

Visual Analytics in the Pharmaceutical Industry

Jeffrey D. Saffer,
Vicki L. Burnett,
and Guang Chen

OmniViz

Peter van der
Spek

*Erasmus Medical
Center*

With recent advances in biology, medicine, chemistry, and informatics, the pharmaceutical industry has the opportunity—and the challenge—to use unprecedented volumes of data across the entire spectrum of business activities. The Human Genome Project and high-throughput sequencing have provided complete sequences (three billion base pairs for humans) for several organisms, and new sequences are being published continuously. DNA chip (or microarray) technologies let scientists study hundreds of thousands of gene expression data points at once. Advances in proteomics (study of the proteins in a cell) are now creating a similar flood of data on the biology of proteins. Metabolomics is providing large-scale perspectives on thousands of cellular metabolites under hundreds of conditions. Combinatorial chemistry and ultra-high-throughput screening are creating a similar data landslide of potential lead compounds, in some cases with companies working with tens of millions of compounds. In addition, researchers are combining these approaches with traditional methods in other areas, such as toxicology and clinical studies, further contributing to the data glut. In addition, the pharmaceutical industry is already faced with lots of data related to marketing, customer information, and other postproduct launch information.

It seems that having so much data in every aspect of the pharmaceutical business should provide a benefit; surprisingly, this is not necessarily the case. Much of the decision process in this realm, as in others, is reductionist in nature.

Consider having information on 1,000 genes and wanting to find the one or few that would serve as a drug's point of action. You can eliminate the genes that aren't relevant and continue to refine the search according to various criteria, thus providing better insight into the mechanism of a drug's action. Now consider 100,000 genes. The efforts required to decide which of those genes can be eliminated become much greater. Similar problems exist in deciding which compounds should become leads, or which drug candidates should go through clinical trials. As a result, the number of new drugs (molecular entities that have novel chemical structures) has fallen in each of the last eight years, despite more than a twofold increase in pharmaceutical R&D spending over that period (see <http://www.fda.gov/oc/initiatives/>

criticalpath/whitepaper.pdf). But, intuitively, we know that more data should lead to more discoveries.

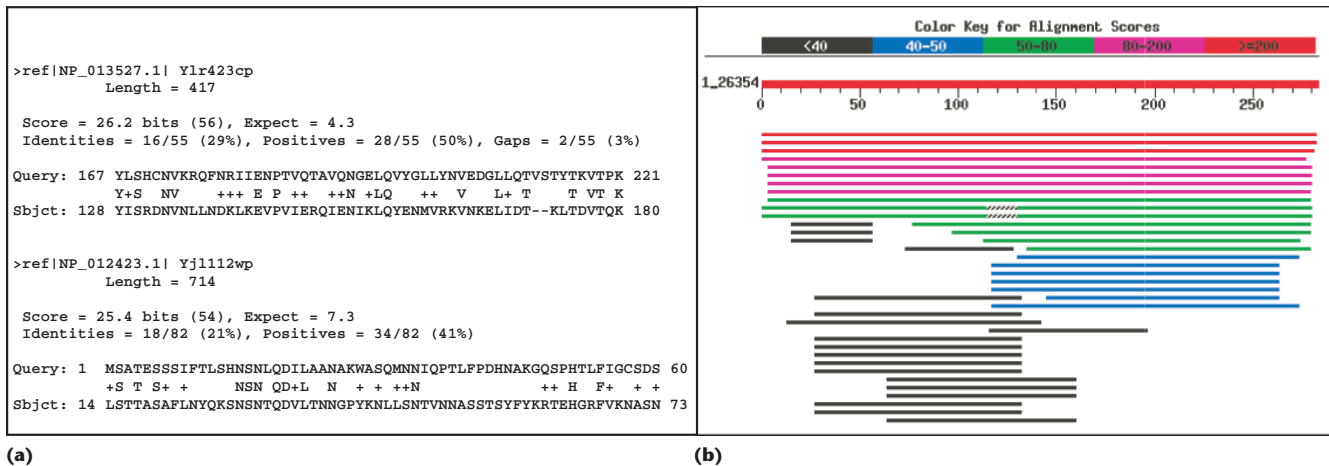
The goal, then, is to find ways to use these large volumes of data effectively—saving time, minimizing risk, and saving opportunity costs. We want to be able to make decisions faster to eliminate unfruitful targets, leads, or trials, while increasing the number of promising discoveries. The financial implications of this effective use of data are huge. For a drug that makes \$365 million per year (certainly not a blockbuster in this industry), each day sooner the drug makes it to market earns \$1 million more, and this isn't counting the savings on not having performed an experiment or trial that wasn't going to be fruitful.

