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 a 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 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 data analysis results to customers and regulatory agencies. On the other hand, the project will allow us, the proposing institution, to further advance our research into interactive visualizations and machine learning, and apply our new approaches to the 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 [25]. Survival models, including the most famous, 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 measure objectively and use as an indicator of the system's state. For example, in cancer, markers may differentiate between patients that respond to the treatment and those who do not.

Based on markers, we can predict the success of treatment and choose the right therapy for an individual patient. Identification of good markers is thus crucial to the development of personalized medicine. Because improving survival benefits the patient directly, it is essential to understand how participants respond to various forms of treatment. The treatment should thus be selected based on the patient's state and characteristics defined through a set of markers. Markers may be clinical, related to the patient's symptoms, or biological, related to some measurements on the molecular level, like concentration of a specific protein or expression of a particular gene or a set of genes. Markers may refer to a single measurement or to a group of measurements possibly related through a prognostic model or a network [30].

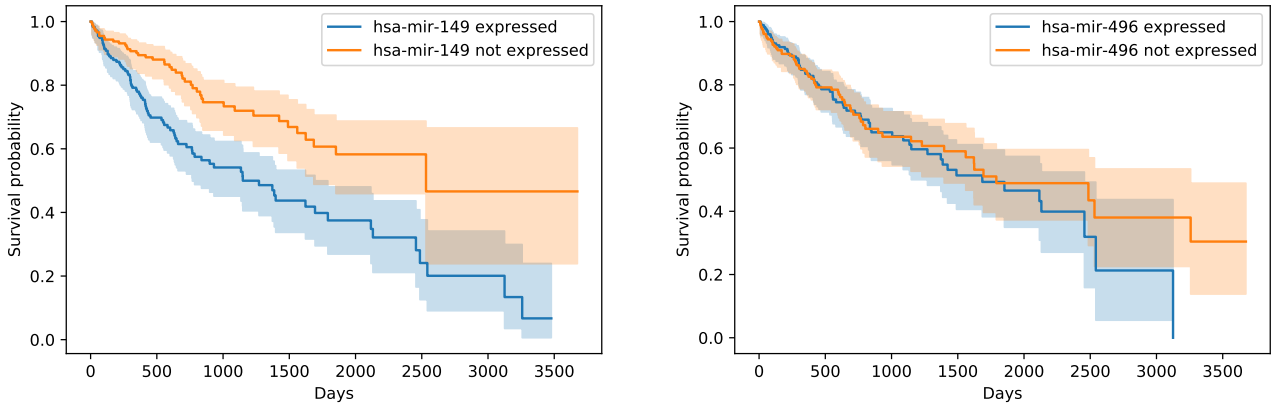

Fig. 1. An example of a Kaplan-Meier plot for two gene expression-dependent conditions associated with patient survival. The survival function is substantially higher for a group of patients with highly expressed microRNA hsa-mir-149 (left panel). The difference is not so evident for microRNA hsa-mir-496 (right panel). 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 survival signatures given gene expression.

High-throughput sequencing can be used to measure the degree of activity of genes in biological samples. In recent years, attention has turned from simple single-gene DNA markers to complex 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 (see Fig. 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 functions. 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 [24].

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 functions. However, there are many problems and challenges in this procedure related to noisy data, low 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 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 address a specific survival problem, related to, say, the 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, sample treatment, and gene expression measurement can be high. This setting can lead to false discoveries and overfitting. The problem is especially exposed when searching for 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 biomarker gene set. For instance, with 20000 protein-coding genes, there are over 1.3 trillion possible gene triples. Even if we managed to examine them all computationally, this would necessarily lead to overfitting. With overfitting, results would apply well to the training data but would not generalize to new cases. Besides noise and overfitting, computational challenges include finding gene expression thresholds (when is a gene expressed?) and aggregation functions (when is a set of genes collectively active?).

Inclusion of 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[1], KEGG[2], CellMarker[3], 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 fascinating results in bioinformatics research [41, 42] but has been insufficiently explored in the domain of survival biomarker discovery.

