van Es, E. Skamene, E. Schurr, *Clin. Invest. Med.* **16**, 285 (1993); M. L. Cohen, *Trends Microbiol.* **2**, 422 (1994).
86. A. T. Haase *et al.*, *Science* **274**, 985 (1996).
87. W. L. Walker and J. Cook, *Bull. Math. Biol.* **58**, 1047 (1996).
88. T. Saga *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 8999 (1995).
89. T. B. Kepler and A. S. Perelson, *ibid.*, p. 8219; J. Jacob, J. Przylepa, C. Miller, K. J. Kelsoe, *J. Exp. Med.* **178**, 1293 (1993); P. E. Seiden and F. Celada, *Eur. J. Immunol.* **26**, 1350 (1996); Z. Agur, G. Mazor, I. Meilijson, *Proc. R. Soc. London Ser. B* **245**, 147 (1991).
90. T. B. Kepler and A. S. Perelson, *Immunol. Today* **14**, 412 (1993).
91. S. Han *et al.*, *J. Exp. Med.* **182**, 1635 (1995).
92. M. A. Fishman and A. S. Perelson, *J. Theor. Biol.* **170**, 25 (1994); B. F. Morel *et al.*, *Bull. Math. Biol.* **58**, 569 (1996); J. Caneiro, J. Stewart, A. Coutinho, *Int. Immunol.* **7**, 1265 (1995); A. Brass, R. K. Grencis, K. J. Else, *J. Theor. Biol.* **166**, 189 (1994).
93. M. Fishman and A. S. Perelson, *J. Theor. Biol.* **160**, 311 (1993); R. J. De Boer and A. S. Perelson, *ibid.* **169**, 375 (1994); S. J. Merrill, R. J. De Boer, A. S. Perelson, *Rocky Mountain J. Math.* **24**, 213 (1994).
94. R. J. De Boer and A. S. Perelson, *J. Theor. Biol.* **149**, 381 (1991); P. E. Seiden and F. Celada, *ibid.* **158**, 329 (1992).
95. J. Gleick, *Chaos: Making of a New Science* (Penguin, New York, 1987).
96. Figure 1 is a frame from the video, D. Deutschman and S. A. Levin, *SORTIE Simulations*, by Cornell and Princeton Universities; computation by L. Buttel; visualization by C. Devine, Cornell Theory Center, Cornell University.
97. L. A. Dugatkin and D. S. Wilson, *Am. Nat.* **138**, 687 (1991).
98. D. Schenzle, *IMA J. Math. Appl. Med. Biol.* **1**, 169 (1984).
99. B. Bolker, *ibid.* **10**, 83 (1993).
100. B. T. Grenfell, A. Kleczkowski, S. P. Ellner, B. M. Bolker, *Philos. Trans. R. Soc. London Ser. A* **348**, 515 (1994).
101. B. M. Bolker and B. T. Grenfell, *Philos. Trans. R. Soc. London Ser. B* **348**, 309 (1995); B. T. Grenfell, B. M. Bolker, A. Kleczkowski, *Proc. R. Soc. London Ser. B* **259**, 97 (1995).
102. N. M. Ferguson, D. J. Nokes, R. M. Anderson, *Math. Biosci.*, in press.
103. M. S. Bartlett, *J. R. Stat. Soc. Ser. A* **120**, 48 (1957); *ibid.* **123**, 37 (1960).
104. M. J. Keeling and D. A. Rand, unpublished data.
105. C. J. Rhodes and R. M. Anderson, *Nature* **381**, 600 (1996).
106. P. E. M. Fine and J. A. Clarkson, *Int. J. Epidemiol.* **12**, 332 (1983).
107. A. D. Cliff, P. Haggett, D. F. Stroup, E. Cheney, *Stat. Med.* **11**, 1409 (1992); B. T. Grenfell, A. Kleczkowski, C. A. Gilligan, B. M. Bolker, *Statistical Methods Med. Res.* **4**, 160 (1995).
108. B. T. Grenfell, A. Kleczkowski, S. P. Ellner, in *Forecasting and Chaos*, H. Tong, Ed. (World Scientific, Singapore, 1994), pp. 321–345.
109. We thank P. Kollman and the participants in the meeting, "Modeling of Biological Systems," which inspired this article, and NSF, which funded the workshop. Supported by NASA grant NAGW 4688 and the Andrew Mellon Foundation (S.L.), NSF grant DEB 9629236 (A.H.), the Wellcome Trust (B.G.), NIH grants RR06555 and AI28433 (A.S.P.), and the Jeanne M. Sullivan and Joseph P. Sullivan Foundation. D. Deutschman provided useful comments. Most of all, we thank A. Bordvik, who brought order to a chaotic sequence of drafts of this manuscript.

An Information-Intensive Approach to the Molecular Pharmacology of Cancer

John N. Weinstein,* Timothy G. Myers, Patrick M. O'Connor, Stephen H. Friend, Albert J. Fornace Jr., Kurt W. Kohn, Tito Fojo, Susan E. Bates, Lawrence V. Rubinstein, N. Leigh Anderson, John K. Buolamwini,† William W. van Osdol,‡ Anne P. Monks, Dominic A. Scudiero, Edward A. Sausville, Daniel W. Zaharevitz, Barry Bunow, Vellarkad N. Viswanadhan,§ George S. Johnson, Robert E. Wittes, Kenneth D. Paull

Since 1990, the National Cancer Institute (NCI) has screened more than 60,000 compounds against a panel of 60 human cancer cell lines. The 50-percent growth-inhibitory concentration (GI_{50}) for any single cell line is simply an index of cytotoxicity or cytostasis, but the patterns of 60 such GI_{50} values encode unexpectedly rich, detailed information on mechanisms of drug action and drug resistance. Each compound's pattern is like a fingerprint, essentially unique among the many billions of distinguishable possibilities. These activity patterns are being used in conjunction with molecular structural features of the tested agents to explore the NCI's database of more than 460,000 compounds, and they are providing insight into potential target molecules and modulators of activity in the 60 cell lines. For example, the information is being used to search for candidate anticancer drugs that are not dependent on intact p53 suppressor gene function for their activity. It remains to be seen how effective this information-intensive strategy will be at generating new clinically active agents.

Drug discovery is being transformed by new developments in molecular cell biology and the information sciences. A case in point is the drug discovery program conducted by the Developmental Therapeutics Program (DTP) of the NCI. Before 1985, the NCI used mice bearing murine leukemia P388 cells to screen new compounds for anticancer activity. That strategy identified agents active against leukemias but relatively few that were effective against solid tumors, including the most common human carcinomas. Hence, the NCI established a primary screen in which compounds are tested in vitro for their ability to inhibit growth of 60 different human cancer cell lines (*1*). Included are melanomas, leukemias, and cancers of breast, prostate, lung, colon, ovary, kidney, and central nervous system origin. A highly schematic view of this portion of the NCI drug discovery–development process is shown in Fig. 1. Compounds for testing have come principally from synthetic chemistry and natural product sources, but combinatorial libraries and products of biotechnology are also being screened.