Visualization is a key component to achieving the goal. Visual representation is the most effective way to present and understand large volumes of data.

But the type of visualization used is critical. Two key problems must be addressed. First, with extreme volumes of information, traditional charts and graphs—even when fully interactive—are not sufficient to provide a clear understanding. Rather, we need more effective visualizations to communicate the value within the data. Second, progression along the pharmaceutical pipeline requires an understanding of multiple types of data (numeric, categorical, chemical structures, genome sequences, and text documents) crossing several domains. Part of the reason more data has not resulted in more drugs is that connections among domains (for example, between biology and chemistry) have largely been inefficient. Visualizations that allow integration at the analytical level will have particular value to the pharmaceutical arena.

Target discovery

Target discovery entails defining what cellular component (usually a protein) or process should be interrupted or augmented by a drug. One of the key steps in this process is to identify the function of genes and their respective proteins. A common approach is to compare the gene (nucleic acid) or protein (amino acid) sequences of one set of objects (possibly those of a known function) to another set (perhaps unknown). One tool for this is the Basic Local Alignment Search Tool (Blast),¹ or more recent variants such as Blast2.² Originally, the output from such



1 (a) The output from Blast analysis showing what sequences are similar and how they align. (b) A visual presentation covering more sequence comparisons summarizes more information in an easy-to-understand interface. (We produced this graphic with the pBlast implementation at the National Center for Biotechnology Information Web site at <http://www.ncbi.nlm.nih.gov/blast>.)

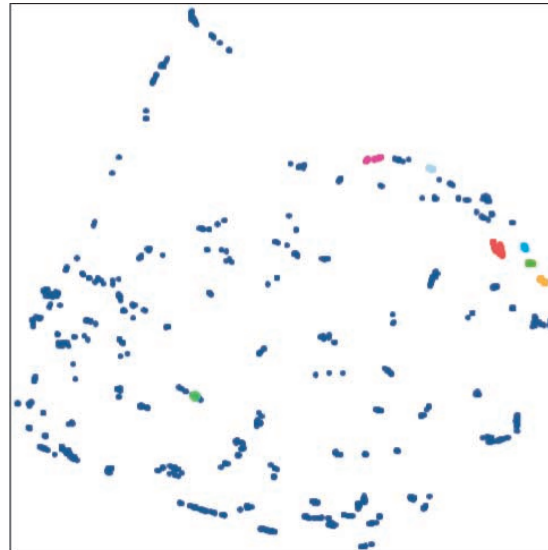
tools was a text output showing the alignment of sequences (see Figure 1a). Contrast this to the more advanced representation (see Figure 1b), in which it's immediately clear where, and to what degree, the alignments exist.

Although the visualization in Figure 1b is nicely linked to the details in Figure 1a, this approach is insufficient for navigating large volumes of data. The presentation is adequate for comparing one sequence against others, but for industrial-scale sequencing, so many of these comparisons are necessary that it's overwhelming.

For example, one class of proteins of particular interest to the pharmaceutical industry is G protein-coupled receptors (GPCRs). This class contains thousands of such proteins and many putative GPCRs as well. To compare all of these sequences against each other, we perform the Blast analysis for each protein against all others. Visualizing those results effectively requires a different approach.

The OmniViz Galaxy visualization (see Figure 2) shows how each protein relates to every other protein in a proximity map. In Figure 2, we used the calculated similarities between all pairs of proteins to create a visualization that highlights all relationships simultaneously. Since the suggested homologies can reflect evolutionary origin and duplication of gene function, this global perspective makes it immediately clear that alpha1A, alpha1B, alpha2B, beta1, and beta2 adrenergic receptors are more similar to D1 dopamine receptors than they are to alpha1D and beta3 adrenergic receptors. Such observations clarify relationships among possible targets and identify potential unwanted side effects.