Data exploration interface. The past two decades have seen an emergence of various methods, statistical tools, and machine learning tools to analyze high-throughput data from molecular biology. Survival analysis, however, lacks an 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. Excellent code libraries for survival analysis in R and Python are available, yet, for systematic use, these are just building blocks that require advanced programming skills to utilize and integrate. Instead, we need 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 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. The project will develop new techniques to discover survival-specific gene interactions, construct biomarker candidate maps, and interpret created visualizations. We will devise means for heuristic search that will use published knowledge-bases on gene function and pathway annotation. The project's principal scientific contribution is thus biomarker interaction discovery, intelligence and interpretable visualization of biomarker space, and integration of knowledge and data to improve interpretability, the robustness of results, and the speed of heuristic search for best biomarker candidates.

The project will empower domain experts and data miners to 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 a graphical user interface. We will enhance our own open-source data mining platform Orange[4] [9, 8, 12] (Fig. 2) with survival analysis capabilities. We will show that the resulting visual programming

[1] `http://geneontology.org`
[2] `https://www.genome.jp/kegg`
[3] `http://biocc.hrbmu.edu.cn/CellMarker`
[4] `http://orangedatamining.com`

platform not only substantially reduces the complexity and time spent on data analysis, but also enhances the collaboration and motivation of domain experts through informative visualizations and the 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.

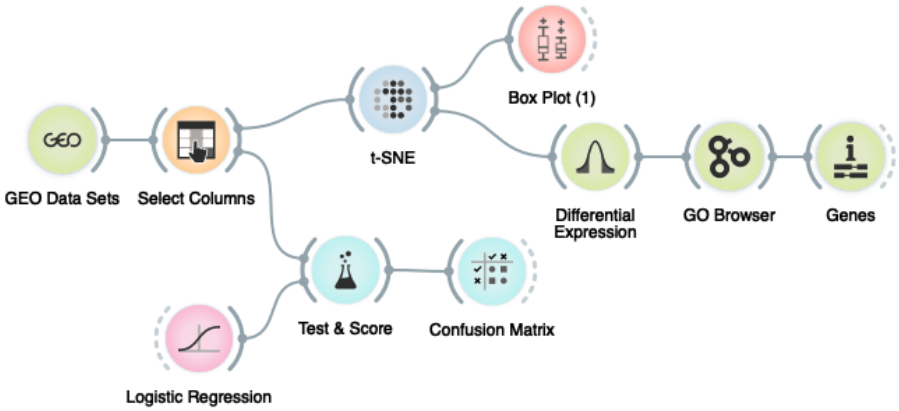

Fig. 2. A typical bioinformatics workflow in Orange. The figure shows an example where we have reanalyzed expression data from peripheral blood mononuclear cells (GDS5363), where the authors investigated if gene expression profiling could detect the onset of osteoarthritis [27]. The workflow loads the data from Gene Expression Omnibus and defines the dependent variable (disease state, *Select Columns* component). The upper branch checks the clustering structure of tissue samples (*t-SNE* component), finds differentially expressed genes for selected samples, and analyzes their common characteristics through Gene Ontology term enrichment. In the lower branch, we check the investigators' hypothesis directly and evaluate the accuracy of the predictions of a logistic regression model through cross-validation (*Test & Score* component). In the proposed project, we will develop similar-style workflows, but with components that will load, process, analyze and display survival data and discover biomarkers.

