# An Optimized Research Process for Real-time Drug Response Analysis

Matthieu-P. Schapranow*, Konrad Klinghammer§, Cindy Fähnrich*, and Hasso Plattner*

*Hasso Plattner Institute
Enterprise Platform and Integration Concepts
August–Bebel–Str. 88
14482 Potsdam, Germany
{schapranow|cindy.faehnrich|plattner}@hpi.de

§Charité – Universitätsmedizin Berlin
Comprehensive Cancer Center
Charitéplatz 1
10117 Berlin, Germany
konrad.klinghammer@charite.de

***Abstract*—Latest medical diagnostics, such as genome sequencing, generate increasing amounts of "big medical data". Healthcare providers and medical experts are facing challenges outside of their original field of expertise, such as data processing, data analysis, or data interpretation. Specific software tools optimized for the use by the target audience, as well as systematic processes for data processing and analysis in clinical and research environments are still missing. Our work focuses on the integration of data acquired from latest next-generation sequencing technology, its systematic processing, and instant analysis for researchers and clinicians in the course of precision medicine. We share our research results on developing specific software tools for drug response analysis built on top of our distributed in-memory computing platform for genome data processing. For that, we present our technical foundations, as well as process aspects of integrating and combining heterogeneous data sources, such as genome, patient, and experiment data in the clinical routine.**

***Keywords-Drug Response Analysis; Genome Data Analysis; Process Integration; In-Memory Database Technology; Precision Medicine; Next-Generation Sequencing.***

## I. Introduction

The Human Genome (HG) project that was officially launched in 1990 involved thousands of research institutes worldwide and required more than a decade to sequence and decode the full HG [1]. Nowadays, Next-Generation Sequencing (NGS) devices enable processing of whole genome data within hours at reduced costs [2]. NGS is used to support precision medicine, which aims at treating patients specifically based on individual dispositions, e.g., genetic or environmental factors [3].

The In-Memory Database (IMDB) technology has proven to have major advances for analyzing big enterprise and medical data, e.g., to support researchers and clinicians in evaluating best therapies for cancer patients [4, 5, 6].

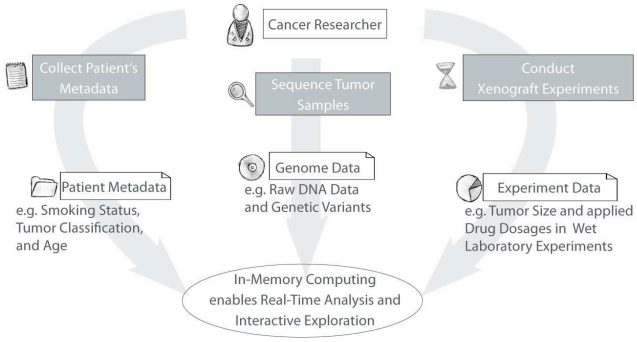


Figure 1. The optimized drug response analysis process involves data in heterogeneous formats from different data sources.

In this work, we present our findings of applying IMDB technology to enable integration of experiment results, real-time analysis, and prediction of drug response in silico in course of precision medicine. We introduce an integrated research process for oncologists built upon our High-performance In-memory Genome (HIG) cloud platform to reduce media breaks and to improve the efficiency of drug response testing [7]. The HIG platform provides services for processing of huge amounts of high-throughput genome data in real-time. In interdisciplinary teams we developed jointly with cancer researchers a special purpose application to evaluate results of conducted Xenograft experiments without significant delay [5, 8]. Figure 1 depicts the optimized research process and the involved data sources.

