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.

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})

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 physico-chemical 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 $\mathbf{A}$ yields three matrices Λ, $\mathbf{U}$ and $\mathbf{V}$ which comprise (i) the singular eigenvalues λ_i of $\mathbf{A}$, organized in ascending order in the diagonal matrix Λ; (ii) the orthonormal transformation matrix $\mathbf{U}$ 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. $\mathbf{A}$ can thus be decomposed, hence the term 'singular value decomposition', into the product of these three matrices

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

where the subscripts denote the dimensions of the two-dimensional matrices and the superscript T indicates the transpose. In general, the columns of $\mathbf{A}$ 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 $\mathbf{A}$, 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 $\mathbf{A}$ 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 $\mathbf{A}$ 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	morpholinodoxorubicin	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 chloro-alkylating 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 topo-isomerase 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
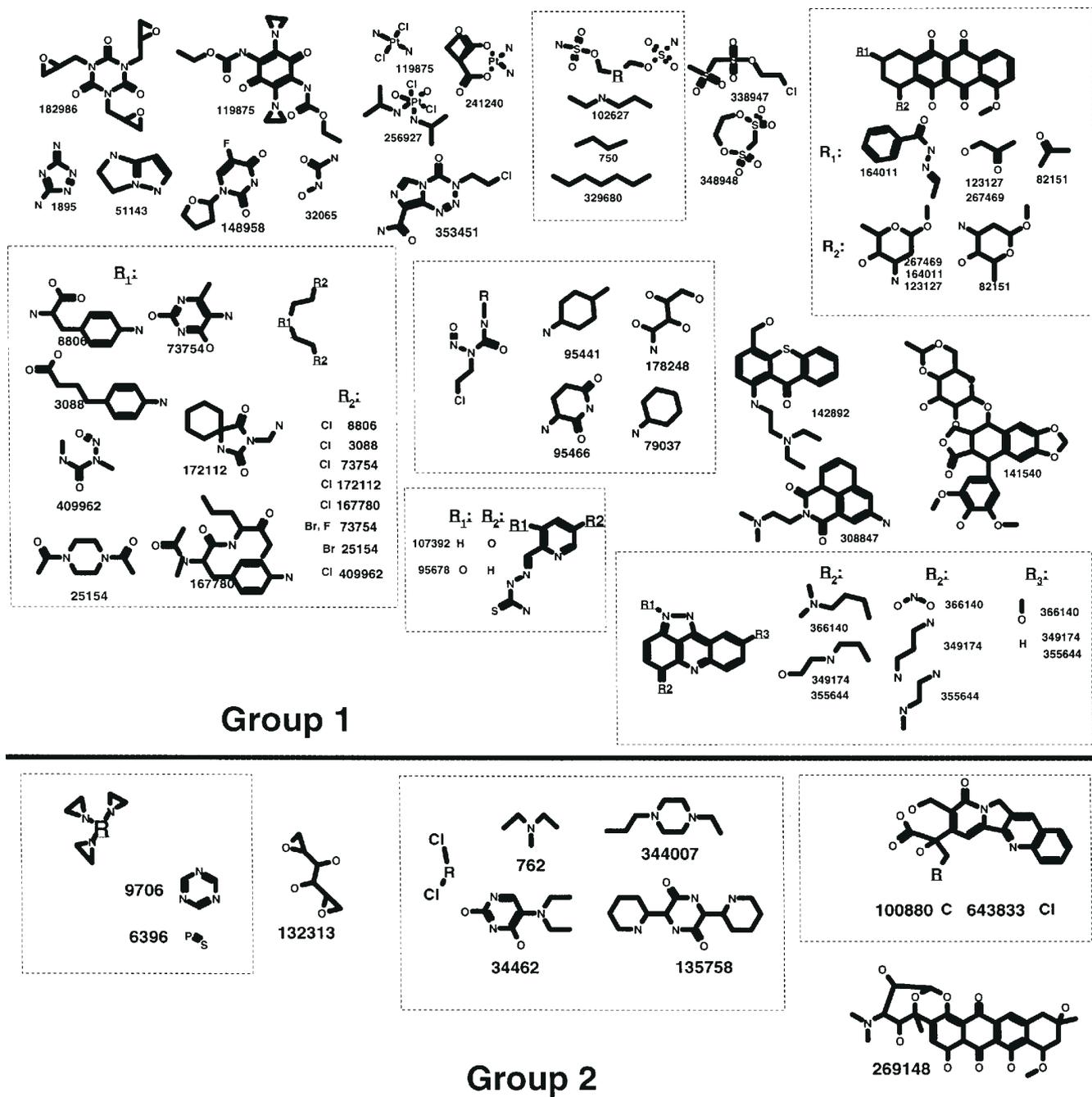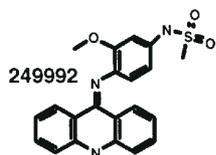
Group 1

