Design and development of a comprehensive data management platform for cytomics: cytomicsDB

Larios Vargas, E.

Chapter 6

Conclusions and Future work

This thesis addresses our research in the design and development of CytomicsDB, a comprehensive data management system for cytomics. This chapter presents the conclusions of the research. First, it addresses the research questions and the problem statements identified in this work. Second, the contibutions of this thesis are highlighted and finally the conclusions are presented, including an outline of future work.

6.1 Research questions and Problem statement

In this section, the evidence collected throughout this dissertation is summarized and used to address the research questions developed in this thesis. The Problem statement (PS) and the four Research questions (RQ) developed in this work are listed as follows:

- **PS**: *How to optimally/flexible organize the data managed for cytomics so as to be able to deal with the data deluge?*
- **RQ1**: *Which components and processes are required to build in a data management platform for cytomics?*
- **RQ2**: *How can be addressed the needs of metadata organization handled in Cytomics?*
- **RQ3**: *How can we prove the consistency of the metadata managed in Cytomics?*
- **RQ4**: *How can we speed up the data processing in Cytomics?*

6.1.1 Designing a software architecture for cytomics

- RQ1: Which components and processes are required to build in a data management platform for cytomics?

In Chapter 2 it was highlighted that given the large amount of different conditions and the readout of the conditions through images, it is clear that the HTS approach requires a proper data management system to reduce the time needed for experiments and the chance of man-made errors. Additionally, it was emphasized that due to the fact that there is different types of data, the experimental conditions need to be linked to the images produced by the HTS experiments with their metadata and the results of further analysis. Moreover, HTS experiments never stand by themselves, as more experiments are lined up, the amount of data and computations needed to analyze these increases rapidly. To that end it was necessary first to design an automated workflow for HTS experiments. Subsequently, it was proposed a software architecture which supports the demands for data management in typical HTS workflows. Finally, it was presented the database design capable to store experiments metadata and image and data analysis results.

6.1.2 Adressing the metadata organization

- RQ2: How can be addressed the needs of metadata organization handled in Cytomics?

In Cytomics, the study of cellular systems at the single cell level, High-Throughput Screening (HTS) techniques have been developed to implement the testing of hundreds to thousands of conditions applied to several or up to millions of cells in a single experiment. Recent technological developments of imaging systems and robotics have lead to an exponential increase in data volumes generated in HTS experiments. This is pushing forward the need for a semantically oriented bioinformatics approach capable of storing large volume of linked metadata, handling a diversity of data formats, and querying data in order to extract meaning from the experiments performed.

For addressing a solution, in Chapter 3 it was necessary to analyze and categorize the metadata managed in HTS experiments. The metadata was organized in five levels according to their role in the automated HTS workflow presented in Chapter 2. These levels are: (1) *Project*, (2) *Experiment*, (3) *Plate-wells*, (4) *Raw images*, and (5) *Measurements*. Finally, a case study is introduced which describes how the metadata is managed in CytomicsDB platform. The proper organization of the metadata facilitates scientist to work with a common data model thus improves the interoperability and the understandability of the metadata managed by the system.

6.1.3 Metadata validation

- RQ3: How can we prove the consistency of the metadata managed in Cytomics?

High-Throughput Screening (HTS) techniques are typically used to identify potential drug candidates. These type of experiments require invest in large amount of resources. The appropiate data management of HTS experiments has become a

key challenge in order to succeed in the target validation. Current developments in imaging systems has to cope with computational requirements due to the significant increment of volumes of data. However, no special care has been taken to ensure the consistency, integrity and reliability of the data managed in HTS experiments. The appropiate validation of the data used in an HTS experiment has turned to be a key success factor in the target validation, thus a mandatory process to be included in the HTS workflow.

For tackling with these issues, in Chapter 4 firstly, it was analyzed which are the validation needs in the automated HTS workflow, then it was described the strategies available for metadata validation and how is the validation workflow developed by CytomicsDB, the workflow is explained using two case studies based on the *Compounds validation* and *siRNA validation*. Furthermore, it was described the implementation of the validation strategies in the system architecture. Finally, it was analyzed and evaluated the validation strategies in order to select the most appropiate ones according to the type of metadata managed by the platform. The strategy "Trust your friends" implemented with the "Most similar" function was chosen for the metadata validation stage.

6.1.4 Speeding up data processing

- RQ4: How can we speed up the data processing in Cytomics?

