FreeFlowDB: Storage, Querying and Interacting with Structure-Activity Information from High-Throughput Drug Discovery

Rahul Singh*, Elinor Velasquez*, Preeti Vijayant*, and Emmanuel Yera*
Department of Computer Science, San Francisco State University, San Francisco CA

Abstract

The state of the art in modern drug discovery involves investigating a large number of drug-like molecules using medium or high-throughput assays, often being conducted against multiple targets. Managing the information generated in such processes requires the ability to deal with complex, multifarious data as well as the development of new user-data interaction paradigms that help glean patterns hidden in the multitude of data by emphasizing exploration and information assimilation. This paper describes our research in developing FreeFlowDB, a drug discovery information database system that is geared towards storing both structural as well as high-throughput assay information generated as part of a typical drug discovery process. FreeFlowDB supports powerful structural querying facilities that subsume within a common algorithmic framework exact structural matching, sub-structure querying, and in-exact matching. Furthermore, the system supports unified visualization-query facilities that allow interacting with assay as well as structure-activity information. This allows efficacious and intuitive query-analysis of large amounts of data for knowledge discovery. Case studies and experimental results demonstrate the capabilities of the system.

1. Introduction

The contemporary paradigm in pharmaceutical drug discovery is built around advancements in Genomics, Combinatorial Chemistry, and High-Throughput screening (HTS). Recent advances in Genomics promise to provide a proliferation of targets that can lead to newer or improved therapeutics. Similarly advances in combinatorial chemistry, and HTS have significantly increased the number of lead compounds that can be synthesized and experimentally studied in the drug-discovery process. Thus it is possible today to simultaneously study a larger number of targets, investigate more putative drug candidates, and conduct significantly more biological assays than ever before. However, translation of this obvious scale-up in data generation capabilities to the ultimate goal of faster and less expensive development of better therapeutics requires addressing many unresolved problems. Of these two critical ones are:

- *Development of information management techniques that can support effective storage of both structure as well as high-throughput screening information:* Specifically, such techniques need to (1) support data storage in high-throughput assays, where the structure of the generated data encodes important biological and/or context dependent information. (2) store and manage heterogeneous data, such as molecular structures, biological assay values, as well as a slew of non-alphanumeric information such as graphical information, data from analytics, and image-based information, and (3) facilitate correlation of semantically related information across highly heterogeneous data such as molecular structure and biological experiments (activity).
- *Development of user-data interaction techniques that allow researchers to formulate hypotheses, explore the data, and assimilate information:* In particular, we are interested in (1) supporting exploratory structure-querying, (2) seamlessly relating structure and activity patterns, and (3) developing novel ways of simultaneously visualizing and querying activity data from high-throughput screens.

In this context, the state-of-the-art today consists of privately held data repositories such as the CAS registry of the American Chemical Society, industrial solutions such as [6, 7, 8], and a small number of academic or public initiatives such as ChemDB [2] and the National Cancer Institute (NCI) open database, that are designed to contain drug-discovery related information. Of these, efforts such as the CAS registry, the NCI open database, and ChemDB are envisaged as large collections of small molecule structural information that may additionally contain simple physico-chemical properties of the molecules. This is distinct from our goals of developing a system that can not only store molecular structural information but also information related to their biological activity obtained

*Equal Contributors

through complex assays. Thus we are interested not only in structure information management but also in assay (activity) information management. In this sense, our efforts closely mirror the goals of industrial solutions in this area. However, three key points distinguish our effort is from similar industrial systems. The first two are technical in nature and relate to the two aforementioned issues of storing structure and activity information as well as developing novel user data interaction paradigms that emphasize exploration and assimilation and extend the limited query retrieval formulation supported in contemporary commercial systems. Finally, we aim ultimately for a publicly available solution that will aid biochemical investigations targeted at discovery of novel therapeutics by researchers around the world. For example, an early version of FreeFlowDB is being currently used by our collaborators at the University of California, San Francisco for anti-malarial drug discovery [4].