Group 2

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
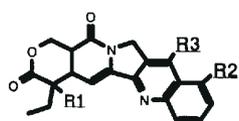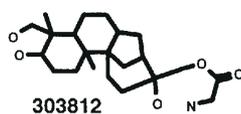
Group 3

249992

26980
56410 C
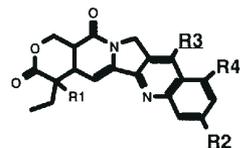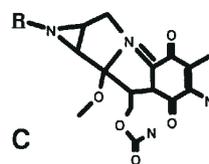
	R1:	R2:	R3		R1:	R2:	R4:
94600	O			603071	O		N
107124	O	O					
295501	O		o—	606172	O		
95382				606985			
606173	O			610457			
364830				610458			
374028				618939			

Group 4

	R1:	R2:	R3:
249910	O		Cl
629971	O	N	
606497			
606499			
610456			
610459			

303812

Group 7

Group 5

	R1:	R2:	R3:
176323	C		
295500			

Group 6

122819

301739

Figure 1 (continued)

Group 8

Group 9

609395
too complex

Group 10

Group 11

Group 12

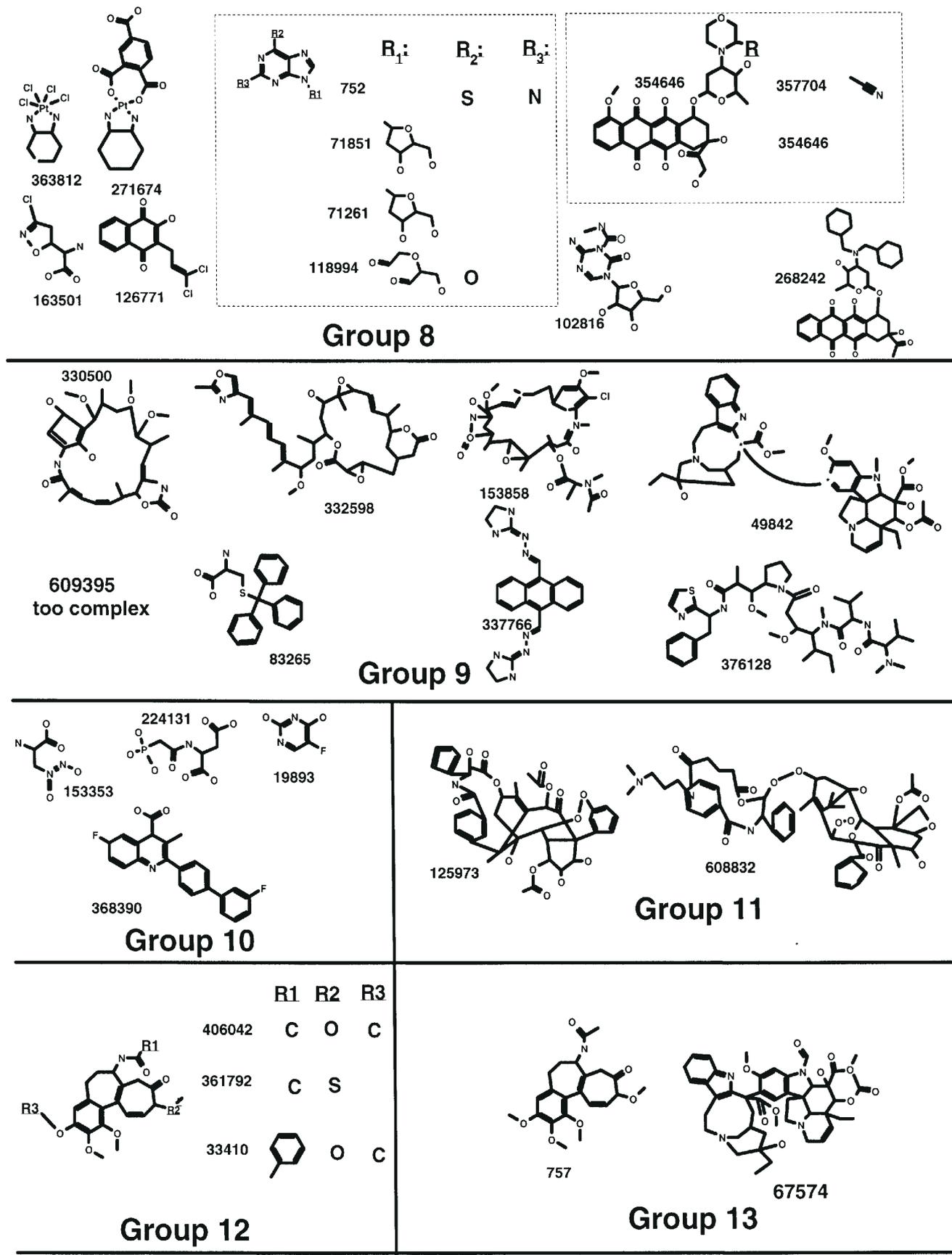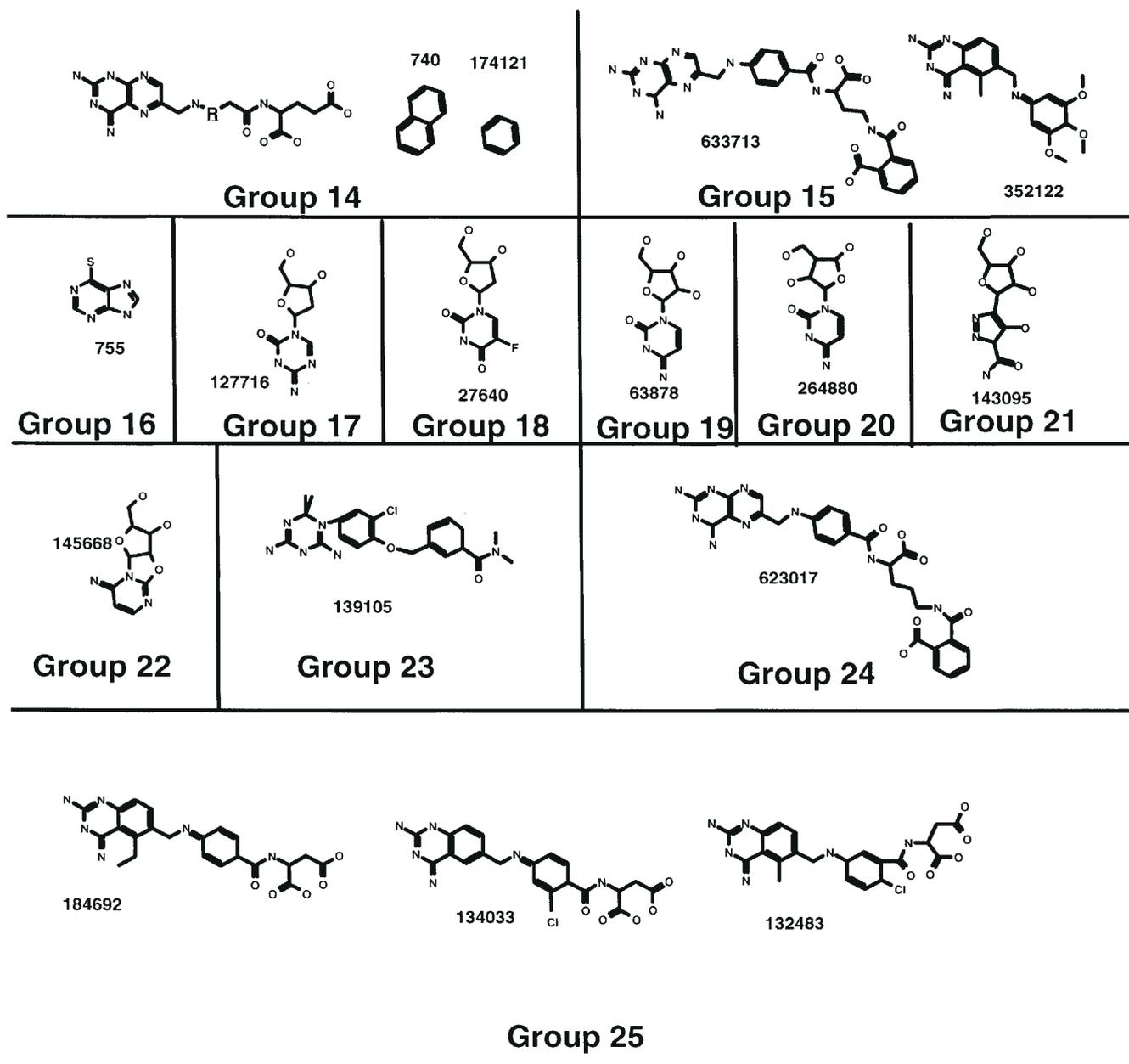|	R1	R2	R3
406042	C	O	C
361792	C	S	
33410	(ring)	O	C