This type of global visualization is also excellent for mapping new sequences within the framework of all GPCRs. For example, the bright green protein toward the middle of the OmniViz Galaxy visualization in Figure 2 is a sequence thought to be a GPCR, but one of unknown function. Since this falls within a group of known chemokine receptors, it's likely that this new sequence shares that function.



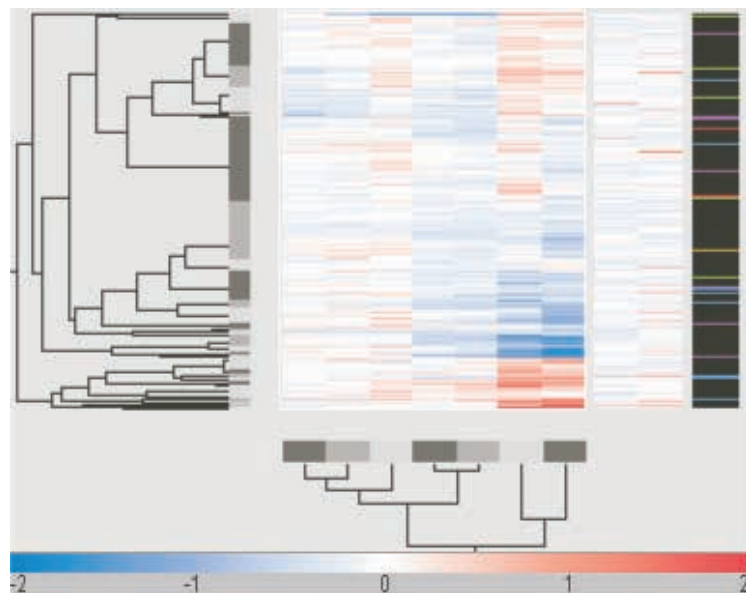
2 Blast all-against-all analysis of 917 full-length GPCRs from multiple species presented in a proximity map. This OmniViz Galaxy representation shows each GPCR as a dark blue dot. Some GPCR classes are color coded (beta3 adrenergic receptors, purple; alpha1D adrenergic receptors, gray-blue; alpha2B adrenergic receptors, red; alpha1A and 1B adrenergic receptors, turquoise; beta1 and beta2 adrenergic receptors, orange; D1 dopamine receptors, dark green). In addition, a probable GPCR of unknown function is highlighted bright green.

Gene expression

The development of microarray techniques has made it possible to study the level of expression of essentially all genes within a cell at once. This powerful tool enables the identification of differentially expressed genes in various disease tissues or under a range of conditions, such as drug treatment over time. Data visualization approaches have gone hand in hand with microarrays since the method was first used for whole organism screening.³

The heat map visualization (see Figure 3, next page)

3 Heat map visualization of gene expression data. We use the dendrograms from the OmniViz TreeScape visualization to show the hierarchical relationships of rows and columns. We've added extra columns of other data to the right to assess possible relationships between the clusters and other attributes.



provides an overview of all the data values and their distribution in the data set. Each gene is a row in the map and each condition tested is a column. This is essentially a spreadsheet of the values, but instead of using a text representation of individual values, a color scale is presented. To provide visible structure to the map, the genes (and, if desired, the columns) are clustered to put similar patterns near each other. In some cases, as shown in Figure 3, additional information can be provided simultaneously such as the hierarchical nature of the relationships.

Many heat map implementations provide interactions that support efficient exploration. Brushing over a region to see gene names, conditions, and values; flipping dendrogram nodes to stress particular arrangements of interest; or adding additional columns of data within the same visualization support rapid understanding and further exploration of the data.

A drawback to the heat map representation is that there is a 1D order forced on the genes. Thus, the complex relationships that truly exist are obscured. Coupling the heat map presentation with other methods, such as the Galaxy, provide a solution to that issue.

