A VERSATILE FRAMEWORK FOR THE ANALYSIS OF HIGH-THROUGHPUT SCREENING DATA

Johannes Hamecher[1], *Thorsten Riess*[1], *Enrico Bertini*[1], *Karol Kozak*[3], *Johanna Kastl*[2], *Thomas U. Mayer*[1,2], *Dorit Merhof*[1,4]

[1] Interdisciplinary Center for Interactive Data Analysis, Modelling and Visual Exploration (INCIDE), University of Konstanz

[2] Molecular Genetics, University of Konstanz

[3] Institute for Biochemistry, ETH Zurich

[4] Visual Computing, University of Konstanz

Email: Johannes.Hamecher@uni-konstanz.de

ABSTRACT

Mitosis is an essential process within the cell life cycle, and research in this field has many applications in medicine. In order to study the proteins involved in mitosis and their function, small molecules (compounds) are required which inhibit the protein under investigation and hence allow to switch off the respective protein. For this purpose, high-throughput screening is performed, where thousands of compounds from commercially available libraries are probed using optical readouts. The vast amounts of data generated in high-throughput screening require dedicated data analysis and visual analytics approaches for evaluation. In this work, a versatile framework is presented which provides data preprocessing and visualization approaches for the analysis of high-throughput screening data.

1. INTRODUCTION

The mechanisms of mitosis and involved proteins are of great interest for various applications in medicine, such as cancer research and research about ageing.

In order to study protein function and the role of proteins in mitosis, small molecules (compounds) are required which inhibit the protein under investigation. This method is denoted as chemical knock-out of the protein, and the term 'chemical genetics' has been coined for approaches that use small organic molecules as probes to study protein functions in cultured cells or whole organisms [1].

Compounds are provided in commercially available libraries which display a high degree of structural diversity, whereas the individual compounds are likely to cross cell membranes, contain substructures resembling known bioactive molecules, and they do not contain 'functional groups' (e.g. highly reactive groups) that are likely to cause cytotoxic effects.

In order to study protein function, a compound needs to be identified which inhibits the protein under investigation and hence allows to switch off the protein. In order to identify a compound with the required properties, high-throughput screening is performed. For this purpose, a small volume of a single compound is added to each well of an assay plate containing protein solution. The binding of protein and compound can be assessed using optical readouts such as luminescence, absorbance or fluorescence (e.g. fluorescence intensity, fluorescence polarization, fluorescence resonance energy transfer).

Due to the vast amounts of data generated in such high-throughput screenings, data analysis and visual analytics approaches are required for evaluation. In this work, a versatile framework is presented which provides dedicated approaches for the analysis of high-throughput screening data.

2. MATERIAL AND METHODS

2.1. Screening Data

The high-throughput screening is performed by an integrated robot system which transfers the protein solution to the 384-well plate, adds the compounds and finally reads out the fluorescence polarization signal.

The readout data is stored in Excel files, where one Excel sheet represents one 384-well plate. Each Excel sheet contains multiple blocks of data that correspond to different measurements (i.e. fluorescence intensity, fluorescence polarization, fluorescence resonance energy transfer). The last two to four columns (depending on the experimental setup) of the well plate contain control values (positive and negative control) which can be used for intensity normalization.

2.2. Data Processing

2.2.1. Software Framework:

The software platform KNIME (The Konstanz Information Miner [2]) is an open-source tool for data integration, processing, analysis and exploration. Essentially, KNIME ist designed to import, transform and visualize large data sets in a convenient and easy to use way. KNIME