This "disease-oriented" strategy for drug discovery was based on the hypothesis that selective activity in vitro against cancer cell lines from a particular organ would predict selective activity against corresponding tumors in humans. That concept is being tested as agents progress through clinical trials, and the answer is not yet clear. However, patterns of activity observed in the screen have proved predictive in an even more powerful way at the molecular level: They provide incisive information on the mechanisms of action of the compounds tested and on molecular targets and modulators of activity within the cancer cells. The cell lines are not fully representative of solid tumors in humans, but their patterns of pharmacological response are rich in information. We refer to this test system as a "screen," but it has also become a way to "profile" or "fingerprint" potential therapeutic agents.

The patterns of activity were first analyzed by the COMPARE algorithm (*2*). Given one compound as a "seed," COMPARE searches the database of screened agents for those most similar to the seed in their patterns of activity against the panel of 60 cell lines. Similarity in pattern often indicates similarity in mechanism of action, mode of resistance, and molecular structure (*2*). This form of analysis has been applied productively to topoisomerase II inhibitors (*3*), pyrimidine biosynthesis inhibitors (*4*), and tubulin-active compounds (*5*), among

other classes of agents. Back-propagation neural networks and predictive methods from classical statistics have also been used to verify that the patterns of activity could predict a compound's mechanism of action (6). More detailed information on the relation between pattern and mechanism has come from additional analyses based on techniques from statistics and artificial intelligence (7, 8). To date, five compounds (spicamycin analog KRN 5500, flavopiridol, UCN-01, a depsipeptide, and a quinocarmycin analog) assessed in the screen and analyzed by the methods described above have been selected for entry into clinical trials (9).

Bioinformatics: The Structure, Activity, and Target Databases

Here we describe a general way in which information on the activity patterns is being combined with other types of information to address problems in drug discovery and molecular pharmacology. A formulation of this approach in terms of three databases is shown in Fig. 1: (**A**) contains the activity patterns already discussed, (**S**) contains molecular structural features of the tested compounds, and (**T**) contains possible targets or modulators of activity in the cells. Portions of these databases can be accessed through DTP's World Wide Web site (http://epnws1.ncifcrf.gov:2345/dis3d/DTP.HTML). Links to these and additional pertinent databases can be found at http://www.nci.nih.gov/intra/lmp/jnwbio.htm. These two Web sites will be updated progressively with additional data and tools of analysis (10).

The chemical structure (**S**) database can be coded in terms of any set of two-dimensional (2D) or 3D molecular structure descriptors. The NCI's Drug Information System (DIS) contains chemical connectivity tables for approximately 460,000 molecules, including the 60,000 tested to date. Three-dimensional structures have been obtained for 97% of the DIS compounds, and a set of 588 bit-wise descriptors has been calculated for each structure by use of the Chem-X computational chemistry package (Chem-DBS-3D module, Chemical Design, Oxford, U.K.) (11). This data set provides the basis for pharmacophoric searches; if a tested compound, or set of compounds, is found to have an interesting pattern of activity, its structure can be used to search for similar molecules in the DIS database (12).

In the target (**T**) database, each row defines the pattern (across 60 cell lines) of a measured cell characteristic that may mediate, modulate, or otherwise correlate with the activity of a tested compound. When the term is used in this general shorthand sense, a "target" may be the site of action or part of a pathway involved in a cellular response. Among the potential targets assessed to date are oncogenes, tumor-suppressor genes, drug resistance–mediating transporters, heat shock proteins, telomerase, cytokine receptors, molecules of the cell cycle and apoptotic pathways, DNA repair enzymes, components of the cytoarchitecture, intracellular signaling molecules, and metabolic enzymes (13).

In addition to the targets assessed one at a time, others have been measured en masse as part of a protein expression database generated for the 60 cell lines by 2D polyacrylamide gel electrophoresis (2D PAGE) (14). The aim is to look for molecules that have not been considered previously as targets. In the process, a link has been established between the molecular pharmacology of cancer and the growing enterprise of proteome research (15). The current database consists of 1014 indexed and quantitated protein spots, of which 151 have been quality controlled over all 60 current cell lines and incorporated into a primary data set for analysis (14). Analogous links to genome research are being established through analyses of gene amplification and mRNA expression patterns. Figure 1 indicates approximately 100 targets, but that number is increasing rapidly.

Relating Molecular Targets to Drug Activity Patterns

The first target analyzed in detail by the COMPARE program was the drug-resistance transporter P-glycoprotein (Pgp), encoded by multidrug resistance gene *MDR-1* (16–18). The result was a list of agents predicted and then experimentally verified to be good Pgp substrates. Related strategies identified Pgp inhibitors (19). We present here a complementary approach for analysis and display of these data, the DISCOVERY program package (20), which maps coherent patterns in the data, rather than treat-

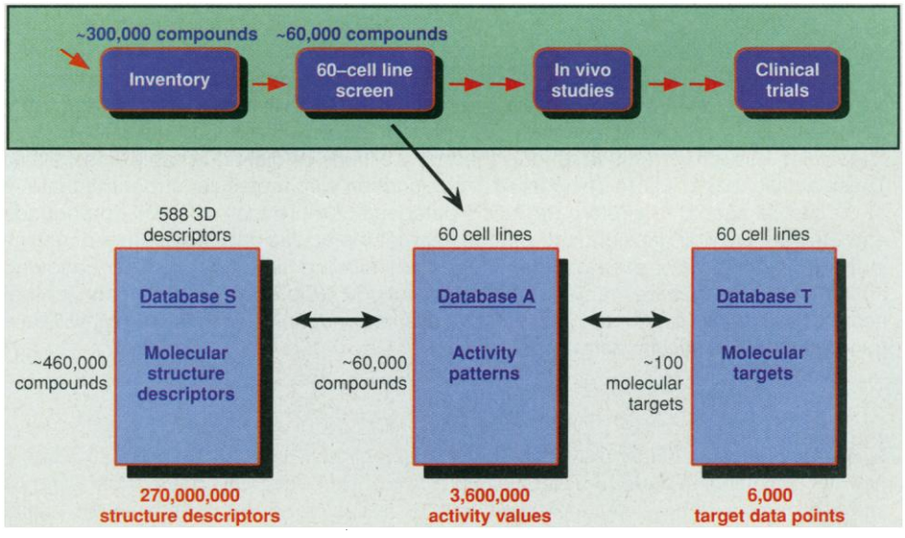

Fig. 1. Simplified schematic overview of an information-intensive approach to cancer drug discovery and molecular pharmacology at the NCI. Each row of the activity (**A**) database represents the pattern of activity of a particular compound across the 60 cell lines. As described in the text, the **A** database can be related to a structure (**S**) database containing 2D or 3D chemical structure characteristics of the compounds and a target (**T**) database containing information on possible molecular targets or modulators of activity within the cells.