Lead discovery

In lead discovery, it's essential to identify, from a large

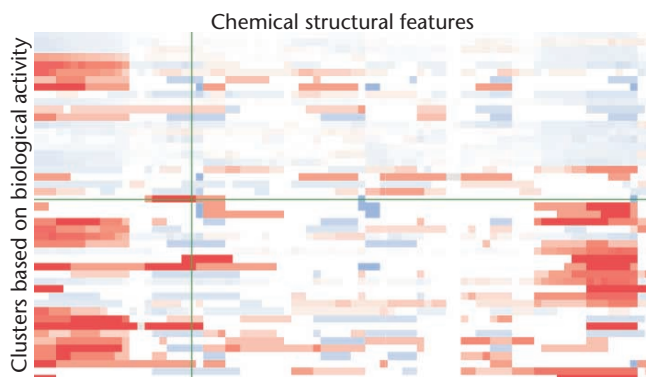
number of compounds, common chemical structures responsible for desired bioactivities. One approach involves clustering (or grouping) the compounds according to their spectrum of activities in biological assays of activity. Traditionally, a chemist will view all of the structures and, by eye, pick out the features of those compounds that might be responsible for the activity. Naturally, this process can work only for small numbers of compounds. To propel this type of analysis further, researchers now usually characterize compounds by a digital representation of their structure (structure fingerprints). Fast computational methods⁴ have been developed to quickly generate fingerprints from hundreds of thousands or more compounds. With

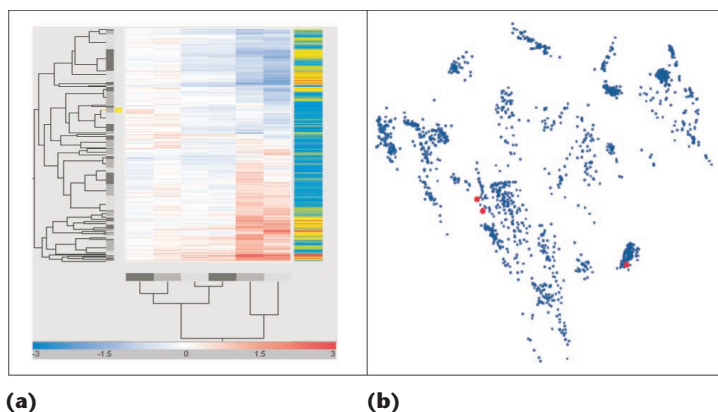
fingerprints or descriptors that can be directly related to structural attributes, chemists can use statistical tests to determine what features are likely responsible for what activities. However, with thousands of fingerprint components, it's impossible to test every possibility. Visualization can help direct the researcher to what is most important on the required global scale.

One useful type of visualization that can guide chemists is a matrix showing the groups of compounds (based on activities) as rows, and the structural fingerprints as columns. This approach color codes the cell at the intersection of a row and column according to some measure of importance. For example, the coloring can show whether that particular fingerprint occurred in compounds with a particular activity spectrum more often than would have been expected by chance (based on a random distribution of all structural features). The chemist can then order rows and columns to provide the greatest structure to the visualization. In this way, the chemist's eye is drawn to relationships across all the features and is guided toward the most useful structures to pursue.

In Figure 4, we clustered a collection of compounds based on their spectra of biological activity across a number of high-throughput screening assays. For each com-

4 OmniViz CoMet visualization of chemical structural features and resulting biological activities. The rows represent clusters of compounds based on high-throughput screening results and the columns represent structural features of the compounds. Red colors indicate a positive association of the feature with the compounds in the screening cluster; blue indicates a negative association.





5 (a) OmniViz heat map visualization of proteomics data linked to (b) an OmniViz Galaxy visualization of compounds compared by chemical structure. Through the linkage between the visualizations, finding proteins of interest (yellow highlights in 5a, the relevant ligands present in the compound library were also selected (red points in 5b).

pound, structural features were previously defined. We then created a matrix with clusters of compounds as rows and structural features as columns. Each cell is color coded depending on how often a structural feature occurs in a cluster (deviation from what would be expected by chance). A greater-than-expected occurrence (implying the structural feature might cause the activity) is shown as red, and lower occurrences are blue. In this particular implementation (OmniViz CoMet), information about each row, column, or cell is immediately available through interactions with the visualization, and clicking on a cell displays all of that cell's compounds (that is, compounds that have the particular activity spectrum and include the structural feature).

Intersection of biology and chemistry

Traditional workflows in the pharmaceutical industry have the biologists discovering targets and then passing these along to the chemists for use in screening assays, with little interaction at the interface between these domains. In part, this junction has been difficult to take advantage of because it requires an understanding that spans both domains. Information visualization is one emerging means to gain the natural synergies between these scientific areas (see the “Intersection of Experimental Data with Historical Information” sidebar on page 15).