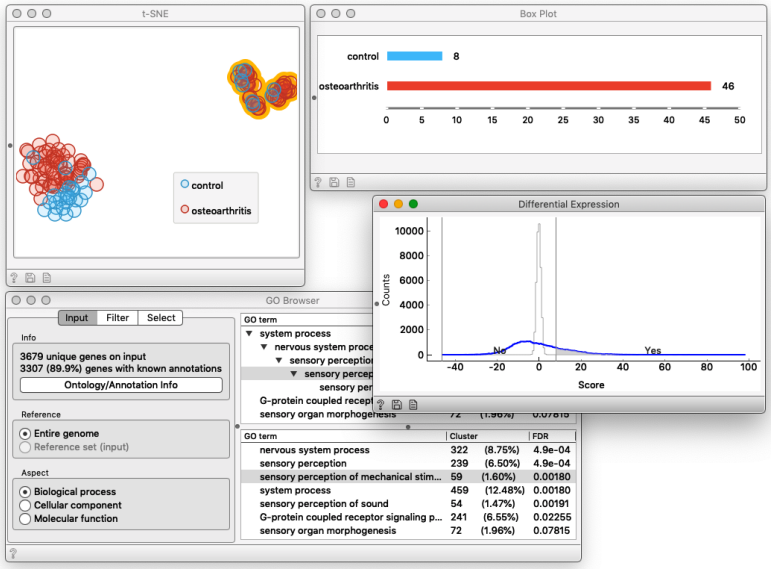

Fig. 3. Most graphical components – widgets – in Orange are interactive. We here show the content of several widgets from the workflow in Figure 2. The user can, for instance, choose a subset of data points from the *t-SNE* visualization (data points with a yellow outline on the top right of the visualization), or data corresponding to a specific bar in the *Box Plot*, or a set of genes associated with a selected term from the *GO Browser*. In the *Differential Expression* widget, we can choose a set of differentially expressed genes in the tails of the gene score distribution plot. Most widgets contain both a control and a visualization part. The project will reuse some of the widgets from this visualization, including the t-SNE widget, and will develop specialized highly-interactive widgets for survival analysis and biomarker discovery.

Anticipated Results

The expected principal results of this project are encompass the following scientific, engineering contributions and practical results:

1. **New methods and approaches for biomarker discovery**, including computational approaches for survival-based biomarker interaction analysis, visualization approaches to map the space of potential biomarkers, and improved heuristic search for groups of biomarkers through the integration of data and knowledge-bases.

2. **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;

3. **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, interactive exploration of data and models (see Fig. 3), and domain knowledge-based guided data exploration;

4. **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.

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

The log-rank test and Cox's proportional hazards model are most commonly used in the literature to compare survival between two different groups of observers [29]. These two methods can be regarded as a baseline to design strategies for detecting predictive marker genes. In a nutshell, the search for marker genes can be split into two subproblems: the grouping, or better, binarization of response variables (e.g., expression of the genes) and the search for promising predictive gene groups.

In clinical studies, biological markers are usually continuous variables obtained by various measurements. Establishing a cut-off point that represents the boundary between high and low gene expression, or more generally, that distinguishes between high and low risk groups, may be essential for their use in clinical decision making [23]. Budczies et al. [3] propose several approaches to selecting cut-off values: according to the distribution of the biological marker, by optimizing the interdependence of the target variable, such as response to treatment, or by finding a minimum p-value. The latter is the most common and selects the cut-off value according to the optimal difference in the survival outcome prediction between the groups [36]. In general, however, finding the optimal value is a difficult problem that also depends on the study or research itself. Proper procedures to find the limit value are very important, as we may overestimate the true effect of the biological marker [1].

Witten et al. [35] highlight the problem of finding predictive features in high-dimensional data. When the number of variables is many times greater than the number of cases, the usual statistical approaches to survival analysis are no longer sufficient. There are many different published approaches to finding marker genes. Some recommend two-stage filtering: first, filtering differentially expressed genes (genes that distinguish well between selected groups), and then, further narrowing the set of possible candidates based on statistical significance in survival analysis [33, 19, 40, 17]. Relator et al. [28] are critical of such approaches because they may leave many possible combinations of genes untested. They suggest a solution that can detect interactions of potential markers that conventional approaches would omit. An important shortcoming of their approach,

as they acknowledge themselves, is computational complexity. They suggest splitting the data into smaller samples, running the proposed solution over the individual samples, and then combining the results. Consequently, also the proposed solution may omit promising gene interactions.

In complex diseases such as cancer, the effects of genomic data on survival are generally nonlinear. To detect nonlinear gene relationships, various approaches with deep learning techniques have emerged recently [13]. Several different models of deep learning have been proposed to predict survival, including the standard Cox model of relative risk (Cox-nnet [6], SurvivalNet [38], DeepSurv [15]). Despite advanced techniques and an increase in the number of potential biological markers, very few have been clinically used [4]. If the markers found are difficult to explain, insufficiently researched, or without known biological functions, they may be discarded despite their promise. With deep learning models, this challenge is all the greater. Hao et al. [13] are taking a step towards finding explicable gene marker groups with deep neural networks by incorporating genomic and clinical data.

