Presented at the IEEE Symposium on Information Visualization (InfoVis 2003), Seattle, October 19-21, 2003

Interactive Poster: Trend Analysis in Large Timeseries of High-Throughput Screening Data Using a Distortion-Oriented Lens with Semantic Zooming

Dominique Brodbeck
Macrofocus GmbH
dominique.brodbeck@macrofocus.com

Luc Girardin
Macrofocus GmbH
luc.girardin@macrofocus.com

Abstract

We present a design study that shows how information visualization techniques and information design principles are used to interactively analyze trends in large amounts of raw data from high-throughput screening experiments. The tool summarizes trends in the data both in space and time, through the use of distortion-oriented magnification as well as semantic zooming. Careful choice of visual representations allows an information-rich yet easily interpretable display of all the data and statistical indicators in a single view. It is used commercially for quality control of measurements in the drug discovery process.

1. Introduction

High-throughput screening is a technique used in the drug discovery process to find lead candidates for further biological screening and pharmacological testing. Biological targets are thereby tested against large chemical compound libraries, and the intensity (e.g. fluorescence) of the chemical reactions with all the compounds measured. Typical libraries contain 100'000 to 1 million compounds. Several hundreds of them are filled into the wells of a microtiter plate and are brought in contact with the target substance. All the reactions in the wells then take place and are measured in parallel at the same time. This is repeated sequentially with as many plates as it takes to test all the compounds. The processing of such an assay is performed automatically by a robot in several screening runs and stretches over hours or days.

For subsequent data analysis, we therefore have to deal with on the order of 10^2 measurements per plate, for 10^3 plates, leading to a total of 10^5 to 10^6 values. In a first step, the quality of the raw data needs to be assessed in terms of signal strength, background noise, and other effects introduced by changes in the environment during the course of the measurements. The result from this quality control leads to the elimination of bad plates and serves as input for the choice of normalization and correction modes. After this assessment, the data is normalized and corrected, and the timing information discarded. Time is only an artefact of the measuring process and not relevant for the identification of lead candidates.

In the following we describe a tool - named TrendDisplay - that supports the quality control process of raw high-throughput screening data. It solves the problem of representing and evaluating large amounts of time-dependent measured data. In particular our design objectives were:

- show the trend of the raw data for all the wells across a plate
- show the trend of the raw data over time, on different time scales
- provide comparison with additional derived statistical values (signal to noise ratio, standard deviation, etc.)
- allow masking of plates based on thresholding of any combination of derived values
- industrial-strength information design and ease-of-use

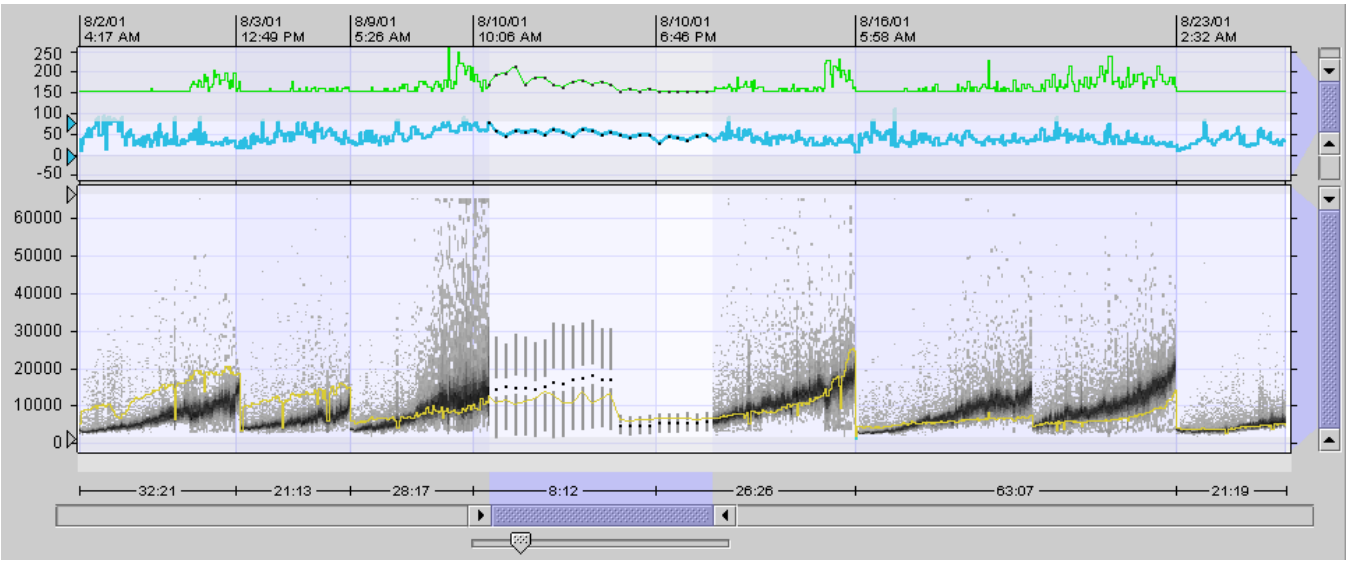

Figure 1: TrendDisplay showing trends both across space and time of in this example 230'400 measurement values, revealing saturation effects, time-dependent drift, as well as outliers. A bifocal lens with semantic zooming allows quick access to and investigation of temporal discontinuities, and anomalies. Derived values such as standard deviation (blue) or number of inhibitor reactions (green) are plotted in the top panel. Thresholds can be set interactively for the active plot (bold blue line) to visually define masking criteria.