The rest of the paper is structured as follows: In Section II, our work is set in context of related work

 ISBN: 978-1-61208-359-9 

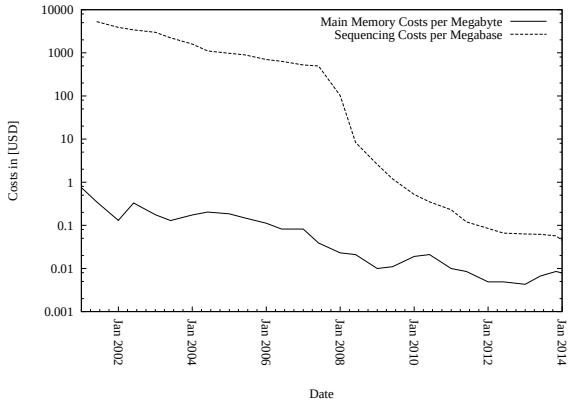


Figure 2. Costs for next-generation sequencing and main memory from 2001 to 2014 adapted from [9, 10].

and in Section III, we introduce our applied research methodology. We present the current drug response process in Section IV and introduce optimized research process in Section V. In Section VI, we discuss our contribution and our work concludes with an outlook in Section VII.

## II. Related Work

Figure 2 provides a comparison of costs for sequencing and main memory modules. Both costs follow a steadily declining trend, which facilitates the increasing use of NGS for whole genome sequencing and IMDB technology for its data analysis. Related work in the field of genome data processing has increased in recent years. However, work focusing on implementing end-to-end processes is still rare. Thus, we focus on implementing innovative processes, e.g., by tight integration of genome data processing and statistical data analysis in course of drug response analysis.

Sun investigated gene regulations in prostate cancer samples combining latest sequencing technology and bioinformatics approaches. We agree that an integrated data processing and analysis approach is also essential for other application fields. Thus, we integrate various heterogeneous data sources to enable multi-modal modeling of diseases. Furthermore, we enable researchers for the first time to perform data analysis a) in real-time without any delay and b) without the need to involve dedicated IT experts, e.g., to prepare analysis reports.

Rossello et al. propose the use of Xenograft models as data sources for preclinical models when primary tumor samples are rare, e.g., for small cell lung cancer. They share very detailed insights into their methodology using state-of-the-art alignment and variant calling tools, such as BWA, GATK, and snpEff [13, 14, 15]. However, they still miss a tight integration of their genome sequencing pipeline and their data analysis pipeline, which consumed a major part of the their experimental time. Our contribution enables tight integration of various experimental data, such as NGS tumor data, and their real-time data analysis as described in Section IV.

## III. Methodology

In the course of this project, we followed the design science methodology to improve the existing research process with the help of software artifacts [16].

For that, we applied the Design Thinking (DT) methodology, which proposes to work in interdisciplinary teams [17]. The idea behind this proposal is that team members from different disciplines, e.g., a software developer and a medical researcher, will have different viewpoints on the same problem domain. Thus, if a team is comprised of members from different disciplines relevant to the problem at hand, chances that an important aspect is forgotten are minimized. Additionally, an interdisciplinary team will not suffer from rivalry between experts of the same field, instead all expertise necessary to implement the solution is already available in the team. Besides suggesting interdisciplinary team compositions, DT provides a process framework as depicted in Figure 3. It asks for constant communication between the developing team and the stakeholders and targeted end users.

Following DT, we conducted user interviews with cancer researchers and physicians to document the existing research process as described in Section IV. Furthermore, we developed an optimized process by integrating heterogeneous data sources and manual process steps within a software prototype. Based on the obtained insights, we iteratively extended software prototypes in short development sprints, evaluated the functionality either in workshops at the users' sites or conducted telephone interviews, while giving end users the chance to use the software artifacts via screen sharing. Based on the input from the workshops and interviews, the next iteration was planned according to the scrum software development methodology [18].

## IV. Current Drug Response Analysis Process

Nowadays, drug response analysis consists of a) conducting drug experiments, e.g., in Xenograft models, and b) the analysis of the obtained experiment results [19]. The following selected data sources are used for drug response analysis as depicted in Figure 1:

- **Patient Metadata** is retrieved from Clinical Information Systems (CISs) and contains specific patient details, such as age, gender, and anamnesis. Its data volume typically ranges from one to 100 MB excluding any diagnostic data, such as imaging data,
- **Genome Data** is obtained by sequencing resected tumor material, e.g., with NGS devices. Its data