J. N. Weinstein, T. G. Myers, P. M. O'Connor, A. J. Fornace Jr., K. W. Kohn, J. K. Buolamwini, W. W. van Osdol, and V. N. Viswanadhan are at the Laboratory of Molecular Pharmacology (LMP), Division of Basic Science, National Cancer Institute (NCI), National Institutes of Health (NIH), Building 37, Room 5C-25, 9000 Rockville Pike, Bethesda, MD 20892, USA. S. H. Friend is at the Fred Hutchinson Cancer Research Center–NCI, Seattle, WA 98105, USA. T. Fojo and S. E. Bates are in the Medicine Branch, Division of Clinical Science, NCI, NIH, Bethesda, MD 20892, USA. L. V. Rubinstein is in the Biometric Research Branch, Cancer Therapy Evaluation Program, Division of Cancer Treatment, Diagnosis, and Centers (DCTDC), NCI, NIH, Bethesda, MD 20892, USA. N. L. Anderson is with Large Scale Biology, Rockville, MD 20850, USA. A. P. Monks and D. A. Scudiero are at the SAIC–NCI-Frederick Cancer Research and Development Center (FCRDC), Frederick, MD 21701, USA. E. A. Sausville is in the Developmental Therapeutics Program (DTP), DCTDC, NCI, NIH, Bethesda, MD 20892, USA. D. W. Zaharevitz and K. D. Paull are in the Information Technology Branch, DTP, DCTDC, NCI, NIH, Bethesda, MD 20892, USA. G. S. Johnson is in the Grants and Contracts Operations Branch, DTP, DCTDC, NCI, NIH, Bethesda, MD 20892, USA. R. E. Wittes is in the Office of the Director, DCTDC, NCI, NIH, Bethesda, MD 20892, USA.

*To whom correspondence should be addressed. E-mail: weinstein@dtpax2.ncifcrf.gov
†Present address: University of Mississippi School of Pharmacy, University of Mississippi, MS 38677, USA.
‡Present address: Alza, 1454 Page Mill Road, Palo Alto, CA 94304, USA.
§Present address: Gensia, 9360 Towne Center Drive, San Diego, CA 92121, USA.

ing the compounds and targets one pair at a time. Because the **S**, **A**, and **T** databases contain, in aggregate, many millions of numbers, the challenge was to compact that information sufficiently for analysis without losing or obscuring important local features of the data. These often contradictory requirements have guided development of DISCOVERY, which integrates the disparate types of information on the compounds and displays them in novel ways suited to human pattern recognition. The same algorithms can be applied to other types of databases, including those generated by screening and profiling systems in which agents are tested in multiple assays—for example, against mammalian cells, yeast mutants, bacteria, or biochemical targets.

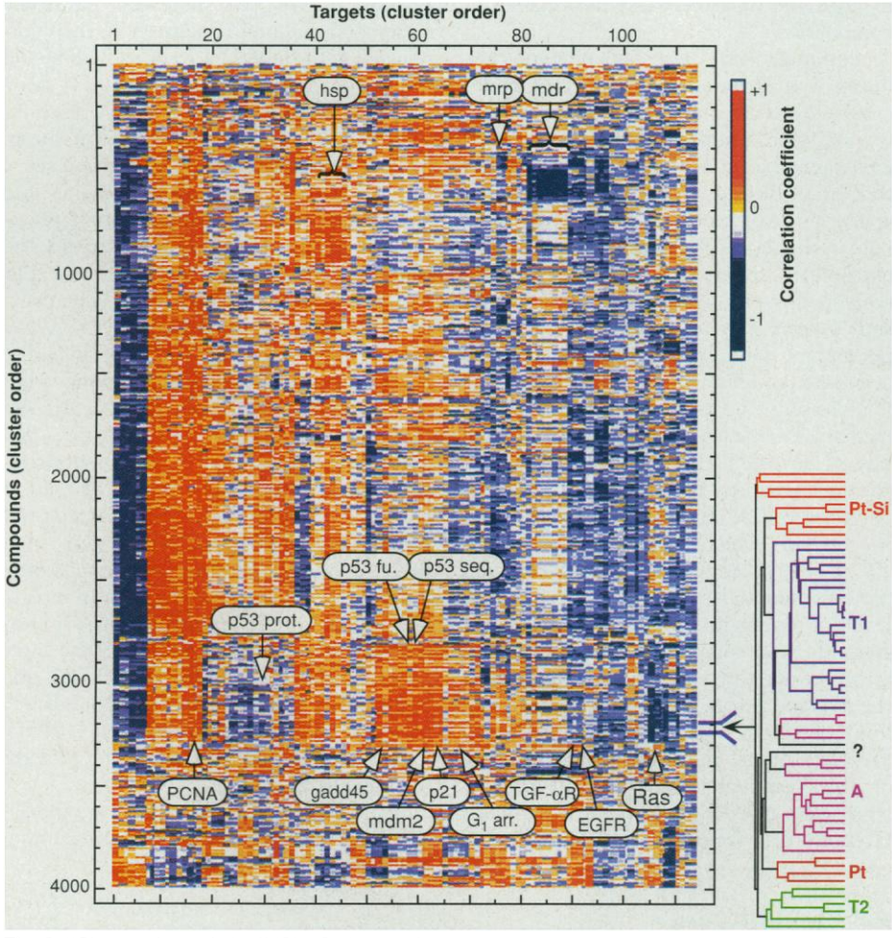

Fig. 2. "Clustered correlation" (ClusCor) map of the relation between compounds tested and molecular targets in the cells. This normalized **A·T′** product matrix (where the prime symbol indicates the matrix transpose) correlates target patterns across the 60 cell lines with patterns of growth inhibition for an important set of 3989 compounds. A red or orange point (high positive Pearson correlation coefficient) indicates that the agent tends to be selectively active against cell lines that express the target in large amounts (or in functional form). A dark blue point (high negative correlation) indicates the opposite tendency (selective potency against cell lines that have less target or function). The 113 columns correspond to 76 distinct target molecules or functions, some represented multiple times in different mathematical transformations. Compounds and targets are cluster-ordered as explained in the text. To the right is shown one 61-leaf "twig" of the overall 3989-leaf cluster tree of compounds. Symbols for mechanisms of action (*6, 8*) are as follows: T1, topoisomerase 1 inhibitors; T2, topoisomerase 2 inhibitors; A, alkylating agents; Pt, platinum compounds (of the cisplatin-carboplatin family); Pt-Si, platinum agents containing a silane moiety; ?, mechanism unknown; PCNA, proliferating cell nuclear antigen determined from 2D gels (column 16) (*14*); p53 seq, p53 sequence, wild-type versus mutant (*30*); p53 fu., p53 function in a yeast-based assay (*30*); p53 prot., p53 protein expression by protein immunoblot (columns 29 and 30) (*30*); hsp, heat shock–related proteins (Hsp60, Hsc70, Hsp90, Grp75, Grp78) from 2D gels (columns 40 to 45) (*14*); gadd45, mdm2, and p21, *GADD45*, *MDM2*, and $p21^{CIP1/WAF1}$ mRNA induction in response to γ-irradiation (columns 54 to 57, 60, and 61 to 64, respectively) (*30*); G_1, G_1 arrest in response to γ-irradiation, assessed by flow cytometry (columns 65 to 69) (*30*); mrp, mRNA expression levels for the *MRP* multidrug resistance transporter (columns 75 and 76) (*18*); mdr, *MDR-1* mRNA (*16*) and function in terms of rhodamine efflux (columns 81 to 88) (*17*); TGF-αR, transforming growth factor–α receptor mRNA (columns 89 to 91); EGFR, epidermal growth factor receptor (column 92) (*37*); and Ras, *RAS* sequence, wild-type versus mutant (*38*).