Besides computational methods, the proposed project will also relate to existing published applications of survival analysis. For example, Xiwen et al. [19] and Wang et al. [18] studied correlations between miRNAs and the prognosis of hepatocellular carcinoma patients (HCC). They both established a five-miRNA signature model that could serve as a potential biomarker in the prognosis of HCC patients. Similarly, Guodong et al. [37] identified a novel five-miRNA signature model as a prognostic biomarker in colorectal cancer patients. Different approaches and methods were tested using miRNA expressions and related clinical data accessible through the TCGA database. Furthermore, Martinez-Ledesma et al. [22] explored the network-based approach and identified a gene expression-based biomarker that can successfully predict the clinical outcome of 12 different types of cancer. Listed studies are a great example of the importance of projects like TCGA. The comprehensive and structured data from the TCGA database may drastically speed-up the development of techniques for discovering (Di et al., [14]) and validating (Chen et al. [5]) promising gene-based biomarkers.

26.3 Detailed Description of the Work Programme

26.3.1 Project Tasks

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

T0 Setting-up of the collaborative environment. We will deposit all the code and documentation on GitHub[5]. 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[6] will be developed as an add-on and will be stored in a separate repository on the GitHub.

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[7] and TCGA, The Cancer Genome Atlas database[8]. 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:

[5] `https://github.com/biolab`
[6] `https://orangedatamining.com`
[7] `https://www.ncbi.nlm.nih.gov/geo`
[8] `https://portal.gdc.cancer.gov`

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 random forests 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 [10] for embedding, and then adapt the embedding to specific survival problem using transfer learning [12], 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 (see Fig. 4).

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,

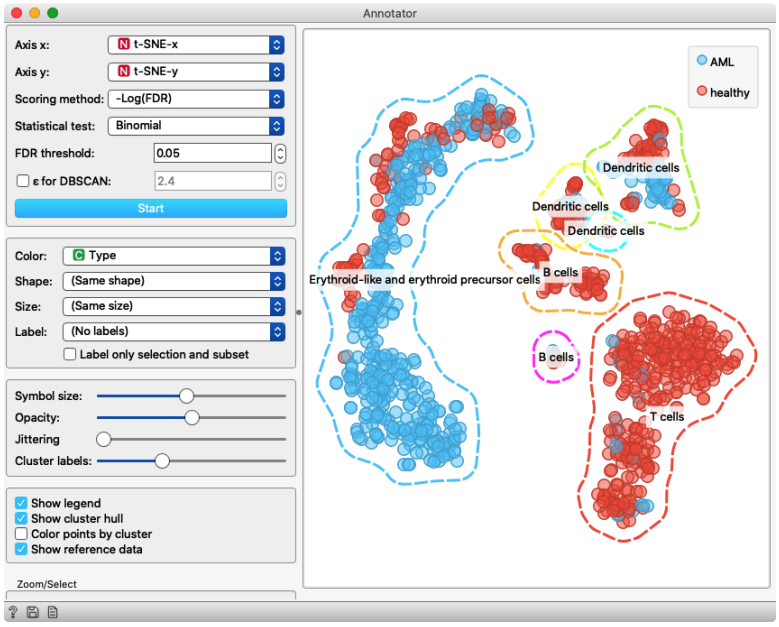

Fig. 4. A prototype of Orange's widget with automatic annotation of point-based visualisation. On the input, this component considers gene expression profiles of single cells and their associated embedding (e.g., t-SNE coordinates), and a list of marker genes for each candidate cell-type. In scope of the proposed project, we will use a similar approach to design a widget that would annotate and thus explain the space of potential markers.

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[9]. 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 evaluation. 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 correctness 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[10], 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 prognostic biomarker discovery.

