Searching for pharmacogenomic markers: The synergy between omic and hypothesis-driven research

John N. Weinstein*
Laboratory of Molecular Pharmacology, National Cancer Institute, Bethesda, MD, USA

With 35,000 genes and hundreds of thousands of protein states to identify, correlate, and understand, it no longer suffices to rely on studies of one gene, gene product, or process at a time. We have entered the "omic" era in biology. But large-scale omic studies of cellular molecules in aggregate rarely can answer interesting questions without the assistance of information from traditional hypothesis-driven research. The two types of science are synergistic. A case in point is the set of pharmacogenomic studies that we and our collaborators have done with the 60 human cancer cell lines of the National Cancer Institute's drug discovery program. Those cells (the NCI-60) have been characterized pharmacologically with respect to their sensitivity to $> 70,000$ chemical compounds. We are further characterizing them at the DNA, RNA, protein, and functional levels. Our major aim is to identify pharmacogenomic markers that can aid in drug discovery and design, as well as in individualization of cancer therapy. The bioinformatic and chemoinformatic challenges of this study have demanded novel methods for analysis and visualization of high-dimensional data. Included are the color-coded "clustered image map" and also the MedMiner program package, which captures and organizes the biomedical literature on gene-gene and gene-drug relationships. Microarray transcript expression studies of the 60 cell lines reveal, for example, a gene-drug correlation with potential clinical implications – that between the asparagine synthetase gene and the enzyme-drug L-asparaginase in ovarian cancer cells.

Keywords: Microarray, genomics, proteomics, omics, cancer, cell line, pharmacology, pharmacogenomics, molecular marker, cancer therapy, clustered image map, MedMiner

*Address for correspondence: Dr. J.N. Weinstein, M.D., Ph.D., Bldg. 37, Rm 4E-28, NIH, 9000 Rockville Pike, Bethesda, MD 20892, USA. Tel.: +1 301 496 9571; Fax: +1 301 402 0752; E-mail: Weinstein@dtpax2.ncifcrf.gov.

Disease Markers 17 (2001) 77–88
ISSN 0278-0240 / $8.00

1. Introduction

A prediction: Future historians of science will refer to the turn of the millennium as a watershed, the start of a Golden Age of biomedical science [1]. They will note – in passing and without much excitement – the half-century prodromal period after Watson-Crick in 1953, during which increasingly powerful techniques were developed to study one gene or gene product at a time and during which the foundations of high throughput molecular biology were laid down. But, they will be distinctly impressed by completion of the DNA sequences of small organisms just before the turn of the century and quasi-completion of the human sequence soon after it. Beyond simply the sequence, they will focus on development at this time of large databases on transcript and protein expression patterns, single nucleotide polymorphisms, chromosomal aberrations, and epigenetic changes. They will appreciate the increasing integration of these massive new molecular biology databases with those from structural and combinatorial chemistry, x-ray crystallography, magnetic resonance spectroscopy, high-throughput screening, two-hybrid and fluorescence energy transfer studies of protein-protein interaction, epidemiological studies, and the clinic.

All of these developments – which are rapidly transforming our ability to identify and use molecular markers of disease – reflect what can be termed "omic" research [1–3]. Omic research includes studies in genomics, proteomics, transcriptomics, CHOmics (for the carbohydrates), kinomics (for the kinases), and methylomics (for epigenetic methylations and imprinting), among many others. It also includes compound forms like pharmacogenomics, functional genomics, structural genomics, and pharmacomethylomics [3]. Notions such as immunomics, metabolomics, toxicomics, literomics, and ecogenomics have been introduced, not entirely in jest. It's not that we really need

more jargon, but, aside from any amusement value, the omic terminology can be a useful shorthand – and it is at least etymologically respectable. Webster's dictionary defines "-ome" as an abstract entity, group, or mass, so omic research in biology is the study of entities in aggregate – DNA, RNA, protein, or other molecular components of a cell, tissue, or organism. The substantive point here is that omic research requires a different mind-set from the more traditional study of one gene, gene product, or process at a time [1–3]. One generally ends up knowing a little about a lot, rather than a lot about a little. Often, the databases of molecular information are generated without knowing what about them will prove most valuable, but that fact in no way obviates the need for careful design and rigorous attention to experimental detail. In a sense, the guiding hypothesis in omic research relates to information and its utility, rather than to biological specifics. But anyone who does omic research quickly realizes its dependence on traditional one-at-a-time hypothesis-driven studies. The former type of research establishes context in a world of 35,000 genes and hundreds of thousands of interesting protein states; the latter identifies what data to generate and which relationships in the final database are worth further pursuit.

This synergy between traditional and omic approaches to biology is reflected in the way we identify and validate molecular markers of disease and molecular markers for therapy. The aim of this article is to illustrate that synergy through our studies with the drug discovery and development program of the National Cancer Institute (NCI). The NCI's cell-based screen, in which $> 70,000$ chemical compounds plus natural products have been tested one at a time and independently over the last 11 years, provides a unique opportunity complementary to the study of clinical tumors. Cancer cell lines clearly are not the same as cancer cells in vivo. Even primary cultures from tumors are artificial in that they have been removed from their natural state and society in the body. But cultured cells do at least circumvent many of the logistical, technical, ethical, and conceptual difficulties that complicate work with clinical materials, and one can step into the same stream multiple times. Most of our present understanding of basic molecular pharmacology has come from studies in cultured cells, not from clinical materials. However, projecting in the other direction – from cultured cells toward the clinic – is more dangerous. One can hope to find clues with which to formulate hypotheses for further study.

2. The NCI-60 panel of human cancer cell lines

In 1990, the NCI Developmental Therapeutics Program (DTP) began operation of what was then considered a rather high-throughput screen, in which compounds are tested for their ability to inhibit growth of 60 different human cancer cell lines (the NCI-60) in culture [4–9]. Included currently are melanomas (8 cell lines), leukemias (6), and cancers of breast (8), prostate (2), lung (9), colon (7), ovary (6), kidney (8), and central nervous system (6) origin. The assay is a simple one. The cells are incubated with various concentrations of drug for 48 hours, and growth inhibition is then assessed using a sulforhodamine B assay for the amount of protein in the well. Fifty percent growth inhibitory concentrations (GI_{50}'s) and other indices of potency are then read from the resulting dose-response curves. The top section of Figure 1 shows a highly schematic view of this part of the NCI drug discovery-development process. The compounds have come largely from synthetic chemistry and natural product sources, but biologicals and combinatorial libraries are also being tested. In recent years, the role of this process has changed progressively from primary screening to secondary testing as compounds have, increasingly, been selected for the assay on the basis of interesting prior information, and as molecular screens have been established in the program.

This cell-based strategy for drug discovery was originally based on the hypothesis that selective activity in vitro against cancer cell lines from a particular organ would predict selective activity against the corresponding tumor types in humans. For present purposes, however, we will avoid the endless arguments about the best way to screen or test for anticancer agents and focus on the screen as a generator of profile data on the potencies of compounds tested and the drug sensitivities of the 60 cell types. Patterns of activity against the NCI-60 have proved predictive at the molecular level; they often provide incisive information on mechanisms of action and also on molecular targets and modulators of activity within the cancer cells.