Group 13
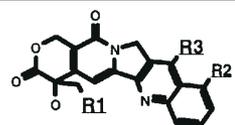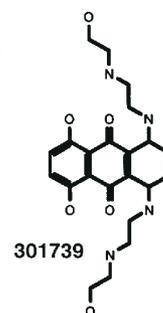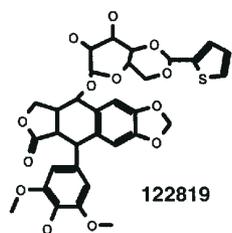
Figure 1 (continued)

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
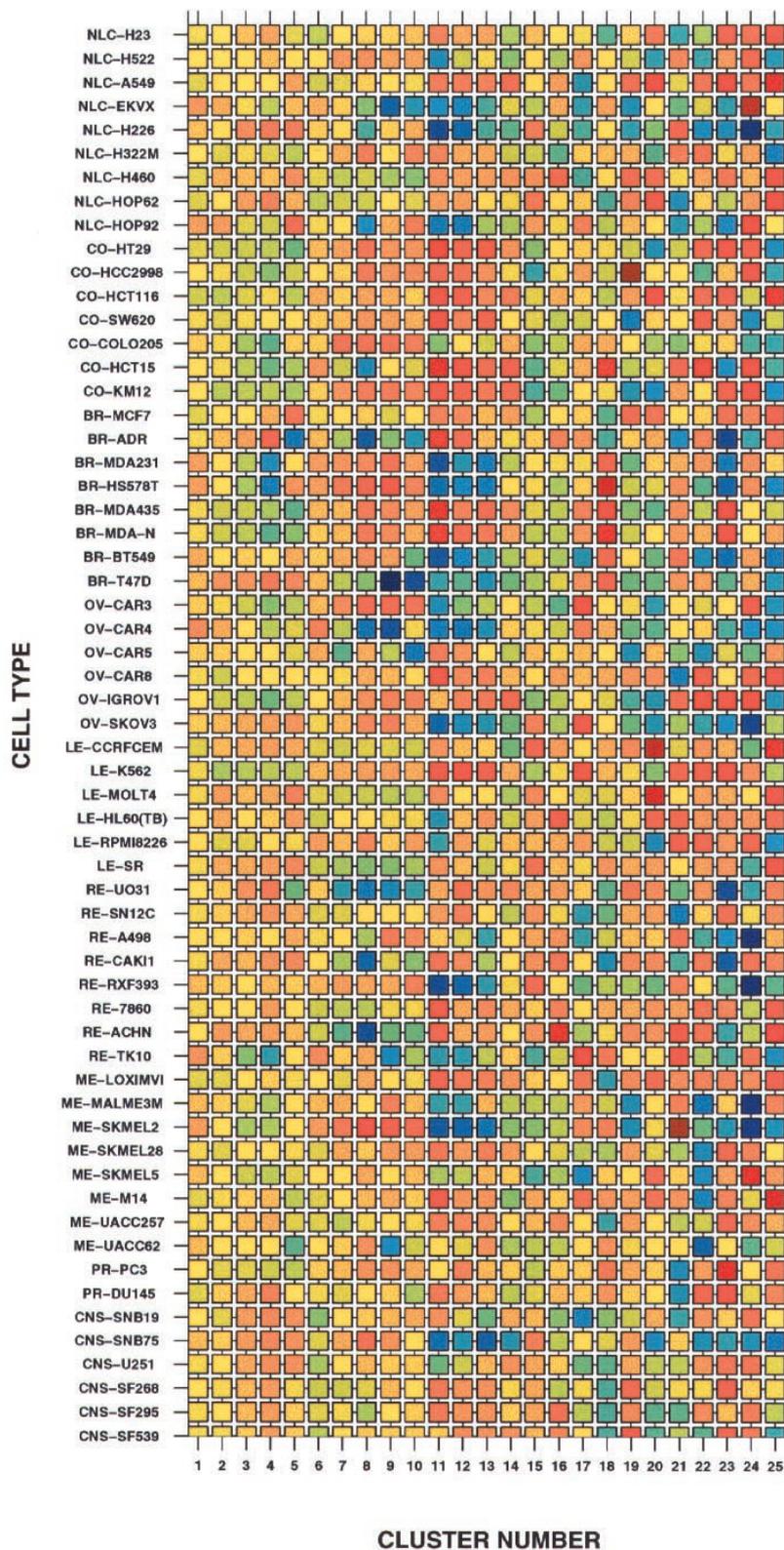
Figure 2