 ISBN: 978-1-61208-359-9 

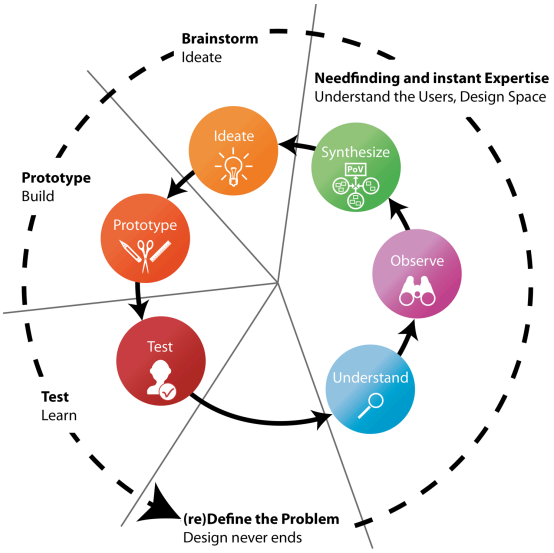


Figure 3. Design thinking process as defined by the HPI School of Design Thinking in Potsdam and Stanford adopted from [17].

volume is in the range of some 100 MB for panel sequencing and up to 500 GB for NGS.

- **Experiment Data** is obtained by wet laboratory assistants, e.g., documenting the individual drug tests in Xenograft experiments. Its data volume is in the range from 10 MB to 1 GB.

The time consumed in wet laboratories can range from days to weeks depending on the conducted experiments. Although the data analysis phase is assisted by use of software, it still takes days up to weeks to perform complex data analysis. The reasons are many-fold, e.g., the absence of specific tools for flexible data analysis, tools limited to a small set of data sources, and transformation of relevant data.

Manual or semi-manual time-consuming process steps, such as the use of Microsoft Excel for complex data analysis, characterize all phases of the existing process. From a software engineering perspective, we focus on all process steps where digital data processing and analysis is conducted. Thus, our work focuses on the data analysis phase of the existing process to optimize the overall research process.

## V. Enabling Real-time Data Analysis

Figure 4 depicts a screenshot of our developed drug response cloud application. It shows details of genetic changes of a specific mama carcinoma tumor sample. Our optimized research process is divided in the following process steps:

- **Computational biology** performs data processing, e.g., raw DNA,
- **Clustering of tumor data** enables real-time classification of results, and

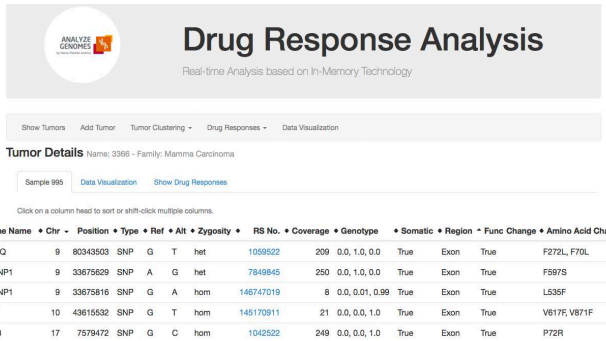


Figure 4. Screenshot of the drug response analysis cloud application of the HIG project.

- **Visual data exploration** supports the interactive testing and verification of research hypotheses.

### A. Computational Biology

In the following, we share insights of our process extensions focusing on processing of raw DNA data.

*1) Open Reading Frame Detection in the In-Memory Database:* The detection of Open Reading Frames (ORFs) builds the foundation of finding potential gene locations within the genetic code [20].

The detection of ORFs is two-fold as follows. In the first phase, we search for start and end codons in all possible reading frames, i.e., three reading frames per strand of the double helix. In the second phase, pairs of corresponding start and end codons within the same reading frame are analyzed to identify ORFs with a minimum length.

Within the first phase, we process the forward and the backward strand in parallel. For example, when searching for the start codon "ATG" on the forward strand, we also search for the reserve-inverted triplet "CAT" to detect the start codon on the backward strand. The reading frame is determined by the position of the first base of the codon modulo three on the forward strand and by adding three for the position on the backward strand. In addition to the reading frame, we store the type and the position of the found codons.

In the second phase, we group the results from the first phase by reading frame and search for a start followed by its corresponding stop codon.