The patterns of activity were first analyzed using the COMPARE algorithm developed by the late K.D. Paull [5,10,11]. Given one compound as a "seed", COMPARE searches the database of agents screened and generates a list of those most similar to the seed in their patterns of activity against the 60 cell lines. Similarity in pattern generally indicates similarity in mechanism of action, mode of resistance, and molecular structure [10–14]. This form of analysis has been

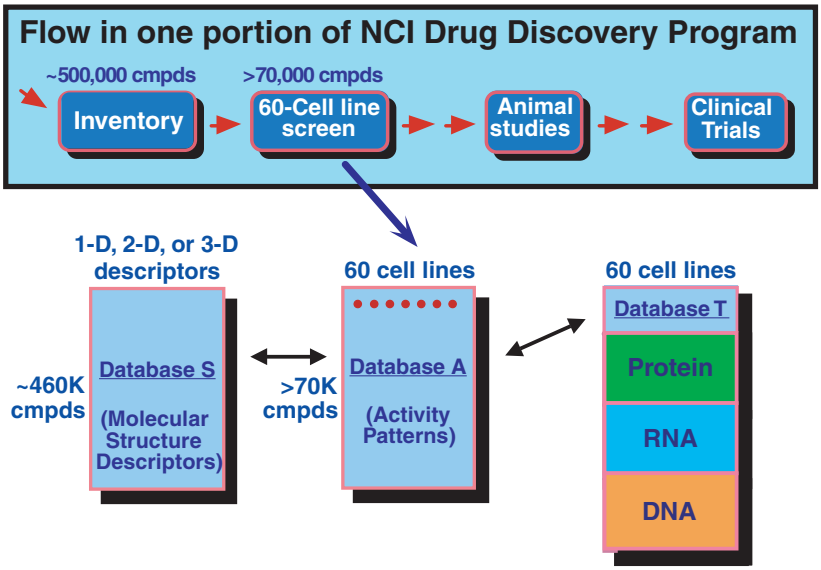

Fig. 1. Schematic of the NCI-60 screen and profiling system, with associated databases of activities (A), molecular structure descriptors of the compounds tested (S), and molecular "targets" in the cells (T). The T-database includes measurements of one target at a time and aggregate (omic) measurements at the DNA, mRNA, and protein levels. Conceptually, there is also a clinical features database (C), not shown here. The informatics challenge is to analyze and understand each of these databases separately, then to integrate them with each other and with public information resources to address pharmacogenomic questions. Modified from [14].

applied productively to topoisomerase 2 inhibitors [15], pyrimidine biosynthesis inhibitors [16], and tubulin-active compounds [17,18], among many other classes of agents. We have used back-propagation neural networks and predictive methods from classical statistics to find ways in which the patterns of activity could indeed predict a compound's mechanism of action [12]. More detailed information on the relationship between pattern and mechanism has come from a variety of other statistical and artificial intelligence techniques [13,14, 19–26].

3. Structure, activity, and target databases

The bottom half of Fig. 1 shows three types of databases that arise from the NCI-60 screen [14]: (A) contains the activity patterns, (S) contains molecular structural features of the tested compounds, and (T) contains characteristics of the cells that may be targets or modulators of drug activity or may be neither.

The chemical structures in (S) can be coded in terms of any set of 1-, 2- or 3-dimensional molecular structure descriptors, or a combination thereof. The NCI's Drug Information System (DIS) contains structural builds for $\sim 500,000$ molecules, including most of the $> 70,000$ tested to date ([27] and D. Zaharevitz, et al., unpublished). This database provides a basis for pharmacophoric searches; if a tested compound is found to have an interesting pattern of activity, its structure can be used to search for similar molecules in the DIS database that have not been tested.

More pertinent for present purposes is the target (T) database, each row of which defines the 60-cell line pattern of a measured cell characteristic [14]. Many laboratories at the NCI and elsewhere have been assessing these targets one at a time (or a restricted class at a time). The list includes oncogenes, tumor suppressor genes, molecules of the cell cycle and apoptotic pathways, drug resistance-mediating transporters, metabolic enzymes, cytokine receptors, heat shock proteins, telomerase, DNA repair enzymes, intracellular signaling molecules, and components of the cytoarchitecture. But a number of years ago, we decided to take a broader brush, omic approach to characterization of these cells – at the DNA, RNA, and protein levels. We started where any molecular pharmacologist would, given a choice: with the proteins.

4. Pharmacoproteomics and the NCI-60

In collaboration with Leigh Anderson (Large Scale Biology, Inc.), we [28] assessed patterns of protein expression by two-dimensional polyacrylamide gel electrophoresis (2-D PAGE) with detection by colloidal