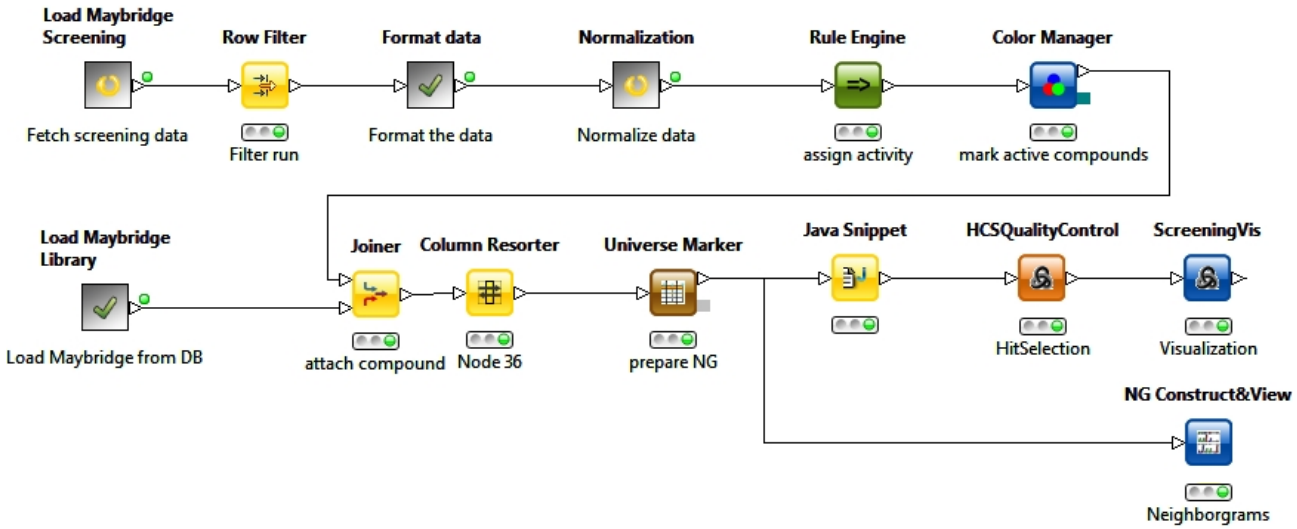

Figure 1. KNIME workflow for high-throughput screening data analysis. The pipeline comprises nodes for data preprocessing (e.g. normalization), interactive plate visualization and for computing neighborgrams.

workflows consist of interacting nodes, which may each represent an algorithm, a single import routine or a visualization tool. The data flow is visually represented by connections between the nodes, typically starting with a node to import the data, followed by one or more processing nodes and finally one or more output nodes. A graphical user interface makes it possible to construct workflows consisting of different nodes and their interconnection via a simple drag-and-drop mechanism.

In this work, KNIME is used as a basis to implement a fully automated data analysis workflow for high-throughput screening. Due to the modular design of KNIME workflows, it is also possible to assess intermediate results at every stage of the processing pipeline.

2.2.2. *Data Preprocessing*

Prior to further data processing, the screening data need to be normalized. The positive and negative controls are used to normalize every value in each plate. Subsequently, abnormally high values resulting from auto-flurescent compounds are thresholded and excluded from the analysis. Finally, the user can choose a threshold value in order to classify compounds into active and inactive ones (hit selection), depending on their fluorescence polarization value.

2.2.3. *z-Factor*

A z-factor is computed per plate and is used to assess the quality of a screening assay. The z-factor is defined in terms of the sample means $\hat{\mu}$ and the sample standard deviations $\hat{\sigma}$ of the positive p and negative controls n:

$$\text{z-factor} = 1 - \frac{3\left(\hat{\sigma}_p + \hat{\sigma}_n\right)}{|\hat{\mu}_p - \hat{\mu}_n|} \qquad (1)$$

According to [3], a z-factor less than 0 indicates too much overlap between positive and negative controls, a factor between 0 and 0.5 refers to a marginal, and a factor between 0.5 and 1 to an excellent result. The z-factor is therefore an important indicator whether a plate was screened successfully.

2.2.4. *Visualization*

A simultaneous visualization of all plates provides an overview of the whole experiment. The normalized an thresholded data values are mapped to color values, where dark colors represent low values and brighter colors represent higher values. Since each plate is screened multiple times, an overview visualization allows to compare the test - retest results obtained from multiple scans of the same plate. In this way, plates can be rejected where the measurements suffer from confounding factors and are less accurate.

After applying the overview visualization for quality control, further visualization approaches are needed in order to identify interesting compounds. This is supported by a comparative view which allows to compare the duplicate screens (of the same compounds), a histrogram view to obtain insight into the count of interesting data values, and most importantly the hit-selection view. In this view, the user can select compounds with an interesting total intensity and activity level. The selected data value is automatically connected with the value obtained in the second run, which allows to compare compound values between both screens.

2.2.5. *Visual Analytics Approaches*

Fingerprints are binary vectors which are used in order to describe the presence or absence of some properties (e.g. fragment substructures) within a compound [4]. The bits in a fingerprint may encode structural information, e.g. simple descriptors such as the numbers of atoms and bonds or the number of rotatable bonds, or distance information between pharmacophoric groups. Compounds are

