: San Diego, CA.

Chabner B.A., Longo D.L. (eds) (1996) *Cancer Chemotherapy and Biotherapy: Principles and Practice*. Lippincot-Raven: Philadelphia, PA.

Chee M., Yang R., Hubbell E., Berno A., Huang X.C., Stern D., Winkler J., Lockhart D.J., Morris M.S., Fodor S.P. (1996) Assessing genetic information with high-density DNA arrays. *Science*, **274**, 610.

Clackson T.,Wells J.A. (1995) A hot spot of binding energy in a hormone-receptor interface. *Science*, **267**, 383.

Cummins D.J., Andrews C.W., Bentley J.A., Cory M. (1996) Molecular diversity in chemical databases: comparison of medicinal chemistry knowledge bases and databases of commercially available compounds. *Journal of Chemical Information and Computer Sciences*, **36**, 750.

Ganesan A. (1998) Combinatorial chemistry in the hunt for medicines. *Nature*, **393**, 727.

Gillet V.J., Willett P., Bradshaw T. (1998) Identification of biological activity profiles using substructural analysis and genetic algorithms. *Journal of Chemical Information and Computer Sciences*, **38**, 165.

Giuliani A., Colosimo A., Benigni R., Zbilut J. (1998) On the constructive role of noise in spatial systems. *Physics Letter A*, **247**, 47.

Golub G., Loan C.V. (1989) *Matrix Computations*. Johns Hopkins University Press: Baltimore, MD.

Good A.C., Lewis R.A. (1997) New methodology for profiling combinatorial libraries and screening sets: cleaning up the design process with HARpick. *Journal of Medicinal Chemistry*, **40**, 3926.

Gordon E.M., Barrett R.W., Dower W., Fodor S.P., Gallop M.A. (1994) Application of combinatorial technologies to drug discovery. 2. Combinatorial organic synthesis, library screening strategies and future directions. *Journal of Medicinal Chemistry*, **37**, 1385.

Gray N.S., Wodicka L., Thunnissen A.M., Norman T.C., Kwon S., Espinoza F.H., Morgan D.O., Barnes G., Clerc S.L., Meijer L., Kim S. H., Lockhart D.J., Shultz P.G. (1998) Exploiting chemical libraries, structure and genomics in the search for kinase inhibitors. *Science*, **281**, 533.

Grever M.R., Schepartz S.A., Chabner B.A. (1992) The

National Cancer Institute: cancer drug discovery and development program. *Seminars in Oncology*, **19**, 622.

Harary F. (1971) *Graph Theory*. Addison-Wesley: Reading, MA.

Hrach K. (1997) Comparison of survival between two groups using software SAS, S-PLUS and STATISTICA. *International Journal of Medical Informatics*, **45**, 31.

Hwang J., Shyy S., Chen A.Y., Juan C.C., Whang-Peng J. (1989) Studies of topoisomerase-specific antitumor drugs in human lymphocytes using rabbit antisera against recombinant human topoisomerase II polypeptide. *Cancer Research*, **49**, 958.

Janin J., Chothia C. (1990) The structure of protein–protein recognition sites. *Journal of Biological Chemistry*, **265**, 16027.

Johnson M.A., Maggiora G. (eds) (1990) *Concepts and Applications of Molecular Similarity*. John Wiley: New York.

Kauver L.M. (1995) Affinity fingerprinting. *Biotechnology*, **13**, 965.

Koutsoukos A.D., Rubenstein L.V., Faraggi D., Simon R.M., Kalyandrug S., Weinstein J.N., Kohn K.W., Paull K.D. (1994) Discrimination techniques applied to the NCI *In Vitro* Anti-tumor Drug Screen: predicting biochemical mechanism of action. *Statistics in Medicine*, **13**, 719.

Lee J.S., Paull K., Alvarez M., Hose C., Monks A., Grever M., Fojo A.T., Bates S. (1994) Rhodamine efflux patterns predict P-glycoprotein substrates in the National Cancer Institute Drug Screen. *Molecular Pharmacology*, **46**, 627.

Lewis R.A., Mason J.S., McLay I.M. (1997) Similarity measures for rational set selection and analysis of combinatorial libraries: the diverse property-derived (DPD) approach. *Journal of Chemical Information and Computer Sciences*, **37**, 599.

Liu K. (1997) Application of SVD in optimization of structural modal test. *Computers and Structures*, **63**, 51.

Marchington T. (1995) From data to drugs. *Biotechnology*, **13**, 239.

Mardia K.V., Kent J.T., Biby J.M. (1979) *Multivariate Analysis*. Academic Press: London.

Martin Y.C., Willet P. (1998) *Designing Bioactive Molecules*. American Chemistry Society: Washington, DC.

Miller M.D. (1994) SQ. A program for producing rapid molecular superimpositions. *Am. Chem. Soc.*, **207**, 27.

Monks A., Scudiero D., Skehan P., Shoemaker R., Paull K., Vistica D., Hose C., Langely C., Cronise P., Vaigro-Wolff A., Grey-Goodrich M., Cambell L., Mayo J., Boyd M.R. (1991) Feasibility of a high flux anticancer drug screen using a diverse panel of cultured human tumor cell lines. *Journal of the National Cancer Institute*, **83**, 757.

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

Nogales E., Downing K.H., Amos L.A., Lowe J. (1998a) Tublin and FtsZ form a distinct family of GTPases. *Nature Structural Biology*, **5**, 451.

Nogales E., Wolf S.G., Downing K.H. (1998b) Structure of the α/β tubulin dimer by electron crystallography. *Nature*, **391**, 199.

Nogales E., Whittaker M., Milligan R.A., Downing K.H. (1999) High resolution model of the microtubule. *Cell*, **96**, 79.

O'Connor P.M., Jackman J., Bae I., Myers T.G., Fan S., Mutoh M., Scudiero D.A., Monks A., Sausville E.A., Weinstein J.N., Friend S., Fornace A.J.J., Kohn K.W. (1997) Characterization of the p53 tumor suppressor pathway in cell lines of the National Cancer Institute Anticancer Drug Screen and correlations with the growth-inhibitory potency of 123 anticancer agents. *Cancer Research*, **57**, 4285.

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

Paull K., Hamel E., Malspeis L. (1995) Prediction of biochemical mechanism of action from the *In Vitro* Antitumor Screen of the National Cancer Institute. In *Cancer Chemotherapeutic Agents*. Foye W.O. (ed.), p. 9. American Chemistry Society: Washington, DC.

Pearlman R.S., Smith K.M. (1998) Novel software tools for chemical diversity. *Perspectives in Drug Discovery*, 339.

Pratt W.B., Ruddon R.W., Ensminger W.D., Maybaum J. (1994) *The Anticancer Drugs*. Oxford University Press: New York.

Randic M. (1997) On characterization of chemical structure. *Journal of Computer Information and Computer Sciences*, **37**, 672.

Sadowski J., Kubinyi H. (1998) A scoring scheme for discriminating between drugs and non-drugs. *Journal of Medicinal Chemistry*, **41**, 3325.

SAS(R) (1992) Technical Report A-108. Cubic Clustering Criterion. SAS Institute, Inc.

Schreiber G., Fersht A.R. (1995) Energetics of protein–protein interactions: analysis of the Barnase–Barstar interface by single mutations and double mutant cycles. *Journal of Molecular Biology*, **248**, 478.

Shemetulskis N.E., Dunbar J.B. Jr, Dunbar B.W., Moreland D.W., Humblet C. (1995) Enhancing the diversity of corporate databases using chemical database clustering and analysis. *Journal of Computer-Aided Molecular Design*, **9**, 407.

Shi L.M., Fan Y., Myers T.G., Waltham M., Paull K.D.,

Weinstein J.N. (1998a) Mining the anticancer activity database generated by the U.S. National Cancer Institute's drug discovery program using statistical and artificial intelligence techniques. *Mathematical Modeling and Scientific Computing*, **38**, 189.

Shi L.M., Fan Y., Myers T. G., O'Connor P.M., Paull K.D., Friend S.H., Weinstein J.N. (1998b) Mining the NCI anticancer drug discovery databases: genetic function approximation for the QSAR study of anticancer ellipticine analogues. *Journal of Chemical Information and Computer Sciences*, **38**, 189.

Shi L.M., Myers T.G., Fan Y., O'Connor P.M., Paull K.D., Friend S.H., Weinstein J.N. (1998c) Mining the National Cancer Institute anticancer drug discovery database: cluster analysis of ellipticine analogs with p53-inverse and central nervous system selective patterns of activity. *Molecular Pharmacology*, **53**, 241.

Sneath P.H.A., Sokal R.R. (1973) *Numerical Taxonomy*. W.H. Freeman & Co.: San Francisco, CA.

Snedecor G.W., Cochran W.G. (1980) *Statistical Method*. Iowa State University Press: Ames, IA.

Stryer L. (1988) *Biochemistry*. W.H. Freeman & Co.: New York.

van Osdol W.W., Myers T.G., Paull K.D., Kohn K.W. Weinstein J.N. (1994) Use of Kohonen self-organizing map to study the mechanism of action of chemotherapeutic agents. *Journal of the National Cancer Institute*, **86**, 1853.

Weininger D., Weininger A., Weininger J.L. (1997) SMILES. 2. Algorithm for generation of unique SMILES notations of combinatorial libraries: the diverse property-derived (DPD) approach. *Journal of Chemical Information and Computer Sciences*, **37**, 599.

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

Wu L., Smythe A.M., Stinson S.F., Mullendore L.A., Monks A., Scuderio D.A., Paull K.D., Koutsoukos A.D., Rubinstein L.V., Boyd M.R., Shoemaker R.H. (1992) Multidrug-resistant phenotype of disease-oriented panels of human tumor cell lines used for anticancer screening. *Cancer Research*, **52**, 3029.

Zhang L., Zhou W., Velculescu V.E., Kern S.E., Hruban R.H., Hamilton S.R., Volgelstein B., Kinzler K.W. (1997) Gene expression profiles in normal and cancer cells. *Science*, **276**, 1268.

Appendix. Survey results from an analysis of available crystal structures complexed with ligands that are structurally similar to the standard anticancer agents analyzed here

Table IV lists the protein complexes identified here for investigating this issue. Our intention here is not to provide a complete list of all structural analogs within the Protein Data Bank (PDB) (Bernstein *et al.*, 1977), but to indicate the range of protein structures that are known to form complexes with the structural analogs to the 122 anticancer agents. The results presented in Table IV were obtained using the SMILES-based searching tools available in the RELIBASE part of the PDB browser (http://www.pdb.bnl.gov). The first column in the table describes the types of enzymes, the second and third give the name and PDB identifier of each enzyme, the fourth column is the ligand bound in the complex, and the fifth column lists the anticancer agents that are either identical or structural analogs to the standard 122 anticancer agents.

The results in Table IV directly indicate the sites of action of many of the agents assigned to Groups 9–25 of our cluster analysis. For example, crystallographic complexes exist for most of the enzymes involved in pyrimidine biosynthesis pathway. This pathway involves six enzymatically catalyzed steps. The CAD gene encodes a trifunctional protein associated with the activity of the first three enzymes in this six-step pathway: carbamoylphosphate synthase (EC 6.3.5.5), aspartate transcarbamoylase (EC 2.1.3.2), and dihydroorotase (EC 3.5.2.3)—also referred to as CPSase, ATCase and DHOase, respectively. Crystallographic complexes exist for acivicin (163501) bound to CPSase, PALA (224131) bound to ATCase and brequinar (368390) bound to DHOase. In addition, the sites of action of methotrexate (740) as well as other folate by-products, include dihydrofolate reductase, thymidylate synthase, AICAR transformylase and GAR transformylase, all of which are included in the set of complexes listed in Table IV. Purine biosynthesis occurs by *de novo* pathways as well as from preformed nucleosides and nucleotides via salvage reactions (Stryer, 1988). Phosphoribosyl kinases and transferases are involved in both processes, and are found in crystallographic complex with many of the nucleoside analogs included in this study. A surprising finding includes the recent dimeric structure of tubulin in complex with a taxane. A nucleoside analog is also bound at the dimer interface between the α and β tubulin subunits (Nogales *et al.*, 1998a,b, 1999). Taken together, these crystallographic complexes indicate that many of the antitumor agents included in these groups target one or in some cases many proteins involved in nucleic acid biosynthesis or mitosis. The cell screening patterns of these compounds, when clustered according to the methods used here, clearly separate the compounds from DNA-damaging agents.

