Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

An Interactive Visualization-Based Approach for High Throughput Screening Information Management in Drug Discovery

Tammy Pui Shan Chan, Preeti Malik, and Rahul Singh
Department of Computer Science, San Francisco State University, San Francisco CA 94132

Abstract - While High Throughput Screening (HTS) techniques are capable of generating large amounts of biologically significant data, assimilating and mining this information can be extremely complex and potentially crucial information patterns can easily be lost in the mounds of data. The predominantly life-science oriented scientific training of the researchers in this area furthermore, precludes their using complex querying or data-mining algorithms. Keeping in account these challenges, our goal in this paper is to provide a highly intuitive environment for storing and interacting with large amounts of HTS assay data. The principal modes of user-data interactions supported in the proposed paradigm are interaction and visualization rich. Moreover, they span the heterogeneous data modalities common to drug discovery, including but not limited to chemical structures, high-throughput assay formats, graphical information, and alpha-numeric data types. Case studies and experiments demonstrate the efficacy of the proposed approach in terms of its ease of use as well as its capability to discern complex information patterns in the data.

I. Introduction

Advances in combinatorial chemistry have resulted in a need for rapid screening of potential drug compounds. High-Throughput Screening (HTS) [1] is a recently adopted technique by the pharmaceutical industry and academia to quickly narrow down a list of potential drug candidates among a number of compounds for further experimentation. High-Throughput Screening enables testing a large number of compounds at a fixed known dose for binding activity or biological activity against target molecules, often in parallel in a multi-well plate. A screening experiment can contain a single or multiple plates. Positive or active results are called hits. After identifying the corresponding lead compounds, an iterative lead optimization process is initiated to locate the optimal concentration and structure of the drug as well as to optimize its Pharmacokinetic and Pharmacodynamic characteristics. Currently, the cycle to research and develop a drug takes 10-15 years and is highly expensive. *Therefore, development of technologies that can shorten this process and make it more effective can have a crucial impact.* So, there exists an urgent need of systems for storage, management and knowledge discovery, especially in context of high-throughput drug discovery data. The FreeFlowDB project [2-3] seeks to address this challenge by developing a seamless and flexible information management system with an emphasis on visualization and interactive queries to interface between users and the complex bio-chemical information. Some of the major challenges and how FreeFlowDB addresses them include:

- *Need to support data capture of the complex data being generated from high throughput screening experiments:* A flexible data model is required to encapsulate the storage of the heterogeneous high throughput assay and chemical data have been the core to the FreeFlowDB database. An extension to this database allows the storage of data at the experiment level (multiple plates) also, along with storing chemical drug conformers.
- *Need for rich graphical and intuitive user interface to provide better user-data interaction to assimilate information:* The system should also support molecular structure-querying and provide intuitive ways of simultaneously visualizing and querying activity data. After spotting similar drug molecules with a targeted hit compound, exploring the Structure Activity Relationship (SAR) is possible. Besides, a comprehensive collection of visual aids are available to detect equipment malfunction by showing hit distribution trend, to validate hit reliability by a scatter-plot of two assay result indicators and to study hit result for instance, by a dosage-response curve.

There has been prior and ongoing work in this field both in academia as well as in the industry. Public databases such as NCI [4] and ChemDB [5] contain activity and chemical data. However, these databases do not provide comprehensive, intuitive and flexible visualization tools, and most importantly, do not consider the problem of assay data management. Commercial products such as Accelrys, IDBS and MDL [6-8] address some of these issues. However their high cost typically precludes their use in academic research or even for small enterprises. Thus at the state of the art there is no publicly available system for management and interaction of data from modern High Throughput pharmaceutical investigations. Towards this goal, this paper extends our previous research [2-3] in providing a platform for storing and processing activity data and performing data analysis of both biological activity data and chemical drug structures simultaneously, by having intuitive user interface to visualize HTS data in tabular report or graph at plate level as well as experiment level (consists of multiple plates). FreeFlowDB is currently being used and validated at the University of California, San Francisco (UCSF) for anti-malarial drug discovery [2].

