Data Storage and Analysis System for Conducting Biotechnological Experiments

Damir Spahić*, Goran Mauša*, Sandra Kraljević Pavelić** and Tihana Galinac Grbac*
* Faculty of Engineering/University of Rijeka, Rijeka, Croatia
** Department of biotechnology / University of Rijeka, Rijeka, Croatia
dspahic@riteh.hr, gmausa@riteh.hr, sandrakp@biotech.uniri.hr and tgalinac@riteh.hr

Abstract – Biomedical research, such as analyses of antiproliferative properties of certain compounds designed to target various human cancer cells, often requires extensive preclinical evaluation to obtain conclusive evidence. The first step in biological evaluation is typically very rigorous in comparing different compounds' concentrations on a number of target cells *in vitro*. Therefore, the experiments yield a set of data that needs to be properly documented, classified, stored and analyzed. This paper presents the Data Storage and Analysis System designed to make that process simpler, faster and less prone to human errors. The importance of such research and its implications lead to high sensitivity to bias. Our system minimizes the chance for human error in the process of data transfer by making the process as automatic as possible and allowing the posterior validation of data by the lead researchers. The analysis most often requires both numerical calculations, such as the *IC_{50}* value, and graphical visualization of compounds' performances. The system offers to researchers a simple interface for simultaneous comparison of results by choosing the appropriate combination of a compound, target cells and date of experiments.

Keywords: data storage, data analysis, visualization

I. Introduction

Antiproliferative evaluation of novel potential or existing anticancer compounds is a standard part of preclinical evaluation of great chemical and pharmacological importance. Those compounds that posses the ability to target specific signalling pathways, *i.e.* in cancer could be used in combined treatment approaches to improve cytotoxicity of drugs and clinical response of patients.

Extensive *in vitro* experiments need to be performed to achieve conclusive first evidence in such studies. For example, a study performed by Gazivoda, Kraljević et al. [1] tested 23 different compounds for antiproliferative activities on 7 cell lines (5 tumour cell lines and 2 human fibroblasts) at various micromolar concentrations. The results showed that five compounds exhibited antiproliferative effects and the one with the strongest concentration-dependent effect was chosen for further biological evaluation. The authors compared its effect with a commercial kinase inhibitor used as standard treatment of Philadelphia positive leukaemia. The new compound presented by the authors, exhibited promising results as it achieved greater antiproliferative effects *in vitro*. Indeed, *in vitro* evaluation of biological data coupled to bioinformatics analyses represents the 'golden standard' in the modern drug development process.

Having in mind a high number of potential compounds' combinations and tested concentrations, as well as requirement to perform the tests on diverse target cells, the amount of data may increase rapidly [2]. The need for a systematic data storage and analysis approach is evident [3] and this paper presents a system designed precisely for that purpose. The system's database is structured to allow multiple users – researchers to upload the data obtained in a standard form by any microplate absorbance reader. The process is made as automatic as possible to avoid human errors, which are typically the dominant reason of data bias. The researchers are required to fill the date related to compounds' names, tested concentrations and names of employed cell lines. The entering of data into the system is a two-step process with the first step based on initial values 'Day zero' and 'Blank' and the second step on final results template load. Since these steps might be slightly prone to human error, the researcher administrator is given the privilege to check and correct the entered data. The system also allows the researcher to automatically calculate the most important parameters like: Values for controls (untreated cells), percentage of growth cells (PG) treated with certain compounds at various concentrations and inhibitory concentration 50 (***IC_{50}***) and to visualize the concentration-response curves for each tested compound and cell lines. In this paper we present *data **S**torage and **A**nalysis for conducting **B**iotechnological **E**xperiments **S**ystem*, SABES. The SABES system was developed by students within the Software Engineering and Information Processing (SEIP) Lab, at the University of Rijeka [4], where we have had the experience of working on other ICT platforms like data collection tool for software defect prediction [5] and ambient assisted living environment [6]. The SABES system accelerates and improves the experimental workflow in early evaluation of potential drugs or biologically active compounds *in vitro*. In particular, it allows comparison of any compound's effect on any cell type *in vitro* by ensuring a simple and systematic storage within the system.

The structure of the paper is as follows: Section II presents the purpose of SABES system and gives a detailed design in five subsections; Section III presents the conclusion and future work intentions.