2. System Overview

Modern drug discovery is a complex process involving multiple stages. Typically in the initial stages a large library of compounds is tested for activity against targets (typically proteins or enzymes) of interest, using high-throughput miniaturized assays. It is not uncommon to see millions of putative drug molecules being tested in this stage. Molecules that interact with the target at this stage are called *hits or leads*. Generally, this stage is followed by more detailed bio-chemical experiments, often involving high to medium throughput cell-based assays. Finally, a cycle of iterative chemical refinement and testing is performed on the *leads* until a molecule with desired potency, pharmacokinetic, and pharmacodynamic profile is obtained. This molecule then undergoes clinical trials and if successful enters the market. Information management efforts in such a setting need to address the following technical challenges:

- *Facilitate automated or semi-automated data capture from HTS equipment*: This includes machine-readable data that is generated from the laboratory automation instrument and metadata that describes key details of an experiment, such as the origin, concentration, and characteristics of the given samples, drug structure data and processed data.
- *Data preprocessing and analysis*: Involving noise reduction and data analysis to determine lead compounds.
- *Design of interfaces that allow users to browse, explore, and query the data*: Specifically, such interfaces should be generic, intuitive, and be able to support efficacious interactions with huge amount of data.
- *Support structure-based query-retrieval and structure-activity correlation and visualization*

As illustrated conceptually in Figure 1, FreeFlowDB provides a unified and seamless environment to address the aforementioned issues, ranging from information upload to advanced facilities for information retrieval and visualization. In the following two sections, we describe the capabilities of the system in context of assay-data management and structural information management in greater details.

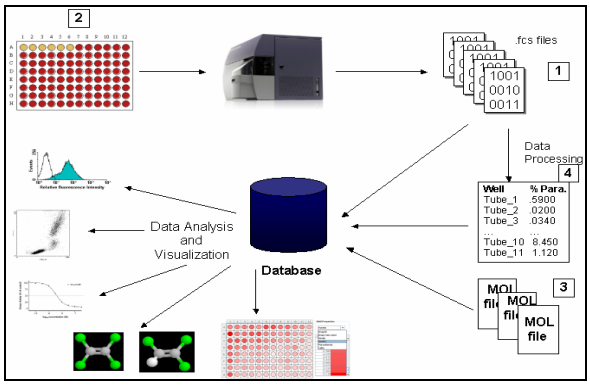

Figure 1. Overview of FreeFlowDB

3. Assay Information Management: Data Entry, Processing, and Visualization

As the first step in high-throughput assay information management, binary files, generated from devices used for initial screening (such as flow cytometers or other types of fluorescence detection devices) are uploaded. This step requires design of device dependent readers. For further details on this step, we refer the reader to [4]. Following the data upload, a key step of *plate-mapping* is executed. This involves creation of mapping file which specifies the association between the binary files and their well locations. Plate mapping is achieved through a powerful visualization environment that allows users to configure the data entry. Once the plate mapping is defined, it is possible to track the specific data in each well through the entire life-cycle of the drug-discovery process.

Following the data upload and plate mapping, FreeFlowDB allows noisy data to be identified and removed through the specification of certain processing and quality control (QC) parameters. At this stage for ease of assimilation, the QC indicators are calculated and displayed (through color-coding) alongside each well, to indicate the validity and

trustworthiness of that data in the well. For example, if the number of cells collected in a well is less that specified by the user, the well data is highlighted in red to alert that the data is being computed based on an insufficient input volume. Users can then either choose to tolerate the deficiency or discarded the data. Once the processing step is finished, a measure of drug effectiveness (such as IC_{50} or the number of cells where the drug interacted with the target) is calculated and stored into the database. This information can later be used to detect trends in the data or visualize drug structure-efficacy relationships.