II. Design Philosophy and System Overview

The importance of usability and user experience is becoming very clear since most bench biologists do not have the database know-how to explore their data directly and thus, they are limited by the options presented to them on the interface. So, intuitive interfaces to aid in HTS experiments where a huge amount of data needs to be addressed are necessary. FreeFlowDB has been developed

as a web-based application running on an Apache web server and Tomcat servlet container configured for hosting dynamic web content. The data is persisted in the backend using MySQL and Berkeley DB XML databases. The system is organized using a three-tier methodology such that the presentation logic, business logic, and data access layers are loosely coupled. Its key modules are:

1. *Data Loader and Data Processor:* Are an important component of the web tier and are implemented in PHP. Data processor can take processing scripts to calculate the necessary result on assay data which is uploaded via the Data Loader, depending upon the study.
2. *Plate Builder/Viewer:* Is implemented as a Java applet to provide a rich set of GUI controls and client-side dynamic behaviors to perform virtual plating.
3. *Flowalyze:* Includes a set of data analysis and visualization tools to help the researchers in understanding their data, by the use of histograms, graphs, scatter plot and other visualization techniques.
4. *DB Interface:* Serves as the middle tier of FreeFlowDB, it acts as the controller in the process flow of the system. It is responsible for all the business logic as well as delegates database calls to the appropriate DB module.
5. *Data Access Modules:* Consist of the data access layer and are implemented using Java JDBC and Java-on-XML technologies.

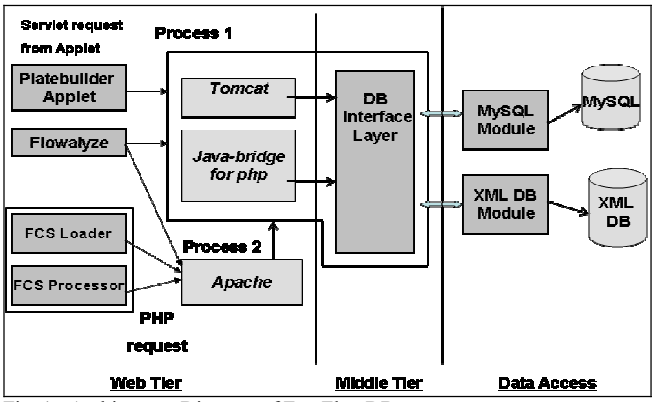

Fig. 1. Architecture Diagram of FreeFlowDB

III. Data Storage and Modeling

Some of the challenges facing database design for HTS systems are:

- Designing an extensible schema to overcome the dynamic nature of data from different HTS assays being tested with different research perspectives.
- Supporting a data model that takes into consideration the dynamic user-workflow that can change from experiment to experiment.
- Storing and querying of molecular structures, including multiple molecular conformations
- Storing structure-activity data.

FreeFlowDB is designed to provide support for all the above mentioned features. An extension to our earlier work is storing the drug conformers as well. Using drug conformers, shape and property similarity can be used to determine potential new leads for biological screening. Most of the data in FreeFlowDB is stored inside the database tables to ensure proper indexing of the different entities, but the raw data from HTS instruments are persisted in the file system. The key entities of the database schemata are: the *Plate entity*, the *Well entity* and the *Drug entity* as discussed in [3]. These entities have static information for an HTS experiment and hence are persisted in the relational database. Other well properties typically have dynamic characteristics that may change at different stages of the drug-discovery experiment process and do not have fixed data types, so they are persisted using an Berkeley DB XML database. These are results either by processing the raw HTS data using some algorithmic scripts or by virtually plating the activity data collected from some outside sources. Such properties can be stored at plate level as well as at experiment level in multiple-plate format. In the context of anti-malaria drug research, these properties include survival rate, cell event count, hit indicator as assay related well properties and log P, solubility as activity related properties and also any other properties that might be of importance to the research.

IV. Interaction and Visualization Support

FreeFlowDB provides a set of visualization tools for user-data interaction after assay data are captured and processed [2]. The Data Loader module allows efficient upload of raw HTS data files and associates each file with its corresponding well location using a mapping file. Afterwards, it is essential to collect and track the experimental data for the corresponding raw data files so that advanced data analysis can be fully explored later. Plate Builder provides an intuitive graphical representation of a multiple-well plate for capturing metadata of an experiment in a similar workflow when compared to an actual one. Researchers can select the entire row/column or a particular well and enter (virtually plate) the metadata values (such as drug, drug concentration, reagent, target used, etc) on the plate, thereby promoting accurate and thorough data entry. Data Processor will then process raw HTS data files along with the experimental data captured to generate quantifiable result to be stored in databases, for later analysis in either Plate Viewer or tabular form. An extended version of Data Processor is implemented to compute processed result of an experiment that screens multiple plates and select potential hits with different criteria, either by fixed capacity (e.g. top 1% of result) or simple statistics (e.g. results that are 2 standard deviation away from sample mean). A comparison of hit selected by these criteria can identify marginal hits which could be a potential drug candidate. Other than this extension, a variety of visualization tools and graphs to study assay information are created to explore the data, such as histogram of count of Fluorescent Intensity (FI) reading from screening devices, histogram of sum of hit distribution, dosage-respond curve, scatter plot of control based assay activity with non-control based indicators and Structure Activity Relationship (SAR) visualization capability which