II. SABES SYSTEM

The usual workflow of *in vitro* experiments is presented in Fig 1. *Groups* of *Experimenters* are formed based on common research interest in research projects and one experimenter may be a part of more than one group. A group is analyzing the growth of a certain number of *Cells* when interacting with certain *Compounds*. In each experiment, the compound effect is analyzed in several levels of *Concentration*. The experimenter may perform a number of experiments using microplate absorbance reader that produces a standardized output table in .xls format for each experiment (*Experiment Table)*. There is one experiment table at the beginning of the experiment and it is called *Day* zero and the experimental end-point table that we refer to as the *Measurement*. Both experiment tables include *Blind Test* values and *Control value* only in the Measurement file. The blind test values refer to absorbance values of test reagents used for assessment of cell growth and the control value stays for untreated cells. These values are important for proper analysis of measurement results. Finally, besides the experimenters, we also distinguish the Role of researchers that perform these analyses. They can view, modify and combine data from several experiments.

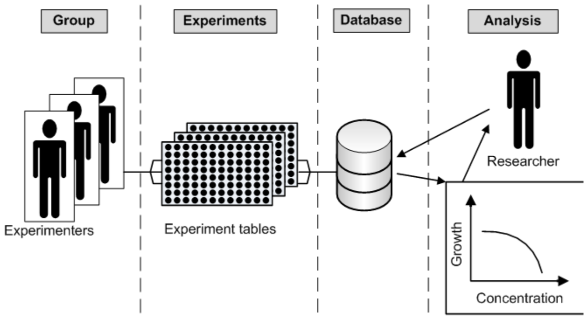

Figure 1. SABES Experiment

Here presented experiment storage and analyses is time-consuming and in particular, major problems are encountered when a need arises for localization of raw data - experimental files upon a certain length of time, or when finding only those files related to effects of the same compound or its behaviour for the same target cells, or even when a need is to additionally compare and group results from various experiments. Furthermore, the experiments may be run by students where unintentional mistakes occur at a higher rate in transfer or analysis of raw data. The previous modality of work included a manual copy-paste activity of raw data into semi-formatted Excel sheets prepared for data analyses. That is why the first task of SABES system was to provide for a controlled import of raw experimental data in a user friendly fashion and less prone to errors. With the SABES system, all the data is stored in a single database, and the user is guided through the data import process, without the need to be an expert in the field. Stored data are easily accessible through the interface. Furthermore, the system allows the supervision of main researchers or mentors and the possibility of additional manual checking of errors. The second task we achieved with the SABES system is comparison of previously performed experiments by the following filters: (1) the person who made the experiments, (2) by tested compound, or by (3) type of cells. The user may choose the data, make a graphical plot, calculate IC_{50} values, and even analyze mean values for a certain experimental point and standard deviations on a graph. In the following section we present the system architecture and technologies employed for its implementation where a detailed explanation of its most important parts is described.

A. Structure

The SABES data storage and analysis system is implemented using standard web development technology tools PHP, MySQL and jQuery. PHP is a server-side scripting language that is used for web development for a long time. Thus, it is supported both in Windows and Linux OS. It contains built-in modules for accessing many of the most popular database servers, making it easy-to-implement. One of the main advantages of PHP is its ability to run the scripts directly on a server, without previous compiling. The system's data storage functionality needs to support large amounts of raw data and that is why we used a relational database designed in MySQL, the most widely used open-source relational database management system. AJAX (Asynchronous JavaScript And XML) is a group of techniques used on the client-side, which allows the development of fast dynamic web applications because it sends and retrieves data asynchronously (in the backend). The jQuery framework is a cross-platform library designed to simplify the client-side scripting. We used it to simplify the implementation of AJAX queries, *i.e.* to achieve the asynchronous communication with server-side, which makes the application resemble to a desktop application from the user perspective [7]. The SABES system architecture is presented in Fig 2. The technologies we used removed the web browser incompatibility so the system can be run in any browser without any design adjustments.