Figure 2 shows a color-coded DISCOVERY pattern map relating a **T** database of 113 target vectors to an **A** database of 3989 nonconfidential compounds deemed sufficiently interesting in the initial screen to be tested more than once. This map was obtained by an algorithm we term "clustered correlation" (ClusCor). Each database was treated as a mathematical matrix, and the following four steps were applied: (i) each row of **A** and **T** was normalized by its mean and standard deviation; (ii) the two matrices were multiplied to obtain **A·T′**, where the prime symbol indicates the matrix transpose; (iii) each entry was divided by $n - 1$, where n (=60) is the number of cell lines, producing a matrix of Pearson correlation coefficients relating activity and target patterns; and (iv) the rows and columns of the product matrix were rearranged into "cluster order." Only with this last step did patterns emerge.

The 3989 compounds were cluster-ordered (*21*) along the ordinate on the basis of their activity patterns across the 60 cell lines. Thus, compounds with the most nearly identical patterns appear side by side. Because this clustering of compounds was done independently of targets, the coherent patterns observed as patches of color validate the hypothesis that the activity patterns and targets are related. The possibility that these patterns were created spuriously by the clustering process is ruled out by the lack of pattern features in Fig. 5A. Figure 5A shows the result when the 60 activity values for each drug were randomly permuted before the calculation and clustering algorithm that had produced Fig. 2 were applied. The 113 targets were cluster-ordered along the abscissa in Fig. 2 on the basis of their apparent effect on activities of compounds in the database. Thus, targets with the most similar columns of correlation coefficients appear side by side.

To illustrate the result of the clustering process, the right-hand side of Fig. 2 shows one small 61-leaf "twig" of the overall 3989-leaf cluster tree. Compounds similar in mechanism of action cluster together. Among the classes that are organized in a coherent way elsewhere in Fig. 2 are the Taxol (paclitaxel) analogs (taxanes): 34 of the 37 taxanes in the database appear side by side (compounds 620 to 653), and the other 3 are found on nearby twigs (compounds 655, 658, and 701). The largest chemically coherent set of compounds is a set of 72 thiosemicarbazones (compounds 1491 to 1579, with small gaps occupied by phenylhydrazones) (*22*). Most of the tin-

containing molecules in the database are contiguous (compounds 2034 to 2062). The closely related clinical agents cisplatin and carboplatin fall side by side (compounds 3260 and 3261) within one cluster of 11 structurally related platinum analogs, whereas the diaminocyclohexyl platinum compounds, which have very different pharmacological behavior (*23*), fall elsewhere in the map (compounds 2838 to 2849). Perhaps more important than the branches with known agents, however, are those that contain no familiar compounds. The DISCOVERY program set, as its name implies, was developed primarily to explore and organize these new classes of compounds.

Although some degree of coherent clustering was expected for families of molecules related by chemistry or mechanism, the precision indicated by the above examples was unexpected; the a priori probability that any given pair of compounds would appear as nearest neighbors along the ordinate in the set of 3989 is only 2 in 3988. An explanation for the observed coherence is suggested by a thought experiment in which the patterns are considered, because of experimental noise, to be binary; that is, one is assumed to know only whether a cell line is more sensitive or less sensitive than the median. Then each compound would have one of $60!/(30!30!) = 1.2 \times 10^{17}$ possible patterns (that is, the number of ways of choosing the 30 out of 60 that fall above the median). The number would increase to $2^{60} = 1.2 \times 10^{18}$ for all possible binary patterns and to $4^{60} = 1.3 \times 10^{36}$ if four levels of sensitivity could be reliably distinguished. Each compound displays a unique "fingerprint" pattern, defined by a point in the 60D space (one dimension for each cell line) of possible patterns. In information theoretic terms, the transmission capacity of this communication channel is very large, even after one allows for experimental noise and for biological realities that constrain the compounds to particular regions of the 60D space. Although the activity data have been accumulated over a 6-year period, the experiments have been reproducible enough to generate the patterns of coherence described here (*24*).

Each patch of color in Fig. 2 suggests a possible correlation between targets and compounds. The dark blue patch for compounds 513 to 667 indicates that these compounds are highly negative in their correlation with targets 81 to 88, which are all indices of Pgp/Mdr-1 expression and function (*16*–*18*). Several lines of evidence indicate the significance of this observation. (i) We analyzed cell screen data for a set of 35 compounds of diverse structure and mechanism that had been reported previously on the basis of transport assays to be Mdr-1 substrates (*17*, *25*). Of these, 18 (51%) fell within the blue patch, a percentage 13-fold greater than the 4% (155/3989) expected by chance alone. The probability (exact binomial) of such an extreme event happening by chance is <0.0001. (ii) Although 18 of 35 reported substrates fell within the patch, 0 of 12 compounds reported not to be substrates (*17*, *25*) did so ($P = 0.0010$ by one-sided Fisher's exact test for the associated 2 by 2 table). (iii) It has been reported (*17*) that Mdr-1 substrates tend to be natural products, high in molecular weight, and often cationic. We find by linear discriminant analysis that these three factors predict with a sensitivity of 78% and a specificity of 84% which compounds will be found in the blue patch ($P < 0.0001$). These findings further validate the patterns seen in Fig. 2.