are going to be elaborated in the next section.

V. Visualization Support for Hit Analysis

Ranging from different granularity, either for a well, a plate or an experiment, the set of visualization tools in *Flowalyze module* provides researchers intuitive interfaces to explore and analyze the data in most aspects:

1. *Structure Query-Analysis and Visualization:* FreeflowDB represents molecular structures as 2D or 3D graphs and supports a graduated assignment-based non-linear optimization technique for graph matching. Among others, exact matching (in the presence of Euclidean transformations), sub-structure matching, and in-exact matching are supported. For details on the structure analysis capabilities, we refer the reader to [3].

2. *Histogram plot of frequency count of Fluorescent Intensity (FI):* By left clicking a well in Plate Viewer, the FI reading in the associated raw HTS data files are counted and plotted to provide a basic understanding of the raw data. Fig. 2(a) shows a negative control well that has lots of low FI reading but few high FI reading while Fig. 2(b) shows a positive control well that has lots of high FI reading. The processed result of these two wells represents two extreme ends of the assay indicator.

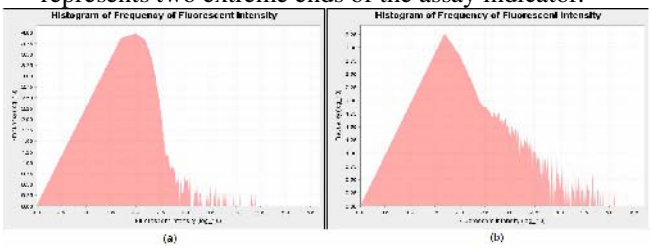

Fig. 2. Histogram plot of FI count to show in this assay the extent of malarial infection of cells by FI readings of negative control well which is shown in Fig. 2(a) and that of positive control well is shown in Fig. 2(b).

3. *Dosage-response curves:* This is useful for researcher to identify the efficacy of a potential drug molecule after an initial screening. The lower the drug concentration needed for the desired biological effect, the better. The graph can be plotted by well data horizontally in certain row/column on a plate, or vertically in certain well(s) of all the plates in an experiment. Fig. 3(a) shows the dosage-respond curve of a highly effective drug molecule that reached 50% of the desired biological effect in a small dosage (250nM) while a less effective one that requires a large dosage (5200nM) is shown in Fig. 3(b). So, the drug molecule that produces the curve of Fig. 3(a) is more preferable to that of Fig. 3(b).

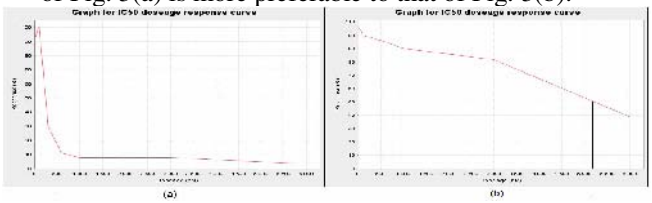

Fig. 3. Dosage-response curves of two drug molecules help to identify the preferable one which requires lower concentration to reach desired biological effect

4. *Sum of hit distribution trend visualization:* By adding up the number of hits selected in each well location for all the plates in an experiment, histograms of sum of hit distribution by row/column and an aggregated plate can be generated to display hit distribution trend. They can help to detect equipment malfunction or calibration problem so as to maintain a better quality control of data and thus increase hit reliability. For instance, if the histogram of sum of hit distribution is particularly low in Row A/Column 1, which is the top/left edge of a plate, an edge effect may be possible for wrong calibration of reading device. On the other hand, if an unexceptional high value of hit of a fixed well location is detected on the aggregated plate, researcher can study and determine if the result is reliable: caused by similar potential drug molecules used in that well of all plates in an experiment or by equipment reading malfunction in a fixed location.