Linkages among visualizations are one means of connecting the biological and chemical domains. For example, from one set of experimental proteomics data, we found a cluster of related proteins to be of interest (see Figure 5a). In this data set, we visualized protein abundance measurements for six different assays (color matrix in the left-hand portion of Figure 5a) along with a column of functional data. Some of these proteins have known ligands (small compounds that bind to the proteins). Those ligands were part of a chemical compound library, which we visualized (see Figure 5b) based on structural fingerprints generated using methods from Barnard Chemical Information (BCI), Sheffield, UK. We linked the two visualizations together, so that when a protein is selected in the proteomics data the corresponding ligands are highlighted in the compound structure visualization (or vice versa). When the proteins of interest were selected (yellow highlight next to the dendrogram in Figure 5a), it's immediately obvious that the ligands known for those proteins (highlighted

red in the Galaxy visualization in Figure 5b) fall into two distinct regions, indicating different structural features. This suggests that additional compounds similar to those in both of these regions would be worthwhile to screen for binding to (and possible inactivation of) the proteins.

Toxicology and preclinical studies

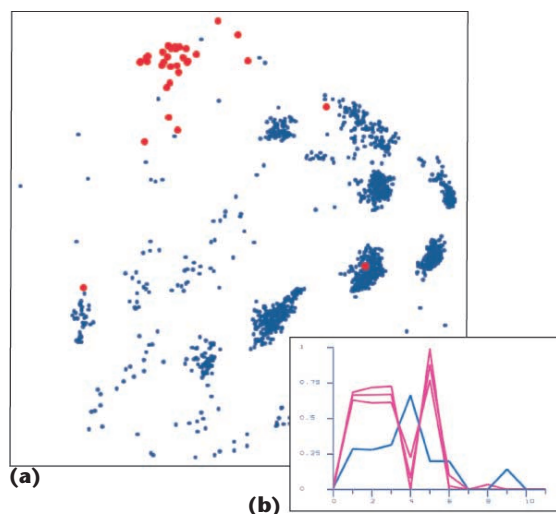
Visualizing data from toxicology and preclinical studies generally present particular challenges since the data are often mixtures of numeric and categorical data. Thus, to create visualizations that use all available data to compare subjects to each other, the underlying methods, such as clustering, must support multiple data types. Enabled by such methods, visualizations of toxicology and preclinical data can provide an easy way for understanding the large volumes of data associated with these areas.

One problem with preclinical data is the often subjective nature of the diagnoses made. When buried in a sea of other information, those missed diagnoses are difficult to spot. In our experience, such errors are made with about 10 percent of the subjects, leading to difficulties in understanding the underlying biology. In one study, a group of 1,561 animals were studied in an investigation covering eight different treatments, including controls. Thirty-five of the animals were diagnosed as having hematological dyscrasia (an imbalance of blood components), shown as red points in the Galaxy visualization in Figure 6 (next page). While most of these animals are in the region around 11 o'clock in the visualization—representing similarity in their clinical parameters—three are in distinct areas. Examination of the available data clearly shows that these three animals are quite different and almost certainly do not have dyscrasia (for example, see Figure 6b comparing three animals with dyscrasia to one of these outliers).

Clinical

Novel means of visualizing clinical data can provide important insights into diagnoses and potential treatment regimens. Researchers applied a recent example of this to acute myeloid leukemia (AML).⁵

AML is not a single disease but a class of neoplasms with diverse generic abnormalities and variable responses to treatments. To elect effective treatments and predict appropriate prognosis for individual

