VISUALIZATION AND INTEGRATED DATA MINING OF DISPARATE INFORMATION

JEFFREY D. SAFFER,* CORY L. ALBRIGHT, AUGUSTIN J. CALAPRISTI, GUANG CHEN, VERNON L. CROW, SCOTT D. DECKER, KEVIN M. GROCH, SUSAN L. HAVRE, JOEL M. MALARD, TONYA J. MARTIN, NANCY E. MILLER, PHILIP J. MONROE, LUCY T. NOWELL, DEBORAH A. PAYNE, JORGE F. REYES SPINDOLA, RANDALL E. SCARBERRY, HEIDI J. SOFIA, LISA C. STILLWELL, GREGORY S. THOMAS, SARAH J. THURSTON, LEIGH K. WILLIAMS, AND SEAN J. ZABRISKIE

OmniViz, Inc., 3350 Q Avenue, Richland, WA 99352, USA.
E-mail: saffer@omniviz.com

Received: 7th July 2000 / Published 11th May 2001

ABSTRACT

The volumes and diversity of information in the discovery, development, and business processes within the chemical and life sciences industries require new approaches for analysis. Traditional list- or spreadsheet-based methods are easily overwhelmed by large amounts of data. Furthermore, generating strong hypotheses and, just as importantly, ruling out weak ones, requires integration across different experimental and informational sources. We have developed a framework for this integration, including common conceptual data models for multiple data types and linked visualizations that provide an overview of the entire data set, a measure of how each data record is related to every other record, and an assessment of the associations within the data set.

INTRODUCTION

Modern methods in the chemical and life sciences are providing data at an unprecedented pace. This is occurring in many areas with multiple types of information. For example, combinatorial chemistry and ultra-high-throughput screening methods are providing incredible numbers of, and information about, chemical compounds. Related screening methods, such as gene chip assays, and the associated expanding world of genome science is also providing information at a very high rate. And data annotations, scientific literature, patents, and a wide range of other documents have text information that is difficult to assimilate due to the sheer volume and complexity.

Given this flood of diverse information, effective and timely use of the results is no longer possible using traditional approaches. With large volumes of information, it is difficult to learn from long lists, tables, or even simple graphs, particularly with multidimensional data. Furthermore, it is clear that more valuable hypotheses can be derived by simultaneous consideration of multiple types of experimental data (e.g. chemical structural information in addition to activity data), a process that is problematic with large amounts of data.

As one solution for moving from large volumes of information to knowledge, we have developed an integrated data visualization and mining framework (OmniViz Pro¶). The primary premise upon which this framework was built is that discovery of the unexpected is a key goal of data mining. That is, in addition to searching for data records of well-defined behavior (testing specific hypotheses),

Published in "Chemical Data Analysis in the Large: The Challenge of the Automation Age", Martin G. Hicks (Ed.), Proceedings of the Beilstein-Institut Workshop, May 22nd – 26th, 2000, Bozen, Italy

http://www.beilstein-institut.de/bozen2000/proceedings/saffer/saffer.pdf

considerable value can often be obtained from assessing all the relationships within the full data set. To this end, there are several full data set overview visualizations that provide value to the analyst. The rationale behind these and the operational issues that have to be dealt with in their implementation are presented here.

CONCEPTUAL DATA MODELS

In working toward an integrated framework for data visualization and mining, we recognized that a common conceptual data model was essential. This conceptual model provides a familiar framework for the analyst and a common view that is independent of data type.

Functionally, this conceptual model can be considered similar to a spreadsheet where each record is a row in the data table and each column contains data describing a distinct attribute. This collection of attributes, or any subset, can be used directly in multivariate analyses. The goal is to use these attributes to define for each record a high-dimensional vector representation that can be used for cluster analysis as well as a common structure for visualization and interaction. Although the mental picture for this paradigm is two-dimensional, functionally the resulting vector space model can be multi-dimensional, providing a framework for integrating different analyses of the same data records.