Table IV

Proteins complexed with ligands similar to anticancer agents

Enzyme class	*Name*	*PDB*	*ligand*	*NSC*
Ligase	carbamoyl phosphate synthase	1jdb	GLN chan	163501
	"	1jdb	ADP	71851,71261
Hydrolase	cytidine deaminase	1aln	3-deazacytidine	102816,143095
	"	1ctt	dihydrozebularine	102816,143095,264880
	"	1ctu	zebularine	148958,264880
Oxidoreductase	dihydroorotate dehydrogenase	2dor	flavin mononucleotide	148958,27640
	"	2dor	orotic acid	148958
	diaminopimelic acid dehydrogenase	1dap	NDP	71851,71261
	"	1dap	DA3	163501
	cyclooxygenase	3pgh	flurbiprofen	368390
	dihydrofolate reductase	1ai9	NDP	71851,71261
	"	1ao8	MTX	740
	"	1dhf	MTX	740
Transferase	thymidylate synthase	1bjg	5-F-deoxyuridine	148958
	"	1bjg	hydrofolic acid	623017,174121
	"	1vzd	dideazafolic acid	134033
	"	2tdd	hydrofolic acid	134033
	"	1tls	5-F-deoxyuridine	148958
	"	1lce	hydrofolic acid	132483
	amidotransferase carbamoyl phosphate synthetase	1a9x	GLN	163501
	"	1a9x	ADP	71851,71261
	"	2tdd	hydrofolic acid	134033
	"	1tls	5-F-deoxyuridine	148958
	"	1lce	hydrofolic acid	132483
	"	1a9x	GLN	163501
	aspartate transcarbamylase	1acm	PALA	224131
	phosphoribosyl transferase	1opr	orotic acid	148958,102816
	"	1sto	orotidine	148958,27640
	carbamoyl transferase	1rai	cytidine	102816,27640
	phosphoribosyglycinamide	1cde	ribonucleotide	102816
	formyltransferase	1gar	U89	118994,71851,71261
	methyltransferase	1v39	homocysteine	71261,71851
	nucleotidyl transferase	1waf	GMP	71261,71851
	thioredoxin	1t7p	guanosine	71261,71851
	nucleoside phosphorylase	1a69	formycin	143095
	"	1a9t	hypoxanthine	71851,71261
	"	1a9t	ribose–1-phosphate	102816
	diphosphate kinase	1be4	guanosine	71261,71851
	diphosphate kinase	1kdn	ADP	71261,71851
	adenylate kinase	1dvr	adenosine	71261,71851
	thymidine kinase	1kim	thymidine	27640
	protein kinase inhibitor	1kpe	adenosine	71261,71851
	purine phosphorylase	1vfn	hypoxanthine	71851,71261
	UMP/CMP kinase	2ukd	ADP C5P	71851,71261
Microtubules	α/β tubulin dimer	1tub	gtp,gdp	71851,71261
	"	1tub	taxotere	125973

JBC Papers in Press. Published on September 25, 2000 as Manuscript M005299200

MAPPING THE BINDING SITE OF COLCHICINOIDS ON -TUBULIN: 2-CHLOROACETYL-2-DEMETHYLTHIOCOLCHICINE COVALENTLY REACTS PREDOMINANTLY WITH CYSTEINE-239 AND SECONDARILY WITH CYSTEINE-354

Ruoli Bai, David G. Covell, Xue-Feng Pei, John B. Ewell,

Nga Y. Nguyen, Arnold Brossi, and Ernest Hamel

Screening Technologies Branch
Developmental Therapeutics Program
Division of Cancer Treatment and Diagnosis
National Cancer Institute
Frederick Cancer Research and Development Center
Frederick, Maryland 21702
(RB, EH)

Laboratory of Experimental and Computational Biology
National Cancer Institute-Frederick Cancer Research and Development Center
Science Applications International Corporation-Frederick
Frederick, Maryland 21702
(DGC)

Laboratory of Structural Biology
National Institute of Diabetes and Digestive and Kidney Diseases
National Institutes of Health
Bethesda, Maryland 20892
(X-FP, AB)

and

Facility for Biotechnology Resources
Center for Biologics Evaluation and Research
Food and Drug Administration
Bethesda, Maryland 20892
(JBE, NYN)

Address correspondence to: Dr. E. Hamel
Building 469, Room 237
National Cancer Institute-FCRDC
P. O. Box B
Frederick, Maryland 21702
FAX: (301) 846-6014
email: hamele@dc37a.nci.nih.gov

SUMMARY

2-Chloroacetyl-2-demethylthiocolchicine (2CTC) and 3-chloroacetyl-3-demethylthiocolchicine (3CTC) resemble colchicine in binding to tubulin and react covalently with β-tubulin, forming adducts with cysteine residues 239 and 354. The adducts at Cys-239 are less stable than those at Cys-354 during formic acid digestion. Extrapolating to zero time, the Cys-239 to Cys-354 adduct ratio is 77:23 for 2CTC and 27:73 for 3CTC. Using energy minimization modeling to dock colchicinoids into the electron crystallographic model of β-tubulin in protofilaments [Nogales et al. (1998) *Nature* **391**, 199-203], we found two potential binding sites. At one, entirely encompassed within β-tubulin, the C2- and C3-oxygen atoms of 2CTC and 3CTC overlapped poorly with those of colchicine and thiocolchicine, but distances from the reactive carbon atoms of the analogs to the sulfur atoms of the cysteine residues were qualitatively consistent with reactivity. The other potential binding site was located at the α/β interface. Here, the oxygen atoms of the analogs overlapped well with those of colchicine, but relative distances of the reactive carbons to the cysteine sulfur atoms did not correlate with the observed reactivity. A significant conformational change must occur in the colchicine binding site of tubulin in the transition from the unpolymerized to the polymerized state.

Despite the interaction of tubulin with a large number of drugs that inhibit or promote its assembly into microtubules, precise definition of drug binding sites on the protein has not been possible. This is a consequence of the lack of success in crystallizing the protein, probably because of its sequence and post-translational heterogeneity, its instability, and its tendency to form oligomers and polymers of highly aberrant morphology in the presence of many of these drugs. The recent electron crystallographic determination of a relatively detailed structure for zinc-induced antiparallel tubulin protofilaments has provided insights into the paclitaxel/docetaxel site on these protofilaments, since docetaxel was used to enhance their stability during data accumulation (1), and provided a scaffold on which to model other drug sites (2). A limitation in such analysis, however, is that drugs that inhibit assembly have limited ability to bind to tubulin polymers containing linear protofilaments.

Alternative approaches to obtain preliminary information about drug binding sites have included "direct" photoaffinity labeling (e.g., ref. 3), analog photoaffinity labeling (e.g., ref. 4), and cross link formation with chemically reactive analogs that retain biological activity (e.g., ref. 5). In the first method, a drug-tubulin complex is exposed to light of an appropriate wave length and ligand-protein cross link formation is evaluated. In the second, an active drug analog containing a photoreactive moiety is prepared, bound to tubulin, and cross link formation induced by exposure of the complex to light of an appropriate wave length. In the third method, an active analog with a chemically reactive moiety is prepared, bound to tubulin, and cross link formation occurs either spontaneously or, in principle, following a rapid change in reaction conditions. Generally, the reactive ligand is radiolabeled to permit quantitation of the reaction and identification of peptides in the protein involved in cross link formation. In all cases there are two major problems. The first is the potential for nonspecific protein alkylation by the ligand, which is generally excluded by demonstrating that excess nonreactive ligand substantially inhibits the covalent reaction. The second is that adequate radiolabel participates in cross link formation to permit identification of the tubulin subunit (α or β), the peptide region of the subunit, and, ideally, the specific amino acid residue(s) involved.

The tubulin-colchicine interaction has attracted a great deal of attention, probably because of its unusual chemical characteristics (for a review, see ref. 6). The drug (structure in Fig. 1) binds exceptionally slowly, but noncovalently, to tubulin, and, once formed, the drug-protein complex is highly stable. Although some have described the association reaction as "essentially irreversible," the dissociation reaction has been carefully studied, and the half-life of the tubulin-colchicine complex varies from 14-77 h at 37 °C, depending on precise reaction conditions. In addition, the interaction of tubulin with colchicine appears to involve significant changes in conformation in both the drug and the protein.

Photoaffinity analogs of colchicine have reacted predominantly with α-tubulin or with both subunits (7, 8). In contrast, direct photoaffinity labeling, with irradiation at 350 nm (the absorbance maximum of the tropolonic C ring of colchicinoids), resulted in strongly preferential labeling of β-tubulin (3, 9). A cross link was formed between radiolabeled colchicine and amino acid(s) in peptide sequence 1-36 or peptide sequence 214-241, but not with both peptides.

Our own approach has been to place the small chloroacetyl group (about 3 Å in length) at various locations in analogs of colchicine and thiocolchicine. When placed in the side chain or the C ring, we observed no significant specific covalent reaction with tubulin. The derivatized A ring analogs 2CTC[1] and 3CTC, however, are active colchicine site compounds that react covalently with β-tubulin (structures in Fig. 1). The reactions are specific in that they are extensively inhibited by colchicine site drugs, and the two reactions have different properties. We previously showed that 3CTC reacted predominantly with cysteine-354 of β-tubulin and noted that there was a minor reaction as well with cysteine-239 (5). In the present study we demonstrate the reactivity of 2CTC predominantly with cysteine-239, but secondarily with cysteine-354 as well. We also attempted to construct a model for the A ring subsite of colchicine based on quantitative differences in the reactivity of 2CTC and 3CTC with the two cysteine residues and on the electron crystallographic model of tubulin (1).

EXPERIMENTAL PROCEDURES

Materials. Preparation of electrophoretically homogeneous bovine brain tubulin (10) and [^{14}C]2CTC (11) were described previously. Specific activity of the [^{14}C]2CTC was 40 cpm/pmol. Decylagarose was from ICN Immunobiologicals; CNBr and NEM from Sigma; podophyllotoxin and formic acid from Aldrich; "sequencing grade" trypsin and "sequencing grade" EP-GC (*Staphylococcus aureus* V8) from Boehringer-Mannheim; and precast Tricine-16% acrylamide polyacrylamide gels and PVDF membranes from Novex. Kodak Biomax MR film was used for preparation of autoradiographs.

Preparation of tubulin derivatized with [^{14}C]2CTC and separation of α- and β-tubulin subunits by decylagarose chromatography. Reaction mixtures contained 25 μM (2.5 mg/ml) tubulin, 25 μM [^{14}C]2CTC, podophyllotoxin as indicated, 1.0 M monosodium glutamate, 0.1 M sodium phosphate (pH 7.0), 0.1 mM GDP, and 0.5 mM $MgCl_2$. Incubation was for 30 min at 37 °C, and the reaction was stopped by adding NEM to a final concentration of 5 mM. The mixture was left overnight at 4 °C, but precipitation of the tubulin was usually incomplete. The mixture was made 5% (v/v) in trichloroacetic acid, and, after an additional 30 min at 0 °C, the precipitated protein was harvested by centrifugation. The pellet was dissolved in a solution containing 4 M guanidine hydrochloride and 2 M NaCl (adjusted to pH 5.0 with HCl) and applied to a column of decylagarose (12). Chromatography and analysis of protein peaks by SDS-PAGE was performed as described previously (5). Stoichiometry of [^{14}C]2CTC associated with the β-tubulin peak was 0.17, and with the α-tubulin peak, 0.03. Only β-tubulin of least 90% purity was used in further studies, except in the podophyllotoxin inhibition study, in which unresolved tubulin was used.

Chemical and enzymatic digestions of β-tubulin cross linked to [^{14}C]2CTC. The [^{14}C]2CTC-β-tubulin at 2.5 mg/ml was digested at 37 °C in the dark either with 75% formic acid for 96 h (13) or with CNBr (20 mg/ml) in 70% formic acid for 24 h. The formic acid and, if present, CNBr were removed by lyophilization, and the residue was washed twice with water, which was removed each time by lyophilization.

For enzymatic digestions the [^{14}C]2CTC-β-tubulin was dissolved in 1.0 M Tris (pH 8.0 with HCl). The tubulin solutions were diluted 10-fold into 50 mM ammonium acetate (pH 4.0) for EP-GC or water for trypsin, and the appropriate enzyme was added at an enzyme to substrate ratio of 1:50. The resulting reaction mixtures were incubated for 24 h at 37 °C in the dark. At the end of the incubation, about 75% of the water in the samples was removed by lyophilization.

Peptide purification. Peptide separation was by SDS-PAGE on Novex precast gels, with the peptide solution to be analyzed dissolved in the Tricine-SDS Sample Buffer solution provided by Novex. Following electrophoresis the separated peptides were transferred from the gel to a PVDF membrane (pore size, 0.2 μm) with an Enprotech semidry transblot system (1 h, 100 v). The membrane was stained with Coomassie Blue R250 and autoradiographed (24-72 h exposure). Radiolabeled peptides were cut from the membrane for sequencing.