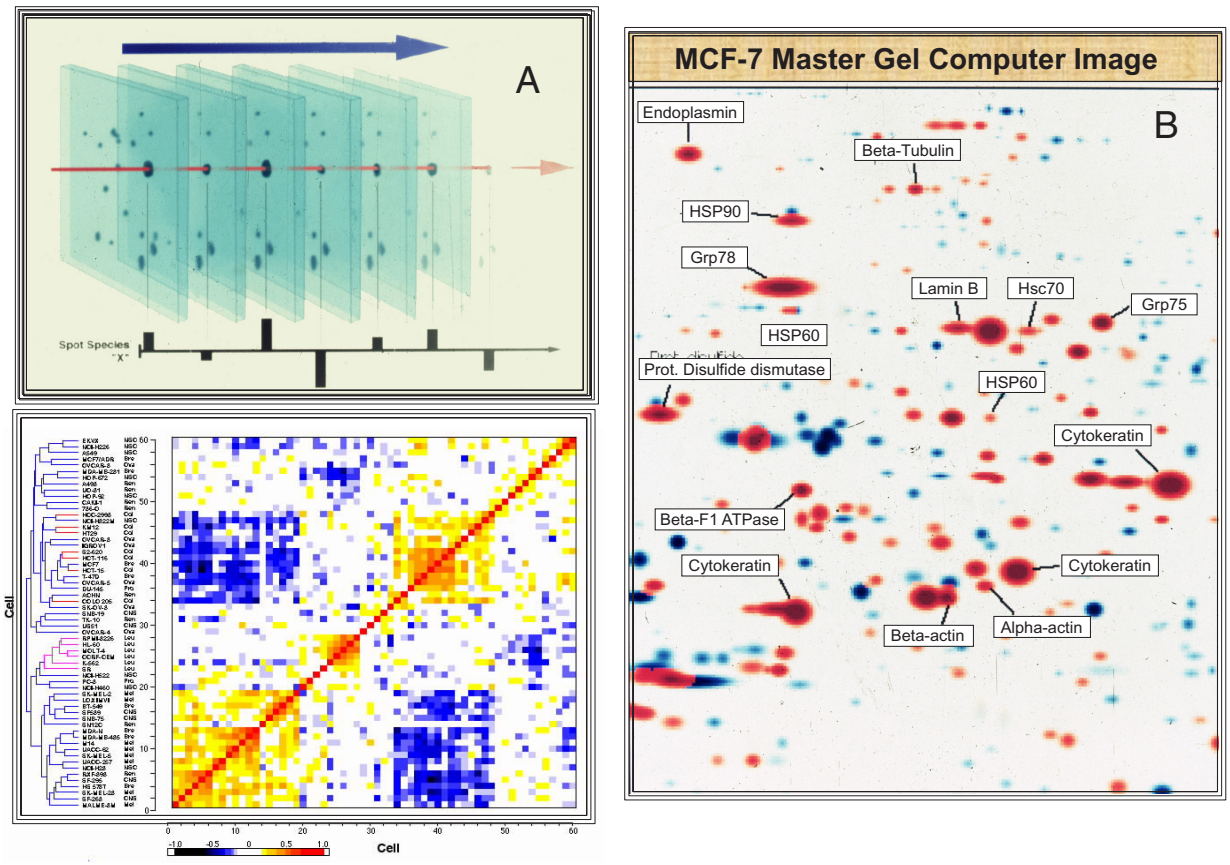

Fig. 2. Proteomic profiling of the NCI-60 cell lines [28]. A: 2-D gels were run for duplicate harvests of the 60 cell lines, the aim being to index spots across the 60 and develop quantitative patterns of protein expression analogous to those for compound activities. B: Computer-processed pseudocolor image of the central section of a master gel based on breast cancer line MCF-7. The spots are modeled as bivariate Gaussians of Coommassie blue intensity, and spot "volumes" are calculated as the integral of intensity over the area of the spot. Red indicates spots in a quality-controlled database of 151; blue indicates additional spots in the overall cross-indexed database of 1,014 spots. C: Clustered image map (CIM) showing cell-cell relationships in terms of patterns of protein expression. Red indicates positive Pearson correlation coefficient; blue indicates negative Pearson correlation coefficient. The cells are clustered in the same order on both axes, so there is, by definition, 100% correlation on the main diagonal. This is a TT.T CIM (see discussion of CIMs later in text). Modified from Buolamwini, et al. (in preparation).

Coomassie blue and image processing by the Kepler program package. Figure 2 summarizes that project, which established an early link between the enterprise of proteome research [29,30] and the molecular pharmacology of cancer. The database generated consisted of 1,014 indexed and quantitated protein spots, of which 151 were quality-controlled over all 60 cell lines and incorporated into a primary data set for analysis [28]. The database was informationally coherent in the sense that different harvests of the same cell line were more highly correlated with each other in expression pattern than were parallel harvests of different cell lines. That is, the signal-to-noise ratio was sufficiently high to permit meaningful clustering of the cell lines on the basis of their patterns of protein expression. For this purpose, the 2-D gel spots were quantitated in terms of spot "volume", intensity of staining integrated over the area of the computer-processed spot image.

The bottleneck in the project turned out to be identification of the spots. It was possible to distinguish meaningful patterns of association between spots or between cell types without knowing the identities of the spots. But for most purposes, including the search for molecular markers, spot identity proved crucial. For the identification, we developed our own version of a rapid MALDI-TOF mass spectrometric technique based on peptide mapping [31]. The essential steps in the method included in-gel digestion of the proteins with combinations of proteases, purification of the peptides, analysis by MALDI-TOF mass spectrometry, and peptide fingerprinting. We used the method to identify a number of spots but soon realized that it was not the

job of a small academic laboratory to identify hundreds of proteins in that way. Accordingly, we decided to move on to mRNA expression profiling and wait for high-throughput proteomics to catch up. The wait has been longer than I expected. Despite numerous promising techniques, most of them based on mass spectrometry for detection, there still does not seem to be a complete solution to the proteomic profiling of mixtures as complex as those of mammalian cells. Even the nature and magnitude of the challenge become harder and harder to define, given the increasing focus on alternative splicings, post-translational modifications, and extensive, complex family relationships among proteins and their domains. We will all await with interest the results of ongoing large-scale proteomic efforts in the public and private sectors.

5. Transcriptomics and the NCI-60

Most drug targets are proteins, and, clearly, proteomic status cannot be inferred or predicted from data on the RNA. Not yet, at least. Complicating factors include the complexities of translational regulation, post-translational modifications, and differing patterns of protein metabolism and degradation. However, mRNA expression levels are a useful second best, and the technology for determining them is considerably more advanced than it is for proteins. Most important, it is easier to establish identities. We have performed gene expression profiling studies of the NCI-60 using cDNA microarrays [32,33] with the Brown/Botstein laboratory at Stanford University and Affymetrix oligonucleotide chips [34] with the Lander/Golub group at the Whitehead Institute. The cDNA microarray studies profiled approximately 8,000 distinct genes using the two-color methodology [32,33]. Figure 3 shows hierarchical clustering of the cells based on gene expression patterns (left) and on drug sensitivities (right). In each case, the cells group in part by organ of origin but in part according to other principles. It was a surprise, though perhaps it should not have been, that the two clusterings are very different. The correlation of correlations between them [33] is only $+0.21$. At least one reason is that particular gene products, for example mdr1/Pgp, can influence the activities of many drugs across organ of origin categories but, being only single genes, have little effect on the clustering by gene expression pattern. We have since gone on to cross-compare the cDNA array and oligonucleotide chip databases gene by gene and establish a robust database of $> 2,000$ transcripts for which results from the two very different technologies are reasonably concordant across the 60 cell types (J.K. Lee, et al., in preparation). This concordance set is as well validated as any gene expression database of which we are aware. Conceptually, it is almost as if one had done northern blots or real-time RT-PCR studies for all of the genes across 60 cell lines to validate the cDNA array results. The drug and cDNA gene expression databases used in this study, along with tools of analysis, can be found at our web site, http://discover.nci.nih.gov. The oligonucleotide chip data will appear there soon. Additional data and the COMPARE program can be found at the DTP's web site, http://www.dtp.nci.nih.gov.

6. Color-coded clustered image maps (CIMS)

One useful and compact way to represent patterns in the data from "high-dimensional" datasets such as gene expression profiles is what we have termed the "clustered image map" (CIM) (sometimes called a clustered "heat map"). The principle is illustrated in Fig. 4 for gene expression over the 60 cell lines. We developed CIMs in the early 1990's for data on drug activities, target expression levels, gene expression values, and proteomic profiles [13,14,28,33]). The clustering of both axes (or sometimes only one if there is another organizing principle for the second axis) puts like together with like and brings out patterns. A red-green color scheme for the CIM has been popularized by our collaborators [35]. A flexible program for producing CIMs can be found at our web site, http://discover.nci.nih.gov.

The gene-cell CIM in Fig. 4 is simple in that, in terms of Fig. 1, it involves only a single database, T. If we want to assess relationships between drug activity and gene expression, it is necessary to map the A database into the T database (which can be done most straightforwardly by multiplying A by the transpose of T and normalizing so that entries in the product matrix ($A \cdot T^T$) are Pearson correlation coefficients [14, 33]. Figure 5 shows such a drug-target CIM. Alternatively, CIMs can be formed by multiplying a database (i.e., matrix) times its own transpose to produce a symmetrical product matrix [13,14,28,36]. For example, the $T^T \cdot T$ CIM expresses the correlation of each cell type with each other cell type in terms of pattern of expression, as in Fig. 2(C).