The system presented provides a highly interactive and intuitive environment for interacting with large amounts of data. The interface provides a graphical representation of a 96-well plate, allowing the researcher to use a similar workflow for data entry as what is used for the tissue culture experiment. Figure 2 shows the plate builder-interface in "configuration mode" allowing entry of four key parameters: drug ID, drug concentration, disease strain (in this case the malaria strain being investigated), and the presence or absence of key indicators of drug activity (in this case red blood cell count) for a particular experiment. The "wells" and column and row headers are selectable for quick and intuitive data entry. The figure shows the drug IDs with a unique color assigned to each sample for easy recognition and validation of the plate configuration. Furthermore the drug concentration in a plate is displayed with a color gradient. This allows researchers to easily identify complex relationships such as how the variation in a drug concentration influences its therapeutic behavior. To further facilitate visualizing structure-property relationships, the chemical structure present in each well can be displayed by clicking on the well. Section 6, elaborates on such capabilities.

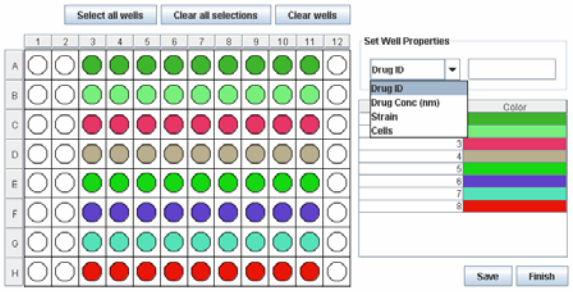

Figure 2. HTS assay visualization capabilities of FreeFlowDB

4. Data Modeling and Storage

Data modeling for pharmaceutical data needs not only to address the challenge of supporting semantics across multiple data types but also support both a data-centric as well as a process-centric view of the information.

Within FreeFlowDB, most of the data is stored inside the database tables. This ensures proper indexing of the entities. One of the exceptions to this rule is the storage of raw data from screening instrumentation, which are archived in the file system. The key sections of the database schemata are shown in Figure 3 and depict the main entities. These entities include: the *Plate entity* which is stored in the relational database with the following properties: a unique plate id, a unique plate name, plate type, number of wells in the plate (96 or 384 wells), plate description and key statistical information about the data in the plate. For instance, such information may include cutoff values used in determining IC_{50} or the percentage of data that needs to be ignored to account for noise. Another key entity, the *Well entity* captures well-level information such as a unique well id, well row and column information, pointer to the binary data file archived in the file system, and a set of assay specific information such as cell culture date, strain present in the well, drug ID, and drug concentration. In FreeFlowDB, structural information about chemical compounds is stored as Mol-files that describe atoms, atomic coordinates, and bond patterns that constitute a molecule. These files are linked to the DrugID and the information contained in them can be used to execute various types of structure search algorithms as described in Section 5.

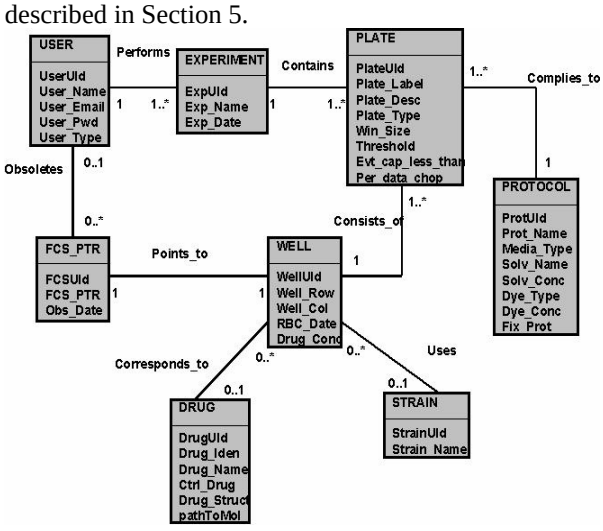

Figure 3. Key components of the FreeFlowDB database schemata

Information on the drug molecule and the target (disease strain) being investigated are treated as static properties for each wells and persisted in the database. Other well properties are typically treated as dynamic

characteristics that may change at different stages of the drug-discovery process (a process centric view). These are persisted using an XML database (Berkeley DB is used by us). The data types of these properties are not fixed for all experiments. After processing the plate the processing parameters, or statistical data, for the plate are stored in the Plate entity, while the results are stored as read-only well properties in the XML database.