Sequence analysis. Automated Edman degradation for determination of amino acid sequence was performed with an Applied Biosystems model 494A Protein Sequenator. Identification of phenylthiohydantoin amino acid derivatives was carried out with an Applied Biosystems model 140C Microgradient System and model 785A Programmable Absorbance Dectector. Identification of radiolabeled amino acid residues was performed by the University of Virginia Biomolecular Research Facility. Following each cycle of Edman degradation the sample stream was analyzed by liquid scintillation counting for radiolabel instead of analyzed by HPLC to identify the derivatized amino acid residue.

Molecular modeling. A two-stage modeling analysis was used, first, to identify candidate binding sites on the tubulin dimer and, second, to dock the molecular structures of colchicinoids into these sites. The first procedure probes the c-alpha coordinates of the tubulin dimer (1) to determine exterior positions that are most likely to be found within a binding interface. This analysis uses information about local geometry and chemical composition of subregions of the target surface for selecting candidate sites. Relative rankings of these potential interaction sites are based on a scoring scheme derived from a statistical analysis of all

known protein-ligand complexes. In applying this method to analysis of new crystal complexes, we have found that the correct ligand interface is found within the top 5% of candidate binding sites (D. Covell, unpublished data). This method has also been shown to correctly identify ligand binding sites for a wide range of proteins and ligands (14, 15). For a complete description of this procedure see ref. 16.

The second stage of the the analysis involves docking the test ligands at candidate binding sites. The initial docking is based exclusively on geometric considerations. This step uses a "geometric hashing technique" that has been found to rapidly determine a family of possible binding geometries for each ligand (17). Each of these possible binding arrangements are further refined to determine those positions with the maximum binding strength between ligand and target protein. A previously published model of ligand binding (18) was used to select the best binding geometries. This model is based on the atomic preferences of adjacent surfaces buried within a binding interface (18). The model has been shown to predict accurately ligand binding strengths and assess the relative contributions of atomic interactions within a binding interface (17). Moreover, the model has been extended as an adjunct to computational docking (19), has proven effective for identifying ligands active against NCp7 targets (20), and has been useful in providing testable hypotheses about the modes of action of candidate inhibitors for a variety of enzymes (21-23).

The final stage of docking was obtained from successive in vacuo molecular dynamics and energy minimization calculations using the CVFF91 force field within Discover97.0 (Molecular Simulations, Inc., San Diego CA), based on the candidate geometries obtained from steps one and two, as outlined above. This final step resulted in small changes in geometries, both in the ligand and in the target protein, primarily to eliminate energetically unfavorable van der Waals interactions. These dynamics and minimization steps were preformed repeatedly to achieve the final geometries. Exploration of these final geometries, which were used for our analysis, indicates trapping in a local energy minimum. The resulting geometries did not significantly alter the starting geometries of each ligand and were acceptable within the 3.6 Å resolution of the electron crystallographic structure of tubulin (1).

RESULTS

In our initial characterization of the interactions of 2CTC and 3CTC with tubulin (24), we found that these interactions were similar to that of colchicine with tubulin (slow, temperature-dependent binding; similar quantitative inhibitory effects on polymerization; similar binding stoichiometries), and both compounds were competitive inhibitors of the binding of [^{3}H]colchicine to tubulin (apparent K_i values, about 3 μM). Unlike colchicine, however, they formed a covalent bond with β-tubulin, and bond formation, as well as the initial binding reaction, was strongly inhibited by podophyllotoxin. The major difference between 2CTC and 3CTC was in the covalent reactions, which were studied most extensively with the compounds at 5 μM and tubulin at 20 μM. With 3CTC the covalent reaction occurred almost simultaneously with binding, and about 57% of the bound drug formed a covalent bond with tubulin. With 2CTC covalent bond formation was much slower than the binding reaction. After 30 min about 26% and after 1 h 30% of the bound drug had covalently reacted with the tubulin. Finally, with superstoichiometric concentrations of both 2CTC and 3CTC the covalent reactions were more extensive, but there was a significant reduction in the apparent specificity of the covalent reactions (i.e., a smaller proportion of covalent bond formation was inhibited by podophyllotoxin).

Because of the more extensive covalent reaction with 3CTC, we initially studied it in detail (5), purifying the alkylated β-tubulin by decylagarose chromatography. Analysis of CNBr peptide digests resolved by HPLC were consistent with alkylation of Cys-354 and Cys-239 in roughly a 2:1 ratio, but formic acid peptide digests resolved by SDS-PAGE indicated a 9:1 ratio.

In the studies presented here we used 75% formic acid digestion for the initial analysis of [^{14}C]2CTC-containing peptides. The primary cleavage site under the condition used is aspartylproline (13), and there are two such sites in β-tubulin (positions 31/32 and 304/305). Rao et al. (4) termed the three resulting peptides A1 (residues 1-31), A2 (32-304), and A3 (305-445), and these peptides migrate as expected upon SDS-PAGE. In our hands there are also secondary cleavage sites of β-tubulin in formic acid

(5), but the amino acid sequences of the less prominent bands suggested that secondary cleavage occurred subsequent to hydrolysis of the aspartylproline bonds (also, see below).

In an initial experiment with unresolved tubulin following its interaction with [^{14}C]2CTC, we observed heavy labeling of both A2 and A3 by [^{14}C]2CTC. Formation of both radiolabeled peptides was abolished in the presence of podophyllotoxin (data not presented), consistent with our previous observations (24).

Thus encouraged, we separated α- and β-tubulin on decylagarose following the reaction with [^{14}C]2CTC prior to formic acid digestion. After formic acid treatment the protein digest was subjected to SDS-PAGE, and the peptides on the gel were electrotransferred to a PVDF membrane, which was stained and autoradiographed. Track I in Fig. 2A shows the Coomassie Blue stain pattern obtained, with peptides A1, A2, and A3 indicated. Track II is the autoradiogram of the stained track I, showing the heavily labeled A2 and A3, with a number of minor radiolabeled bands between A2 and A3 and between A3 and A1. These presumptive assignments were confirmed by sequential Edman degradation for 15 cycles in the case of A2 and 10 cycles for A3 (Table I). In addition, sequence analysis was performed on minor radiolabeled bands running between A2 and A3 and between A3 and A1. In the former case, these were found to have the same amino terminal sequence as A2; and in the latter, the same amino terminal sequence as A3 (data not presented). For comparison, we also include as Track III in Fig. 2A an autoradiogram of formic acid digested β-tubulin following an identical incubation with [^{14}C]3CTC. As previously (5), there was little radiolabel in A2, in contrast to the heavy radiolabel in A3 and A3 fragments. Densitometric analysis of gels II and III indicated that the ratio of A2+fragments:A3+fragments was about 0.7:1 following reaction with 2CTC and 0.1:1 following reaction with 3CTC (see Table II).

To better define the two reactive amino acids following the covalent interaction of β-tubulin with [^{14}C]2CTC, we next digested the decylagarose-isolated protein with CNBr, with subsequent PAGE and electrotransfer of the peptides to PVDF. An autoradiogram from a typical experiment is shown as Track I in

Fig. 2B. Reproducibly, only two radiolabeled bands were observed, with the upper "peptide a" more heavily labeled than the lower "peptide b". Sequential Edman degradation of peptides a and b yielded sequences consistent with CNBr-derived peptides spanning residues 234-257 (for 10 cycles) and 331-363 (17 cycles), respectively. The former peptide is entirely within the A2 peptide obtained with formic acid, and the latter within the A3 peptide.

We also re-evaluated β-tubulin following its reaction with [^{14}C]3CTC and CNBr digestion by the PAGE-electrotransfer methodology. A typical autoradiogram is shown as Track II in Fig. 2B. Note that the same two bands were radiolabeled with 3CTC as with 2CTC (confirmed by sequential Edman degradation, data not presented). The relative amounts of radiolabel in the a and b peptides in the two Tracks shown in Fig. 2B were 2:1 for the 2CTC sample and 0.4:1 for the 3CTC sample, as determined by densitometry (summarized in Table II).

In our previous study we had clearly established that the bound [^{14}C]3CTC reacted primarily with Cys-354 of β-tubulin, but we had also obtained preliminary data consistent with a secondary reaction of bound 3CTC with Cys-239 (5). However, the autoradiogram of the electroblot shown as Track II in Fig. 2B and the sequence data obtained from the peptide suggested that the secondary reaction was more substantial than we had previously thought. At the same time, the apparent identity of the CNBr peptides obtained from the [^{14}C]3CTC-reacted β-tubulin and the [^{14}C]2CTC-reacted β-tubulin strongly indicated that the two colchicine analogs alkylated the same cysteine residues in different proportions.

The microsequencing Edman degradation procedure uses too little material for direct identification of the radiolabeled amino acid residue on peptides embedded in PVDF membranes (the degradation steps are performed directly on the membrane slice). In previous studies (5, 25) we have found it possible to perform an appropriate digestion of the entire polypeptide, subject the peptide mixture to sequential Edman degradation, and count the outflow of each cycle to obtain evidence to identify the specific amino acid residue that had been alkylated. Considering the two radiolabeled peptides generated by CNBr digestion

(β234-257 and β331-363), a total β-tubulin-[^{14}C]2CTC cyanogen bromide digest should yield a peak of radiolabel at the sixth degradation cycle if the expected Cys-239 were radiolabeled (radiolabel cross linked to Cys-354 would not appear for 23 cycles). Such an experiment was performed, with the outflow of 10 cycles counted. We obtained the expected result (Fig. 3A) and conclude that the major alkylation of β-tubulin by 2CTC occurs at Cys-239.

To confirm that 2CTC alkylation of β-tubulin also occurred at Cys-354 we turned to enzymatic digestion of β-tubulin that had reacted with the drug. A single experiment with EP-GC was performed. The peptide digest was subjected to SDS-PAGE and electrotransfer to PVDF, and the autoradiogram obtained is presented in Fig. 2C. Two closely spaced radiolabeled peptides were observed, and the sequences obtained from the minor, upper peptide "a" and the major, lower peptide "b" are presented in Table I. Consistent with EP-GC digestion, the sequence of the minor peptide, through 11 cycles, resulted from cleavage between Glu-343 and Trp-344 (the carboxy terminus of this peptide could be at Asp-355, but most likely is at Glu-376, Glu-383, or Glu-401, based on the apparent size of the peptide). This result further narrowed the location of the secondary alkylation site of 2CTC to the β-tubulin sequence spanning amino acid residues 344-363 (the amino terminus of the CNBr secondary peptide).

The major EP-GC peptide, however, yielded an unexpected sequence, beginning at Thr-199, indicating cleavage after Glu-198. Three potential downstream cleavage sites (Asp-203, Glu-205, and Asp-209) are included in the sequence obtained, and an additional potential site at Asp-224 must have also been skipped for this peptide to include Cys-239. Based on peptide size, we assume the carboxy cleavage site was Asp-249, since the next potential residue is Glu-288.

We next examined trypsin, but only very small peptides were generated from β-tubulin that had reacted with [^{14}C]2CTC, and resolution by SDS-PAGE was minimal. On most electroblots a single radiolabeled peptide band was observed, but sequence analysis indicated it was very heterogeneous. However, when the entire β-digest was subjected to sequential degradation and the outflow stream counted,

a dramatic radiolabeled peak was observed at the fourth of 14 cycles (Fig. 3B). This is the expected result for radiolabel at Cys-354, for tryptic cleavage should occur at Lys-350. (The tryptic peptide containing Cys-239 should begin with Leu-217, so no radiolabel should derive from this peptide during the 14 cycles examined.)

Molecular modeling. The electron crystallographic coordinates of the αβ-tubulin dimer (1) were analyzed for candidate binding sites for colchicine. As described above, the procedure consists of scoring the solvent accessible surface of the tubulin dimer for cavities with strong ligand binding features. This method was developed by a detailed examination of multiple crystallographically available ligand-receptor complexes, with subsequent determination of residue types most likely to be found in a binding interface. Subsequent testing of this method against newly available crystal complexes has shown that the correct ligand binding site is found within the topmost candidate sites identified by the method. Regions on the tubulin dimer that were identified as candidate binding sites were then subjected to computational docking with colchicine, thiocolchicine, 2CTC, and 3CTC. The final docking arrangement was obtained by sampling the lowest energy geometries from successive in vacuo molecular dynamics and minimization calculations. In these simulations both ligand and target were allowed to relax to their lowest energy configurations.

In the initial step the entire surface of the tubulin dimer was scanned for candidate binding sites. The analysis yielded sites that were located at the plus and minus ends of the dimer, in a region entirely within β-tubulin near the paclitaxel binding site, and at the α/β interface. The two highest scoring sites for colchicinoid binding were the latter two, and both were located near the Cys-239/Cys-354 region of β-tubulin. The sites at the plus and minus ends of the dimer were not examined further, since they were distant from the two reactive cysteine residues. The site contained within β-tubulin we term "Site A" (cf. ref. 2), and the interface site we term "Site B." Fig. 4 shows a colchicine molecule docked into each of these alternative potential binding sites (*we emphasize that this figure does not mean to imply two independent*

binding sites for colchicine), with the left-hand panel showing the solvent accessible surface and the right-hand panel the peptide backbone as a ribbon diagram. Note that Site B is most consistent with data obtained with photoactive colchicine analogs, where covalent interactions with both α- and β-tubulin were observed (7, 8).