Each point and each patch of color in a CIM (such as that in Fig. 5) represents a possible story. But how can one determine whether a patch represents a causally

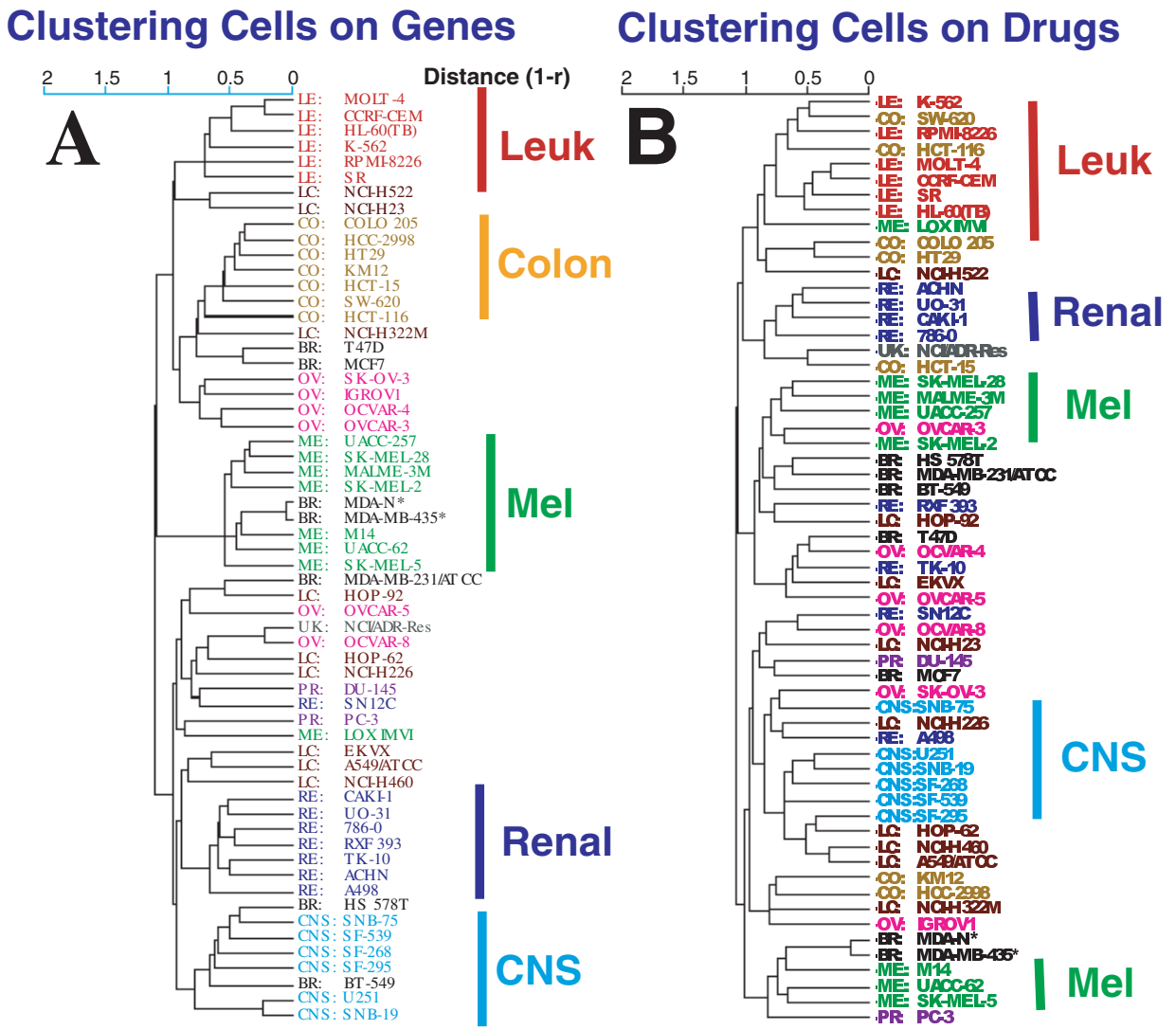

Fig. 3. Clustering of cells in two ways: Based on patterns of cDNA microarray gene expression patterns (A) and on drug sensitivity patterns (B). The two clusterings are very different, the overall "correlation of correlations" being only +0.21.*Indicates parental and transfectant cell lines from the pleural effusion of a breast cancer patient but expressing the proteins and transcripts characteristic of melanoma (as discussed in [32, 33]). Average linkage clustering and a correlation coefficient similarity metric were used in this analysis. Modified from [33].

interesting story, an epiphenomenal correlation (which still may identify a useful molecular marker), or statistical coincidence? The statistical robustness of association can be assessed in various ways, for example by using the bootstrap [37] to obtain approximate confidence limits on the estimated correlation coefficient and to test the null hypothesis that the true correlation is zero. But Fig. 5, which represents a small set of drugs and a relatively small set of genes, still reflects about 160,000 drug-gene pairs. By definition, 5% of these pairs (i.e., 8,000 of them) would appear to be statistically significant at the $P = 0.05$ level even if the data were just noise. There are too many false-positives. If this "multiple comparisons" problem is taken into account by making a Bonferroni correction (which assumes statistical independence), then almost all of the true correlations will be thrown out. There are too many false negatives. Other, more sophisticated corrections can be made but, ultimately, in this type of situation, the statistics can take one only so far. We are left with a long list of gene-drug (or gene-gene) correlations, each of which must be assessed for its biological sense. This problem is most acute for database associations such as those considered here, but it also

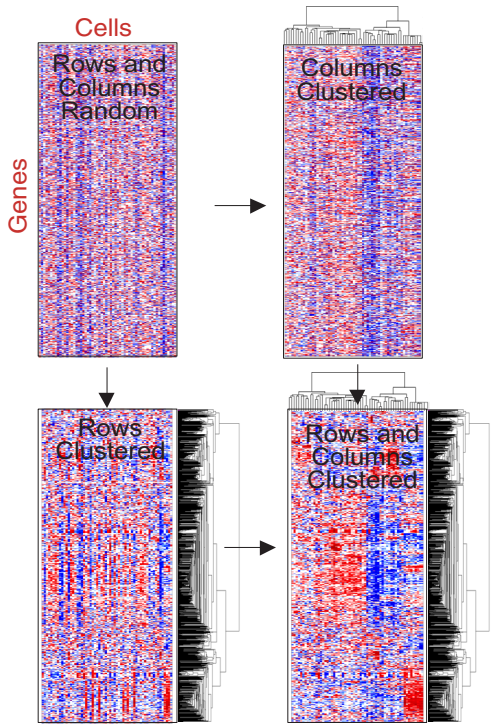

Fig. 4. Illustration of the principle of the clustered image map (CIM) for compact representation of high-dimensional data. Red in this case indicates high gene expression; blue indicates low expression. The clustering of like with like on both axes brings out pattern. (In some other cases, only one axis may be clustered if there is some other organizing principle for the other axis.) See [13,14,28,33,35].

pertains to the simplest binary experiments in which, for example, a malignant cell type or tissue is compared with its normal counterpart. Even with enough replicates to obviate the question of statistical significance, such experiments typically produce lists of hundreds of genes that differ in expression, and one is left to figure out which differences have biological plausibility.