5. Structure-Based Query-Retrieval

Molecular structures constitute a key component in drug discovery databases. Hence, it is essential for such databases to support powerful structure-based query-retrieval capabilities. A variety of structure descriptors representing varying trade-offs between representational complexity and modeling fidelity are possible [5]. However, for reasons of query efficiency, 2D or 3D graph based molecular descriptors are most commonly used. In such settings, the problem of structure matching reduced to finding labeled graphs which are topologically similar up to a specific class of transformations. In the following, we formally define the graph-based structure matching problem and identify three specific cases of this problem.

Let *Q* be a query molecule, modeled as the graph *G(Q)* with vertices *V(Q)* and edges *E(Q)*. Let *B* be another molecule modeled as the graph *G(B)* with vertices *V(B)* and edges *E(B)*. If the number of vertices is same between the two graphs ($|V(Q)| = |V(B)|$), the problem becomes that of finding a one-to-one mapping, $f: V(Q) \rightarrow V(B)$ such that *(v1, v2)* belongs to *E(Q)* iff *(f(v1), f(v2))* belongs to *E(B)*. This specific problem may be termed exact graph matching and can match molecules that are arbitrarily rotated or translated with respect to each other. If $|V(Q)| < |V(B)|$, the problem is that of sub-graph matching. Otherwise, finding a one-to-one map between *Q* and *B* addresses the problem of inexact matching. Unfortunately, the problem of graph matching has a combinatorial nature. Therefore, most approaches constrain the search formulation. For instance, the commercial system ISIS [8] uses a screening technique with 933 predefined sub-elements as structure keys. The presence or absence of these sub-elements in a molecule is denoted in a vector and molecular comparisons are carried out by comparing the corresponding vectors. This however compromises the efficacy of structure search.

To address this challenge, we have developed a graph-based molecular matching algorithm that works in two steps. Given a query and database molecule, in the first step, a rough correspondence is calculated. Then based on this correspondence as the starting point, a branch-and-bound strategy is utilized to find the final atom/bond correspondences to score a match. We compute the initial correspondence by using the *graduate assignment algorithm* (GAA) [3]. This step has a complexity of *O(np)*, in which n and p are the vertex size of each of the two graphs, respectively. Briefly, the algorithm works as follows: A correspondence matrix *M* between the molecules is found by optimizing a nonlinear energy function, that explicitly depends on *M* and implicitly dependent on the adjacency matrices of the two graphs. It is assumed that *M* is a doubly stochastic matrix with values either *0* or *1*. In general, a doubly stochastic matrix is a matrix in which both the rows sum to 1.The energy function used for optimization is defined as:

$$E(M) = \frac{-1}{2} \sum_{a=1}^{V(Q)} \sum_{i=1}^{V(B)} \sum_{b=1}^{V(Q)} \sum_{j=1}^{V(B)} M_{ai} M_{bj} C_{aibj} \qquad (1)$$

with $C_{aibj} = 0$ if $A(Q)_{ab}$ or $A(B)_{ij}$ is null, and $1 - 3|A(Q)_{ab} - A(B)_{ij}|$, for *A(Q)* and *A(B)*, the adjacency matrices for *G(Q)* and *G(B)*, respectively. The process starts with an initial value for *M* (called M^0), and the energy function *E(M)* is expanded in a Taylor series about M^0 and a partial derivative with respect to *M* is taken, evaluated at M^0, with the intention of minimizing *E(M)*. Our investigations showed that the GAA by itself, gives good results for exact graph matching, but does not work well for sub-graph matching. However, when the GAA is complemented by a branch-and-bound algorithm (see below), excellent results for sub-graph matching, as well as inexact graph matching, are obtained.