Amino acid residues of β-tubulin in closest contact with colchicinoids bound in Site A were His-227 and Phe-270 (A ring), Val-23 and Ala-231 (B ring), and Asp-26, Tyr-36, and Phe-242 (C ring). For Site B β-tubulin residues in closest contact with bound colchicinoids were Tyr-36 (A ring) and Arg-2 (C ring), and α-tubulin residues were Asp-76 (B ring) and Thr-73 (C ring).

In Fig. 5 we show the relationship of the Site A (panel A) and Site B (panel B) docked colchicine molecules to the peptides identified by Uppuluri et al. (9) following direct photoaffinity labeling of tubulin by colchicine. In both panels the backbone of the peptide containing residues 1-36 is shown in white, while the peptide containing residues 214-241 is shown in orange. In addition, the side chains of Cys-239 and Cys-354 are shown in both panels, with the sulfur atom of the former colored light blue and of the latter colored yellow. In Table III we present the distances for both potential binding sites between the C2 and C3 oxygen atoms of colchicine and the Cys-239 and Cys-354 sulfur atoms. Although the backbone of the 214-241 residue peptide appears to be in closer contact with colchicine docked in Site A than in Site B, van der Waals distances measured from colchicine docked in the sites to any atom of the peptide, including the side chains, did not allow us to choose between Sites A and B. Colchicine docked in Site A was 2.8 Å from residue 33 and 2.1-2.5 Å from residues 231, 234, and 239. Colchicine docked in Site B was 1.6-2.6 Å from residues 2, 3, and 36 and 2.1-2.3 Å from residues 240 and 241.

In addition, in examining this region of the model in detail, we noted that for potential binding Site A several amino acid side chains formed a significant barrier between the S atoms of Cys-239 and Cys-354 and between colchicine and the S atom of Cys-354, and, conversely, for potential binding Site B these side chains formed a barrier between the bound colchicine and the S atom of Cys-239 (not shown). This aspect

of the model appears to be inconsistent with the ready cross linking of Cys-239 and Cys-354 in unpolymerized tubulin that occurs with EBI and the extensive inhibition of this intercysteine cross link formation by colchicine site drugs (26, 27). A significant conformational change may therefore occur in this region of the molecule when tubulin αβ-dimers polymerize into protofilaments.

When thiocolchicine was modeled into both sites by energy minimization, its position did not differ greatly from that of colchicine (Table III), consistent with its similar properties in binding to tubulin (28, 29). However, neither the C2-oxygen atom of 2CTC nor the C3-oxygen atom of 3CTC overlapped closely onto the C2- and C3- oxygen atoms, respectively, of colchicine when these molecules were positioned into Site A by energy minimization (Fig. 6). These shifts in position, however, did bring the reactive carbon atoms of the chloroacetyl groups relatively close to the two sulfur atoms, and the relative distances (shown in Table III) did correlate qualitatively with the relative reactivities of the two cysteines that we have observed. However, in this modeling method the amino acid side chains blocking access to the sulfur atom of Cys-354 were not substantially repositioned, so that this sulfur atom remains shielded from the chloroacetyl moieties and therefore should not react with them.

For potential binding Site B, the energy minimization modeling again showed negligible differences between colchicine and thiocolchicine (Table III), and in this binding site the C2-oxygen atom of 2CTC and the C3-oxygen atom of 3CTC more closely overlapped the corresponding oxygen atoms in colchicine (Fig. 7). However, the distances from the reactive carbons of both analogs to the sulfur atoms of Cys-239 and Cys-354 were all nearly identical (Table III). In addition, the Cys-239 sulfur atom remained shielded from the chloroacetyl moieties of 2CTC and 3CTC by amino acid side chains. In Fig. 7 we also show a portion of the α-tubulin peptide backbone, and it is notable how the B ring side chain of the colchicinoids has similar proximity to both tubulin subunits. Note also that in Site B the colchicinoid C ring overall is in closer contact with α-tubulin than with β-tubulin.

The poor overlap of the colchicinoids in Site A raised the question whether the iterative energy

minimizations had resulted in a significant change in drug conformation in fitting the compounds into this site. Comparison of in vacuo drug structures with those bound in Sites A and B demonstrated that this was not the case, as shown in Fig. 8 for colchicine, 2CTC, and 3CTC. Each compound is shown in energy minimized unbound conformations superimposed with the conformations of the compound bound in Site A and in Site B. There thus appears to be a shift of the 2CTC and 3CTC molecules relative to colchicine when energy minimized structures are bound in Site A.

This poor overlap of the C2 and C3 oxygen atoms of 2CTC and 3CTC with those of colchicine in Site A remained of concern to us. Our original rationale in undertaking this study was that the reactivity of the analogs with tubulin amino acid residues should provide insight into the location of these two oxygen atoms of colchicine in its binding site, having made the assumption that the molecular conformations of the unbound drugs would be almost identical. We therefore confirmed this original assumption by performing an energy minimization comparison of unbound 2CTC, 3CTC, thiocolchicine, and colchicine. Fig. 9 shows the nearly complete overlap of common structural elements, including the C2 and C3 oxygen atoms.

DISCUSSION

Our goal in this project was to attempt to model colchicine (and closely related structural analogs) into a theoretical binding site on tubulin. Ludueña and his colleagues have shown that EBI cross links β-Cys-239 and β-Cys-354 with high specificity. Longer carbon bridges between the two iodoacetamide moieties almost eliminated cross link formation, leading to the conclusion that the two S atoms were about 8 Å apart; and formation of this cross link is potently inhibited by colchicine site drugs (26, 27). In our previous study with [^{14}C]3CTC (5) we had concluded, largely based on results of formic acid digestion, that β-Cys-354 was the primary alkylation site in tubulin for this analog, although we also noted a minor reaction at β-Cys-239. In our current studies with [^{14}C]2CTC alkylation was greater at Cys-239 than at Cys-354, and this was most apparent following digestion with CNBr or endoproteinases. The reactivity of 2CTC with Cys-239 was apparently reduced following formic acid digestion compared to other methods, and this caused us to re-evaluate the reactivity of 3CTC. The apparent relative reactivity of the two cysteine residues with both colchicine analogs is summarized in Table II, which summarizes both the autoradiograms shown in Fig. 2 and average results obtained from multiple formic acid and CNBr digestions.

It is clear that with both analogs the cross link to Cys-239 is relatively labile in the formic acid digestion, and we assume that this is due to the prolonged 4-day incubation. Since the 1-day incubation with CNBr occurs at nearly the same formic acid concentration, it is not unreasonable to extrapolate the ratio of Cys-239 to Cys-354 modification back to zero time by drawing straight lines through the 1- and 4-day time points. When this is done (not shown), one obtains a 77/23 distribution of Cys-239/Cys-354 modification for 2CTC, and a 27/73 distribution for 3CTC.

The obvious conclusion from these results is that the C2-oxygen atom of colchicinoids is closer to Cys-239 and the C3-oxygen atom is closer to Cys-354, assuming that in all cases the covalent reactions involve a nucleophilic attack of electrons of the cysteine S atom on the chloroacetyl group, with displacement of the chlorine atom. There are additional quantitative aspects of the two drug-tubulin

interactions that should be considered.

Our initial observations that 3CTC reacted covalently with tubulin virtually as fast as it bound while the 2CTC covalent reaction lagged behind binding (24) was reflected in the stoichiometry of covalent drug bound to β-tubulin isolated by preparative decylagarose chromatography for the sequencing studies. The average value for [^{14}C]3CTC was 0.32 mol drug/mol tubulin (5) and for [^{14}C]2CTC was 0.17 mol/mol (see Experimental Procedures). Multiplying these values by the zero time distributions, one can conclude that with 2CTC there is 0.12 mol/mol of [^{14}C]-labeled adduct at Cys-239 and only 0.05 mol/mol at Cys-354; and with 3CTC 0.11 mol/mol at Cys-239 as compared with 0.21 mol/mol at Cys-354 (Table II).

There is, however, a further complication in these calculations, in that there are four isotypes of β-tubulin in bovine brain. One of these, β_{III}, has a serine residue instead of cysteine at position 239, and this isotype represents about 25% of total brain β-tubulin (30). Assuming that 2CTC and 3CTC bind equivalently to all isotypes and that the covalent reactions occur equally (however, see below), then the relative reactivity of Cys-354 needs to be corrected for the absence of Cys-239 in β_{III}-tubulin. This reduces the relative stoichiometry of 2CTC cross linked to Cys-354 to 0.04 and of 3CTC to 0.16.

Let us assume these relative stoichiometries correlate with distances between the C2- and C3-oxygen atoms and the cysteine sulfur atoms. If this is true, then i) the C2-oxygen is about 3 times as far from the Cys-354 sulfur as the Cys-239 sulfur, ii) the C3-oxygen about 50% further from the Cys-239 sulfur as the Cys-354 sulfur, iii) both oxygens are nearly equidistant from the Cys-239 sulfur, and iv) the C2-oxygen should be about 4 times further than the C3-oxygen from the Cys-354 sulfur.

However, when we used energy minimization programs to model colchicine and the analogs into the electron crystallographic model of β-tubulin (1), we obtained entirely unexpected results, largely inconsistent with the above predictions (see Table III). For potential binding Site A, colchicine and thiocolchicine were relatively close to the components of the two peptide sequences (one of which included Cys-239) that reacted with [^{3}H]colchicine following direct photoaffinity labeling (9), but Cys-354 was

somewhat more distant, and direct access to its S atom was blocked by several amino acid side chains. Moreover, the A rings of 2CTC and 3CTC differed significantly in location from the A rings of colchicine and thiocolchicine following modeling by energy minimization into Site A (despite the near-identical biochemical properties of the four colchicinoids and the equivalent conformations of the unbound molecules). This could indicate a remarkable degree of plasticity in the colchicine site in its accommodation of structurally similar compounds, and the larger Site A could probably accommodate the significant structural variability in ligands known to bind at the colchicine site. Finally, the reactive carbon atoms of 2CTC and 3CTC were relatively close to the sulfur atoms of Cys-239 and Cys-354.

For Site B, colchicine remains in close proximity to components of the peptides cross linked to the drug in the direct photoaffinity study (9), and Site B is consistent with the published subunit reactivity in the analog photoaffinity studies (7, 8). Moreover, in Site B there is greater similarity in the binding footprints of 2CTC and 3CTC to that of colchicine. In this model, however, the reactive carbon atoms in the chloroacetyl groups are more distant from the cysteine sulfur atoms, and it is now the sulfur atom of Cys-239 that is shielded by amino acid side chains. In addition, binding Site B is smaller and would appear to be more restrictive in the structural diversity it would tolerate in potential ligands.

Alternatively, the colchicine site may undergo significant conformational change from αβ-dimer to protofilament, suggested by the failure of polymer to bind colchicine (31, 32), the failure of dimer to bind paclitaxel (33), and the ease with which EBI cross links Cys-239 and Cys-354 (26, 27) in the dimer.

The above discussion, as well as the electron crystallographic model (1), assumes no differences between the different isotypes of β-tubulin. This is probably not justified, since immunopurified tubulin containing each of the three major brain β-isotypes (β_{II}, 58%; β_{III}, 25%; and β_{IV}, 13%; data from ref. 30) reacts with colchicine (34, 35) and colchicine analogs (36, 37) with different kinetics and different affinities. It is even possible, although probably far-fetched, that only one isotype (e.g., tubulin containing β_{III}) reacts at Cys-354 and another (e.g., β_{II}) reacts at Cys-239.

Our findings with [^{14}C]2CTC and [^{14}C]3CTC allow us to make one additional speculation. Although thiocolchicine (28, 29) and 2CTC and 3CTC (24) bind to tubulin more rapidly than colchicine, these analogs share with colchicine a relatively slow binding reaction. Since, within the measured time frame, 3CTC reacts covalently with β-tubulin as fast as it binds, it is likely that the spatial relationship between Cys-239 and Cys-354 and the C3-substituent is relatively static, at least under the reaction conditions used. However, the covalent reaction of 2CTC with β-tubulin increases with time of incubation, lagging well behind the binding reaction. Perhaps this is due to the changes in the conformation of drug or tubulin previously documented by other workers (6). This implies that such conformational changes would include moving the C2-substituent closer to Cys-239 and/or Cys-354.

