Computational Toolbox for Discovery of Prognostic Markers in Survival Analysis from Gene Expression Data: Description of the Research Project

We propose a project to design and develop an interactive, visualization-based exploratory analysis toolbox to assist in finding molecular prognostic biomarkers from high-throughput molecular data and survival data obtained in clinical trials. The project will devise computational and machine learning methods to search for biomarkers, encapsulate them within interactive components with graphical user interface, and provide visual programming to stitch these components into data analysis pipelines. The constructed methods and toolbox will support collaborations between the data scientists and domain experts – physicians, biomedical or pharma researchers – to sift through the molecular cell-response data of thousands of genes to find those that correlate most with the survival. The proposed tool will access existing models, ontologies, and knowledge bases to speed-up the interpretation and provide semi-automatic explanations of results.

This is an applied project where we are teaming up with Genialis, a data science company specializing in computational support for precision medicine. Genialis is currently in the process of registering with the FDA (US Food and Drug Administration) a first ever machine learning model that utilizes transcription data to predict cancer patients' response to treatment. To remain at the cutting edge of biomarker research, Genialis needs methods, tools, and visualizations to speed-up the discovery and improve communication of their data analysis results to their customers and regulatory agencies. On the other hand, the project will allow us, the proposing institution, to further advance our research into a combination of interactive visualizations and machine learning, and apply our new approaches to a challenging field of biomarker discovery.

26.1 Scientific background, problem identification and objectives of the proposed research

Scientific Background

The project will contribute new methods and practical data exploration approaches to the field of survival analysis, and study how many covariates (genes with their expression) jointly affect survival. Survival analysis is a set of statistical methods aimed at determining the life expectancy of the investigated population. Survival analysis studies the expected duration of time until an event, say, a cancer relapse after chemotherapy, or a recurrence of disease [**?**]. Survival models, including the most famous one, the proportional hazards model, relate the time until the event to one or more covariates. In biomedicine, the covariates with a significant impact on survival are potential *markers*, a characteristic of a biological system we can objectively measure and use as an indicator of the system's state. For example, in cancer, markers may differentiate between patients that respond to treatment and those where the treatment has no effect.

Based on markers, we can predict the success of treatment and choose the right therapy for individual patient. Identification of good markers is thus essentials for advancements of medicine and treatment of diseases, and crucial to the development of personalized medicine. Because improving survival is a direct benefit to the patient, it is very important to understand how participants respond to various forms of treatment. The treatment should thus be selected based on patient's state and characteristics, which are defined through a set of markers. Markers may be clinical, related to patient's symptoms, or biological, related to some measurements on molecular level, like concentration of specific protein or expression of particular gene or set of genes. Markers may refer to a single measurement of state, or to a group of measurements possibly related through a prognostic model or a network [**?**].