The branch-and-bound algorithm (BBA) uses a breadth-first traversal and then the maximum value in each column is chosen. Given a maximum value associated to vertex v on the sub-graph, the GAA match matrix dictates that vertex v maps to v' on the second graph. The question becomes, to what other vertices is v connected? Once these set of vertices on the first graph is established, it is necessary to confirm that v' is connected to the images in the neighborhood of the first graph. If the image of v is not connected to the other images, it is disregarded (reddened) and the next lesser maximum is chosen. Once the vertex has been found, it is blackened, and the next column is inspected for the maximum and the vertices are un-reddened. This process is reiterated until all the vertices of the sub-graph graph have been matched. For the case of inexact graph matching, the algorithm proceeds as previously described except that only a certain number of vertices are blackened. Once the correspondence is established, a distance measure between the two graphs is computed to assess the

degree of similarity between the two molecules. The Tanimoto measure is used here because it is known to give the best overall results [1]. The Tanimoto measure is defined as:

$$S_{Q,B} = \frac{\sum_{j=1}^{n} x_{jQ} x_{jB}}{\sum_{j=1}^{n} (x_Q)^2 + \sum_{j=1}^{n} (x_{jB})^2 - \sum_{j=1}^{n} x_{jQ} x_{jB}} \qquad (2)$$

Where Q and B are the two molecules, and x_{jQ}, x_{jB} are the corresponding attributes for molecule A and molecule B, respectively.

6. Case Studies and Experimental Analysis

We present two sets of results to provide readers with an impression of the working of the system as well as its efficacy both for interacting with assay data as well as with structural information.

For interacting with assay (drug activity) data, two modes are possible and supported. These include:

- A traditional tabular view of the numeric results that include well-level information that a user may select to view. For example, in an anti-malarial drug discovery experiment, the user may select to view the percentage of censored data for each well, the average threshold activity value, and a numerical exponent describing the influence of the drug molecule.
- FreeFlowDB also supports interactive visualizations of the above information that are much more capable of helping users browse through large volumes of data and detect patterns that may be hidden.

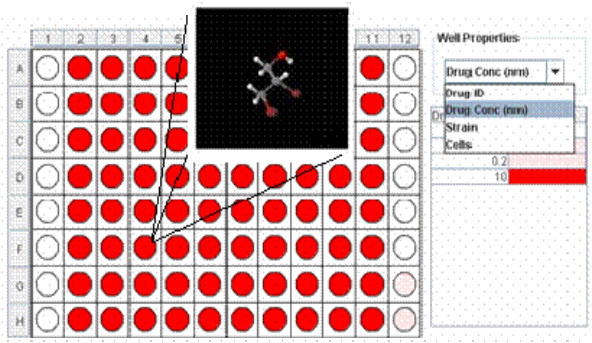

Figure 4. Visualization of drug concentration and structure-activity relationships

An example of the latter capability is presented in Figure 4, which shows results from an anti-malarial assay run on a 96-well plate. As shown in Figure. 4, events or cell counts are visualized using a color gradient. The correlation between the number of living cells present in a well and the position of the well on the plate can thus be quickly detected. In such visualization, the efficacy of a particular drug compound can be rapidly detected through the color gradient corresponding to the degree of parasitized red blood cells. Visualizing data in this manner provides a quick way to screen samples or identify compounds fir which more extensive follow-up is required. This example also illustrates how users can look-up drug molecule that induces then specific activity by clicking on a well. Using this feature, for example, users can click on wells that show activity of interest and compare the structural motifs leading to the activity and examine possible structural similarities amongst such molecules. These molecules can also be then used as a query to systematically search the database from a structural perspective and retrieve all relevant biological assay information.

Another powerful capability of such visualization is depicted in Figure 5, where the lower right-hand corner of the plate indicates poor information registration. In this case, such a result was totally unexpected and on follow-up was found to be due to experimental error. This example indicates the power of such user-data interaction functionality.

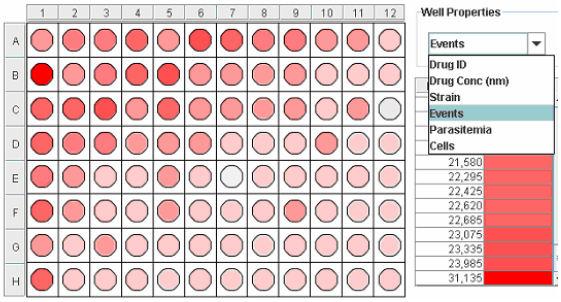

Figure 5. Plate Viewer interface showing correlation between cell growth and well position.