Multiple data types can be used as attributes in this conceptual model, as with a spreadsheet, providing great flexibility. Numeric data (e.g., screening assay results), categorical data (e.g., functional classification or structure descriptors), genomic sequence (protein or nucleic acid), or even free text can be used. Some of this data can be used directly in high-dimensional vector representations. Other types of data may require the definition of specific descriptors or features, leading to the generation of a new collection of attributes. That is, a column of the data table is translated to a new set of one or more columns. As a result, each data record can ultimately be considered as a vector, whose dimensions are the attribute columns chosen for comparison. Some examples of how this might be accomplished are shown in Figure 1.

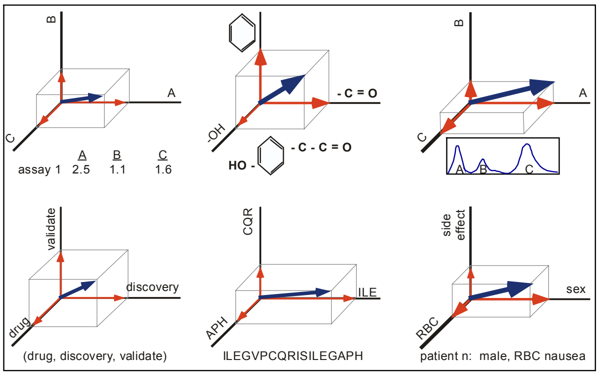


Figure 1: Examples of high-dimensional vector representations for several data types – numeric, chemical structure, chromatographic, text, genomic sequence, and mixed mode (numeric and categorical).

The methods for defining attributes or features for many data types are well known and will not be presented here. However, because of the relatively recent application of these approaches to genomic sequences, it is worth mentioning that a variety of sequence descriptors have been used that in many ways parallel the approaches used for chemical descriptors. For example, van Heel [1] has used a sequence-based method in which each protein sequence is represented the collection of amino acid dimers present in the sequence, somewhat analogous to using contiguous atom pairs for small molecule comparison. More diverse sequence properties have been employed by Hobohm and Sander; [2] in this case, protein sequences were translated to 144 attributes that included sequence components (amino acid composition and a subset of dimers) and several physical-chemical properties. More recently, as for chemical compounds, structural descriptors have been derived for comparing proteins. [3,4]

DATA VISUALIZATION – BASIC CONCEPTS

Exploratory data analysis requires a framework in which

1. the data can be organized along the lines of interest to the analyst and
2. a collection of tools is available for pursuing specific inquiries.

For both, the methods need to handle large volumes of data, with reasonable speed, and provide linkage among complementary views and to other tools.

Presenting data in an organized fashion requires appropriate data overviews, especially those that allow inference by comparison. For this, we have adopted visualization methods since they offer unequalled facility in presenting large volumes of data. In addition, the structure within a well-designed visualization can suggest relationships that might otherwise be overlooked. In that regard, it should be clear that data visualization methods assist, but cannot replace the analyst.

A key component of this approach is to use all the relevant attributes simultaneously for deriving the comparisons. With very large data sets, such as high-throughput screening, it is not possible for the analyst to examine the behavior of the data records a few columns at a time and be able to assess the overall behavior. The selection of attributes for comparison can be useful for testing specific hypotheses, but do not facilitate discovery of the unexpected. Hence, cluster-based methods that utilize all the appropriate data attributes simultaneously are preferred.

Even with mathematical methods that use all the data, no single visualization method can convey all of the information likely to be needed by the analyst and several complementary approaches are necessary. In that spirit, these should not be viewed as stand-alone entities, but linked together for continuity in data analysis. This becomes particularly important in an integrated analysis across different experimental data sets, for example, where distinct visualizations are used to organize the data from separate experimental regimens. The data overviews also need to be supported by complementary tools that support access to and, in many cases, visualization of the details of the data. The easy access to these tools is the foundation for progressing from visualization to data mining.

Given that the data exploration is necessary in the first place since the volume of data is too large to assimilate at once, the key features of the visualization methods are speed and progressive disclosure. Speed is essential since iterative analyses are necessary. Progressive disclosure is a specific type of iteration that is needed frequently. This goes beyond simply zooming in, but rather needs to allow a finer resolution based on comparison of a subset of data records. For example, the relationships uncovered from a subset may be driven by a very different set of attributes than in a full data set comparison.