We implemented the ORF detection algorithms directly within IMDB. For that, we used the programming languages SQLScript and L [21, 22]. As a result, our implementation can directly incorporate advances of in-memory computing by performing all data processing directly on top of the genome data stored within our HIG platform without the need for data transformations.

*2) Evaluation of Genetic Variants to Detect Functional Changes in In-Memory Database:* The detection of functional changes is an essential step of genome

 ISBN: 978-1-61208-359-9 

| NAME | ABBREV | POS | BASE | VARIANT |
|---|---|---|---|---|
| Alanine | A | 1 | G | 1 |
| Alanine | A | 2 | C | 1 |
| Alanine | A | 3 | A | 1 |
| Alanine | A | 3 | T | 1 |
| Alanine | A | 3 | C | 1 |
| Alanine | A | 3 | G | 1 |
| Arginine | R | 1 | A | 2 |
| Arginine | R | 1 | C | 1 |
| Arginine | R | 2 | G | 1 |
| Arginine | R | 2 | G | 2 |
| Arginine | R | 3 | A | 1 |
| Arginine | R | 3 | A | 2 |
| Arginine | R | 3 | T | 1 |
| Arginine | R | 3 | C | 1 |
| Arginine | R | 3 | G | 1 |
| Arginine | R | 3 | G | 2 |

Figure 5. The columnar database implementation of the amino acid coding sun [23].

data analysis. Each genetic variant needs to be analyzed according to its potential impact on the Amino Acid (AA) built from the genetic code. As a result, a potential change in the product built from the AA is an indicator for a genetic variant that describes a harmful mutation within the DNA [23].

Our algorithm for processing of the genetic variants is described in the following. Firstly, we check if a concrete variant is located within the range of a known gene. For that, we join the variant's location consisting of chromosome and position with a database table containing a list of known genes [24]. If the current location is outside of a known gene range, we consider its impact as minor, since the current medical knowledge about variants outside the range of genes is very limited. However, once a new gene is added to the list of known genes, it is automatically analyzed.

If the variant is located on a gene, all splicing variants of the gene are analyzed in parallel to derive the individual impact per splicing variant.

Each splicing variant consists of introns and exons [25]. Thus, the reading frame of the splicing variant is determined by the codon of the genetic variant. This allows us to identify the position of the changed nucleotide within the affected codon. The codon of the variant and corresponding codon of the reference genome are translate into AAs using our database table implementing the AA coding sun as depicted in Figure 5 [23].

An AA change is documented by the expected and the detected AA followed including the position of the affected triplet, e.g., S12Y donating an AA change of Serine to Tyrosine at triplet location 12.

If the resulting AAs of the reference genome and the variant differ, we refer to it as functional genetic change.

The reading frame around the variant remains the same for all splicing variants of the gene. Thus, the reading frame for a variant is only calculated once independent of the number of splicing variants of the concrete gene.

Our algorithm can be applied to the forward and the backward strand of the DNA. If a gene is located on the backward strand of the DNA the list of exons needs to be considered in the reverse order as well. However, all other steps of the algorithm can be reused.

In our case, a specific version of the stored procedure within the IMDB is executed to translate codons on the corresponding strand.

### B. Real-time Clustering of Tumor Data

In the following, we share our process extension for automatic classification of processed research data.

*1) Tumor Data Association Rules:* Association Rules Mining (ARM) requires a set $S$ of item sets $S_i$ as its data basis: $S = \{S_1, ..., S_m\}$. Every item set $S_i$ consists of several items $i_i$ from the list $I$ of distinct items: $I = \{i_1, ..., i_n\}$. These item sets are processed to detect reliable rules of type: $A \Rightarrow B$ where $A \subseteq I \wedge B \subseteq I$. $A$ is called prior whereas $B$ is called posterior.

In our use case, items are all distinct variants found in the library of available tumors. Item sets correspond to the set of variants found for one tumor together with respective drug response classes determined by Xenograft experiments. In the context of our current work, we only focus on functional changes. The goal in our use case are rules of type $A \Rightarrow B$, where $A$ is a set of functional changes and $B$ is a specific drug response class. We investigate the impact of single functional changes on drug responses to limit the problem space, i.e., we restrict $|A| = 1$.