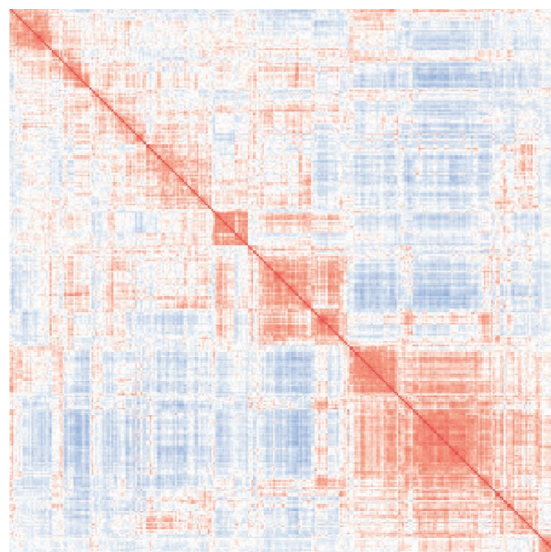


6 (a) OmniViz Galaxy visualization using all available clinical data from a preclinical study of 1,561 animals. Each blue point represents a different animal and those with similar clinical parameters are near each other. Thirty-five animals diagnosed as having hematological dyscrasia are color coded red. (b) Parallel coordinate plot of 12 clinical parameters from three animals representative of those with dyscrasia (magenta) and one of the outliers diagnosed as having dyscrasia (blue).

patients, it's clinically important to identify such genetic subgroups. Valk et al. developed a molecular classification method based on DNA expression profiles from blood and bone marrow samples of AML patients.⁵ For each of 285 patients diagnosed with primary AML, a DNA microarray analysis was performed on the blood or bone marrow sample using the Affymetrix U133A gene chip. Of the approximately 13,000 genes whose expression was measured, 2,856 were differentially expressed and used for identifying the mechanistically distinct subgroups with diagnostic relevance.

Although clustering is a widely used technique in gene expression analyses for identifying unknown groups, that method always finds the predetermined number of clusters even if that number is not appropriate or even if no cluster structure exists in the data. Therefore, Valk et al. deployed a novel approach based on a visualization of correlations provided in OmniViz software (see Figure 7).

Figure 7 displays all the pair-wise correlations between the patients (based on their gene expression profile) with color indicating strength and directions of the correlations. An applied ordering rearranges the patients so that correlated individuals are close to each other. As a result, a structure emerges if it exists in data. By clearly displaying the strength of associations within and between the blocks, this visualization shows whether a viable cluster structure actually resides in the data and what diagnostically relevant genetic heterogeneity exists amongst these patients. Valk et al. used this approach to find patients sharing similar mechanistic defects with the finding of significantly expressed biomarkers allowing detection of these clinically rele-



7 OmniViz Correlation visualization displaying all pair-wise correlations among acute myeloid leukemia patients. Using 2,856 differentially expressed genes determined at the Erasmus Medical Center, the pair-wise correlation between each pair of patients was calculated and displayed with patients (both rows and columns) ordered to provide the maximal structure and place the most similar patients next to each other. A positive correlation is red and a negative correlation is blue. Since each patient has a correlation of 1 (highest score) to themselves, the diagonal is bright red.

vant distinct subgroups. The correlation visualization confirmed all monocytic-type leukemias grouped together based on characteristic gene expression changes. The two major drivers of these correlation plots were observed to be phenotype (or morphology) and cytogenetic abnormalities.

Conclusion

With the flood of data across all aspects of the pharmaceutical industry, information visualization is emerging as a critical component of discovery, development, and business. But it's a new class of visualizations that is having the greatest impact. Higher-level summaries that can provide a framework for understanding the immense volumes of data and that reveal unexpected relationships have come to the forefront. And, the ability to use these visualizations to cross domains and data types provides the ability to integrate analyses and support fast, effective decisions. ■

References

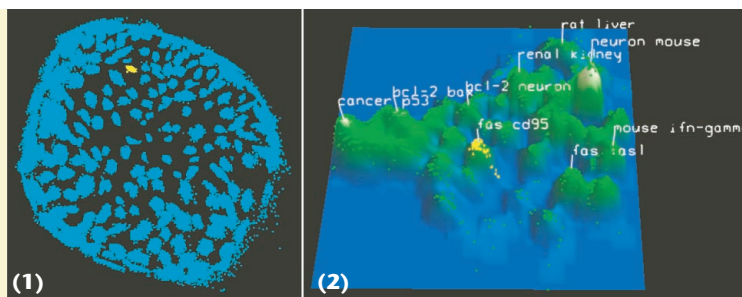
1. S.F. Altschul et al., "Basic Local Alignment Search Tool," *J. Molecular Biology*, vol. 215, no. 3, 1990, pp. 403-410.
2. T.A. Tatusova and T.L. Madden, "BLAST2 Sequences—A New Tool for Comparing Protein and Nucleotide Sequences," *FEMS Microbiology Letter*, vol. 174, no. 2, 1999, pp. 247-250.
3. J.L. DeRisi, R.I. Vishwanath, and P.O. Brown, "Exploring