Finally, recognizing that no exploratory data analysis package can do everything, the visualizations and tools need to provide easy access to external databases and analytical methods. For example, in the bioinformatics realm, the collection of public domain tools is enormous and rather than attempt to duplicate these, all that is necessary is easy export of data from a visualization into these tools and *vice versa*.

DATA OVERVIEW VISUALIZATIONS

As noted above, complementary data overviews are needed to address different aspects of a large data set. We classify these overviews into four types:

- overviews of the data itself,
- overviews of the relationship of each data record to every other record,
- overviews of the associations within the data set, and

- overviews specific to a particular data type.

To enable the discovery process, each of these visualizations must provide ready access to the underlying information and appropriate analytical tools. With these, it becomes possible to explore prior hypotheses as well as the unexpected relationships often suggested by the structure of visualizations of complex data sets.

CorScape

As one approach for viewing an entire data set, we have created the CorScape visualization (Figure 2A). Here, each data record is a row in the visualization and each attribute in the data table a column. Each cell in this visualization is color-coded to represent the actual data. The color-coding can be defined by continuous variables using a color gradient or specific colors for categorical data or missing values. Thus, this is like a spreadsheet with the individual cells color-coded and then shrunk to make it all visible in a single glance.

The rows in the CorScape are ordered for better recognition of the types of behaviors in the data set. Specifically, the records are first clustered (with cluster membership indicated by the alternating gray bars on the left), then the clusters are correlation ordered, and finally the records within each cluster are ordered using a Euclidean distance measure. The result of this layered ordering is the ability to see structure in the data. Furthermore, with large numbers of records (greater than the number of pixels available for the visualization), the ordering allows smoothing with minimal loss of ability to recognize types of behavior.

In addition to the record (row) ordering, the CorScape allows the columns to be ordered in a variety of ways as well. As for the rows, this provides useful structure in the visualization, but moreover provides an analytical tool. For example, consider a visualization of a number of compounds (rows) tested in several HTS assays (columns); arranging the assays by similarity, the analyst can immediately determine which assays may be providing redundant information and allow future screens to be done in a more cost effective manner.

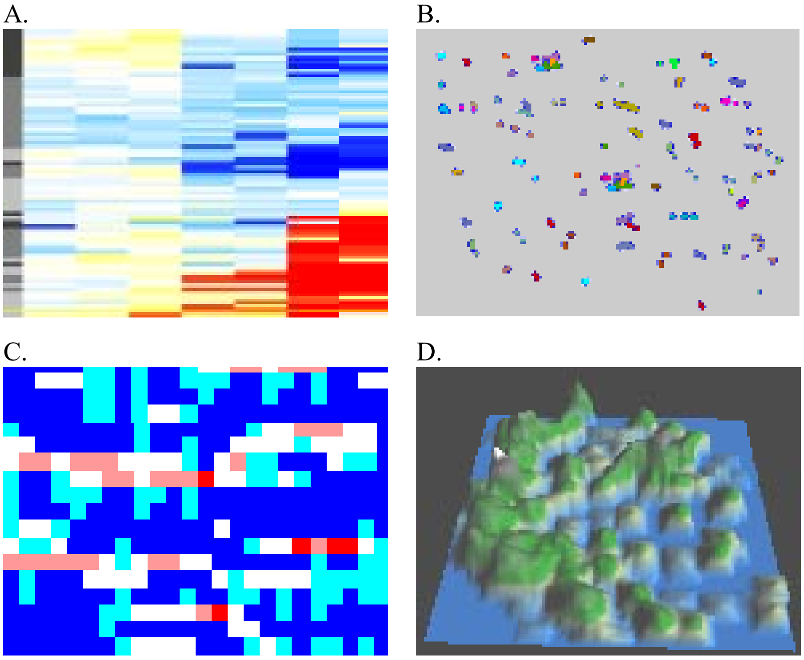


Figure 2: Visualization schemes. A. CorScape. B. Galaxy. C. CoMet. D. ThemeMap.

The CorScape simultaneously provides both a 'far view' which shows the entire data set in one frame and a 'near view' which provides a close-in view of a region of interest in a separate frame. Thus the far view provides the overall context for a data set and the near view allows detailed probing of the data. This two-tiered approach is particularly important for very large data sets.