Two important measures for association rules are support and confidence. Support $supp(A)$ represents the relative frequency of $A \subseteq I$ in all items sets $S$ as defined by Equation 1.

$$supp(A) = \frac{|\{S_i | S_i \in S \wedge A \subseteq S_i\}|}{|S|} \quad (1)$$

$$conf(A \Rightarrow B) = \frac{supp(A \cup B)}{supp(A)} \quad (2)$$

Confidence $conf(A \Rightarrow B)$ is the relation of the number of item sets where $A$ and $B$ occur to the number of items sets where only $A$ can be found defined by Equation 2.

Support defines how important a found rule is, with respect to all data, while confidence shows how reliable it is. We applied the Apriori algorithm for ARM by using the PAL integrated in the IMDB and by using the implementation provided by the R package `arules` [26, 27, 28].

For tumor classification, either the Tumor/Control (T/C) value or the RECIST value can be used. Therefore, cancer researchers can decide individually per analysis run. In the remainder of this paper, we investigate both measures.

 ISBN: 978-1-61208-359-9 

In order to use Apriori ARM for classification, we need to add drug response classes to each item set. Possible classes are Partial Response (PR), Stable Disease (SD), and Partial Disease (PD). The thresholds, by which a drug response is classified in one of the classes, can be adapted individually.

### C. Visual Data Exploration

We developed specific visualizations to enable researchers to work a) interactively with the data instead of having statically generated charts and b) with commonly used graphical representations.

In the following, we share details about our extensions for interactive exploration of tumor data.

*1) Real-time Classification of Tumor Experiment Data with Support Vector Machines:* Classifying tumor data can be used to identify similarity measures in an unsorted set of data. Thus, an automatic classification of tumor data can be used to generate hypotheses, e.g., to identify new tumors subtypes.

Our implementation builds on Support Vector Machine (SVM) as machine learning algorithm. SVM is available in many popular statistical frameworks, such as R [26]. To leverage the complete performance advance of in-memory computing, we built on the implementation directly integrated within our IMDB as part of the Predictive Analysis Library (PAL). Since the algorithm can directly access experiment data without the need to export/import data. It improves the existing process by eliminating media breaks. Furthermore, the SVM implementation within the IMDB incorporates technology advances and performs faster than the aforementioned implementations, e.g., due to reduced disk I/O.

Our tumor classification algorithm incorporates the steps configuration, preparation, and execution.

*Configuration Stage:* During configuration stage, the researcher is guided through a web page to configure relevant SVM parameters, e.g., drugs to predict and experiment data to be used as training data.

The result of a SVM prediction depends on the criteria selected during the configuration phase, e.g., a concrete Tumor/Control (T/C) or a RECIST value for a specific pharmaceutical based on the selected tumor attributes. SVM uses a regression mode also known as Support Vector Regression (SVR) to estimate correlation between attributes of the train data and applies these correlations to the data points to predict [29].

In addition to prediction of concrete values, we focus on classification of data in response classes as introduced in Section V-B1. For our concrete use-case, we classified values from 0.0 to 0.7 as PR, representing a reduction in tumor growth by administering the drug, 0.7 to 1.2 is as SD, which represents no significant change, and values greater than 1.2 as PD, representing a negative response and thus growth of the tumor.

SVM in classification mode calculates class membership probabilities instead of drug efficiency values [30].

*Preparation Stage:* Input and output tables are created, training data is selected, and the database procedure is prepared to process SVM model during preparation stage.

The input database table for SVM is constructed according to the chosen tumor attributes, with each column representing a specific tumor attribute. A table record in the input table represents all relevant data for a single tumor entity based on the configured attributes.

*Execution Stage:* The train formula in the R procedure is "**drug .**", telling R that the drug column of the data frame is the depending variable, whereas the rest are deciding variables, indicated by a dot. SVM decides whether to perform classification or prediction by investigating the dependable variable in the SVM training formula, i.e., if a numeric value is encountered, SVM will run in regression mode otherwise in classification mode.

The execution of SVM for multiple drugs is done by running preparation and execution stages individually per drug isolated from each run ensuring that no side effects influence the SVM algorithm.