Columns 76 and 77 in Fig. 2 are indices of messenger RNA (mRNA) expression for

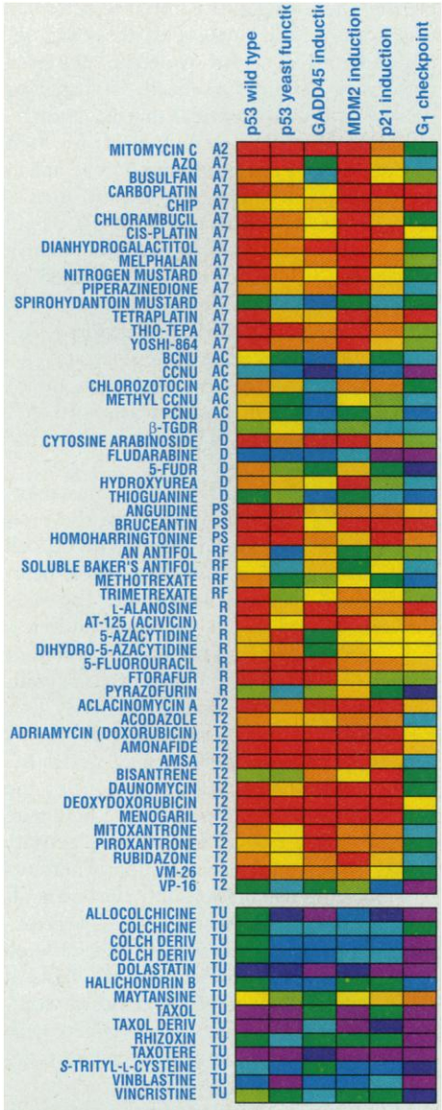

Fig. 3. Relation between *p53* pathway molecular targets and patterns of activity for clinically evaluated anticancer agents. The compounds have been grouped by their presumed principal mechanisms of action. A number of additional antitubulin agents have been added to increase representation of that category. Color coding indicates the Pearson correlation coefficient relating agent to target. A2, guanine-N2 alkylator; A7, guanine-N7 alkylator; AC, chloroethylating alkylator; D, DNA-RNA antimetabolite; PS, protein synthesis inhibitor; R, RNA antimetabolite; RF, antifolate RNA antimetabolite; T2, topoisomerase II inhibitor; TU, antitubulin (antimitotic) agent. The data on p53 pathway parameters are from (*30*).

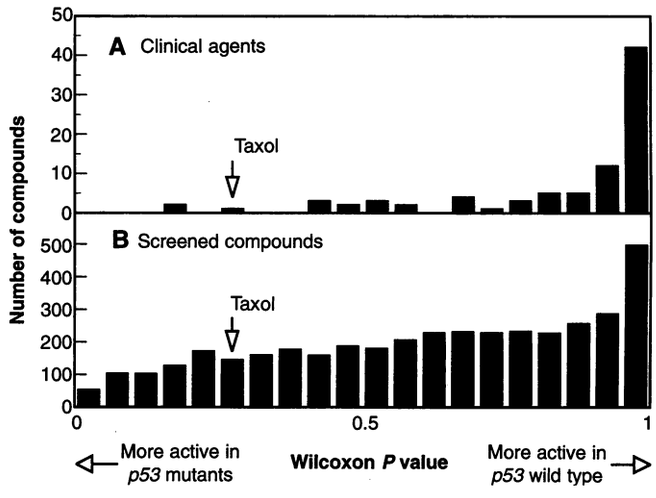

Fig. 4. Histograms showing the relation between *p53* status and patterns of growth inhibition in the screen (**A**) for a set of 86 phase II–evaluable clinical agents and (**B**) for a set of 3989 multiply tested compounds. Most of the clinical agents appear more active in the presence of wild-type *p53*; the other compounds show a lesser trend in the same direction. The parameter calculated for each drug has the form of a Wilcoxon rank sum *P* value. $P > 0.5$ indicates a compound that tends in this screening assay to be more active in the cells with wild-type *p53*; $P < 0.5$ indicates the opposite tendency. Values >0.975 or <0.025 would be required to reject the null hypothesis of equal median activities in p53 wild-type and mutant cells for any single compound. The data on *p53* sequence are from (*30*).

Mrp, another transport molecule associated with multidrug resistance (*18*). There is only a slight overlap between the Mdr-1– and Mrp-sensitive families of compounds. As indicated by columns 40 to 45, high basal levels of heat shock proteins (Hsp60, Hsp90, Hsc70, Grp75, and Grp78) correlate positively with activity for a large set of agents, including some of those in the group sensitive to Mdr-1. This type of analysis makes it possible to cross-compare multiple targets for their expression levels and for their apparent impact on the activities of different classes of agents (*26*).

Activity Patterns and p53 Pathway Status

The *p53* tumor-suppressor gene is mutated in more than 50% of human tumors, more than any other gene examined to date (*27*). p53 functions as a transcriptional regulator with the ability to both transactivate and suppress gene transcription (*28*). It is activated in response to DNA damage and can orchestrate a number of cellular responses to genotoxic stress, including G_1 arrest and apoptosis (*29*). A large cluster of compounds (numbers 2802 to 3309) is positively correlated with intact p53 pathway status (as indicated by a large red patch in Fig. 2). The indices of p53 status assessed in the cells include *p53* sequence, basal p53 protein level, p53 function in a yeast-based assay, G_1 checkpoint integrity, and γ-ray induction of the *p53*-regulated genes $p21^{CIP1/WAF1}$, *MDM2*, and *GADD45* (*30*). The activity patterns of most of these compounds are inversely correlated with expression levels of p53 protein, as would be expected given that the protein is overexpressed in most *p53*-mutant cell types (*29*).

Compounds 2802 to 3309 include a large percentage of the familiar cytotoxic antitumor agents. Of 86 agents considered evaluable on the basis of phase II clinical trials (*31*), 45 appear in this relatively small region of the map, giving an odds ratio of (45/41)/(463/3440) = 8.2:1 ($P < 0.0001$ by Fisher's exact test). This odds ratio substantially understates the enrichment of this region of the map with clinical agents because the region is artificially enlarged by the many analogs synthesized on the basis of the clinical molecules (*21*).

The correlation of p53 pathway factors with activity patterns for a subset of the clinical agents with defined mechanisms of action (*6*, *8*) is shown in Fig. 3. Most, although not all, of the agents damage DNA, and in this assay they tended to be more potent in p53 wild-type cells than in p53 mutant ones (*32*). The principal exception was the set of antimitotic tubulin-active agents, including Taxol, which generally do not show any clear correlation with p53 status. Examination of a previously defined set (*6*, *8*) of 123 standard anticancer agents (which overlaps with the set of clinical agents studied here) yields similar results (*30*).

The large majority of clinical agents appear in this assay to be more active on average in the p53 wild-type cells (Fig. 4A). In contrast, the p53 association is much less pronounced for the set of 3989 multiply tested molecules (Fig. 4B) or for all compounds tested. We examined compounds at the left of Fig. 4B for agents that might be effective in p53-mutant human tumors. In this search for "p53-inverse" (or at least "p53-indifferent") compounds, we used the COMPARE and DISCOVERY program sets to generate lists of candidates on the basis of various sets of explicit criteria (*20*, *33*). Selected compounds are being tested in p53-isogenic human cell sets (*34*), and lead compounds that perform favorably will be further evaluated in vivo.