2. TrendDisplay

TrendDisplay is composed of two panels: the main panel at the bottom shows all the measured values in one view, and the top panel shows various derived statistical values (Figure 1). The two panels share the same timeline (x-axis) along which the plates are positioned according to when they were measured. The background shading (light/dark) highlights the boundaries of the individual screening runs that make up the whole assay. The time axis at the top shows date and start time for each screening run, whereas the axis at the bottom shows their respective duration. The time gaps between screening runs are removed, to keep the representation contiguous and to save screen space. In addition to this "relative" time mode, the axis can also be switched to show the sequence number of the plates only.

An individual plate is represented as a perceptually linear greyscale density distribution of all the measured values that it contains. In order to avoid the visual activation of empty space between plates, the density distributions are drawn in such a way that they appear as a contiguous band along the horizontal direction, i.e. each individual band is connected to its neighbors to the left and right. We do insert a break for large gaps however, in order to prevent the bands from becoming overly asymmetric. This makes it easy to spot places with highly irregular time stamp distributions.

To cope with the large number of plates and to provide access to details on different time scales, we make use of a distortion-oriented magnification technique, namely a bifocal lens [Apperley et al. 1982]. The lens can be opened and its position manipulated by using the two handles at the bottom of the display. Alternatively an area of interest can be chosen by rubberbanding the desired interval directly in the display, or by double-clicking on a screening run, in which case the lens boundary is positioned at the boundaries of the screening run.

There are various ways to represent a set of measured values and their statistical characteristics, each with their own properties. We therefore implemented the lens as a semantic zoom [Bederson and Hollan 1994], choosing the appropriate representation depending on the amount of available screen space per plate at a certain magnification factor. There are four different levels of detail (from lowest to highest magnification): greyscale density distributions, thin box plots [Tufte 1983], box plots plus individual outliers, bar histograms (Figure 2). The magnification factor inside the lens is controlled by the zoom slider just below the lens position controls.

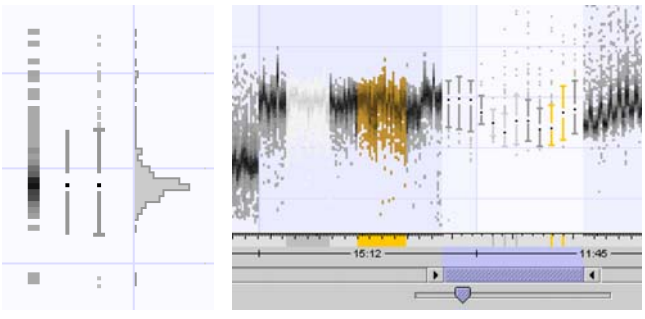

Figure 2: The four different levels of detail: density distributions, thin box plots, box plots plus outliers, bar histograms (left). Brushing and linking: plates can be masked or marked without loosing the representation of the underlying data.

Both panels can also be magnified in the vertical direction independently by using the range sliders on the right side of the panels, or by rubberbanding the desired interval directly in the display. The vertical magnification is implemented as a standard linear zoom, because the y-axis represents a physical scale on which metric comparisons need to be performed, and where geometric distortions would lead to misinterpretations. We use gesture recognition to automatically detect if the desired rubberband interval should be applied to the vertical or horizontal direction, freeing the users from having to learn special keystrokes. All zooming and lens positioning transitions are smoothly animated, to guarantee object constancy and avoid change blindness effects.

In addition to the measured reaction signals, there are several control signals (e.g. neutral reaction signal) and various derived statistical values that need to be visualized and correlated with the compound data. Selected control signals can be overlaid directly over the density distributions in the form of a line plot. In the upper panel, any number of derived statistical values can be plotted. We use different plotting styles that are optimized for the different time scales. Outside the lens, values are plotted in histogram style, to avoid aliasing problems caused by quasi-vertical lines. Inside the lens, values are represented as black dots that are connected by straight lines.

If multiple derived statistics are selected concurrently, then they are overplotted in the same panel on different layers. Each of them is equipped with its own adjustable coordinate system, so that users can freely scale and shift the plots in the vertical direction in order to arrange or overlay them appropriately. In addition there is an upper and a lower threshold for each of the derived statistics that can be set to visually define certain masking criteria (e.g. mask all plates whose standard deviation is above *r*). Thresholds are represented by semi-transparent "curtains" that extend into the panel from the top and bottom.

TrendDisplay supports brushing and linking. Plates can be selected, marked, or masked, which is indicated by different coloring in the main panel, and by little flags in the status strip along the bottom of the display (Figure 2).

3. Conclusion

TrendDisplay is embedded as a component in a comprehensive data analysis suite for biotechnology applications. It receives enthusiastic feedback from customers and enjoys commercial success. We envision similar applications of the approach described here and the techniques used, in timeseries-heavy areas such as finance, event scheduling, or project management.

4. References

Apperley, M.D., Tzavaras, I. and Spence, R. 1982. A Bifocal Display Technique for Data Presentation. In *Proceedings of Eurographics'82*, Conference of the European Association for Computer Graphics, pp. 27-43.

Bederson, B. B. and Hollan, J. D. 1994. Pad++: A Zooming Graphical Interface for Exploring Alternate Interface Physics. In *Proceedings of UIST'94, ACM Symposium on User Interface Software and Technology*, Marina del Rey, CA, pp. 17-26.

Tufte, E. R. 1983. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, Connecticut