Intersection of Experimental Data with Historical Information

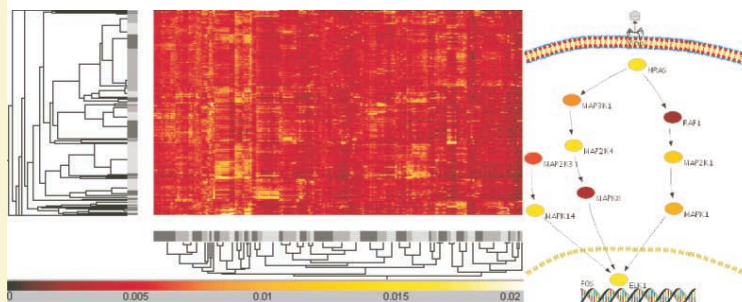
One of the results of the improved pharmaceutical data analysis provided by visualizations is a plethora of targets and leads that researchers need to consider for follow-up. This situation emphasizes that the goal of analysis is not just, what does the data tell me? but, what do I do next? Thus, researchers must simultaneously consider other sources of information, particularly the historical perspective. Two sources of historical perspective critical to pharmaceutical researchers are the scientific literature (publications and patents) and biological pathways. Visualizations can facilitate use of both of these sources of information.

Literature analysis

Visualization of the pharmaceutical literature is perhaps the only means for understanding the vast number of publications relevant to any particular problem area. Such approaches can help us understand the background information—what's known, where the gaps are, and who is working in what area. But perhaps more importantly, visualization of the literature can provide a framework for interpreting data. The investigator needs to know, What is the basis for my observation? Is my observation novel? Are there likely problems? (that is, Should I kill this lead?), and more. Figure A shows an example of connecting chemistry visualization to literature. We created an OmniViz Galaxy visualization of a set of compounds ($\sim 3.3 \times 10^5$) based on their chemical structure (Figure A1). We identified a cluster of those compounds (yellow highlight) deemed worthy of further investigation. We visualized a set of 1,100 patents potentially relevant to the area of interest using an OmniViz ThemeMap landscape representation of the major themes or concepts within the patents (see Figure A2). Linking the two visualizations so that selecting compounds in the Galaxy caused the relevant patents to be highlighted (yellow points in Figure A2), it became obvious that a previous patent position on this class of compounds existed and that efforts were likely better spent in other areas.



A (1) OmniViz Galaxy visualization of a set of $\sim 3.3 \times 10^5$ compounds linked to (2) an OmniViz ThemeMap landscape representation.



B Interactive pathway visualization (using OmniViz Pathway Enterprise) showing a cell signaling pathway (right). The objects (proteins) represented by ovals are color coded according to the experimental values represented in one column of the displayed heat map (left).

Pathway analysis

Biological pathways representing metabolic steps or the processes of cell signaling are another historical perspective that can provide a framework for interpreting experimental data. Integrating visualization of the research data with visualization of the relevant pathways can, for example, identify new members of pathways, determine whether an observed coregulation is novel, or verify known pathways. One visual approach for connecting experimental data to pathways is to color code pathway objects according to the experimental data (see Figure B). By playing through a number of experimental conditions, such as treatment with agonists or antagonists over time, this visual approach can rapidly identify the biologically significant and relevant genes.

the Metabolic and Genetic Control of Gene Expression on a Genomic Scale," *Science*, vol. 278, no. 5338, 1997, pp. 680-686.

4. J.M. Barnard and G.M. Downs, "Chemical Fragment Generation and Clustering Software," *J. Chemical Information and Computer Sciences*, vol. 37, no. 1, 1997, pp. 141-142.
5. P.J.M. Valk et al., "Prognostically Useful Gene-Expression

Profiles in Acute Myeloid Leukemia," *New England J. Med.*, vol. 350, no. 16, 2004, pp. 1617-1628.

Readers may contact Jeffrey D. Saffer at saffer@omniviz.com.

Contact editor Mike Potel at potel@wildcrest.com.