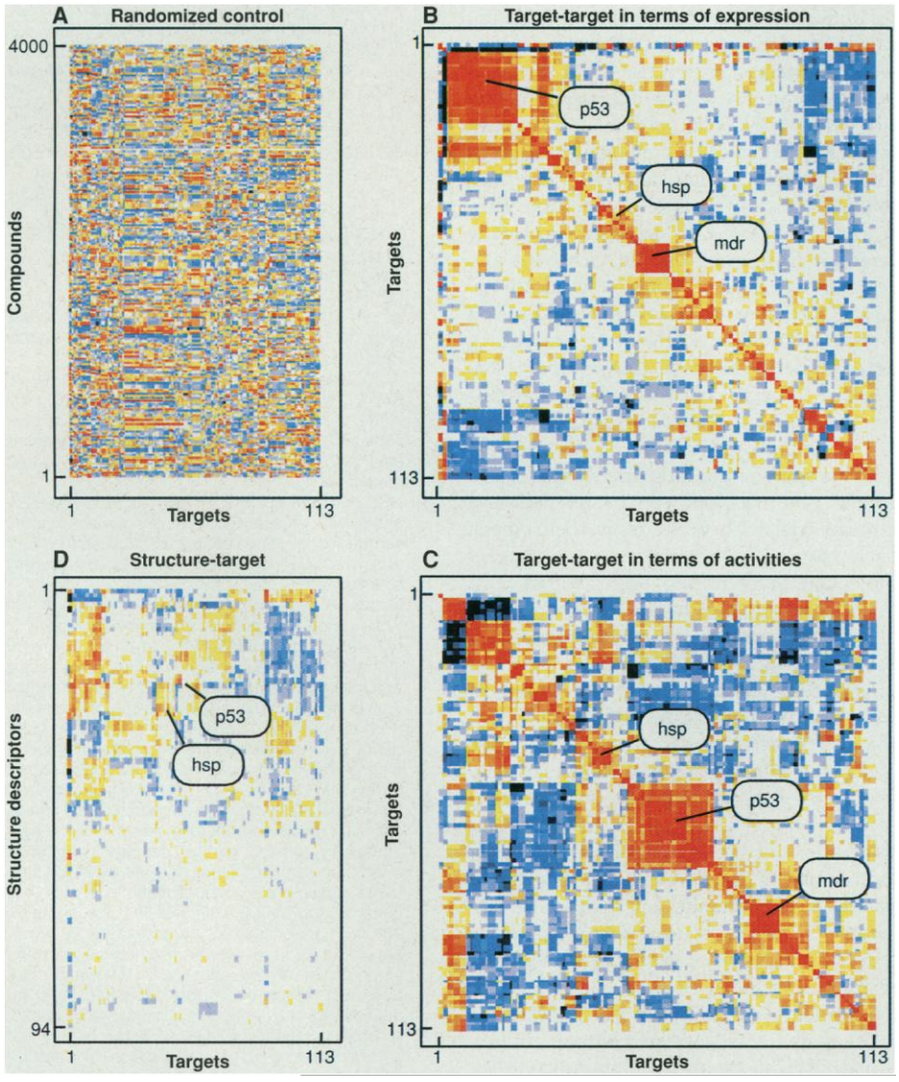

Fig. 5. Four types of "clustered correlation" (ClusCor) matrices involving the **S**, **A**, and **T** databases. (**A**) Activity vectors of the compounds were randomly permuted, and all calculations (including clustering) were then done exactly as for Fig. 2. The lack of apparent pattern verifies that clustering did not spuriously create the patterns seen in Fig. 2. (**B**) A normalized **T**·**T**′ database, which cross-correlates patterns of target expression. Targets with the most similar patterns of expression appear side by side. Because a target's expression is 100% correlated with itself by definition, all values on the principal diagonal are color-coded red. Because of the clustering, targets positively correlated in their expression produce red patches straddling the diagonal. (**C**) A normalized (**A**·**T**′)·(**A**·**T**′) database, similar to (B) except that targets are characterized, not in terms of their expression levels, but in terms of their correlations with activity patterns of the 3989 compounds. (**D**) An **S**′·(**A**·**T**′) database. This database relates 2D substructures of the compounds (*20*) to targets through the activity patterns of the compounds.

Target-Target and Target-Structure Correlations

As indicated by the ClusCor matrices shown in Fig. 5, the databases on activity, molecular structure, and targets have implications for basic biology and pharmacology as well as for drug discovery per se. The correlation of each target's pattern of expression across the 60 cell lines with that of each other target is shown in Fig. 5B. Values of the correlation coefficient on the main diagonal are, by definition, unity because each target is 100% correlated with itself. The red patches of high correlation straddling the diagonal appear because the targets are listed on both ordinate and abscissa in cluster order on the basis of patterns of expression. Clusters of targets related to Mdr-1, heat shock proteins, and p53 function show high degrees of internal correlation. In many instances (for example, that of *p53* and induction of the downstream genes $p21^{CIP1/WAF1}$, *GADD45*, and *MDM2*) this observation reflects the known biochemical relationships (*27, 29*), further validating the significance of patterns seen in Figs. 2 and 5.

A similar pattern of correlation is shown in Fig. 5C, which relates targets to each other, not in terms of their levels of expression, but in terms of their relation to activity profiles for the 3989 compounds in the database. Again, the same three families of targets are highly correlated. As the cells are characterized with respect to more and more targets, these correlations will generate an increasing number of testable cell biological hypotheses for further study. The relation of targets to chemical substructures of compounds through the database of activity patterns is shown in Fig. 5D. Although nonrandom patterns are apparent, they are less pronounced, and other, nonlinear methods of analysis (including ones based on genetic algorithm and neural networks) may prove to be better suited for analysis of this type of relationship.

Hypothesis Generation in the Molecular Pharmacology of Cancer

The approach to drug discovery and molecular pharmacology presented here serves a number of functions. (i) It suggests novel targets and mechanisms of action or modulation. (ii) It detects inhibition of integrated biochemical pathways not adequately represented by any single molecule or molecular interaction. (This feature of cell-based assays is likely to be more important in the development of therapies for cancer than it is for most other diseases; in the case of cancer, one is fighting the plasticity of a poorly controlled genome and the selective evolutionary pressures for development of drug resistance.) (iii) It provides candidate molecules for secondary testing in biochemical assays; conversely, it provides a well-characterized biological assay in vitro for compounds emerging from biochemical screens. (iv) It "fingerprints" tested compounds with respect to a large number of possible targets and modulators of activity. (v) It provides such fingerprints for all previously tested compounds whenever a new target is assessed in many or all of the 60 cell lines. (In contrast, if a battery of assays for different biochemical targets were applied to, for example, 60,000 compounds, it would be necessary to retest all of the compounds for any new target or assay.) (vi) It links the molecular pharmacology with emerging databases on molecular markers in microdissected human tumors—which, under the rubric of this article, constitute clinical (**C**) databases (*35*). (vii) It provides the basis for pharmacophore development and searches of an **S** database for additional candidates. If an agent with a desired action is already known, its fingerprint patterns of activity can be used by COMPARE, DISCOVERY, neural networks, and other pattern-recognition technologies to find similar compounds.