Finally, we should point out an apparent limitation to molecular modeling approaches that was revealed as our studies progressed. Although there is no evidence for more than one high affinity site for colchicine binding to tubulin (6), we initially observed four candidate binding sites that merited exploration. Only two could be eliminated because they were inconsistent with our biochemical data, and neither remaining site is entirely satisfactory without postulating conformational changes that might occur as tubulin alternates between the polymerized and unpolymerized states and/or responds to binding of the ligand. While modeling methods similar to those we used here have found widespread application in areas of rational drug discovery and design, it is clear that additional validation steps for predicted docking geometries will be necessary. Computational results can be especially valuable, however, in providing a scaffold for hypothesis generation and data interpretation.

FOOTNOTES

[1]The abbreviations used are: 2CTC, 2-chloroacetyl-2-demethylthiocolchicine; [^{14}C]2CTC, 2-(chloromethyl-[^{14}C]carbonyl)-2-demethylthiocolchicine; 3CTC, 3-chloroacetyl-3-demethylthiocolchicine; [^{14}C]3CTC, 3-(chloromethyl-[^{14}C]carbonyl)-3-demethylthiocolchicine; NEM, *N*-ethylmaleimide; HPLC, high-performance liquid chromatography; SDS-PAGE, sodium dodecyl sulfate-polyacrylamide gel electrophoresis; PVDF, polyvinylidene difluoride; EBI, *N,N'*-ethylene(bis)iodoacetamide; CNBr, cyanogen bromide; Tricine, *N*-[2-hydroxy-1,1-bis(hydroxylmethyl)ethyl]glycine; EP-GC, endoproteinase Glu-C.

ACKNOWLEDGEMENT

This project has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under contract No. N01-CO-56000. The content of this publication does not necessarily reflect the views of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the United States government.

REFERENCES

1. Nogales, E., Wolf, S. G., and Downing, K. H. (1998) *Nature (London)* **391,** 199-203
2. Downing, K. H., and Nogales, E. (1998) *Eur. Biophys. J.* **27,** 431-436
3. Wolff, J., Knipling, L., Cahnmann, H. J., and Palumbo, G. (1991) *Proc. Natl. Acad. Sci. U.S.A.* **88,** 2820-2824
4. Rao, S., Krauss, N. E., Heerding, J. M., Swindell, C. S., Ringel, I., Orr, G. A., and Horwitz, S. B. (1994) *J. Biol. Chem.* **269,** 3132-3134
5. Bai, R., Pei, X.-F., Boyé, O., Getahun, Z., Grover, S., Bekisz, J., Nguyen, N. Y., Brossi, A., and Hamel, E. (1996) *J. Biol. Chem.* **271,** 12639-12645
6. Hastie, S. B. (1991) *Pharmac. Ther.* **51,** 377-401
7. Williams, R. F., Mumford, C. L., Williams, G. A., Floyd, L. J., Aivaliotis, M. J., Martinez, R. A., Robinson, A. K., and Barnes, L. D. (1985) *J. Biol. Chem.* **260,** 13794-13802
8. Floyd, L. J., Barnes, L. D., and Williams, R. F. (1989) *Biochemistry* **28,** 8515-8525
9. Uppuluri, S., Knipling, L., Sackett, D. L., and Wolff, J. (1993) *Proc. Natl. Acad. Sci. U.S.A.* **90,** 11598-11602
10. Hamel, E., and Lin, C. M. (1984) *Biochemistry* **23,** 4173-4184
11. Boyé, O., Getahun, Z., Grover, S., Hamel, E., and Brossi, A. (1992) *J. Labelled Compounds Radiopharm.* **33,** 293-299
12. Bai, R., Lin, C. M., Nguyen, N. Y., Liu, T.-Y., and Hamel, E. (1989) *Biochemistry* **28,** 5606-5612
13. Landon, M. (1977) *Meth. Enzymol.* **47,** 145-149
14. Bewley, C. A., Gustafson, K. R., Boyd, M. R., Covell, D. G., Bax, A., Clore, G. M., and Gronenborn, A. M. (1998) *Nature Struct. Biol.* **5,** 571-578
15. Caffrey, M., Cai, M., Kauffman, J., Stahl, S., Wingfield, P. T., Covell, D. G., Gronenborn, A. M., and Clore, G. M. (1998) *EMBO J.* **17,** 4572-4584

16. Young, B. L., Jernigan, R. L., and Covell, D. G. (1994) *Prot. Sci.* **3,** 717-729

17. Covell, D. G., Jernigan, R. L., and Wallqvist, A. (1998) *J. Mol. Struct.* **423,** 93-100

18. Wallqvist, A., Jernigan, R. L., and Covell, D. G. (1995) *Prot. Sci.* **4,** 1881-1903

19. Wallqvist, A., and Covell, D. G. (1996) *Prot. Struc. Func. Genet.* **25,** 403-419

20. Rice, W. G., Turpin, J. A., Schaeffer, C. A., Graham, L., Clanton, D., Buckheit, Jr., R. W., Zaharevitz, D., Summers, M. F., Wallqvist, A., and Covell, D. G. (1996) *J. Med. Chem.* **39,** 3606-3616

21. Rice, W. G., Turpin, J. A., Clanton, D., Buckheit, Jr., R. W., Summers, M. F., McDonnell, N., DeGuzman, R. N.,Wallqvist, A., Covell, D. G., Zalkow, L., Bader, J. P., Sausville, E. A., and Haugwitz, R. D. (1997) *Nature Med.* **3,** 341-345

22. Huang, M., Turpin, J., Maynard, A., Graham, L., Janini, G. M., Covell, D. G., and Rice, W. G. (1998) *J. Med. Chem.* **41,** 1371-1381

23. Turpin, J. A., Song, Y., Inman, J. K., Huang, M., Wallqvist, A., Maynard, A., Covell, D. G., Rice, W. A., and Appella, E. (1999) *J. Med. Chem.* **42,** 67-86

24. Grover, S., Boyé, O., Getahun, Z., Brossi, A., and Hamel, E. (1992) *Biochem. Biophys. Res. Commun.* **187,** 1350-1358

25. Bai, R., Ewell, J. B., Nguyen, N., and Hamel, E. (1999) *J. Biol. Chem.* **274,** 12710-12714

26. Ludueña, R. F., and Roach, M. C. (1981) *Biochemistry* **20,** 4444-4450

27. Ludueña, R. F., and Roach, M. C. (1991) *Pharmac. Ther.* **49,** 133-152

28. Kang, G.-J., Getahun, Z., Muzaffar, A., Brossi, A., and Hamel, E. (1990) *J. Biol. Chem.* **265,** 10255-10259

29. Chabin, R. M., and Hastie, S. B. (1989) *Biochem. Biophys. Res. Commun.* **161,** 544-550

30. Banerjee, A., Roach, M. C., Wall, K. A., Lopata, M. A., Cleveland, D. W., and Ludueña, R. F. (1988) *J. Biol. Chem.* **263,** 3029-3034

31. Lee, Y. C., Samson, F. E., Jr., Houston, L. L., and Himes, R. H. (1974) *J. Neurobiol.* **5,** 317-330

32. Wilson, L., and Meza, I. (1973) *J. Cell Biol.* **58,** 709-719

33. Parness, J., and Horwitz, S. B. (1981) *J. Cell Biol.* **91,** 479-487

34. Banerjee, A., and Ludueña, R. F. (1991) *J. Biol. Chem.* **266,** 1689-1691

35. Banerjee, A., and Ludueña, R. F. (1992) *J. Biol. Chem.* **267,** 13335-13339

36. Banerjee, A., Kasmala, L. T., Hamel, E., Sun, L., and Lee, K.-H. (1999) *Biochem. Biophys. Res. Commun.* **254,** 334-337

37. Banerjee, A., D'Hoore, A., and Engelborghs, Y. (1994) *J. Biol. Chem.* **269,** 10324-10329

38. Yeh, H. J. C., Chrzanowska, M., and Brossi, A. (1988) *FEBS Lett.* **229,** 82-86

Table I

Amino acid sequence analysis of the major radiolabeled peptides derived from tubulin cross linked to [^{14}C]2CTC

Peptides sequenced are those shown in Fig. 2. X indicates the cysteine positions in the sequences, and XX the tryptophan residues. These residues cannot be identified following Edman degradation, and in the actual sequence studies no definitive amino acid assignment could be made. Other positions for which no definitive assignment could be made are indicated by the absence of an entry. Sequencing was performed by automated Edman degradation on an Applied Biosystems model 494A Protein Sequenator. Identification of phenylthiohydantoin amino acid derivatives was performed with an Applied Biosystems model 140C Microgradient System and model 785A Programmable Absorbance Detector.

	Formic acid peptides		CNBr peptides		EP-GC peptides	
Cycle #	A2	A3	a	b	a	b
1	Pro-32	Pro-305	Ser-234	Leu-331	XX:Trp-344	Thr-199
2	Thr	Arg	Gly	Asn	Ile	Tyr
3	Gly	His	Val	Val	Pro	X
4	Ser	Gly	Thr	Gln	Asn	Ile
5	Tyr	Arg	Thr	Asn	Asn	Asp
6	His	Tyr	X	Lys	Val	Asn
7	Gly	Leu	Leu	Asn	Lys	Glu
8	Asp	Thr			Thr	Ala
9	Ser	Val	Phe		Ala	Leu
10	Asp	Ala	Pro	Tyr	Val	Tyr
11	Leu			Phe	X	Asp
12	Gln			Val	Asp	Ile
13	Leu			Glu		X
14	Glu			XX		Phe
15	Arg			Ile		Arg
16				Pro		
17				Asn		

Table II

Relative apparent reactivity of Cys-239 and Cys-354 with 2CTC and 3CTC

Densitometry data:

	[^{14}C]2CTC	[^{14}C]3CTC
Digestion method	[^{14}C] in Cys-239 peptide/[^{14}C] in Cys-354 peptide	
Formic acid (Fig. 2A)	0.7	0.1
Formic acid (overall)	1.2 ± 0.5	0.18 ± 0.1
CNBr (Fig. 2B)	2	0.4
CNBr (overall)	1.8 ± 0.4	0.43 ± 0.02

Calculated[a] stoichiometry:

	Total	At Cys-239	At Cys-354
Analog	pmol ligand/pmol β-tubulin (corrected for β_{III})		
2CTC	0.17	0.12	0.05 (0.04)
3CTC	0.32	0.11	0.21 (0.16)

[a]See text (Discussion).

Table III

Calculated intermolecular distances from best-fit models

	To Cys-239 S atom	To Cys-354 S atom
	Distance in Å	
Potential binding Site A:		
Colchicine[a]		
From C2 O atom	5.6	9.0
From C3 O atom	8.3	11.7
Thiocolchicine[a]		
From C2 O atom	5.6	9.2
From C3 O atom	8.3	11.9
2CTC[a]		
From C19	5.1	5.4
3CTC[a]		
From C20	5.9	4.7
Potential binding Site B:		
Colchicine[a]		
From C2 O atom	11.1	9.4
From C3 O atom	10.0	11.1
Thiocolchicine[a]		
From C2 O atom	11.1	9.4
From C3 O atom	10.0	11.1
2CTC[a]		
From C19	13.8	10.9
3CTC[a]		
From C20	11.3	11.2

[a]Numbering as shown in Fig. 1.

FIGURE LEGENDS

Fig. 1. Structures of colchicine, thiocolchicine, 2CTC, and 3CTC. In the diagrams of 2CTC and 3CTC, the radiolabeled carbons are indicated by the arrows. The compounds are shown in the preferred *a*S-7S configuration (38).

Fig. 2. Autoradiograms of sequenced peptides. A. Formic acid digestion. Tracks I and II display, respectively, the Coomassie blue stained PVDF electroblot and its autoradiogram of a digest of β-tubulin cross linked to [^{14}C]2CTC following SDS-PAGE. Track III displays the PVDF electroblot of a digest of β-tubulin cross linked to [^{14}C]3CTC. B. CNBr digestion. Tracks I and II display, respectively, the autoradiograms of digests of β-tubulin cross linked to [^{14}C]2CTC or [^{14}C]3CTC following SDS-PAGE. C. EP-GC digestion. The track displays the autoradiogram of a digest of β-tubulin cross linked to [^{14}C]2CTC following SDS-PAGE.

Fig. 3. Radiolabel recovered following sequential Edman degradation of a CNBr digest (A) or a trypsin digest (B) of β-tubulin cross linked to [^{14}C]2CTC.

Fig. 4. Modeling of colchicine into the electron crystallographic model of αβ-tubulin (1). The α-tubulin subunit is shown in olive, the β-subunit in gray. In the colchicine diagrams carbon atoms are shown in green, oxygen in red, nitrogen in blue, and hydrogens (if shown) in white. The left-hand image represents a solvent accessible surface rendition of the αβ-tubulin structure, showing colchicine in the two alternative proposed binding sites. The right-hand image represents a ribbon diagram of the polypeptide backbone of the αβ-tubulin structure, with colchicine in the two alternative proposed binding sites. Fig. 5. Modeling of colchicine into the electron crystallographic model of αβ-tubulin (1), showing the relationship of the bound colchicine to peptides containing residues 1-31 (white) and 214-241 (orange) and to Cys-239 (S atom, light blue) and Cys-354 (S atom, yellow). In the cysteine and colchicine structures carbon atoms are shown in green, oxygen in red, and nitrogen in blue (hydrogens not shown). A. Colchicine bound in Site A. B. Colchicine bound in Site B.