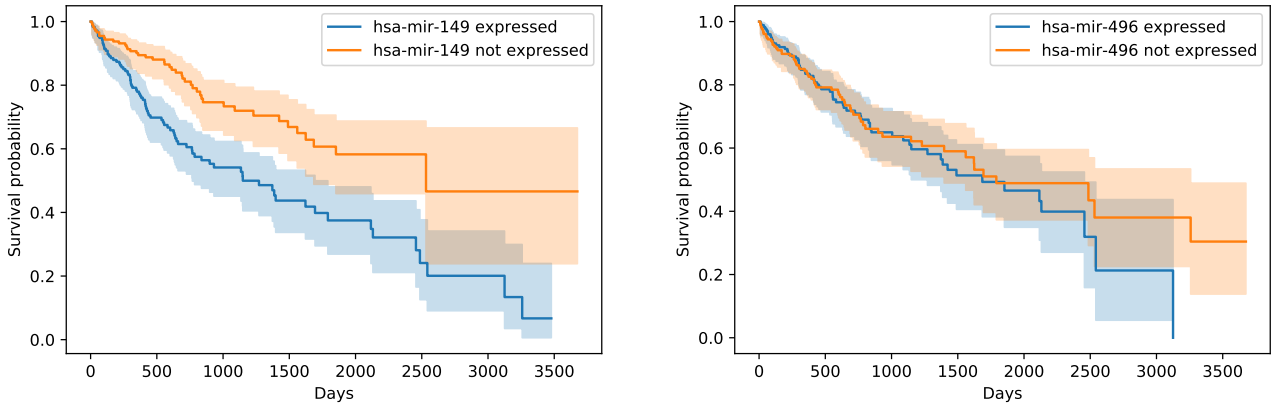

Fig. 1. An example of a Kaplan-Meier plot for two gene expression-dependent conditions associated with patient survival. In panel a), the survival function is substantially higher for a group of patients with highly expressed microRNA hsa-mir-149. The difference is not so evident in the panel b) and microRNA hsa-mir-496. We could say that hsa-mir-149 is hence a better biomarker for survival. In biomarker discovery, one of the tasks is to rank genes and RNA molecules according to the degree of separation between corresponding survival signatures when a gene is expressed and not expressed.

High-throughput sequencing can be used to measure the degree of activity of genes in biological samples. In the recent years, attention has turned from simple single-gene DNA markers to complext multi-gene markers of gene expression. The amount of mRNA in a biological sample that corresponds to a particular gene is correlated to gene's activity and is referred to as gene expression. High-throughput sequencing allows us to determine the expression of all genes in the organism, and hence provide a tool to assess the state of the biological system. A gene can be considered as a biomarker of survival if the survival function is substantially different in a subpopulation where gene is expressed compared to when it is not 1. Note that this definition is vague, as it requires the threshold for gene expression, and quantification and a subsequent threshold for the difference between survival function. Moreover, sets of genes rather than single genes are typically used to capture diverse *biologies* in a sample, e.g. a tendency for blood vessel formation, or priming for immune response. A validated and effective gene expression-based biomarker discovery process can be an incredibly valuable and often a necessary tool in drug discovery, development, and diagnostic research [**?**]. It has shaped the discovery of biomarkers in disease such as cancer [**?**], XXX.

Ideally, therefore, data-driven discovery of new biomarkers would only require survival data with corresponding gene expression profiles. The discovery algorithms would then sift through all the genes and find those that best define groups with different survival function. But there are many difficulties and obstacles in this procedure, related to noisy data, small datasets in terms of the number of investigated samples, higher-order gene interactions, and inclusion of available additional knowledge. We examine these more closely next.

Problem Description

There are three categories of problems and challenges we would like to address in the proposed project, affecting the computation methods to infer potential biomarkers, approaches to data fusion, and implementation:

Computational challenges, noise and overfitting. The experimental data that addresses a specific survival problem, related to, say, impact of a new treatment or drug, is often expensive and hence small in sample size. A typical Phase 1 clinical trial often has less than 30 patients, and Phase 2 clinical trials with more than a hundred patients are an exception rather than the norm. Experimental noise related to sample

collection and treatment, and to measurement of gene expression can be high. This setting can lead to false discoveries and overfitting. The problem is especially exposed when finding sets, or networks of genes that could serve as biomarkers, as the number of candidates (different sets of genes) grows exponentially with the desired size of the gene set. For instance, with 20.000 protein coding genes, there are over $1,3$ trillion possible gene triples; even if we would computationally manage to examine them all, checking so many combinations will necessary lead to overfitting, where our results would apply well to the training data, but not generalize well to new cases. Besides noise and overfitting, computational challenges include those of finding gene expression thresholds (when is a gene expressed?) and aggregation functions (when is a set of genes collectively active?).

Inclusion ob background knowledge. Genes participate in molecular pathways, perform functions, and are associated to diseases and responses to chemicals and drugs. Knowledge about these and other gene annotations is stored in data bases such as GeneOntology [], KEGG [], PathwayCommons [], CellMarker [], and other. Examining gene sets as candidates for biomarkers could and should use these valuable sources of information, both for restricting the biomarker search space and interpreting the sets of best candidate genes (gene set enrichment). Such fusion of data and knowledge bases has generated some fascinating results in bioinformatics research [] but has been insufficiently explored in our target domain of survival biomarker discovery.