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 ($\Delta x = [\log GI_{50}] - <\log GI_{50}>$) 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
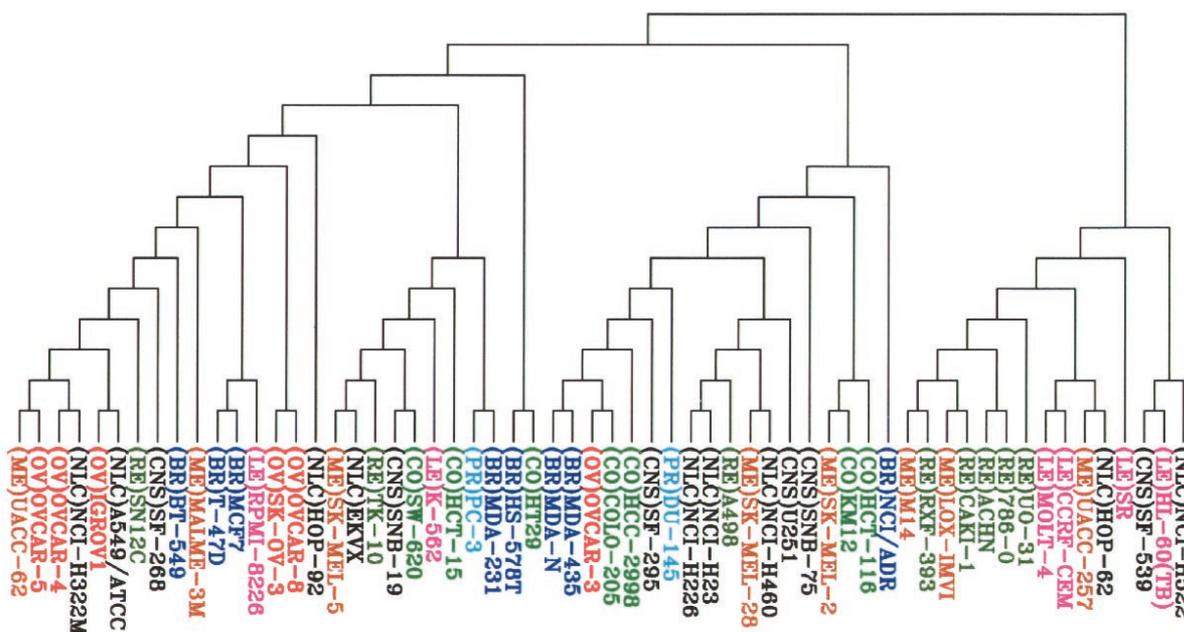
Figure 3

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

$$\langle a*\rangle_{MOA} = \Sigma_i a_i*/N_{MOA} \qquad (2)$$

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 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

$$\Delta a_i*_{MOA} = a_i* - \langle a*\rangle_{MOA} \qquad (3)$$

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_o 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
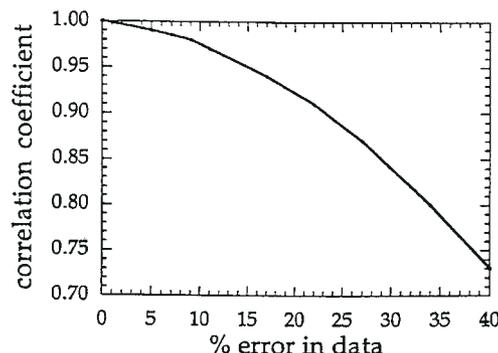
Figure 4

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

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