We present results from two experiments to show the efficacy of the structure search algorithm and analyze its efficiency. The first experiment studies the ability of the algorithm to correctly retrieve results, given a set of molecules used as queries. We used the algorithm on a set of 376 molecules randomly selected from the NCI open database. Each of these molecules was considered both for exact and substructure search. For substructure search, the retrieved results were manually verified for precision and recall. In the case of exact search, the proposed algorithm performed with 100% accuracy. For the sub-structure search, every query resulted in retrieval of all molecules that had the relevant sub-structure motif. Furthermore, in no case was a molecule containing the query sub-

structure not retrieved. Essentially, this implies perfect precision and recall. Figure 6 presents partial results from a sub-structure search on this dataset that was initiated using the molecule shown in top left.

In the second experiment, we studied the influence of using the graduate assignment (GAA) algorithm to obtain an approximate correspondence before applying the branch-and-bound (BBA) step, in terms of the complexity of obtaining the final match. It should be noted that the BBA algorithm, can by itself obtain a correspondence, so this experiment underlines the change in performance (defined as the number of iterations of the BBA required to achieve the best correspondence), introduced due to the use of the GAA. For comparison purposes, the initializations were done using a random correspondence and the GAA, before invoking the BBA step. For sub-graph matching, the GAA resulted in a 7-fold reduction in the number of BBA iterations. For exact match, the speedup was 16-fold. However, for inexact matching, a significant difference between the two initialization strategies was not observed. This observation can be explained by the fact that by definition, inexact matching requires examining all possible vertex correspondences.

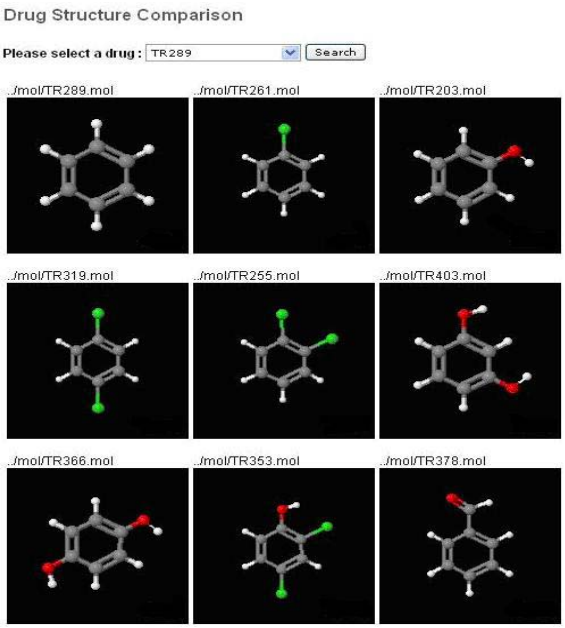

Figure 6. Example showing some of the molecules retrieved from a substructure search initiated using the top left molecule

7. Conclusion

The use of combinatorial chemistry and high-throughput screening technologies has brought about a paradigm change in how the process of modern drug discovery is conducted. This change introduces significant challenges for databases and information management systems that are designed for medicinal and pharmaceutical drug discovery. Among others, such challenges include managing multifarious heterogeneous data, supporting semantics across different data types, supporting data as well as process centric access to information, and facilitating information finding and assimilation through novel user-data interaction paradigms. This paper presents our research in designing FreeFlowDB, a drug discovery information management system that seeks to address these challenges. The system introduces and incorporates techniques spanning information modeling and storage, information retrieval, information visualization, and user interfaces to address these challenges. Case studies and experiments underline the efficacy of the proposed system.

References

[1] Jürgen Bajorath, Ed., Chemoinformatics: Concepts, Methods, and Tools for Drug Discovery, Humana Press (Totowa, NJ: 2004).

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

[3] Steven Gold and Anand Rangarajan, A Graduated assignment algorithm for graph matching, IEEE Transactions on Pattern Analysis and Machine Intelligence, 1996 (Vol. 18, No. 4), pp. 377 - 388.

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

[5] R. Singh, “Reasoning About Molecular Similarity and Properties”, *Proc. IEEE Computational Systems Bioinformatics Conference (CSB), 2004*

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

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

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