Data exploration interface. The past two decades have seen an emergence of a wide array of methods and statistical and machine learning tools to analyze high-throughput data from molecular biology. For survival analysis, however, there is no elegant toolbox with an intuitive user interface that would assist in biomarker discovery, support on-the-fly interactive exploratory data analysis, and offer easy construction of analytical pipelines. Available are excellent code libraries for survival analysis in R and Python, yet, for a systematic use, these are just building blocks that require advanced knowledge of programming to utilize and integrate. What we need instead are intuitive tools with flexible and exciting interactive interfaces to engage the end-users and data scientists in productive communication, data exploration and modeling.

In the project, we will address these three challenges through development of techniques and tools that will support interactive search for and exploration of potential biomarkers. We aim to democratize the field of data-driven biomarker discovery by creating a versatile tool with interactive interface for intelligent analysis of survival data.

Project Aims

The project will develop and apply a set of computational tools for inference of biomarkers from survival data. We will integrate existing approaches to survival data-based biomarker scoring, survival modeling, and gene set enrichment analysis, and propose new techniques for survival-specific gene interactions, construction of biomarker candidate maps, and interpretation of constructed visualizations. We will devise means for heuristic search that will use published data bases on gene function and pathway annotation.

The project will empower domain experts and data miners use these tools in real-life applications - in real time and without the need to write computer code. The project will embed computation methods into components with graphical user interfaces. We will enhance our own, open-source data mining platform Orange[1] with survival analysis capabilities. We will show that the resulting visual programming platform not only substantially reduces the complexity and user time spent on data analysis, but also enhances the collaboration

[1] `http://orangedatamining.com`

and motivation of domain experts through informative visualizations and ability to steer the discovery process using domain knowledge.

Finally, we would like to showcase the utility of the constructed toolbox. In the application of project's approaches we will use a set of published and privately-owned (from participating SME) datasets of gene expression and corresponding clinical outcomes. The success of the project will be judged on use cases carried out by participating SME, and our ability to train them to independently use the results of the project.

Anticipated Results

The expected principal results of this project are:

1. **A bioinformatics library for biomarker discovery from survival data**. The library will be developed in Python and will be published in open-source on GitHub, together with documentation, unit-tests, and working examples;

2. **A biomarker discovery toolbox featuring visual programming interface, interactive visualizations, and interpretation and explanation of results**. The toolbox will support integration of external knowledge-bases, on-the-fly construction of analytical pipelines, and domain knowledge-based guided data exploration;

3. **A set of use-cases developed in close collaboration with Genialis d.o.o.**, a participating SME. The use-cases will demonstrate the applicability of our software, showcase the power of toolbox's intuitive interface, and provide for instruction and educational material in dissemination of project's results.

Preliminary Results and Studies

gene network discovery, gene interactions, intelligent data visualization, embedding, Orange.

26.2 State-of-the-art in the proposed field of research and survey of the relevant literature

Intro paragraph.

Computation Methods for Gene Markers Identification in Survival Analysis

Visual Data Analysis

Toolboxes

26.3 Detailed Description of the Work Programme

26.3.1 Project Tasks

The project will be organized around a following set of tasks:

T0 Setting-up of the collaborative environment. We will deposit all the code and documentation on GitHub[2]. The repository will store project documentation and meeting minutes, tasks management through creation and tracking of issues, Python library code, unit-test, and examples. Data files will be stored on a separate web server. Extensions of Orange[3] will be developed as an add-on and will be stored in a separate repository on the GitHub.

[2] `https://github.com/biolab`
[3] `https://orangedatamining.com`

T1 Data acquisition and organization. The project will use a number of different data sets coming from published studies and databases such as NCBI's Gene Expression Omnibus[4] and TCGA, The Cancer Genome Atlas database[5]. In addition, we will also create a set of synthetic data sets of varying size and complexity. The compiled data sets will be stored in our own dataset repository created in task T1.