This is where synergy between omic research and hypothesis-driven studies of particular genes and drugs becomes necessary. To figure out where to look in the massive databases that arise from the former, we generally need to make use of the latter. That can mean experiments done after the fact, it can mean plumbing rich public databases such as those of the NCI's Cancer Genome Anatomy Project [38,39], or it can mean laboriously searching the extant literature. Because literature searching quickly becomes tedious, we developed web-based text-mining and literature-organizing tools, MedMiner [40] and EDGAR [41], to facilitate the process.

7. Organizing the literature on gene-gene and gene-drug correlations: MedMiner and EDGAR

MedMiner, which is publicly available at our web site (http://discover.nci.nih.gov), can be used for gene, gene-gene, gene-drug, drug-drug, or more general literature queries. Input can include gene accession numbers, gene names, drug NSC numbers, drug names, and/or free text (e.g., "apoptosis" or "transport"). In the case of microarray analysis, the user can specify a list of arrayed genes. MedMiner uses a combination of GeneCards from the Weizmann Institute, PubMed from the National Library of Medicine (NLM), syntactic analysis, truncated-keyword filtering of relationals, and user-controlled sculpting of a Boolean query to generate key sentences from the pertinent abstracts. Those sentences are then organized so that the user can access the most pertinent ones directly by clicking on a relevance-term. Whole abstracts deemed to be of interest can then be accessed fluently and dropped into a "shopping basket" for display or for automated entry into an EndNote library. Experienced users have estimated that MedMiner speeds up 5- to 10-fold the process of capturing and organizing the literature from PubMed searches on lists of gene-gene and gene-drug relationships [40].

MedMiner is fast enough and transparent enough for real-world use on the Web, but it by no means captures all of the information that is theoretically available in the free text of an abstract. Natural language processing (NLP) is one of the great intellectual challenges, and a number of attempts are being made to harness NLP principles for omic studies. Our own effort in this direction is EDGAR, (Extraction of Data on Genes and Relationships), a software tool for semantic analysis and organization of the literature relevant to our studies in the molecular pharmacology of cancer [41]. Many different approaches can be used to the extract factual assertions from biomedical text. Methods used include syntactic parsing, processing of statistical and frequency information, and rule-based decision-making (reviewed in [41]). EDGAR draws on all of these, using a stochastic part of speech tagger in support of an underspecified syntactic parser. Fully general semantic analysis is unrealizable, so we had to develop suitable restricted ontologies and controlled vocabularies. The goal was to extract factual assertions in the form of first order predicate calculus statements about the relationships between genes and drugs in cancer therapy. EDGAR is strong on the identification of

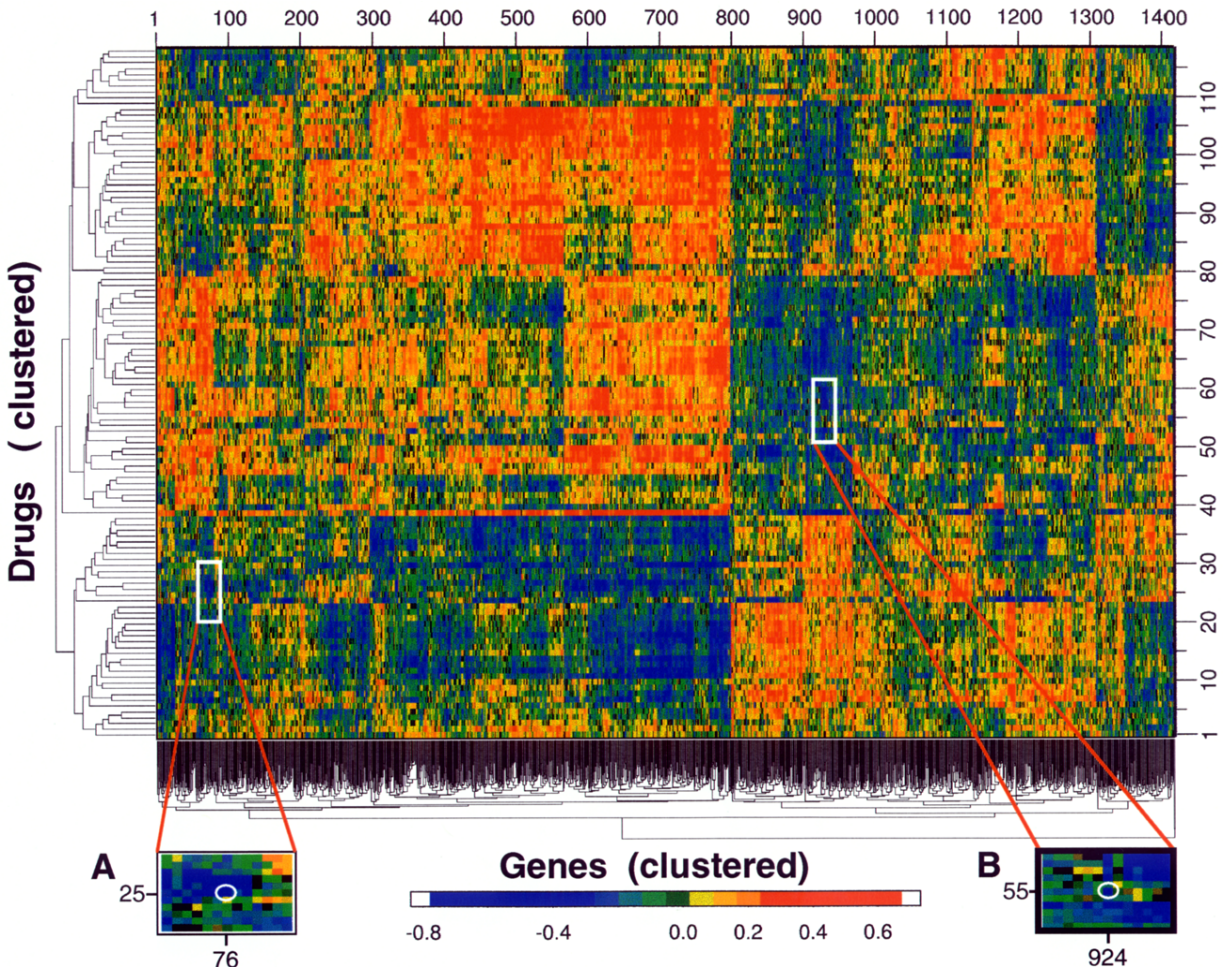

Fig. 5. Clustered image map (CIM) relating activity patterns of 118 tested compounds to the expression patterns of 1,376 genes in the 60 cell lines. Included in addition to the gene expression levels are data for 40 molecular targets assessed one at a time in the cells. A red point (high positive Pearson correlation coefficient) indicates that the agent tends to be more active (in the two-day assay) against cell lines that express more of the gene; a blue point (high negative correlation) indicates the opposite tendency. Genes were cluster-ordered on the basis of their correlations with drugs (mean-subtracted, average-linkage clustered with correlation metric); drugs were clustered on the basis of their correlations with genes (mean-subtracted, average-linkage clustered with correlation metric). Sharp edges of the colored patches reflect deep forks in the corresponding cluster tree. Insert A shows a magnified view of the region around the point (white circle) representing the correlation between the dihydropyrimidine dehydrogenase gene and 5-fluorouracil. Insert B is an analogous magnified view for the asparagine synthetase gene and the drug L-asparaginase. Modified from [33].