The approach used in the CorScape visualization is similar in concept to methods employed by Eisen [5] and Weinstein, [6,7] but is done in a manner that is fully interactive. The end result, as implemented in the OmniViz Pro software, is a visualization that allows the analyst to understand the overall nature of the data, discern groupings of records and attributes, and explore the details quickly.

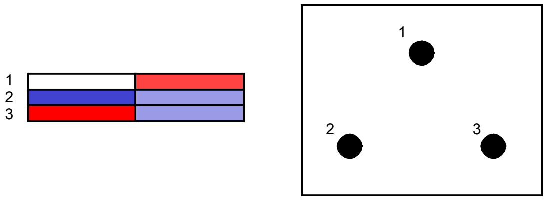
Linking the Familiar with the Useful

Besides providing a useful overview of all the data, the CorScape provides a link from the data table that is familiar to analysts and the higher-dimensional realm of multivariate data. It shows the information in what is essentially a data table, yet adds information about cluster membership. Thus, the CorScape along with tools, such as the NumericRecordViewer, which shows a portion of the data table with both the color code and numeric values, and familiar analytical tools, such as simple plots, provides a natural transition to higher order analyses.

Galaxy

Although the CorScape provides a ready overview of the overall data set, there is a limitation to the one-dimensional ordering in this type of view. Consider a group of three records in a CorScape, ordered 1-2-3 according to some measure of similarity (Figure 3). It may be that objects 1, 2, and 3 are in fact equally related, as in the diagram to the right. In this case, any order of the three records is correct, a complex relationship that can only be indicated in a higher-dimensional view.

Figure 3:



We have created such a visualization, the Galaxy view (Figure 2B), which is a projection of the data records from the high-dimensional space where the cluster analysis takes place to a two-dimensional view in which interactions can take place. In this view, a point represents each data record and a circle represents each cluster centroid. In particular, the Galaxy view shows how each data record is related to every other data record, with emphasis on the natural groups or clusters that occur within that information space. Thus, this visualization is a representation of the information space that allows the analyst to become oriented rapidly and assess global features of the information.

The Galaxy visualization has some features, such as the representation of an entire data set in single view, in common with other methods that have been applied in the chemical and biological sciences. For example, Sammon maps [8] have been used to compare protein sequences [9] and self-organizing maps [10] have been used for gene expression analysis . [11]

The Galaxy visualization also has several unique attributes to assist the analyst. To help with orientation in the Galaxy view, iconic representation of the behavior of the records in each group, for example, miniature plots for numeric data, provide immediate landmarks on the overall map and allow the user at a glance to see how many records have what types of behavior. As implemented in OmniViz Pro, the Galaxy, as with the other visualizations, is fully interactive with ready access to data mining tools.

CoMet

To complement the insights about the data and the relationships of data records as gained from the above visualizations, a separate view of how the attributes associated with each data record are distributed is critical. This can be an assessment of how one or more attributes correlate with clusters of records (associating attributes with the group's behavior) or an assessment of how one set of attributes correlate with another set (independent of record to record relationships).

We have created the CoMet visualization (Figure 2C). This view is a data matrix with the rows and columns representing objects or attributes of interest. For example, if the association of a set of attributes with behavior (clusters) is of interest, the rows would represent each cluster (e.g., compounds grouped by biological activity) and the columns would represent the categorical values for each attribute (e.g., structural descriptor). Each cell in this matrix represents the records in that cluster that contain the attribute in that column and is color-coded according to raw occurrence frequency, percent occurrence, or, usually most valuable, the deviation from expected occurrence. In this way, it is easy to see which attributes contribute to the observed behavior. As with the CorScape, additional value in the visualization is derived by appropriate ordering of the rows and columns. For clusters as rows, these are presented in the same correlation order as in the CorScape. The columns can also be ordered (e.g., correlation) to add structure to the view.

Alternatively, the association of attributes with other attributes can be done by selecting the rows to be other attributes - for example, in a preclinical trial, the association of outcome (categorical attributes) with treatment (a separate categorical value). In this case, each cell in the matrix represents how many records contain the attribute in the row and the attribute in the column, with color-coding using the same statistical options as above.