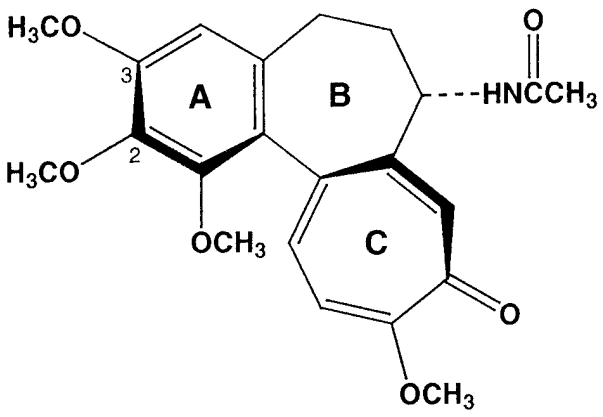
Fig. 6. Modeling of colchicine, 2CTC, and 3CTC into Site A. The β-tubulin polypeptide backbone of the colchicine-bound model is indicated by the continuous dark green thin strand. Shifts in this backbone were not extensive, as indicated by the reiterated positions of Cys-239 and Cys-354. In the cysteine residues, the oxygen atoms are red, the nitrogen atoms blue, the carbon atoms green, the sulfur atoms as described below, and the hydrogen atoms not shown. All atoms in colchicine are in magenta, except that hydrogen atoms are not shown and the C2 and C3 oxygen atoms, as labeled, are in orange. The sulfur atoms in the two cysteine residues in the colchicine-bound structure are also shown in orange. All atoms in 2CTC are in light blue, except that the C2-oxygen atom is shown in garnet (arrow), the reactive C19 atom (see Fig. 1) is in white, and hydrogen atoms are not shown. The sulfur atoms in the two cysteine residues in the 2CTC-bound structure are also shown in white. All atoms in 3CTC are in dark blue, except that the C3-oxygen atom is shown in yellow (open arrow), the reactive C20 atom (see Fig. 1) is in black, and hydrogen atoms are not shown. The sulfur atoms in the two cysteine residues in the 3CTC-bound structure are also shown in black.

Fig. 7. Modeling of colchicine, 2CTC, and 3CTC into Site B. The β-tubulin polypeptide backbone of the colchicine-bound model is indicated by the continuous dark green thin strand, and that of α-tubulin by the yellow strand. Shifts in this backbone were not extensive, as indicated by the reiterated positions of Cys-239 and Cys-354. In the cysteine residues, the oxygen atoms are red, the nitrogen atoms blue, the carbon atoms green, the sulfur atoms as described below, and the hydrogen atoms not shown. All atoms in colchicine are in magenta, except that hydrogen atoms are not shown and the C2 and C3 oxygen atoms, as labeled, are in orange. The sulfur atoms in the two cysteine residues in the colchicine-bound structure are also shown in orange. All atoms in 2CTC are in light blue, except that the C2-oxygen atom of 2CTC is shown in garnet (arrow), the reactive C19 atom (see Fig. 1) is in white, and hydrogen atoms are not shown. The sulfur atoms in the two cysteine residues in the 2CTC-bound structure are also shown in white. All atoms in 3CTC are in dark blue, the C3-oxygen atom of 3CTC is shown in yellow (open arrow), the reactive

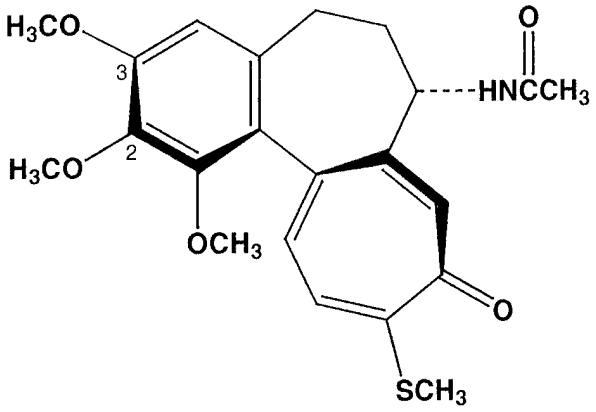
C20 atom (see Fig. 1) is in black, and hydrogen atoms are not shown. The sulfur atoms in the two cysteine residues in the 3CTC-bound structure are also shown in black.

Fig. 8. Superposition of unbound drugs, as indicated, with drugs bound in Site A and Site B. In each case the unbound drug is shown with carbon atoms in green, oxygen in red, nitrogen in blue, sulfur in yellow, chloride in light green, and hydrogen not shown, that bound in Site A enitrely in white, and that bound in Site B entirely in blue.

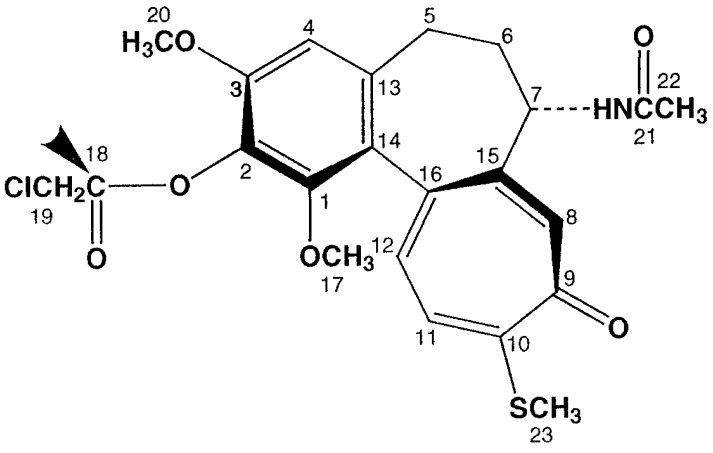
Fig. 9. Superposition of common structural elements of colchicine, thiocolchicine, 2CTC, and 3CTC when the unbound compounds are subjected to energy minimization modeling. Carbon atoms are shown in green, oxygen in red, nitrogen in blue, and sulfur in yellow. Hydrogen atoms are not shown.