*2) Clustered Heat Map using Hierarchical Clustering:* Hierarchical clustering is a clustering method that builds a hierarchy of clusters from the given data by iteratively merging the closest data points to one cluster (agglomerative hierarchical clustering). In order to identify the clusters that should be combined, the clustering algorithm needs a measure of dissimilarity between sets of observations. Figure 6 shows a clustered heat map based on a hierarchical clustering algorithm. For hierarchical clustering, the measure is formed by combining an appropriate metric for distance calculation between data points and a linkage criterion for calculating the distance between merged data points [31].

We used row- and column-wise vectors, the Euclidean distance function as distance metric, and single link clustering to create the heat map.

The result of any hierarchical clustering is a dendrogram, i.e., a binary tree with data points as leaves. It represents the clustered data points and the nested clusters at certain similarity levels as depicted in Figure 7. The dendrogram is used to rearrange / reorder the heat map and identify positions for adding cluster gaps to the heat map. For example, rearranging rows in the heat map according to the clustering result can be done by traversing the dendrogram tree and leveraging the order of the leaves, which are by our definition the row vectors. Thus, similar rows will be side by side or at least close to each other.

 ISBN: 978-1-61208-359-9 

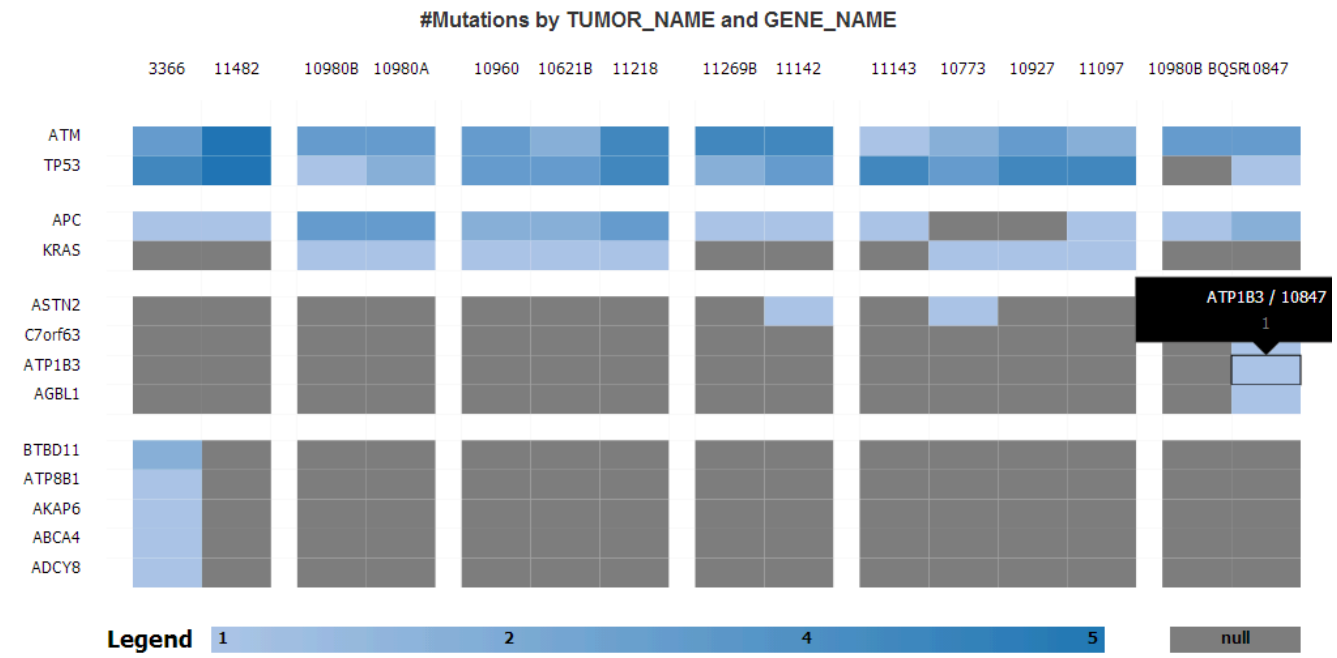


Figure 6. Clustered heat map using hierarchical clustering comparing mutation count, a subset of genes, and tumor samples.