B. Database

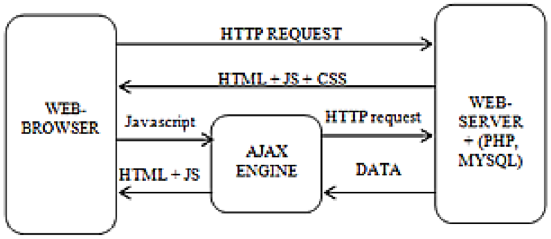

Figure 2. SABES system architecture

The database has two functions in the SABES system, the administrative function of assigning roles to the users and the data storage function. The Entity-Relationship (ER) diagram of the database is presented in Fig 3. Authentication of users is done based on the data from table User. It consists of the username, password and the

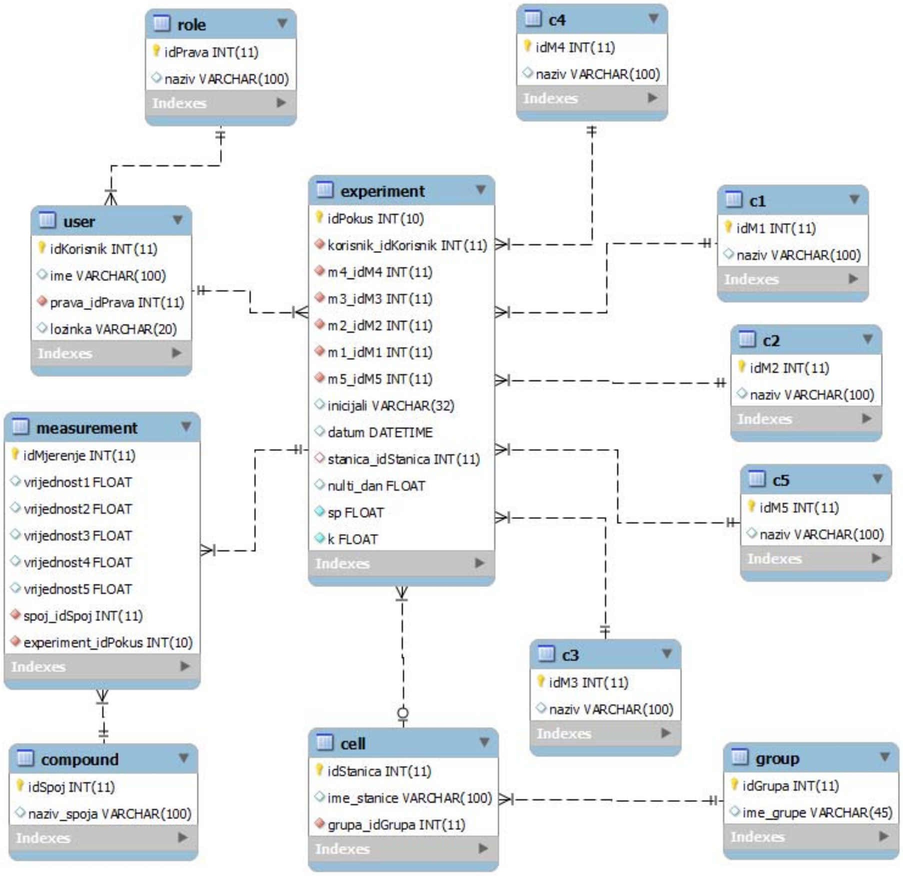

Figure 3. SABES system database

role assigned to each user by the administrators. The Role table divides the users into two groups: experimenter and researcher, as presented in Fig 1.

Experimenters are the users that conduct experiments, import the data into the system and can view and edit only their own data. Researchers have the administrative role and they can import, view and edit any data stored within the system.

The data import is done through the *Experiment table* obtained at the output of the device for cellular growth measurement, so it has the major influence on database structure. General pieces of information are stored within the table named Experiment, the most important table that consists of information about the experimenter, date of experiment, target cell (and indirectly about the group that conducted the experiment through table Group), the day zero, the blind test and the control value for an experiment. The measurement values obtained from the *experiment table* are stored within table Measurement. One *experiment table* can contain measurements for up to 4 compounds on 1 target cell. Each compound, stored within table Compound, has its own row in table Measurement. Each compound can be applied in up to 5 different concentrations. Thus, each compound concentration is stored in adequate tables (c1, c2, c3, c4, c5). Since the compound concentrations usually have recurring values, we avoided duplicated entries by separating them into such, additional tables. The mere form of *experiment table* does not allow the entry of more than 4 measurements and that limitation was not needed in the database design. The integrity of the data obtained for each research project (named *Group* in Fig 1) is achieved with table Group. One group can analyze more than one target cells. If more groups decide to analyze the same target cell, they should name it differently. In that way, no unintentional overlap of research between different groups will occur. Furthermore, the presented ER diagram supports both demands given by researchers-users:

- View the experiment results for a single Cell from a defined Group
- View the experiment results for all Cells tested with one Compound

C. *Data Import*

The user interface contains the data import form, presented in Fig 4. The *experiment table* obtained as output of an experiment is structured in a standardized xls format, which allowed us to create an automatic data import module. It is designed to make the automatic data import process less prone to human error than the previously accustomed manual process. Thus, it guides the user through the process by several steps. It is also designed to be easily understandable and usable by experimenters that are not experts in this field. The only restriction is not to change the default structure of the *experiment table*.

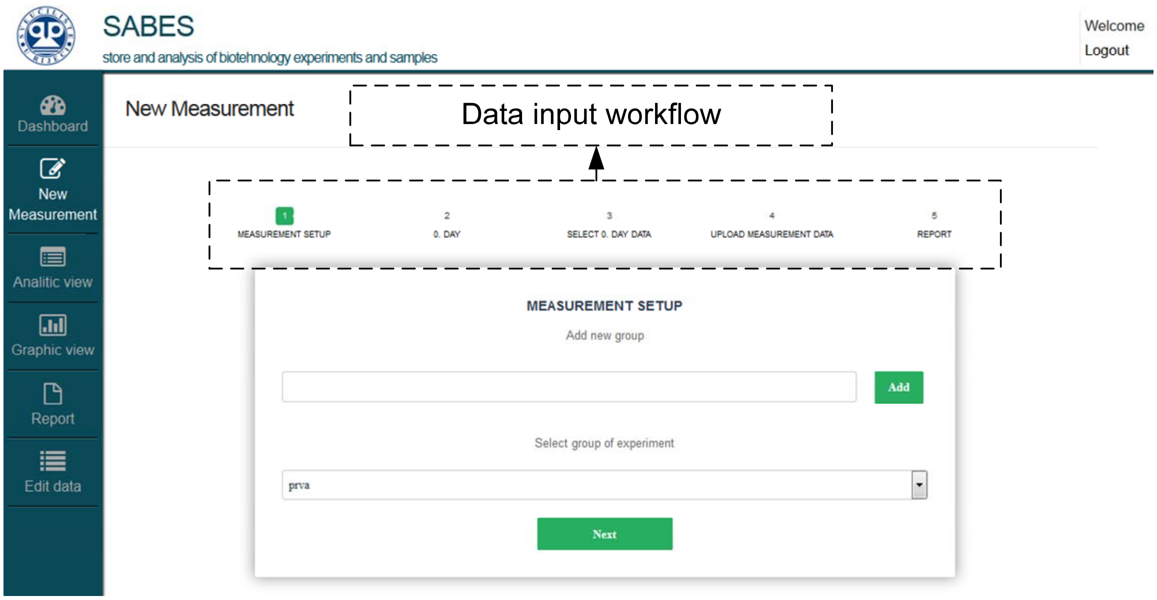

Figure 4. SABES system

The data import process initiates with the selection of the *Group*. The system offers the user to select from all the previously entered Group names. If the Group is a new one, the user is given the option to enter its name for the first time. The second step is to browse the location of the two *experiment tables* in xls format that are stored locally on the computer. The first *experiment table* that the user imports should contain the 0^{th} *day* values. The *experiment table* is then sent to server for processing. The relevant data are then structured and returned to the user for examination. The user then needs to select the column that represents the 0^{th} *day* values for the cell that is currently under examination. The second *experiment table* for import is the one that contains the *measurement* values. All the data are then stored in the database and the user is presented with the data import status. If there was an error in data import process, the user is informed that he should repeat the process. The complete data import sequence diagram is presented in Fig 5. Its steps are:

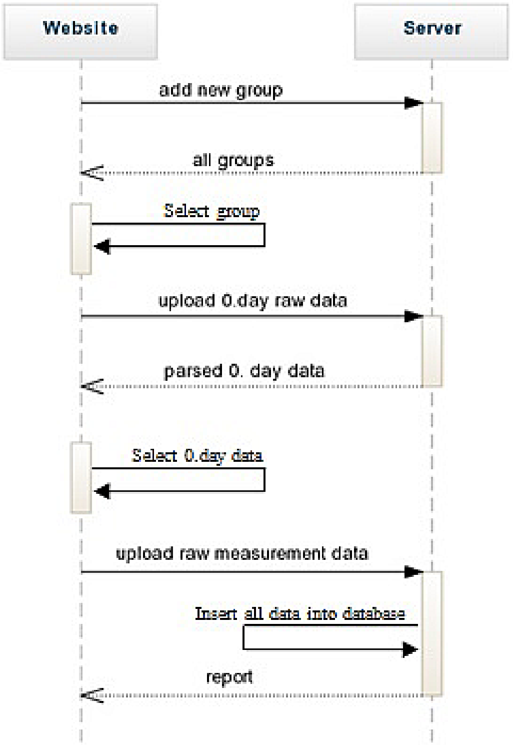

Figure 5. Data import sequence diagram

1. Selecting an existing *Group* or entering the name of a new one
2. Browse the *experiment table* with 0^{th} *day* values
3. Select the column of the *cell* under examination
4. Browse the *experiment table* with *measurement* values
5. Report the user of the data import success

D. Data Analysis

The SABES system automatically performs data analysis of imported data. Measures that are important for data analysis are maximum, minumum, mean and standard deviation values for each *Compound-Cell-Concentration* (*CCC*) combination and the IC_{50} value. The *CCC* combination represents a set of values that analyze the impact of the same *Concentration* of a *Compound* on a certain *Cell*. One *experiment table* contains four values for one *CCC* combination and their mean value and standard deviation for further analysis and visualisation must be calculated. The IC_{50} value represents the most important calculation within the experiment. It represents the *half-maximal inhibitory concentration*, *i.e.* the level of *Compound Concentration* that achieves 50% inhibition of the maximum cellular growth for the analyzed *Cell*.

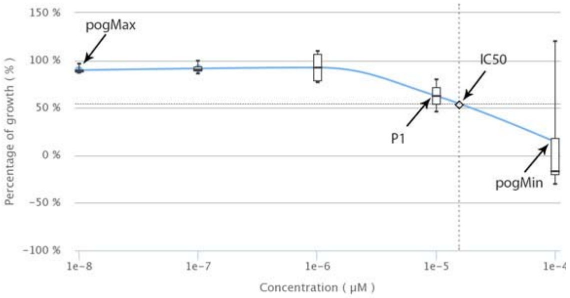

Figure 6. Određivanje intervala za izračun IC_{50}

The estimation of the IC_{50} value is done with linear approximation between two neighbouring levels of concentration that inhibit the cellular growth above and beyond 50%, as presented in Fig 6. The maximum percentage of cellular growth for one *Compound-Cell* combination is presented with the *pogMax* value and the mimum percentage is presented with the *pogMin* value. The 50% cellular growth (pog50) is then obtained by formula (1):

$$pog50 = pogMax - 0.5 \cdot (pogMax - pogMin) \qquad (1)$$

The pog50 value is then used to find the two neighbouring values of concentrations that are used in the experiment and stored in the database using the general expression (2):

$$\dots pog_{n-1} < pog_n < pog_{50} < pog_{n+1} < pog_{n+2} \dots \qquad (2)$$

The values of pog_n and pog_{n+1} are the two neighbouring values we are looking for. For the example given in Fig 6, the pogMin value is the pog_{n+1} value. The IC_{50} value is then calculated using a standard linear equation between two points given by Equation (3):

$$y = y_a + (x - x_a)\frac{y_b - y_a}{x_b - x_a} \qquad (3)$$

where y_a is the pog_n value, y_b is the pog_{n+1} value, x_a is the *Compound concentration* for pog_n value and x_b is the *Compound concentration* for pog_{n+1} value. IC_{50} value is equal to x, when pog50 is inserted as y in (3).