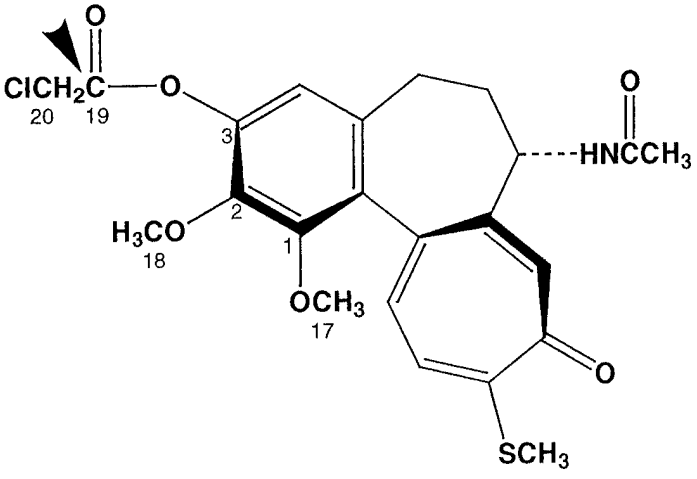
Colchicine



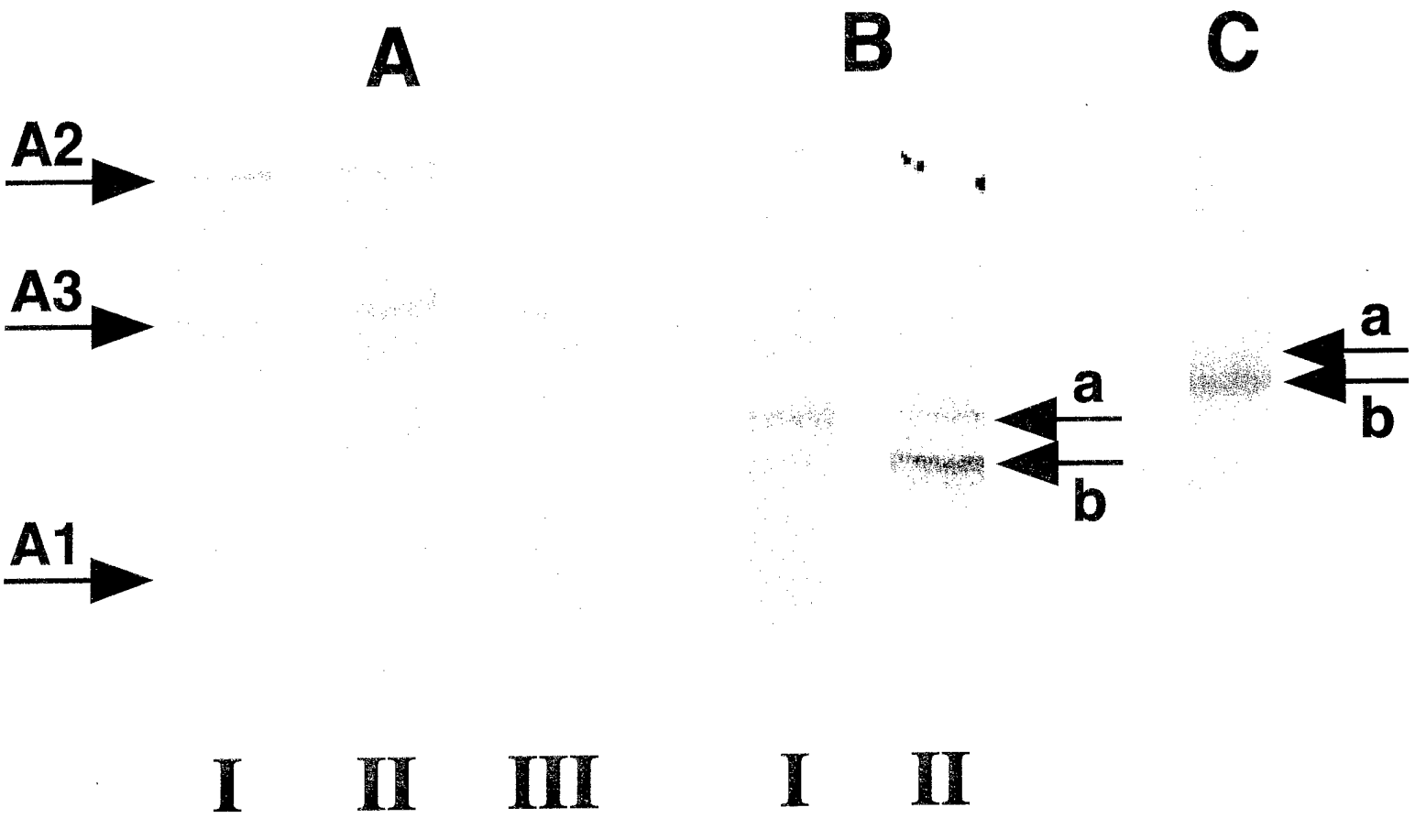
Thiocolchicine

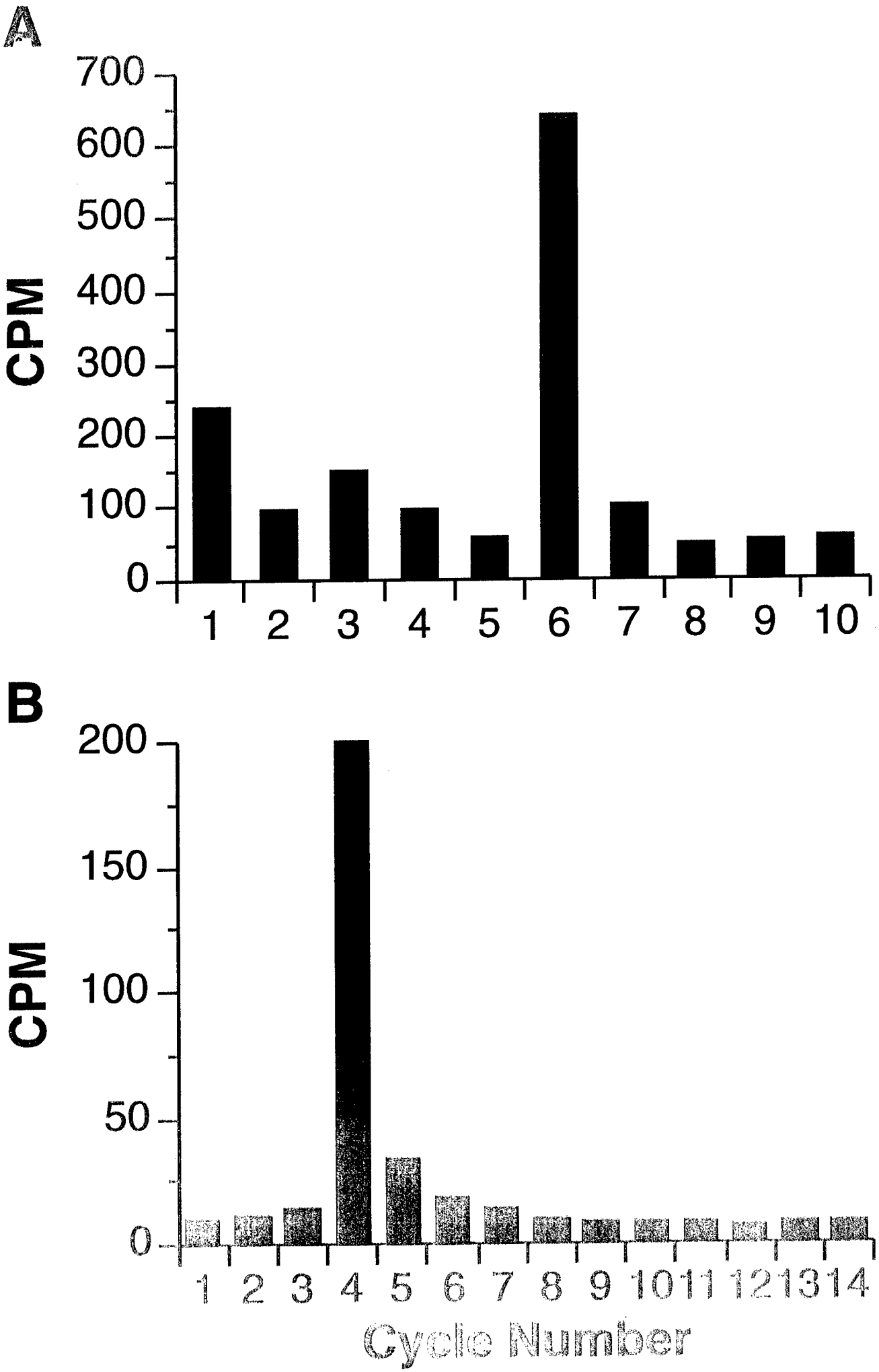


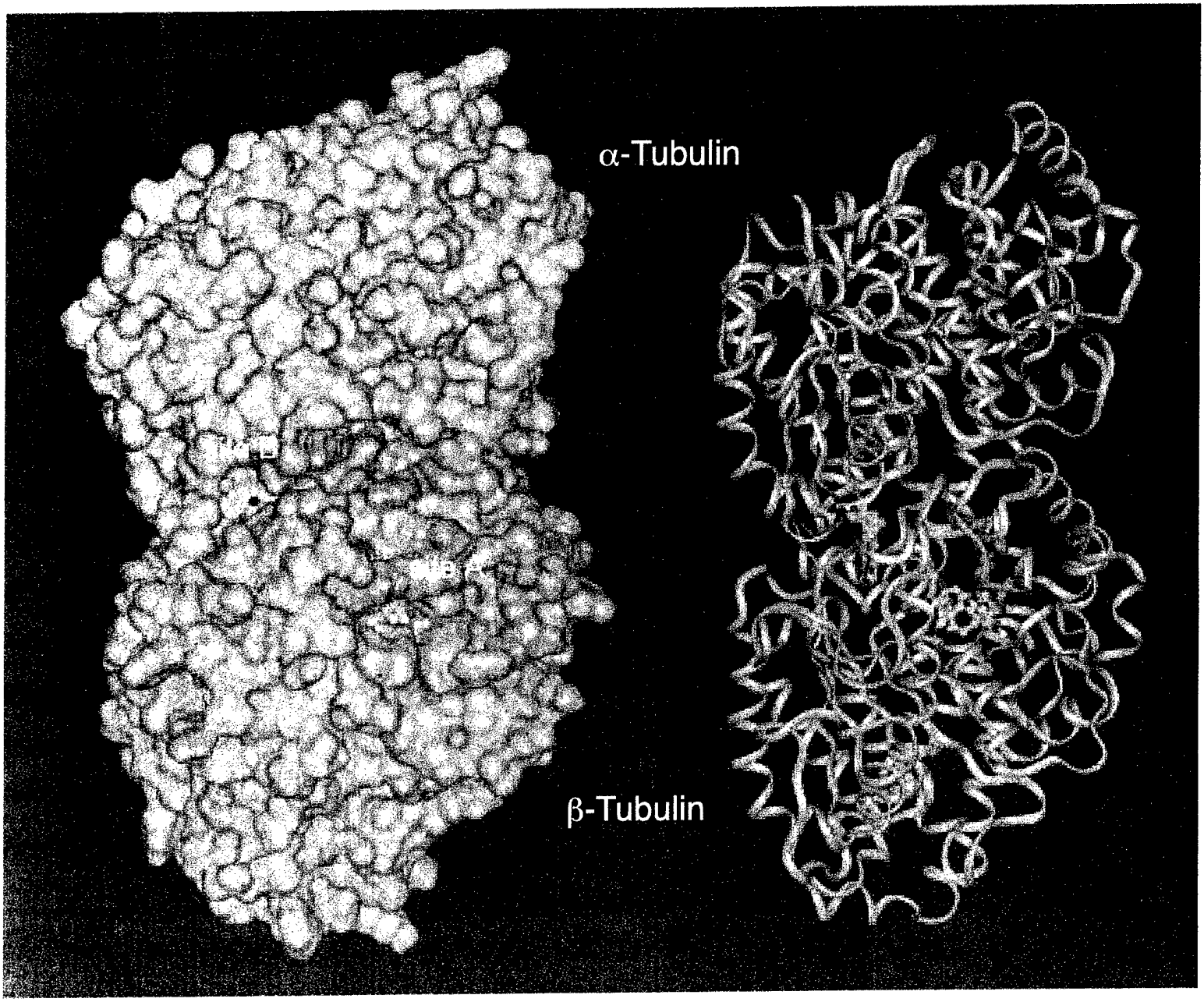
2CTC

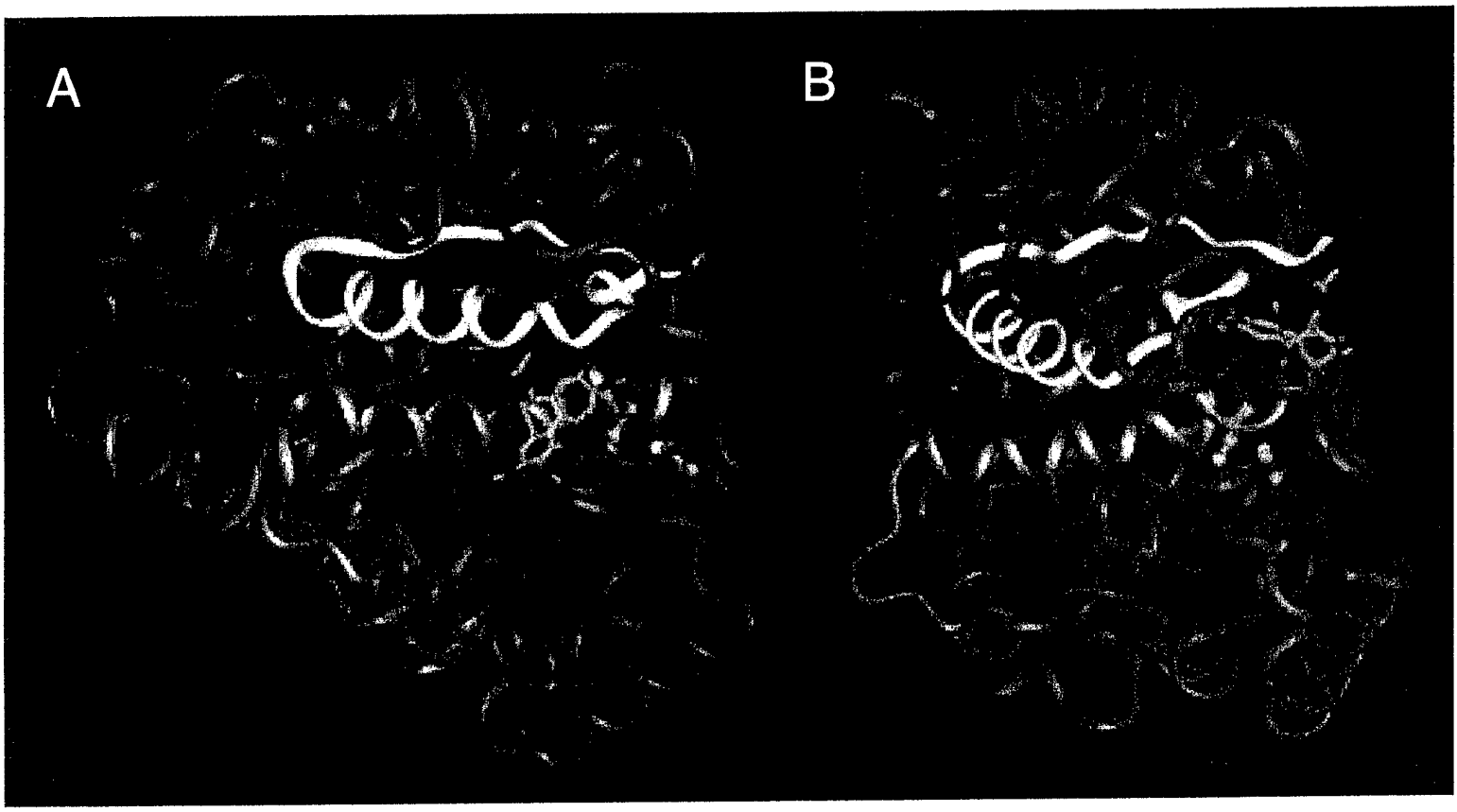


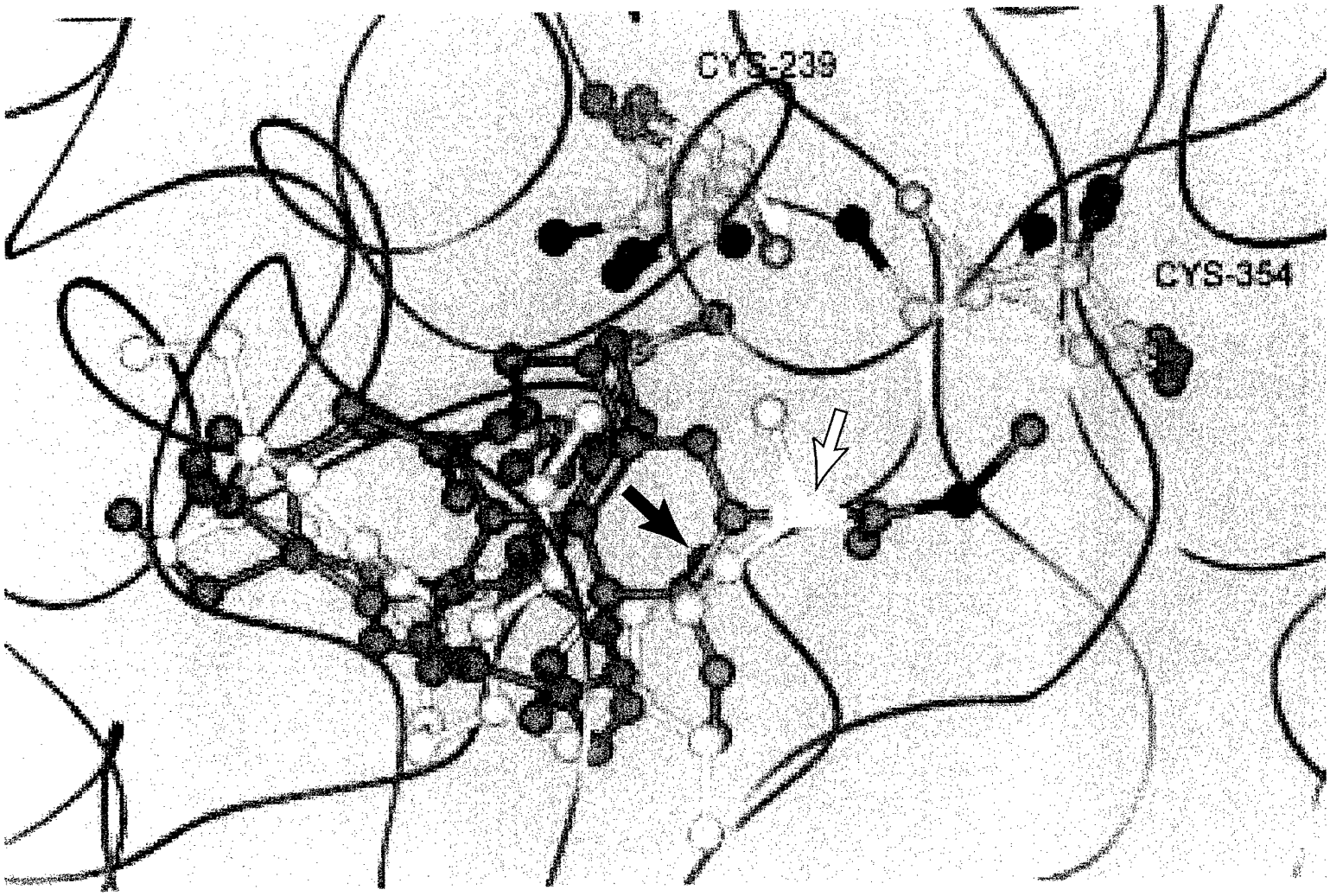
3CTC

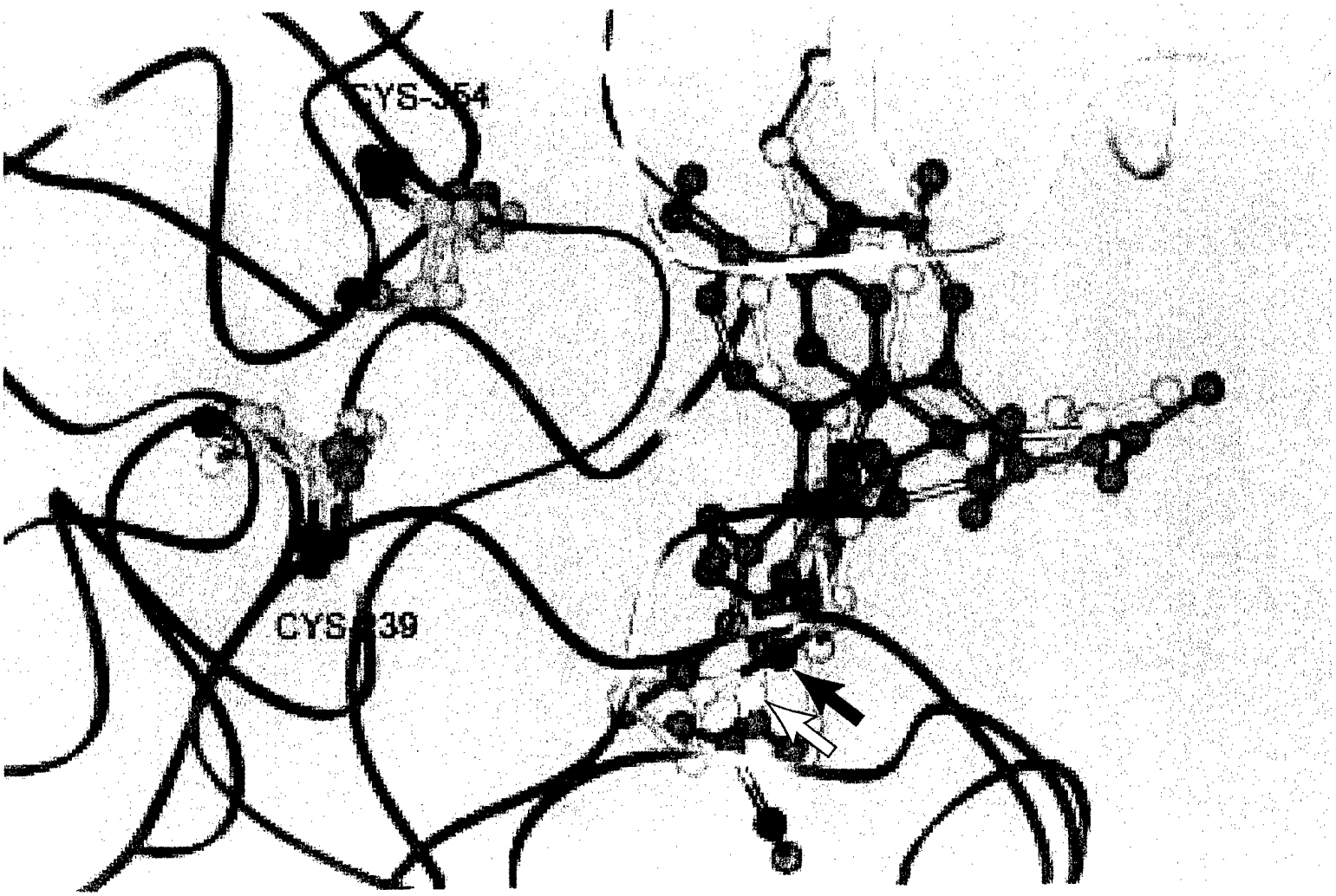

A
B
C
A2
A3
A1