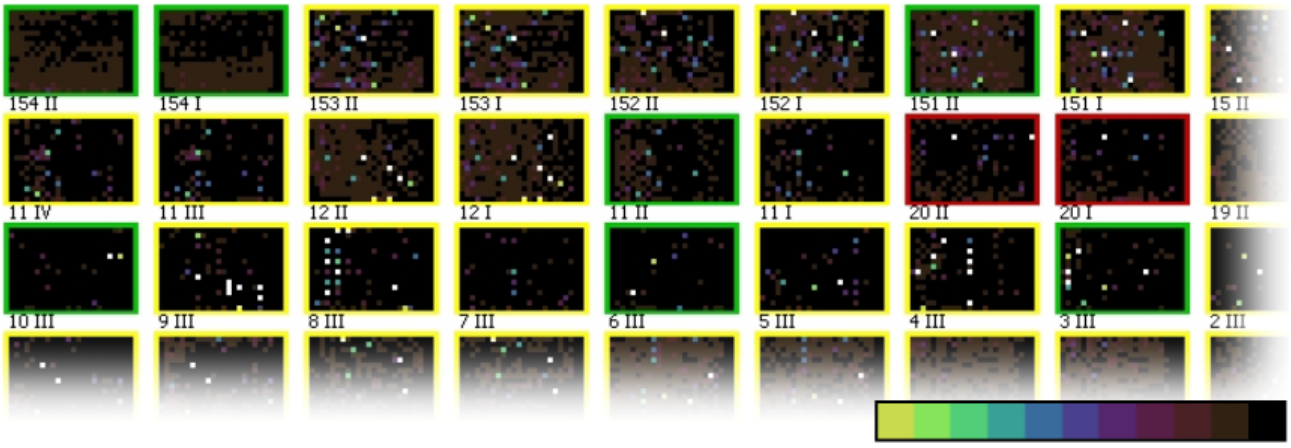

Figure 2. Section of a plate overview visualization. The z-factor for each plate is represented by a colored rim around the plate, where red indicates a z-factor below 0, yellow between 0 and 0.5 and green between 0.5 and 1.

expected to be structurally similar, if they have many of such bits in common.

The distance between compounds is usually described using the Jaccard-Tanimoto coefficient [5, 6] of their fingerprints. Based on these distance measurements, neighborgrams [7] can be reconstructed which allow identifying compounds of interest. For each compound, a neighborgram is constructed with the selected compound as reference (centroid). The n compounds which are closest to the centroid in terms of the Tanimoto coefficient are mapped into the neighborgram. Active compounds are displayed in green, whereas inactive compounds are represented in blue.

3. RESULTS AND DISCUSSION

The previously presented methods for analysis and visualization of high-throughput screenings were integrated into the KNIME workflow shown in Figure 1. As an initial step, the workflow loads both the screening data and the library containing structural and other information about the compounds. The screening data is then normalized and combined with the library. The visualization node which provides a plate overview and a neighborgram viewer conclude the workflow.

In Figure 2, the visualization approaches to simultaneously display all plates along with corresponding z-factors are shown. The z-factor for each plate is represented by a colored rim around the plate, where red indicates a z-factor below 0 (corrupt), yellow between 0 and 0.5 (marginal) and green between 0.5 and 1 (excellent). The color map to visualize the data values in the Excel sheets is chosen such that low data values are represented by dark colors and higher values are represented by brighter colors.

The structural analysis of compounds based on neighborgrams is shown in Figure 3, the structure of individual compounds is displayed on demand. A biologically interesting configuration occurs e.g. if an active compound is surrounded by inactive ones, which indicates that this compound must have a specific structure element which makes it active.

Since fingerprints are represented by long bit vectors (at least 4096 digits), they are elements of a high dimensional space. Visualizing elements of such a space in 2D or 3D whilst keeping basic properties is a major challenge, and neighborgrams are a possible choice. However, they are limited to local neighborhoods of selected compounds and never describe the entire global situation. Also, data analysis via neighborgrams is highly dependent

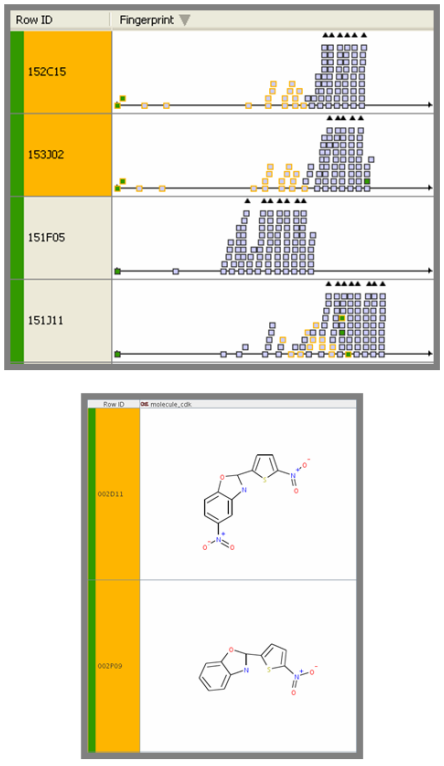

Figure 3. Neighborgram *(top)* and structural view *(bottom)* of selected compounds.