"referential" (i.e., noun-related) relationships, weaker with respect to "relational" (i.e., verb-related) ones. Interpretation of the referential vocabulary in EDGAR is based on NLP tools and knowledge sources developed at NLM. The primary knowledge source supporting EDGAR is the Unified Medical Language System (UMLS) Metathesaurus, a compilation of $> 600,000$ concepts from controlled vocabularies in the biomedical sciences. We tested EDGAR's capability by applying it to a set of 383 literature abstracts related to drug resistance mechanisms. The results, expressed in a cluster tree with 383 leaves, showed considerable coherence by drug and mechanism of action [41]. That was achieved without the manual reading of a single abstract. EDGAR is Web-based but not yet fast enough or transparent enough for public use. It illustrates, however, both the potential and the challenges of automated literature analysis in omic studies.

8. Pharmacogenomic markers

The two white rectangles on the gene expression vs. drug sensitivity CIM in Fig. 5 indicate stories with

likely causal significance on the basis of literature information.

8.1. Dihydropyrimidine dehydrogenase and 5-fluorouracil

5-Fluorouracil (5-FU), an antimetabolite drug often used against colorectal and breast cancer, can inhibit both RNA processing and thymidylate synthesis. Dihydropyrimidine dehydrogenase (DPYD), the rate-limiting enzyme in uracil and thymidine catabolism, is also rate limiting to 5-FU catabolism. Hence, high DPYD levels might be expected to decrease the activity of 5-FU. Consistent with this hypothesis, we found a highly significant negative correlation (-0.53) between DPYD gene expression and 5-FU potency against the 60 cell lines [33]. On closer examination, we found that 14 of the 18 low-expressers of DPYD (> 4-fold lower than the reference pool) are sensitive or highly sensitive to 5-FU. Perhaps not coincidentally, given the clinical use of 5-FU against colon cancer, all of the colon-derived cell lines (7 out of 7) were sensitive to 5-FU and low in DPYD expression. Previous studies of DPYD correlations in clinical materials have been difficult to interpret, but these microarray data suggest further study of DPYD as a pharmacogenomic marker [33].

8.2. Asparagine synthetase (ASNS) and L-asparaginase

Many acute lymphoblastic leukemias (ALL) lack asparagine synthetase (ASNS) and therefore must scavenge exogenous L-asparagine to survive (see Fig. 6). This dependence is exploited by treating ALL and other lymphoid malignancies with bacterial L-asparaginase, which depletes extracellular L-asparagine and selectively starves the cancer cells. As shown in Fig. 7, we found a moderately high negative correlation (-0.44; bootstrap 95% confidence interval -0.59 to -0.25) between expression of the ASNS gene and L-asparaginase sensitivity in the 60 cell lines [33]. But we also knew to look specifically at the leukemic subpanel, and there the correlation was a striking -0.98 (bootstrap 95% confidence interval -1.00 to -0.93). This value survived even a Bonferroni correction for the statistical multiple comparisons problem. Furthermore, the two ALL-derived lines expressed the lowest levels of ASNS mRNA and were the most sensitive to L-asparaginase, as might have been expected. These results supported the possible use of ASNS as a marker for clinical decisions about L-asparaginase therapy [33].

The next question was obvious: Would any other cell line panel show similar correlation. The answer was "yes", though not as strongly. The correlation coefficient for the ovarian lines was -0.88 (confidence interval -0.23 to -0.99) [33]. Early clinical trials done with a scattering of solid tumors showed occasional responses to L-asparaginase in melanoma, chronic granulocytic leukemia, lymphosarcoma, and reticulum cell sarcoma but not in other tumor types (see [33] for references). The microarray findings support a closer look at L-asparaginase therapy for solid tumors, particularly for a low-ASNS subset of ovarian cancers. The preferred material for a clinical trial would be the polyethylene glycol-modified forms of L-asparaginase, which shows much better pharmacokinetic and immunological properties than does the native bacterial form of the enzyme. Studies of asparaginase/L-asparaginase correlations in clinical materials are underway in collaboration with D. Von Hoff and his research group at the Arizona Cancer Center.

9. Concluding remarks

As indicated by the foregoing examples, omic and hypothesis-driven research should be seen as synergistic, not mutually exclusive. But there is a paradox: the easiest associations to identify in an omic database are the least interesting: ones that have been identified previously. Next easiest to identify are those that, with hindsight, make biological or pharmacological sense. Hardest are those that would be most exciting: the unexpected, the paradigm shifters. These tend to get lost among the multitude of false-positives. The problem is most acute for cross-database comparisons, less so but still considerable for binary experimental designs and time-course studies. In this paper, I have emphasized the effort to find markers of sensitivity to a treatment. One can also ask a complementary question about the molecular consequences of therapy. Both omic and hypothesis-driven studies to address the latter type of question are ongoing in our own and many other laboratories [42].

Another type of synergy deserves at least brief mention. Gene expression profiling is in vogue at the moment, but, clearly, no single type of molecular information can capture all of the pharmacological and toxicological phenomena relevant to drug discovery and selection of therapy. Data on DNA sequence, transcript

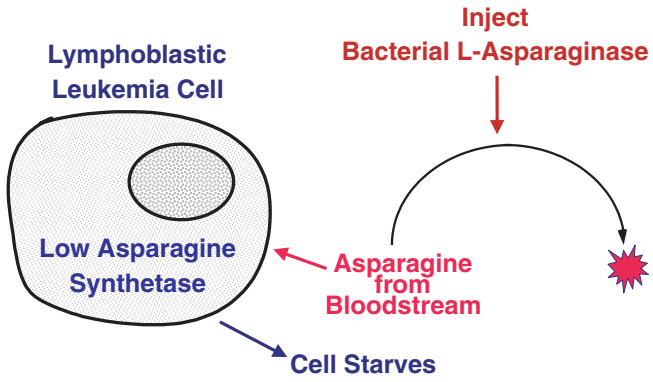

Fig. 6. Schematic of the mechanism of L-asparaginase activity in acute lymphoblastic leukemia. See explanation in text.