As implemented in OmniViz Pro, the CoMet visualization is also fully interactive, allowing ready access to the underlying information and the relevant analytical tools.

Data type specific visualizations

For some data types, there are specific visualizations that are needed to convey aspects of the information space. In the case of text, we have created the ThemeMap visualization. The landscape visualization metaphor for the major themes within the text provides a rapid means for getting oriented in the two-dimensional Galaxy projection. To this visualization, we have added a suite of tools that facilitate analysis, discovery, and presentation.

Integration

Each of the visualizations described above provides unique value, but should not be viewed in a vacuum. In the course of data exploration, the complementary views need to be linked together so that assessment across separate analyses, different experiments, or even different data types is facilitated. This linkage must essentially be universal within the information space defined by the data set so that examination of subsets of data (e.g., in progressive disclosure) or different subsets of the data attributes can be fully integrated.

Our method for implementing this unified approach is to provide active linkage of records throughout the visualizations and tools. Using an event-driven model, each visualization and each interactive tool displays the selected records from any other visualization. Thus, records selected in a CorScape view are immediately highlighted in the Galaxy view to link the data overview with the better presentation of record-record relationships. Similarly, records clustered by one set of attributes (e.g., chemical structure descriptors) in one visualization are automatically linked to records in another view clustered by another set of attributes (e.g., biological activity). Linkage from experimental data sets with literature analysis is also possible, through integrated query capabilities. The integration across data sets and data types is facilitated by the common visualization schemes and interactive tools used for all data. This is made possible by the common data table concept; most visualizations and tools access record information through the same underlying data structures.

SUMMARY

As the methods being employed in chemical and life sciences continue to evolve and produce even greater volumes of information, exploratory data analysis will become increasingly dependent on visualization methods. In addition to analysis of specific high-throughput experiments, the integration of multiple experiments across the discovery and development process can be approached. This integration extends across data types to analysis of internal and external data repositories, including historical information such as literature and patents, bringing a new level of continuity to the data mining process.

REFERENCES AND NOTES

[1] van Heel, M. *J Mol Biol.* **1991**, 220, 877.

[2] Hobohm, U.; Sander, C. *J Mol Biol.* **1995**, 251, 390.

[3] Holm, L.; Sander, C. *Science* **1996**, 273, 595.

[4] Holm, L.; Sander, C. *Nucleic Acids Res.* **1998**, 26, 316.

[5] Eisen, M. B.; Spellman, P. T.; Brown, P. O.; Botstein, D. *Proc Natl Acad Sci USA* **1998**, 95, 14863.

[6] Weinstein, J. M.; Myers, T. G.; O'Connor, P. M.; Friend, S. H.; Fornace Jr., A. J.; Kohn, K. W.; Fojo, T.; Bates, S. E.; Rubinstein, L. V.; Anderson, N. L.; Buolamwini, J. K.; van Osdol, W.W.; Monks, A. P.; Scudiero, D. A.; Sausville, E. A.; Zahaevitz, D. W.; Bunow, B.; Viswanadhan, V. N.; Johnson, G. S.; Wittes, R. E.; Paull, K. D.; *Science* **1997**, 275, 343.

[7] Shi, L. M.; Myers, T. G.; Fan, Y.; O'Connor, P. M.; Paull, K. D.; Friend, S.; Weinstein, J. N. *Mol. Pharma*. **1998**, 52, 241.

[8] Sammon, J. W. *IEEE Trans Comp C* **1969**, 18, 401.

[9] Agrafotis, D. K. *Protein Science* **1977**, 6, 287.

[10] Kohonen, T., Self-organizing Maps, *Series in Information Science, vol. 30*, Springer-Verlag, Heidelberg, **1997**.

[11] Tamayo, P. ; Slonim, D. ; Mesirov, J. ; Zshu, Q. ; Kitareewan, S. ; Dmitrovsky, E. ; Lander, E. S. ; Golub, T. R *Proc Natl Acad Sci USA* **1999**, 96, 2907.

¶ OmniViz Pro, CorScape, Galaxy, CoMet, and ThemeMap are registered trademarks of OmniViz Inc.