E. Data Selection and Visualisation

The SABES system also offers the graphic view after analyzing the imported data. There are two possible ways to select the data for analysis:

6. Filter the data by *Group* and by *Cell*
7. Filter the data by *Compound*

The graphical user interface of the graphic view window is presented in Fig 7. Filtering the data by *Group* and by *Cell* can be done using the filter elements above the cellular growth graph and filtering the data by *Compound* can be done using the filter elements of the left side of the graph. Sequence diagram of filtering the data by *Group* and by *Cell* is presented in Fig 8. The *Group* filter offers a drop-down menu that is automatically filled with all the *Group* names within the database. The server then returns only the names of *Cells* that were analyzed by the chosen *Group*. After selecting the name of *Cell*, the server gives a list of all the experiments that were conducted for selected *Group* and *Cell*. The user then selects the experiment he wants to analyze and the graph is plotted in the cellular growth graph area, as presented in Fig 6. Each graph generated with the SABES system exhibits the effect of four different *Compounds* used in several levels of concentration with a target *Cell*. The x-axis of the graph represents the levels of concentration and the y-axis of the graph represents the percentage of cellular growth. Filtering the data by *Compound* is simpler and hence we omit its sequence diagram. The data filtering starts with the selection of the *Compound* and then the server returns the names of all *Cells* that were analyzed with it in *Cell* filter. The number of *Cells* that will be plotted on the same graph is arbitrary and left to user's decision.

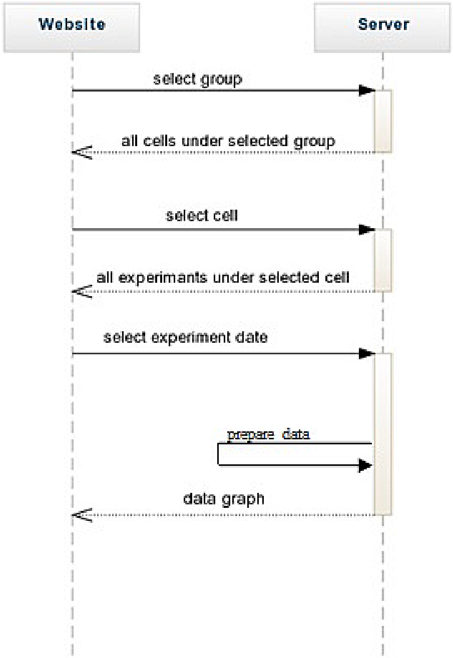

Figure 7. Data selection sequence diagram - filter by *Group* and *Cell*

There are two curve plotting styles implemented in the SABES system. The first one is based on plotting a continuous "*spline*" curve that is calculated based on mean values for a certain *CCC* combination. This is the standard presentation style of presented *in vitro* experiments. The other style is based on drawing of a box & whisker plot. The box represents the inter-quartile range, the whiskers represent the complete range of data, and the middle line presents the median. As we mentioned earlier, each experiment contains at least four measurements for the same *CCC* combination and these values are the basis for the box & whisker plot. If the data are filtered by *Compound*, then we can have a large number of experiments that have the same *CCC* combinations and the box & whisker plot is done for all of them. This style is optional and it can be omitted from the graph.