According to the automated workflow for HTS experiments presented in Chapter 2, the last two stages corresponds to the image and data analysis. These stages require high performance computing resources. Moreover due to the large volume of data involved in the analysis it is mandatory to take into account the location of the data for processing and how to adapt the algorithms so as can be used in a cluster environment.

It was demostrated in Chapter 5 that based on the two case studies for segmentation and object tracking algorithms, the performance of the parallelized version of both algorithms has been improved using a cluster environment. The processing duration has been reduced from aprox. 1 day to the range between 2 and 4 hours. Finally, it has been highlighted that in the data analysis stage is relevant to consider where to process the data. Thus, it has been also described how the MonetDB.R package has been used in CytomicsDB in order to avoid the overhead of shuffling data between different systems.

6.1.5 Adressing the Problem Statement

- PS: How to optimally/flexible organize the data managed for cytomics so as to be able to deal with the data deluge?

With reference to the answers of the four questions I may conclude that it is needed to design and develop a comprehensive data management platform for *Cytomics*. A

platform capable to adapt to the continuos changes in Cytomics environment. Our platform called CytomicsDB relies on an architecture which is based on components which facilitate the integration with other systems in the cytomics domain, the experiments metadata has been properly organized so scientist can have a common data model which promote data collaboration and sharing, and special care has been taken for ensuring the consistency of the metadata managed by the platform. Finally, data processing in cytomics is becoming a challenge due to the large volume of information generated, thus CytomicsDB architecture has been designed to easily integrate the advantages of clustering computing for the image analysis and to speed up the data analysis avoiding the movement of huge amounts of data for processing.

6.2 Contributions

To address my problem statement it was necessary to collect data from the Toxicology Group at Leiden University. The data was obtained by analyzing the protocols, workflows, tools, software and harware environment where High-Throughput Screening experiments are developed. Based on the research findings across this work, it is possible to highlight the following main contributions of this study:

1. The design of an automated workflow system for HTS experiments.
2. The organization of experiments metadata for providing a common data model for data discovery, understandability, interoperability and data sharing.
3. Ensure metadata consistency by including a strict validation procedure in the workflow system of HTS experiments.
4. An integrated platform to automate the data management, capable to speed up the image and data analysis stage in the HTS experiments workflow by integrating a computational cluster to CytomicsDB architecture and including MonetDB.R package for managing the execution of the data analysis requests.

6.3 Future work

During the development of this thesis it was possible to identify other directions for further research. These other directions are summarized as follows:

1. **Reduction of the volume of data generated in HTS experiments**. One of the main challenges for HTS workflow systems is the management of huge amounts of data, each stage of the HTS workflow generate continuously more data. However, during the image analysis this amount of data is increased even more due to the generation of auxiliary image files e.g. binary masks and trajectories. These

image files are relevant for further analysis of the behavior of the cells exposed to different types of treatments during the experiment. Avoiding the use of binary data as an output of the image analysis reduces significantly the resources needed for its storage and further consultation.

2. **Data persistence** Other criteria to take into account is related to the data persistence. Relational Databases have been the dominant technology for data persistence for many years. Most of the databases in life sciences are based on a relational approach, even with the proliferation of other technologies such as NoSQL databases (Pok11). There is the possibility that relational databases may become an issue with the constantly demanding requirements due to the growing data size and computational resources needed for its processing.

3. **Efficient storage of data** In the universe of big data there is a tendency to use specialized distributed file systems e.g. Hadoop (Whi09) instead of typical Relational Database Management Systems (RDBMS). Unlike RDBMS it is not necessary to pre design an schema or prepare a sophisticated structure for the data before using it. The balance between scalability, standards for querying, availability, interoperability should be analyzed for structured and unstructured data. This will open the possibility to design hybrid solutions or to evaluate the impact and costs of turning data from unmanaged to managed, from disconnected to connected, from invisible to findable and from single use to reusable.

In Cytomics environment, scientist has to continuosly deal with a large volume of structured and unstructured data, this condition in particular, makes a challenge the interoperability for any platform developed for cytomics. *CytomicsDB* approach is an effort for developing a framework which takes care of the standardization of the unstructured data, providing a common data model layer for HTS experiments. This model as well is suitable for the integration with other systems in Cytomics, in special other repositories, which allow the validation of key metadata used in the experiments, thus ensure reliability of the data stored. Other possible solutions for cytomics data management, should take special care in the use of data model standards for enhancing the collaboration and data sharing in the scientific community.