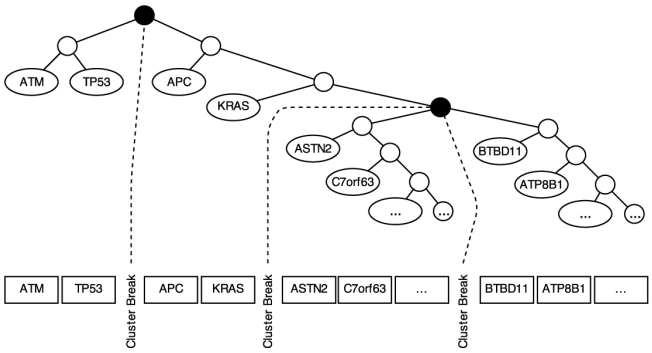


Figure 7. Dendrogram tree as the result of applying hierarchical clustering. Cluster borders are the turning points in the tree that split up in two clusters denoted as black nodes.

## VI. Evaluation and Discussion

With the help of oncologists, we have been able to verify that our introduced research process improves their daily work. We focused primarily on improving aspects of data processing and analysis to create an integrated and reproducible research process. The incorporated IMDB technology provides a data integration platform, which minimizes the need for additional third-party tools. As a result, we were able to reduce media breaks, provide flexible and individual real-time analysis of acquired data, and establish a single source of truth, which holds all acquired data and enables a consistent and iterative research process.

In the following, the advantages and enhancement of our proposed research process are summarized:

- Integration of heterogeneous data sources,
- Elimination of media breaks and improved processing time,
- Automated data processing by integration of algorithms for computational biology,
- Flexible data analysis built on latest in-memory technology, and
- Interactive graphical data exploration enabling direct verification of research hypotheses.

## VII. Conclusion and Outlook

In our contribution, we shared details about our proposed research process for drug response analysis. We incorporated latest in-memory technology as the key enabler for real-time analysis and exploration of experiment data and the integration of various data sources. We outlined the applicability of our HIG platform for processing of genome data and the specific drug response analysis application to optimize existing research processes in this specific field of cancer therapy. Furthermore, we shared detailed insights in our applied research methodology, which involved experts from interdisciplinary teams.

Our future work will focus on applying the research process to additional fields of cancer research in course of precision medicine. Furthermore, we will investigate how a huge library of tumor samples can be used as training data to create more stable prediction models to discover new medical insights.

 ISBN: 978-1-61208-359-9 

## References


[1] F. S. Collins *et al.*, "New Goals for the U.S. Human Genome Project," *Science*, vol. 282, no. 5389, pp. 682–689, 1998.

[2] W. J. Ansorge, "Next-generation DNA Sequencing Techniques," *New Biotechn*, vol. 25, no. 4, pp. 195–203, 2009.

[3] K. Jain, *Textbook of Pers Med*. Springer, 2009.

[4] M.-P. Schapranow, H. Plattner, C. Tosun, and C. Regenbrecht, "Mobile Real-time Analysis of Patient Data for Advanced Decision Support in Personalized Medicine," in *Proceedings of the 5th Int'l Conf on eHealth, Telemedicine, and Social Medicine*, 2013, pp. 129–136.

[5] M.-P. Schapranow, F. Häger, and H. Plattner, "High-Performance In-Memory Genome Project: A Platform for Integrated Real-Time Genome Data Analysis," in *Proceedings of the 2nd International Conference on Global Health Challenges*. IARIA, Nov 2013, pp. 5–10.

[6] M.-P. Schapranow, *Real-time Security Extensions for EPCglobal Networks: Case Study for the Pharmaceutical Industry*. Springer, 2013.

[7] ——, "Analyze Genomes Project Page," http://we.analyzegenomes.com/ [retrieved: Jun, 2014], 2014.

[8] R. S. Kerbel, "Human Tumor Xenografts as Predictive Preclinical Models for Anticancer Drug Activity in Humans: Better than Commonly Perceived But They Can Be Improved," *Cancer Biology & Therapy*, vol. 2, no. 4, pp. 134–139, 2003.