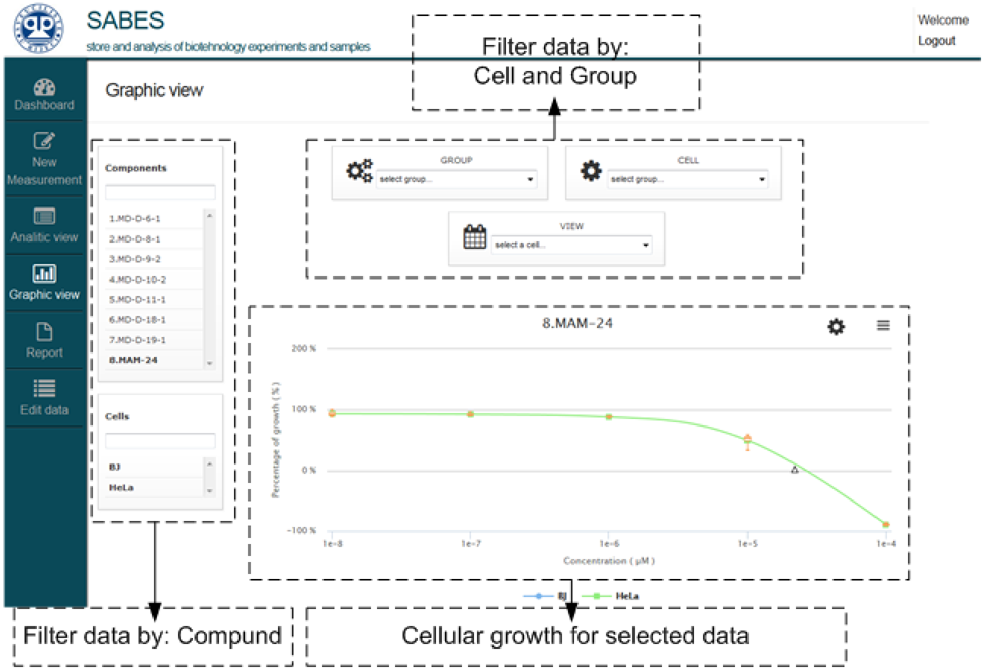

Figure 8. Graphic view for filtering and plotting the experiment data

III. Conlusion and Future Work

One of the biggest obstacles in advancements of any scientific field is limited data storage and analysis capacities. Although, there are vast technologies ready for data generation that are very well supported with powerful and relatively cheap processing equipment for data analysis, there is still a large area of intervention in smaller, specialized domains, *i.e.* biological experimental research where automatization and optimisation of processes might be designed in line with customers/users' needs. Importantly, scientists should understand deeply the phenomena under investigation on large number of data samples to generate generalising theories. And here the key obstacles remain data analysis, data presentation and systematization. This gap might be filled through interdisciplinary collaboration among researches from natural science and computer science. Their interaction indeed, may be of vital importance for future evolution in both disciplines. This paper presents an endeavour in that direction. Here we present a simple IT based solution to error prone process of manual data collection, data aggregation and data analysis. We focused on issues how to simplify the data storage, data aggregation and data analysis process and remove as much as possible human bias from datasets in analysis by taking care of an easy to use approach through the IT system called SABES system that is jointly developed by students and researchers from the University of Rijeka.

One of the major concerns of data management systems today is in their ability to deal with large data sets. Structured data collection process simplifies data accumulation. Data aggregation and analysis rules and policies that may be developed per each individual research experiment may be fault prone and computational intensive task. Future work should consider development of new technologies and programming abstractions that would be able to cope with these issues. We started to create a data management platform for scientific research purposes and future work will be on future evolution of SABES system by integrating more data structures, developing an environment for collaborative and interdisciplinary research as well as extending its abilities to support big data.

Acknowledgment

The work presented in this paper is supported by the University of Rijeka, Croatia research grants (numbers 13.09.2.2.16 and 511-21).

References

[1] T. Gazivoda Kraljević, N. Ilić, V. Stepanić, L. Sappe, J. Petranović, S. Kraljević Pavelić, and S. Raić-Malić, “Synthesis and in vitro antiproliferative evaluation of novel N-alkylated 6-isobutyl- and propyl pyrimidine derivatives”, Bioorganic & Medicinal Chemistry Letters, Vol. 24, No. 13, 2014, pp. 2913-2917

[2] E. G. Stephan, K. R. Klicker, M. Singhal and H. J. Sofia, "Problem Solving Environment Approach to Integrating Diverse Biological Data Sources", *in* 'CSB Workshops', IEEE Computer Society, 2005, pp. 47-50.

[3] C. S. Ang, "Data Management and Analysis Architecture for a More Efficient and Productive Bioinformatics Environment", *Proceedings of HICSS '05.*, 2005 pp.279 - 279.

[4] SEIP Lab: http://www.seiplab.riteh.uniri.hr/

[5] G. Mauša, T. Galinac Grbac and B. Dalbelo Bašić, "Software defect prediction with bug-code analyzer - a data collection tool demo", In: Proceedings of SoftCOM '14, Split, Croatia, 2014

[6] A. Grguric, G. Brestovac, D. Marin, T. Oroz, A. Vidović, S. Bozóki, M. Mošmondor, T. Galinac Grbac, "Ambient orchestration in assisted environment", Engineering Review, May 30, 2014

[7] J. Li and C. Peng, "jQuery-based Ajax general interactive architecture," In Proceedings of Software Engineering and Service Science (ICSESS), 2012, pp.304 - 306