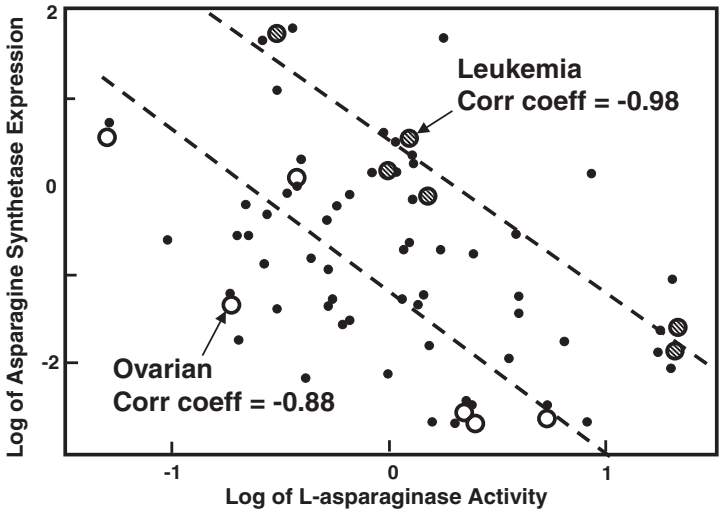

Fig. 7. Relationship between asparagine synthetase transcript expression and chemosensitivity of the NCI-60 to L-asparaginase. Each point represents one of the 60 cell types. Leukemia and ovarian points are larger, open circles. Main effects have been removed for both cells and drugs. Hence, a -log(GI_{50}) value of 1 for sensitivity indicates a 10-fold higher than average sensitivity of the cell line to the agent. The asparagine synthetase expression level is plotted as the relative $\log_2$ abundance of the asparagine synthetase transcript. A value of $+2$ indicates 4-fold higher expression than in the reference pool.

expression, protein expression, chromosomal aberrations, chromosomal copy number changes, single nucleotide polymorphisms, promoter methylation, and molecular interactions, inter alia, can all contribute to our understanding. But each provides only partial insight. As our laboratory and collaborators combine these different classes of information for the NCI-60, it becomes progressively more apparent that they are synergistic.

Acknowledgments

I am very grateful to the many members of my research group, the Genomics and Bioinformatics Group in the Laboratory of Molecular Pharmacology, NCI for their individual and collective contributions to the projects touched on here. Included prominently have been U. Scherf, M. Waltham, W.C. Reinhold, T.G. Myers, Y. Zhou, L.H. Smith, L. Tanabe, J.K. Lee, S. Richman, F. Gwadry, S. Kim, Ajay, H. Kouros-Mehr, J. Alexander, S. Daoud, S. Nishizuka, and K. Bussey. Other principal collaborators in various facets of the proteomic and transcriptomic work have included Y. Pommier, K.W. Kohn, D. Ross, M. Eisen, P.O. Brown, D. Botstein, D. Shalon, D. Lashkari, E. Liu, L. Miller, E. Lander, T. Golub, D. Slonim, H. Coller, P. Tamayo, J. Staunton, and N.L. Anderson. I am especially grateful to members of the NCI Developmental Therapeutics Program, who have made this molecular targets effort

work. Prominent in this regards are E.A. Sausville, M. Grever, A. Monks, D.A. Scudiero, D. Zaharevitz, R. Camalier, S. Holbeck, and K.D. Paull. In particular, the late Kenneth Paull laid the foundations for the informatics of this program, and our subsequent contributions simply follow in his footsteps.

References

[1] J.N. Weinstein and J.K. Buolamwini, Molecular targets in cancer drug discovery: Cell-based profiling, *Current Pharmaceutical Design* **6** (2000), 473–483.

[2] J.N. Weinstein, Fishing Expeditions, *Science* **282** (1998), 627–628.

[3] J.N. Weinstein, Pharmacogenomics: Teaching old drugs new tricks, *New England J. Med.* **343** (2000), 1408–1409.

[4] M.R. Boyd, The future of new drug development, in: *Current Therapy in Oncology*, J.E. Neiderhuber, ed., Philadelphia, Decker, 1992, pp. 11–22.

[5] M.R. Boyd and K.D. Paull, Some practical considerations and applications of the National Cancer Institute in vitro anticancer drug discovery screen, *Drug Development Research* **34** (1995), 91–109.

[6] M.C. Alley, D.A. Scudiero, A. Monks, M.L. Hursey, M.J. Czerwinski, D.L. Fine, B.J. Abbot, J.G. Mayo, R.H. Shoemaker and M.R. Boyd, Feasibility of drug screening with panels of human tumor cell lines using a microculture tetrazolium assay, *Cancer Res.* **48** (1988), 589–601.

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

[8] M.R. Grever, S.A. Schepartz and B.A. Chabner, The National Cancer Institute: Cancer drug discovery and development program, *Seminars in Oncol.* **19** (1992), 622–638.

[9] S.F. Stinson, M.C. Alley, W.C. Kopp, H.H. Fiebig, L.A. Mullendore, A.F. Pittman, S. Kenney, J. Keller and M.R. Boyd, Morphological and immunocytochemical characteristics of human tumor cell lines for use in a disease-oriented anticancer drug screen, *Anticancer Res.* **12** (1992), 1035–1053.

[10] K.D. Paull, R.H. Shoemaker, L. Hodes, A. Monks, D.A. Scudiero, L. Rubinstein, J. Plowman and M.R. Boyd, Display and analysis of patterns of differential activity of drugs against human tumor cell lines: Development of mean graph and COMPARE algorithm, *J. Natl. Cancer Inst.* **81** (1989), 1088–1092.

[11] K.D. Paull, E. Hamel and L. Malspeis, Prediction of biochemical mechanism of action from the in vitro antitumor screen of the National Cancer Institute, in: *Cancer Chemotherapeutic Agents*, W.E. Foye, ed., American Chemical Soc. Books, 1993, pp. 1574–1581.

[12] J.N. Weinstein, K.W. Kohn, M.R. Grever, V.N. Viswanadhan, L.V. Rubinstein, A.P. Monks, D.A. Scudiero, L. Welch, A.D. Koutsoukos, A.J. Chiausa and K.D. Paull, Neural computing in cancer drug development: predicting mechanism of action, *Science* **258** (1992), 447–451.

[13] J.N. Weinstein, T.G. Myers, J.K. Buolamwini, K. Raghavan, W. van Osdol, J. Licht, V.N. Viswanadhan, K.W. Kohn, L.V. Rubinstein, A.D. Koutsoukos, A.P. Monks, D.A. Scudiero, N.L. Anderson, D. Zaharevitz, B.A. Chabner, M.R. Grever and K.D. Paull, Predictive statistics and artificial intelligence in the US National Cancer Institute's drug discovery program for cancer and AIDS, *Stem Cells* **12** (1994), 13–22.

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

[15] R.J. Sapolsky and R.J. Lipshutz, Mapping genomic library clones using oligonucleotide arrays, *Genomics* **33** (1996), 445–456.

[16] E.S. Cleveland et al., *Biochem. Pharmacol.* **49** (1995), 947.

[17] K.D. Paull, C.M. Lin, L. Malspeis and E. Hamel, *Cancer Res.* **52** (1992), 3892.