[9] National Human Genome Research Inst, "DNA Sequencing Costs," http://www.genome.gov/sequencingcosts/ [retrieved: Jun, 2014], Apr 2013.

[10] J. C. McCallum, "Memory Prices (1957-2014)," http://www.jcmit.com/memoryprice.htm [retrieved: Jun, 2014], Apr 2014.

[11] Y. Sun, "Identification of REST Regulated Genes in Prostate Cancer via High-throughput Technologies," Master's thesis, Instituto Superior Técnico, Universidade Técnica, Lisboa, Portugal, Dec 2012.

[12] F. J. Rossello *et al.*, "Next-Generation Sequence Analysis of Cancer Xenograft Models," *PLoS ONE*, vol. 8, no. 9, p. e74432, Sep 2013.

[13] H. Li and R. Durbin, "Fast and Accurate Short Read Alignment with Burrows-Wheeler Transformation," *Bioinform*, vol. 25, pp. 1754–1760, 2009.

[14] A. McKenna *et al.*, "The Genome Analysis Toolkit: A MapReduce Framework for Analyzing Next-Generation DNA Sequencing Data," *Genome Research*, vol. 20, no. 9, pp. 1297–1303, 2010.

[15] P. Cingolani *et al.*, "A Program for Annotating and Predicting the Effects of Single Nucleotide Polymorphisms," *Fly*, vol. 6, no. 2, pp. 80–92, 2012.

[16] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design Science in Information Systems Research," *MIS Quarterly*, vol. 28, no. 1, pp. 75–105, 2004.

[17] H. Plattner, C. Meinel, and L. Leifer, Eds., *Design Thinking Research*, ser. Understanding Innovation. Springer, 2012.

[18] R. Pichler, *Agile Product Management With Scrum: Creating Products that Customers Love*, ser. Agile Software Development. Addison-Wesley, 2010.

[19] S. Oesterreich, A. M. Brufsky, and N. E. Davidson, "Using Mice to Treat (Wo)men: Mining Genetic Changes in Patient Xenografts to Attack Breast Cancer," *Cell Rep*, vol. 4, pp. 1061–1062, Sep 2013.

[20] N. Rani, R. Singh, and G. Arora, "Detection of ORF Frames Using Data Mining," *International Journal of Computer Science and Telecommunications*, vol. 2, no. 3, pp. 90–94, Sep 2011.

[21] SAP AG, "SAP HANA SQLScript Reference," help.sap.com/hana/hana_dev_sqlscript_en.pdf [retrieved: Jun, 2014], 2012.

[22] A. Hannan, "The L Programming Language & System," http://home.cc.gatech.edu/tony/uploads/61/Lpaper.htm [retrieved: Jun, 2014], Jan 2005.

[23] C. Bresch and R. Hausmann, *Klassische und molekulare Genetik*, 3rd, Ed. Springer-Verlag, 1972.

[24] F. Hsu *et al.*, "The UCSC Known Genes," *Bioinformatics*, vol. 22, pp. 1036–1046, 2006.

[25] B. Alberts *et al.*, *Molecular Biology of the Cell*, 5th, Ed. Garland Science, 2007.

[26] SAP AG, "SAP HANA Database R Integration Guide Version 1.1," https://help.sap.com/hana/SAP_HANA_R_Integration_Guide_en.pdf [retrieved: Jun, 2014], Mar. 2014.

[27] ——, "SAP HANA Database Predictive Analytics Library (PAL) Reference Version 1.1," https://help.sap.com/hana/SAP_HANA_Predictive_Analysis_Library_PAL_en.pdf [retrieved: Jun, 2014], Mar. 2014.

[28] M. Hahsler, B. Grün, and K. Hornik, "Introduction to arules: Mining Association Rules and Frequent Item Sets," in *Special Interest Group on Knowledge Discovery and Data Mining*, 2007, pp. 1–28.

[29] A. J. Smola and B. Schölkopf, "A tutorial on Support Vector Regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, 2004.

[30] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[31] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data Clustering: A Review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264–323, 1999.


 ISBN: 978-1-61208-359-9 