T2 Development of data mining and bioinformatics for survival biomarker discovery. In particular, we will develop and implement techniques for:

T2.1 Gene ranking and selection based on the survival function, where we will implement standard techniques from the field, including the log rank test [] and the ranking based on the inference of Cox proportional hazards model. We will also include more recent and advanced modeling approaches based on bootstrap [] and deep learning [], and infer gene ranking through studying the sensitivity of the models [].

T2.2 Feature construction, where we will use predictive models on a smaller subset of genes to aggregate gene expression and with this aim to increase the robustness of so-inferred biomarker. We will employ $\ell 1$ regularization in combination with Cox and derived models, and network-based approaches where biomarker is composed of a small number of genes from the same regulatory network or metabolic pathway [].

T2.3 Gene interaction analysis, where we expect that a group of genes can interact in a non-linear way to form a more robust and informative biomarker. We will adapt the approaches for finding feature interactions [] to address survival data, and the approaches to visualize the results of interaction analysis [].

T2.4 Knowledge-infused biomarker discovery, where we will restrict the search space of survival-effecting gene interactions to groups of genes with shared functional annotations from knowledge libraries on gene annotations, pathways, established panels for measuring gene expression (e.g. nanoString) and known markers.

T2.5 Deep and transfer learning, where our aim is to find gene embeddings for their profiling in low-dimensional space. We will use auxiliary dataset to train (tissue-specific, disease-specific) variational autoencoders [] for embedding, and then adapt the embedding to specific survival problem using transfer learning [], that is, modifying only a small part of the deep model. We will use the embedded, latent profiles of genes for visualizations of gene maps and in heuristics to restrict the search space.

T2.6 Gene interaction maps, where we would like to represent genes – potential biomarkers – in a gene map where vicinity of genes on the map suggest increased joint effect on the survival function. Building on the knowledge and tools for co-expression analysis, these constructed interaction plots will serve for mapping of interaction space and presentation of the space of solutions to biomarker discovery problem.

T2.7 Automatic annotation of point-based visualizations, where points are genes and a primary example of such visualizations are gene interaction maps. We will devise algorithms that search for visualization neighborhoods with enriched gene function or pathways, and annotate visualizations accordingly. This research will follow our prior work on annotation of gene maps for single-cell analysis [].

[4] `https://www.ncbi.nlm.nih.gov/geo/`
[5] `https://portal.gdc.cancer.gov`

T3 Design of visual interfaces for exploratory analysis of survival data and biomarker discovery. In a close collaboration with the partnering SME, we will lead a thorough requirements analysis and product discovery process that will ensure a proper design of components and pipelines for interactive, domain-knowledge driven mining of biomarkers from survival-related gene expression data. The design will include the planning of a set of computational components to address all aspects of survival analysis and biomarker discovery. We will design the graphical interface of the components, their visual presentation, interactive visualizations, and possible data analysis pipelines to combine the design components. The deliverable will include sketches of graphical user interface (in Balsamiq Mockups) and wire-frames. The design will emphasize the quality of user experience, access to advanced computational techniques, and ability to combine components in a Lego-brick way to devise possibly complex and powerful analysis pipelines.

T4 Implementation and Integration. The developed computational techniques will be implemented within the open-source data mining environment Orange[6]. The implementation will use the library of methods from task T2 and graphical user designs from T3. Implementations will be released as an separate add-on to Orange, and will follow implementation guidelines which refer to documentation, and unit testing with near 100% code coverage.

T5 Experimental validation. The developed functionality will be be thoroughly tested in collaboration with Genialis, our project partner. Synthetic and real data sets (prepared under Task 1) will be used, and results compared to those from the literature. The validation will confirm the validity and corectness of developed procedures and will serve to collect case studies to be published on GitHub, Orange's web site, and in planned publications and possibly patents.