[18] R. Bai, K.D. Paull, C.L. Herald, L. Malspeis, G.R. Pettit and E. Hamel, Halichondrin B and homohalichondrin B, marine natural products binding in the vinca domain of tubulin. Discovery of tubulin-based mechanism of action by analysis of differential cytotoxicity data, *J. Biol. Chem.* **266** (1991), 15882–15889.

[19] B.A. Chabner, J.N. Weinstein, K.D. Paull and M.R. Grever, Cell line-based screening for new anticancer drugs, in: *Cancer Treatment*, P. Banzet, F. Holland, D. Khayat and M. Weil, eds, an Update, Springer-Verlag, Paris, 1994, pp. 10–16.

[20] A.D. Koutsoukos, L.V. Rubinstein, D. Faraggi, S. Kalyandrug, J.N. Weinstein, K.D. Paull, K.W. Kohn and R.M. Simon, Discrimination techniques applied to the NCI in vitro antitumor drug screen: Predicting biochemical mechanism of action, *Statistics in Medicine* **13** (1994), 719–730.

[21] L.M. Shi, T.G. Myers, Y. Fan and J.N. Weinstein, Application of Genetic Function Approximation to the QSAR Study of Anticancer Ellipticine Analogs. Fifth Conference on Current Trends in Computational Chemistry, 1996, pp. 1–5.

[22] L.M. Shi, T.G. Myers, Y. Fan, P.M. O'Connor, K.D. Paull, S.H. Friend and J.N. Weinstein, Mining the National Cancer Institute's Anticancer Drug Screen Database: cluster analysis of ellipticine analogs with p53-inverse and central nervous system-selective patterns of activity, *Mol. Pharmacol.* **53** (1998), 241–251.

[23] L.M. Shi, Y. Fan, T.G. Myers, K.D. Paull and J.N. Weinstein, Mining the NCI anticancer drug discovery databases: Genetic function approximation for the quantitative structure-activity relationship study of anticancer ellipticine analogs, *J. Chem. Inf. Comput. Sci.* **38** (1998), 189–199.

[24] W.W. van Osdol, T.G. Myers, K.D. Paull, K.W. Kohn and J.N. Weinstein, Use of the Kohonen self-organizing map to study the mechanisms of action of chemotherapeutic agents, *J. Natl. Cancer Inst.* **86** (1994), 1853–1859.

[25] W.W. van Osdol, T.G. Myers and J.N. Weinstein, Neural network techniques for the informatics of cancer drug discovery, *Methods in Enzymology* **321** (2000), 369–395.

[26] S.E. Bates, A.T. Fojo, J.N. Weinstein, T.G. Myers, M. Alvarez, K.D. Paull and B.A. Chabner, Molecular targets in the National Cancer Institute drug screen, *J. Cancer Res. Clin. Oncol.* **121** (1995), 495–500.

[27] G.W.A. Milne, M.C. Nicklaus, J.S. Driscoll, S. Wang and D. Zaharevitz, National Cancer Institute drug information system 3D database, *J. Chem. Inf. Comput. Sci.* **34** (1994), 1219–1224.

[28] T.G. Myers, M. Waltham, G. Li, J.K. Buolamwini, D.A. Scudiero, L.V. Rubinstein, K.D. Paull, E.A. Sausville, N.L. Anderson and J.N. Weinstein, A protein expression database for the

molecular pharmacology of cancer, *Electrophoresis* **18** (1997), 647–653.

[29] N.L. Anderson, J.-P. Hofmann, A. Gennell and J. Taylor, Global approaches to quantitative analysis of gene-expression patterns observed by use of two-dimensional gel electrophoresis, *Clin. Chem.* **30** (1984), 2021–2036.

[30] N.G. Anderson and N.L. Anderson, Twenty years of two-dimensional electrophoresis: Past, present and future, *Electrophoresis* **17** (1996), 443–453.

[31] G. Li, M. Waltham, E. Unsworth, A. Treston, J. Mulshine, N.L. Anderson, K.W. Kohn and J.N. Weinstein, Rapid protein identification from two-dimensional polyacrylamide gels by MALDI mass spectrometry, *Electrophoresis* **18** (1997), 647–653.

[32] D.T. Ross, U. Scherf, M.B. Eisen, C.M. Perou, P. Spellman, V. Iyer, C. Rees, S.S. Jeffrey, M. Van de Rijn, M. Waltham, A. Pergamenschikov, J.C.F. Lee, D. Lashkari, D. Shalon, T.G. Myers, J.N. Weinstein, D. Botstein and P.O. Brown, Systematic variation in gene expression patterns in human cancer cell lines, *Nature Genetics* **24** (2000), 227–235.

[33] U. Scherf, D.T. Ross, M. Waltham, L.H. Smith, J.K. Lee, K.W. Kohn, W.C. Reinhold, T.G. Myers, D.T. Andrews, D.A. Scudiero, M.B. Eisen, E.A. Sausville, Y. Pommier, D. Botstein, P.O. Brown and J.N. Weinstein, A gene expression database for the molecular pharmacology of cancer, *Nature Genetics* **24** (2000), 236–244.

[34] J.E. Staunton, D.K. Slonim, H.A. Coller, P. Tamayo, M.J. Angelo, J. Park, U. Scherf, J.K. Lee, J.N. Weinstein, J.P. Mesirov, E.S. Lander and T.R. Golub, Chemosensitivity prediction by transcriptional profiling, *Proc. Natl. Acad. Sci. USA*, in press.

[35] M.B. Eisen, P.T. Spellman, P.O. Brown and D. Botstein, Cluster analysis and display of genome-wide expression patterns, *PNAS* **95** (1998), 14863–14868.

[36] P.S. Shenkin and Q. McDonald, Cluster analysis of molecular conformations, *J. Comput. Chem.* **15** (1994), 899–916.

[37] B. Efron and G. Gong, A leisurely look at the bootstrap, the jackknife, and cross-validation, *Amer. Statistician* **37** (1983), 36–38.

[38] R.L. Strausberg, C.A. Dahl and R.D. Klausner, New opportunities for uncovering the molecular basis of cancer, *Nature Genetics* **15** (1997), 415–416.

[39] R.L. Strausberg, The Cancer Genome Anatomy Project: Building a new information and technology platform for cancer research, in: *Early Detection of Cancer*, E.A. Srivastava, ed., Elsevier, 1998.

[40] L. Tanabe, L.H. Smith, J.K. Lee, U. Scherf, L. Hunter and J.N. Weinstein, MedMiner: An internet tool for mining information, with application to gene expression profiling, *BioTechniques* **27** (1999), 1210–1217.

[41] T.C. Rindflesch, L. Tanabe, J.N. Weinstein and L. Hunter, EDGAR: Drugs, genes and relations from the biomedical literature, *Pac. Symp. Biocomput.* (2000), 571–528.

[42] Y. Zhou, F.G. Gwadry, W.C. Reinhold, L. Miller, L.H. Smith, U. Scherf, E. Liu, K.W. Kohn, Y. Pommier and J.N. Weinstein, Transcriptional regulation of mitotic genes by camptothecin-induced DNA damage: Microarray analysis of dose- and time-dependent effects, *Proc. Natl. Acad. Sci. USA* submitted.