5. *Scatter plot of control based assay activity with non-control based indicators:* Selection of hit strategy is divided into two main categories: control based and non-control based. The former depends on positive and negative control values to produce normalized processed result for hit selection while the latter relies on statistics to select outliers from a pool of sample data without control. A drawback of the former strategy is that the normalized result will lead to a higher false positive/negative rate if the control well readings are unreliable. The latter strategy has its own pitfalls as it can identify outliers but there is no guarantee that those outliers are hits. So, a cross comparison of both strategies in a scatter plot for all sample data values can further validate hit reliability on top of studying hit distribution trend. Fig. 4 shows a scatter plot of a normalized processed assay indicator by positive and negative controls verses the well-known statistics Z-score. If the hits selected by both strategies agree with one another, high correlation should be seen in the scatter plot. Fig. 4(a) shows an ideal case while Fig. 4(b) shows some disagreements in the lower left quadrant.

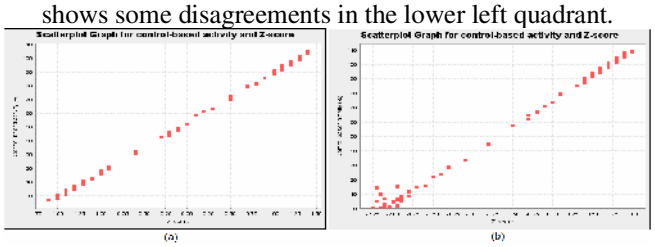

Fig. 4. Scatter plot of two assay indicators, with a highly correlated one shown in Fig. 4(a) and a less correlated one as shown in the lower left quadrant of Fig. 4(b)

6. *Structure Activity Relationship visualization capability:* Our previous research work can successfully identify similar drug molecules from a target by exact structural matching, sub-structure querying or in-exact matching [3]. Now, we embrace this similarity search for researchers to visualize the Structure Activity Relationship (SAR) as shown in Fig. 5. After a similarity search of a hit drug molecule is done, all the wells (B6, C4, C7, D3, E9 & E11) that have similar drug molecules found are highlighted in Plate Viewer. Researcher can then toggle between different assay activities indicators

to study the impact of chemical structures on the assay activity. For instance, to check if these highlighted drug compounds are selected as hits or not and study the reason why or why not. Also the system allows structure viewing of multiple conformers of any drug being tested to find out if any other low energy conformation of the drug exists as an alternative potential drug to be used. This is shown in Fig. 6. A tabular view of the conformer can also be displayed.

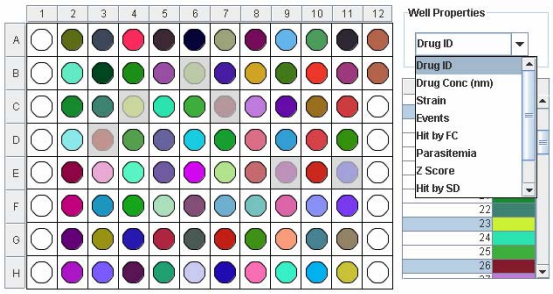

Fig. 5. The targeted drug molecule and similar one found are highlighted in well B6, C4, C7, D3, E9 and E11. Researchers can then toggle between different indicators to study the relationship of activity data alongside with the drug molecule structure.

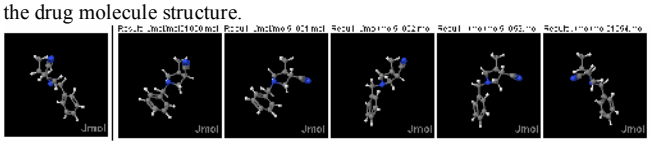

Fig. 6. Multiple drug conformers are displayed alongside with the target drug molecule to help visualize conformational flexibility of the molecules.

VI. EXPERIMENTS

For the evaluation of the system, we use HTS data available from the Harvard Medical School ChemBank [9] and from anti-malarial HTS screens at UCSF. From the first set, we use the data that consists of over forty 384-well plates and found 25 lead drug molecules for inhibition of the BH3 domain of Bcl-2 family members. Using the structure analysis capabilities, similar molecules among the hits can be determined and visualized as shown in Fig. 7.