This approach to drug discovery and molecular pharmacology can be likened to a clinical trial with 60 patients (cell types), each profiled with respect to a variety of molecular markers and each treated with 60,000 different agents, one at a time. It can also be considered as a hypothesis generator based on a set of $60{,}000 \times 60 = 3.6$ million pharmacology experiments. The important word here is "hypothesis." Information from the cell lines is fundamentally correlative and subject to confounding influences. Hypotheses generated must be tested by means of biochemical assays or isogenic systems that differ, insofar as possible, with respect to just one factor. Conversely, hypotheses based on experiments with particular isogenic cell sets can be assessed for generality according to whether they correctly predict responses for most of the 60 cell lines in the screen. For example, the overall impact of p53 function on cellular chemosensitivity can be affected by multiple genotypic and phenotypic factors that determine the balance between p53-mediated apoptosis on the one hand and G_1 arrest and DNA repair on the other (*29*); results obtained for one parental cell type can be misleading if generalized to others. The target and activity databases have, increasingly, provided us with a basis for rational choice of parental and transfected cell pairs to use in experiments addressing particular biological questions.

REFERENCES AND NOTES

1. M. R. Boyd, *Princ. Pract. Oncol. Update* **3**, 1 (1989); A. Monks *et al.*, *J. Natl. Cancer Inst.* **83**, 757 (1991); M. R. Grever, S. A. Schepartz, B. A. Chabner, *Semin. Oncol.* **19**, 622 (1992); S. F. Stinson *et al.*, *Anticancer Res.* **12**, 1035 (1992); M. R. Boyd, in *Current Therapy in Oncology*, J. E. Neiderhuber, Ed. (Mosby, St. Louis, MO, 1992), pp. 11–22.
2. K. D. Paull *et al.*, *J. Natl. Cancer Inst.* **81**, 1088 (1989); H. N. Jayaram, *Biochem. Biophys. Res. Commun.* **186**, 1600 (1992); K. D. Paull, E. Hamel, L. Malspeis, in *Cancer Chemotherapeutic Agents*, W. E. Foye, Ed. (American Chemical Society, Washington, DC, 1993), pp. 1574–1581; M. R. Boyd and K. D. Paull, *Drug Dev. Res.* **34**, 91 (1995).
3. M. Gupta *et al.*, *Mol. Pharmacol.* **48**, 658 (1995); E. Solary *et al.*, *Biochem. Pharmacol.* **45**, 2449 (1993); F. Leteurtre, G. Kohlhagen, K. D. Paull, Y. Pommier, *J. Natl. Cancer Inst.* **86**, 1239 (1994); F. Leteurtre *et al.*, *Biochem. Pharmacol.* **49**, 1283 (1995).
4. E. S. Cleveland *et al.*, *Biochem. Pharmacol.* **49**, 947 (1995).
5. R. Bai *et al.*, *J. Biol. Chem.* **266**, 15882 (1991); K. D. Paull, C. M. Lin, L. Malspeis, E. Hamel, *Cancer Res.* **52**, 3892 (1992); S. C. Kuo *et al.*, *J. Med. Chem.* **36**, 1146 (1993); E. Hamel, C. M. Lin, H. K. Wang, K. H. Lee, K. D. Paull, *Biochem. Pharmacol.* **3**, 53 (1996).
6. J. N. Weinstein *et al.*, *Science* **258**, 447 (1992).
7. J. N. Weinstein *et al.*, *Stem Cells* **12**, 13 (1994); B. A. Chabner, J. N. Weinstein, K. D. Paull, M. R. Grever, in *Cancer Treatment, an Update*, P. Banzet, J. F. Holland, D. Khayat, M. Weil, Eds. (Springer-Verlag, Paris, 1994), pp. 10–16; A. D. Koutsoukos *et al.*, *Stat. Med.* **13**, 719 (1994); S. E. Bates *et al.*, *J. Cancer Res. Clin. Oncol.* **121**, 495 (1995).
8. W. W. van Osdol, T. G. Myers, K. D. Paull, K. W. Kohn, J. N. Weinstein, *J. Natl. Cancer Inst.* **86**, 1853 (1994). A Kohonen neural network was used to generate self-organized 2D maps in which the distances among compounds reflected the differences in their patterns of activity.
9. NSC numbers 638850, 607097, 630176, 650426, and 649890, respectively. These compounds were selected for development in part because their patterns of activity in the screen were unlike those of any agent already in the clinic.
10. Database resources currently available on the World Wide Web include the following. (i) http://epnws1.ncifcrf.gov:2345/dis3d/cancer_screen/stdmech.html: mechanism of action assignments and chemical structures for a set of 122 standard agents based on (*6*) and (*8*). Clicking on a mechanism of action displays a list of the relevant compounds. (ii) ftp://helix.nih.gov/ncidata/canscr/stdagnt.tar.Z: activity patterns for standard agents based on (*6*) and (*8*). (iii) http://epnws1.ncifcrf.gov:2345/dis3d/itb/stdagnt.html: searching by chemical name or NSC number in a set of 175 standard agents. (iv) http://epnws1.ncifcrf.gov:2345/dis3d/cancer_screen/nsc4.html: retrieval by NSC number of "mean graph" representations (*2*) of activity patterns and COMPARE lists for approximately 20,000 nonconfidential compounds. (v) http://epnws1.ncifcrf.gov:2345/dis3d/cancer_screen/cmpmatrix.html: generation of an A·A′ matrix (nonclustered) for any choice of nonconfidential compounds. (vi) http://epnws1.ncifcrf.gov:2345/dis3d/itb/pubtarget.html: published molecular target measurements, along with the ability to use the measurements as seeds in COMPARE searches of the activity databases for synthetic compounds and for natural product extracts.
11. G. W. A. Milne, M. C. Nicklaus, J. S. Driscoll, S. Wang, D. W. Zaharevitz, *J. Chem. Inf. Comput. Sci.* **34**, 1219 (1994) and World Wide Web page by D. W. Zaharevitz (http://epnws1.ncifcrf.gov:2345/dis3d/3Ddatabase/dis3d.html).
12. See, for example, S. Wang *et al.*, *J. Med. Chem.* **37**, 4479 (1994).
13. More specifically, indices for p53, G_1 and G_2 checkpoint integrity, Gadd45, $p21^{Cip1/Waf1}$, Mdm2, H-Ras, K-Ras, N-Ras, Raf, Src, Nm23, Pgp/Mdr-1, Mdr-2, Rb, Mrp, Lrp, telomerase, telomere length, alkylguanine transferase, metallothionein, phospho-

inositol 3-kinase, thioredoxin, aldehyde dehydrogenase, epidermal growth factor receptor, transforming growth factor–α, c-ErbB2, fibroblast growth factor receptor, vascular endothelial growth factor receptor, human growth factor receptor, transforming growth factor–β receptor type II, Bcl-2, Bax, Bcl-X, DNA methylation, DT diaphorase, p450 reductase, b5 reductase, thymidylate synthetase, mismatch repair defects, topoisomerase II, galectin, p15, p16, and glutathione transferase isoenzymes.