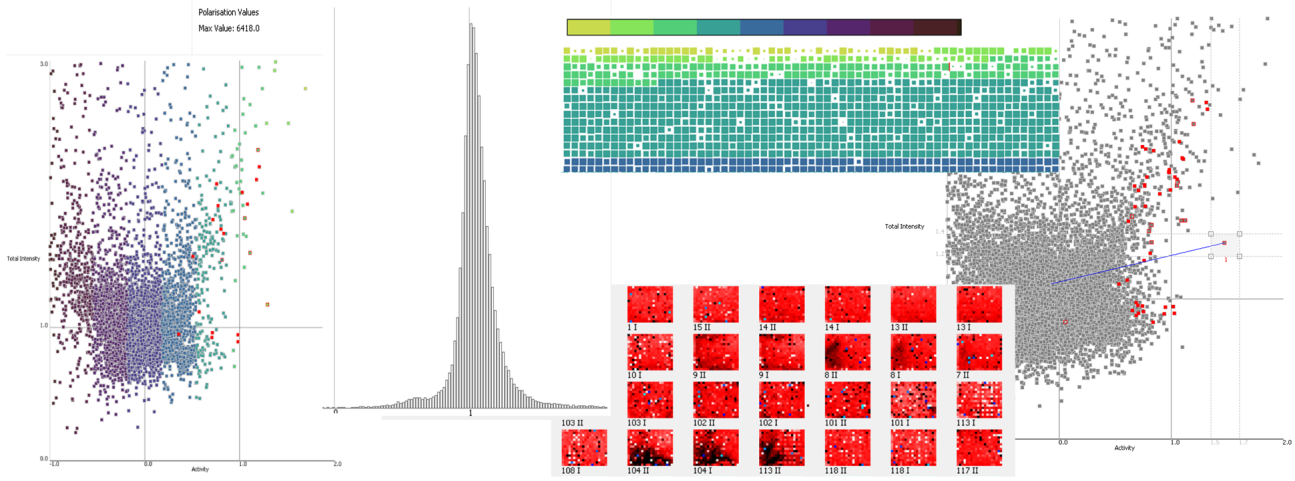

Figure 4. Visualization methods for high-throughput screening data analysis. Illustrative examples for the quality control view, the histogram view and the hit selection tool.

on the choice of fingerprints and the distance metric between fingerprints. For example, if the activity level of a compound is determined by its surface polarization, a fingerprint based on structural information will not provide sensible information, and vice versa.

4. CONCLUSION

The presented framework for data analysis of high-content screenings is a versatile processing pipeline which comprises various analysis tools. The visualization node (plate overview and z-factor visualization) allows for quality control and enables the user to check if any major biases have been introduced in the experiment. The hit selection allows to manually define the data values representing active compounds, which proved to be very useful in combination with the neighborgram analysis. Finally, neighborgrams help to understand the relationships between the compounds and allow to explain the observed behavior based on their chemical properties. Overall, the presented techniques proved to be highly valuable for identifying relevant compounds from high-throughput screening data.

5. REFERENCES

[1] T. Mayer, "Chemical genetics: tailoring tools for cell biology," *Trends in Cell Biology*, vol. 13, no. 5, pp. 270–277, 2003.

[2] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, and B. Wiswedel, "KNIME: The Konstanz Information Miner," in *Proc. Data Analysis, Machine Learning and Applications*, 2008, pp. 319–326.

[3] J. Zhang, T. Chung, and K. Oldenburg, "A simple statistical parameter for use in evaluation and validation of high throughput screening assays," *Journal of Biomolecular Screening*, vol. 4, no. 2, pp. 67–73, 1999.

[4] N. Nikolova and J. Jaworska, "Approaches to measure chemical similarity – a review," *QSAR & Combinatorial Science*, vol. 22, no. 9-10, pp. 1006–1026, 2003.

[5] P. Jaccard, "Distribution de la flore alpine dans le bassin des Dranses et dans quelques regions voisines," *Bulletin del la Societe Vaudoise des Sciences Naturelles*, vol. 37, pp. 241–272, 1901.

[6] D. Rogers and T. Tanimoto, "A computer program for classifying plants," *Science*, vol. 21, pp. 1115–1118, 1960.

[7] M. R. Berthold, B. Wiswedel, and D. E. Patterson, "Interactive exploration of fuzzy clusters using neighborgrams," *Fuzzy Sets and Systems*, vol. 149, no. 1, pp. 21–37, 2005.