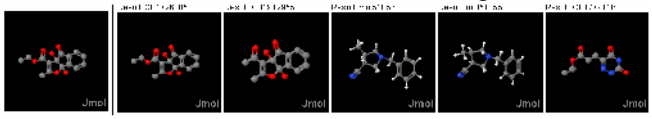

Fig. 7. Similar drug molecules are identified in the hits selected of a fluorescence polarization screening experiment

To demonstrate how FreeFlowDB facilitates rapid assimilation of complex information, four figures are captured from a dosage-response experiment of six drug molecules (put in Row A to F) from the UCSF anti-malarial screen. Each molecule is represented by a unique color in Fig. 8(a) and cells are put in all wells except for Row G-H and Column 12 which are grayed in Fig. 8(b). The drug concentration of each molecule (10nM-6000nM) decreases from left to right and is shown in a gradient scale in Fig. 8(c). The biological effect of the drug molecule and concentration is shown in Fig. 8(d), where more light color wells appear from left to right as concentration decreases in Row C, indicating death of infected cells. Thus, one can rapidly deduce that this drug molecule inhibits the target at lower concentrations (is a more effective one) than the other candidates.

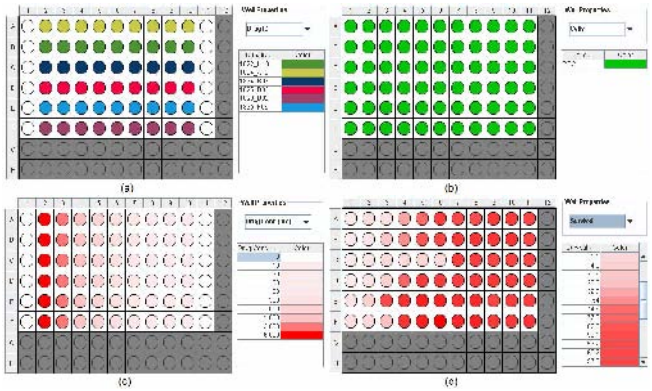

Fig. 8. Six drug molecules are put in each row in Fig. 8(a) and cells are put in 66 wells in Fig. 8(b). The drug concentration decreases from left to right starting from Column 2 in Fig. 8(c) and the biological effect in each well is visualized in Fig. 8(d). It is easy to note that the molecule in Row C is more efficacious as concentration is reduced.

VII. CONCLUSION

With the advancements in automation and combinatorial chemistry, the amount of data collected in drug discovery experiments using HTS approaches is increasingly becomes significant. Assimilating this information is complicated by two factors: (1) the complexity of the captured information and (2) the fact that researchers trained to work with such information are typically not trained to conduct computational querying/data mining operations. This paper presents FreeFlowDB, a high-throughput drug discovery information management system. Its design emphasizes the role of intuitive and flexible visualization-interaction to mediate user-data interactions and information assimilation. The system also provides effective management of heterogeneous data, typical to drug discovery as well as techniques to query and analyze such information. Case studies presented in this paper illustrate the effectiveness of system towards assimilating complex drug discovery data.

REFERENCES:

[1] Hertzberg, R. P., & Pope, A. J. (2000). High-throughput screening: new technology for the 21st century. *Current Opinion in Chemical Biology*, *4*, 445-451.

[2] P. Malik, T. Chan, J. Vandergriff, J. Weisman, J. DeRisi, and R. Singh, "Information Management and Interaction in High-Throughput Screening for Drug Discovery", Database Modeling in Biology: Practices and Challenges Z. Ma, and J. Chen, eds., Springer Verlag, 2006

[3] R. Singh, E. Velasquez, P. Vijayant, and E. Yera, "FreeFlowDB: Storage, Querying and Interacting with Structure-Activity Information for High-Throughput Drug Discovery", *Proc. IEEE Computer based Medical Systems (CBMS), 2006*

[4] National Cancer Institue http://www.cancer.gov

[5] J. Chen, SJ Swamidass, Y. Dou, J. Bruamd, and P. Baldi, "ChemDB: A Public Database of Small Molecules and Related Chemoinformatics Resources", Bioinformatics, 2005

[6] Accelrys: http://www.accelrys.com/chemicals/doe/

[7] IDBS: http://www.idbs.com/solutions/

[8] MDL: http://www.mdli.com/

[9] ChemBank The Broad Institute's Chemical Biology Program. http://chembank.broad.harvard.edu/