T6 Dissemination of results. This task includes publishing of the implementation of the developed methods under General Public License (GPL), writing and web-publishing of relevant documentation with working examples, publishing video tutorials with use cases on Orange's YouTube channel[7], and dissemination in terms of presentation at relevant conferences and journal publications. We will target bioinformatics journals, such as *Bioinformatics*, *Nature Methods*, and *Artificial Intelligence in Medicine*, and related conferences including the top-rated AIM and ISMB. Also, we are planning to file a joint patent with Genialis in the field of a procedure for survival marker discovery.

26.3.2 Research Design and Methods

Overview. The overview of the research in the proposed project is presented in Figure **??**. The figure shows how we will integrate existing and new experimental data on genetic interactions, whole genome gene expression data sets, and data on cytotoxicity to find: 1. which genes are functionally related to a particular drug, or to a class of drugs with common characterization, 2. what is the mechanism through which a set of drugs causes cytotoxic effects, 3. which other experiments could further improve the reliability of the inferred hypothesis on drug action. While all of the above are clearly the issues to be addressed by a biomedical engineer, due to sheer volume of data and additional information in available data and knowledge they can only be addressed through means of computational analysis and discovery approaches. The principal challenge of the project is how to combine these different sources, and use modern data mining approaches to support knowledge discovery to provide interpretable and operational hypotheses to biomedical researchers and drug developers. In the description below, we first

[6] `http://orangedatamining.com`)
[7] `http://youtube.com/orangedatamining`

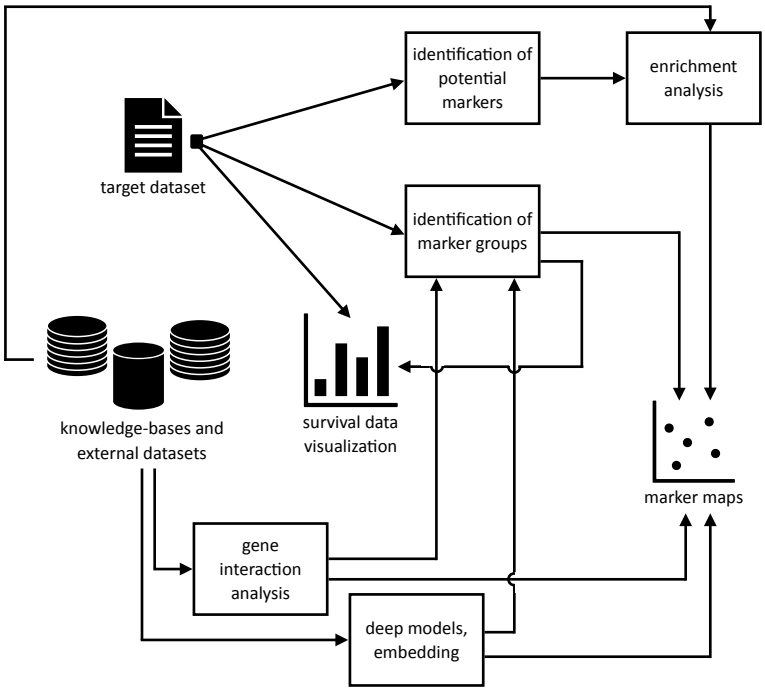

Fig. 2. An example of a Kaplan-Meier plot for two gene expression-dependent conditions associated with patient survival. In panel a), the survival function is substantially higher for a group of patients with highly expressed microRNA hsa-mir-149. The difference is not so evident in the panel b) and microRNA hsa-mir-496. We could say that hsa-mir-149 is hence a better biomarker for survival. In biomarker discovery, one of the tasks is to rank genes and RNA molecules according to the degree of separation between corresponding survival signatures when a gene is expressed and not expressed.

describe the data on which our drug characterization and discovery process will be based, then enlist a set of computational approaches which we will develop and use, comment on their implementation within an existing visual programming-based data mining framework, and integration in robotic experimental chemical genomics platform.

Material and Data Sets. Text here.

Computational Approaches, Data Mining and Bioinformatics. Text here.

Software Implementation. Text here.

Experimental Validation. Text here.

26.4 Available research equipment over 5.000 €

26.5 Project management