a
b
a
b
I
II
III
I
II

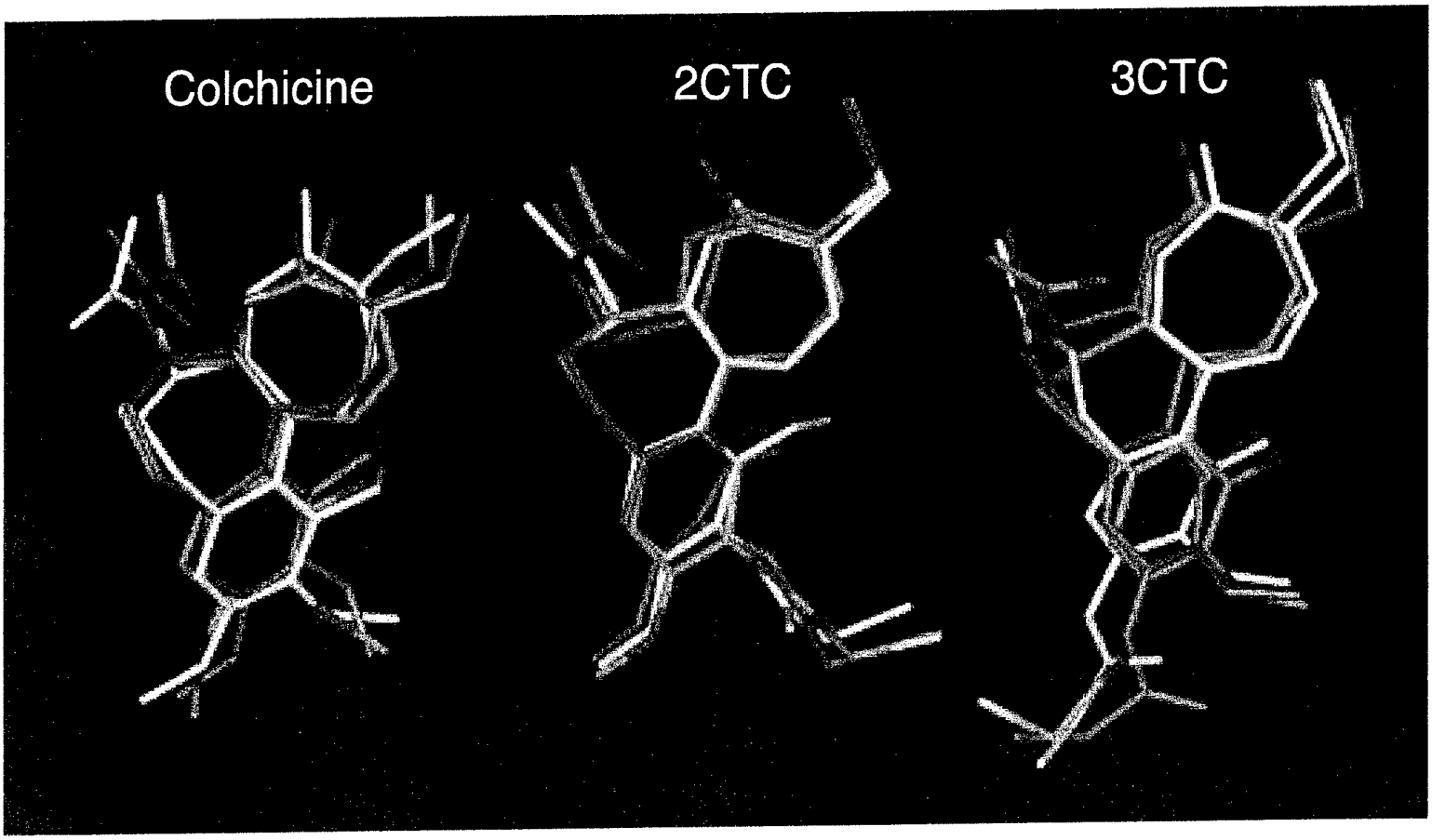

A
CPM
700
600
500
400
300
200
100
0
1 2 3 4 5 6 7 8 9 10
B
CPM
200
150
100
50
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14
Cycle Number

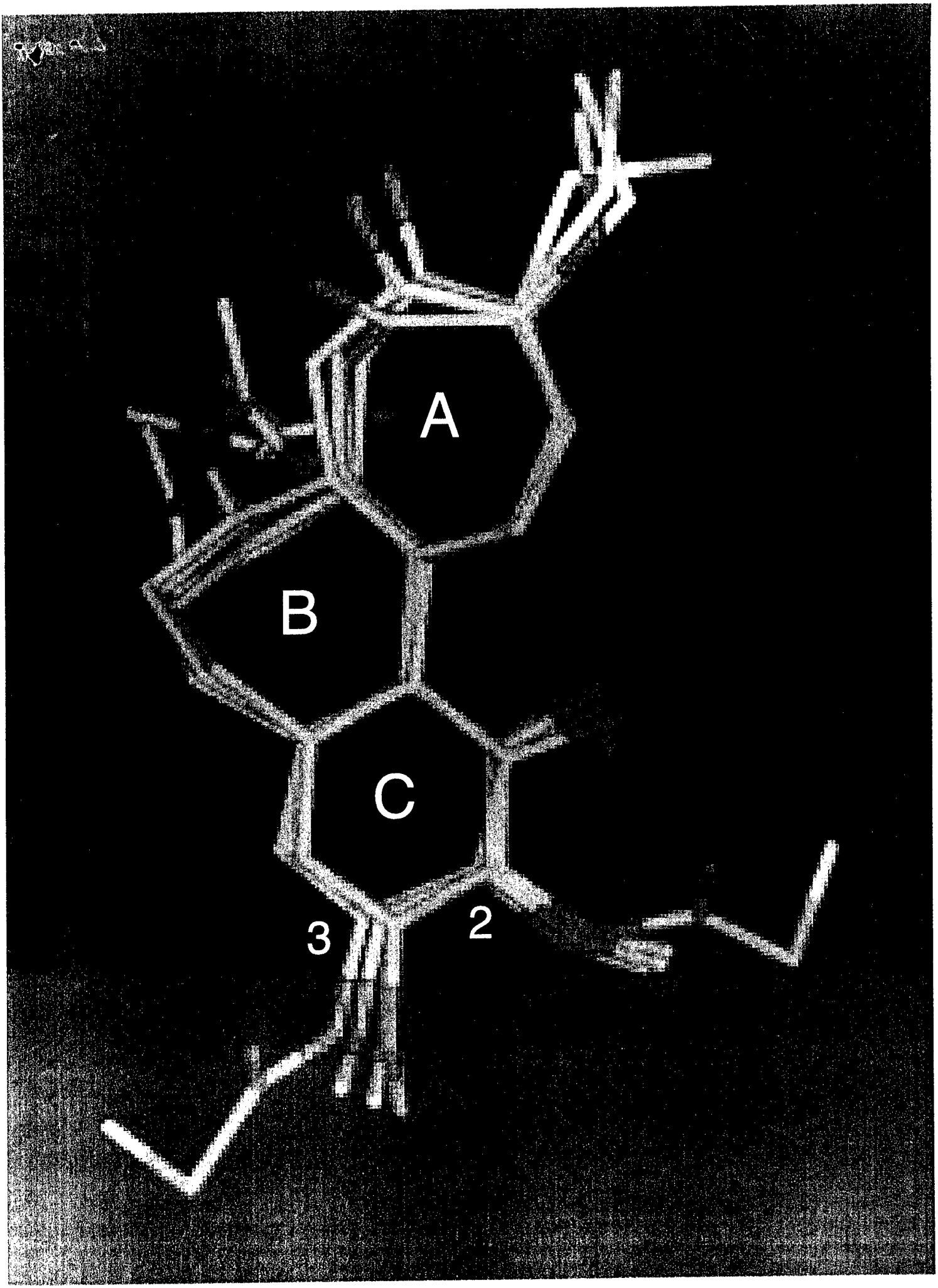

α-Tubulin
β-Tubulin


A
B


CYS-239
CYS-354


CYS-364
CYS-439


Colchicine
2CTC
3CTC


A
B
C
3
2