14. J. K. Buolamwini *et al.*, in preparation; T. G. Myers, *et al.*, *Electrophoresis*, in press. The 2D-PAGE approach can identify correlations with drug activity for cellular proteins not previously recognized to be functionally important (for example, p53, whose significance was not known until long after the molecule was discovered in 1979). Proteins identified to date on ISO-DALT gels (Large Scale Biology, Rockville, MD) for our database include Hsp60, Hsc70, Hsp90, Grp75, Grp78, protein disulfide isomerase, lamin B, β-actin, γ-actin, β-tubulin, three cytokeratins, two myosin light chains, tropomyosin, proliferating cell nuclear antigen, β-F1ATPase, calmodulin, endoplasmin, and calreticulin.
15. M. R. Wilkins *et al.*, *Biotechnol. Gene Eng. Rev.* **13**, 19 (1995). The term "proteome" was introduced, by analogy with "genome," to denote in an aggregate sense the protein complement of a cell or organism.
16. L. Wu *et al.*, *Cancer Res.* **52**, 3029 (1992).
17. J.-S. Lee *et al.*, *Mol. Pharmacol.* **46**, 627 (1994); M. Alvarez *et al.*, *J. Clin. Invest.* **95**, 2205 (1995).
18. M. A. Izquierdo *et al.*, *Int. J. Cancer* **65**, 230 (1996).
19. S. Scala *et al.*, *Proc. Am. Assoc. Cancer Res.* **37**, 325 (1996).
20. T. G. Myers *et al.*, *ibid.* **36**, 305 (1995); *ibid.* **37**, 299 (1996).
21. All cluster analyses were performed by the average linkage method with Pearson correlation coefficient as metric. Other methods (single linkage, complete linkage, centroid-based algorithms) and metrics (Euclidean) also yield coherent patterns but emphasize different features. DISCOVERY uses SAS (Statistics Analysis Systems Institute, Carry, NC) or S-Plus (MathSoft, Seattle, WA) scripting and routines for some of these calculations. Neither the 60 cell lines nor the compounds tested represent random samples from defined underlying populations; hence, all of the statistical parameters used in this article should be considered as heuristic indices.
22. This set appears to be mechanistically different from the well-known anticancer α-formylpyridine thiosemicarbazones, 3-HP and 5-HP (3- and 5-hydroxypyridine-2-carboxaldehyde thiosemicarbazone), which are ribonucleotide reductase inhibitors. The latter two compounds and four of their analogs appear side by side in a separate group (compounds 3037 to 3042). Two structural features distinguish the group of 72 from the HP family: Most are synthetically derived from α-acetyl (rather than α-formyl) heterocyclics and generally have a large, lipophilic moiety at the end distal to the imine linkage.
23. L. Pendyala and P. J. Creaven, *Cancer Res.* **53**, 5970 (1993).
24. Autocorrelation analysis was performed on a set of 43 consecutive screenings of doxorubicin (which is used as a routine control in each experiment) performed over a period of 8 months. The Pearson correlation coefficient for any pair of tests was essentially independent of time elapsed between the tests (*36*).
25. K.-V. Chin, I. Pastan, M. M. Gottesman, *Adv. Cancer Res.* **60**, 157 (1993); M. M. Gottesman and I. Pastan, *Annu. Rev. Biochem.* **62**, 385 (1993).
26. The linear methods described here cannot capture nonlinear, interactive aspects of the biological phenomena. However, the matrix multiplication used to obtain Fig. 2 can be replaced by any chosen nonlinear mathematical operator, statistical operator, or artificial intelligence–based algorithm. The various matrices summarize patterns of information, but the robustness of correlations is also examined on a number-by-number basis, and nonparametric bootstrap confidence limits for the correlation coefficient estimates are calculated when desired.
27. M. Hollstein, D. Sidransky, B. Vogelstein, C. C. Harris, *Science* **253**, 49 (1991); A. J. Levine, J. Momand, C. A. Finlay, *Nature* **351**, 453 (1991); M. S. Greenblatt, W. P. Bennett, M. Hollstein, C. C. Harris, *Cancer Res.* **54**, 4855 (1994).
28. G. P. Zambetti and A. J. Levine, *FASEB J.* **7**, 855 (1993).
29. See, for example, S. Friend, *Science* **265**, 334 (1994); L. H. Hartwell and M. B. Kastan, *ibid.* **266**, 1821 (1994); M. L. Smith and A. J. Fornace Jr., *Mutat. Res.* **340**, 109 (1996); C. C. Harris, *Carcinogenesis* **17**, 1187 (1996).
30. P. M. O'Connor *et al.*, in preparation.
31. Modified from S. Marsoni *et al.*, *Cancer Treat. Rep.* **71**, 71 (1987).
32. This finding is consistent with observations by S. Fan *et al.* [*Cancer Res.* **54**, 5824 (1994)] in Burkitt's lymphoma cells. Interestingly, the small, highly Mdr-1–susceptible subgroup seen in Fig. 2 as compounds 3043 to 3069 includes nine anthracyclines, among them doxorubicin, deoxydoxorubicin, daunomycin, and rubidazone. The apparent differences between antimitotics and DNA-damaging agents are increased somewhat by confounding correlations with Mdr-1, as indicated when cell lines high in Mdr-1 were excluded from the p53 analyses.
33. One such set of criteria included the Pearson correlation coefficient and Wilcoxon *P* value (as defined for Fig. 4) with respect to p53 sequence, the median difference in sensitivity between p53 wild-type and p53-mutant cell lines, and the mean potency of the compound. Additional triage was done on the basis of availability of the compound for testing and uniqueness of activity pattern and molecular structure. To cite one example, a subset of the ellipticines scored well with respect to those criteria.
34. The p53 isogenic sets used here include p53 wild-type parental cells, p53-disrupted derivatives, and control cells [see, for example, S. Fan *et al.*, *Cancer Res.* **55**, 1649 (1995)]. The results obtained can depend markedly on the cell type and genotypic context.
35. Molecular characterization of clinical tumor cells is central to the NCI's recently announced Cancer Genome Anatomy Project (CGAP).
36. T. G. Myers and J. N. Weinstein, data not shown.
37. K. Wosikowski *et al.*, in preparation.
38. H.-M. Koo *et al.*, *Cancer Res.* **56**, 5211 (1996).
39. For continuing collaboration on the screening endeavors, we thank the staff of the DTP and SAIC NCI-FCRDC, especially J. Mayo, V. Narayanan, R. Schultz, R. Camalier, J. Johnson, K. Hite, A. Chiausa, P. Svetlik, and D. Segal. LMP was part of the DTP when much of this work was done. The molecular target data discussed here are the contributions of many others, including J. Jackman, I. Bae, M. Alvarez, and K. Wosikowski. We thank L. Muenz for critique of the statistical issues, and M. R. Boyd, R. Shoemaker, M. Alley, B. A. Chabner, G. Vande Woude, M. R. Grever, and S. A. Schepartz for developing the current screening program and supporting the molecular targets enterprise.