26.3.2 Research Design and Methods

Overview. The overview of the research in the proposed project is presented in Fig. 5. The figure shows how we will integrate the target gene expression data and clinical metadata with additional knowledge bases and other available data sets. It exposes the key scientific approaches we will address:

[9] `http://orangedatamining.com`
[10] `http://youtube.com/orangedatamining`

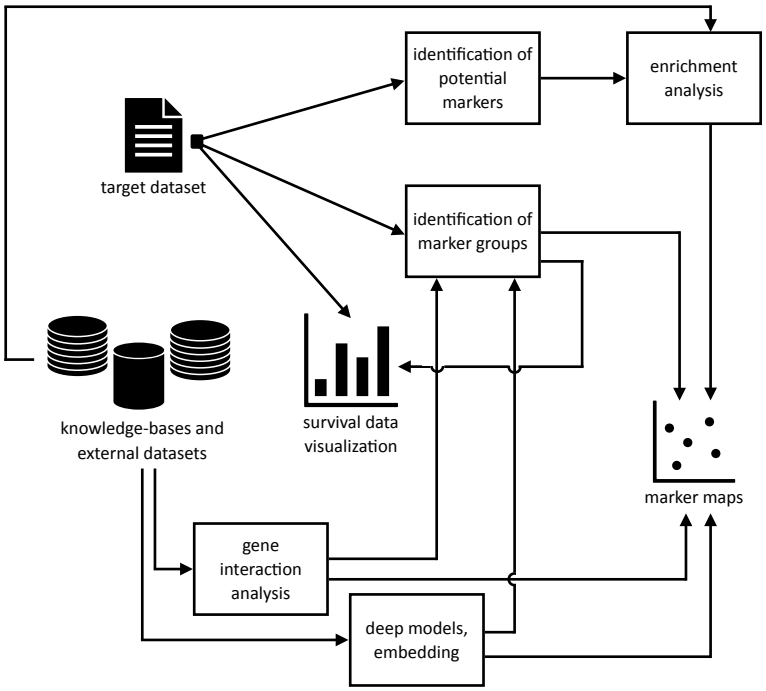

Fig. 5. Knowledge-based analysis to biomarker discovery. We aim to combine our target data set with transcription and survival data from clinical trials with other available data sets and knowledge bases to gain in speed, accuracy and interpretation of results. Substantial part of our project deals with exploratory data interfaces, which combine visualization of data, identified biomarkers and models. Together, computational approaches for biomarker discovery and visualization approaches for making these findings explicit and placing them in the context of entire search space (e.g., *marker maps*) will assist us in explainable, semi-automatic discovery process. The methods will provide hypotheses to be evaluated by a domain expert, and they will be able to work with the assumptions and curated public knowledge rather than machine-generated black-box decisions.

1. which, on their own, are the key molecular markers for survival,
2. which are the combinations of molecular entities (gene sets) that jointly correlate with survival,
3. which characteristics of marker genes can help us interpret the results of the analysis.

While all of the above are clearly the issues that could be individually and manually addressed by a molecular biologist, due to the sheer volume of data and additional information in available data and knowledge bases they can only be tackled through means of computational analysis and data-driven 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 list the data sources on which we will apply our marker discovery process, then list a set of computational approaches which we will develop and use, and comment on the feasibility of their implementation within the existing visual programming-based data mining framework.

Material and Data Sets. Four types of data sets will be gathered, organized and used in the project:

DS1 Transcription data from The Cancer Genome Atlas database[11] [34] and Gene Expression Omnibus[12] [2].

DS2 Survival analysis data from clinical trials from The Cancer Genome Atlas database.

DS3 Transcription and clinical trial data managed by Genialis.

DS4 Simulated data sets.

The data from DS1 and DS2 are fragmented; datasets have to be integrated, so that the clinical trial data is aligned with a corresponding transcription data. Methodological reports on development of computational techniques, including those we have cited in the related work, would refer to data repositories but would seldom publish the reorganized data ready to be used in off-the-shelf software packages. Our

[11] `https://www.cancer.gov/tcga`
[12] `https://www.ncbi.nlm.nih.gov/geo/`

project aims to break with this practice and compile a data repository with aligned clinical and transcription data sets ready for benchmarking and comparing biomarker discovery techniques.

Genialis d.o.o. already has a collection of such align datasets which comes from their existing partnerships with some major pharmaceutical companies. The data is private, but will be shared with us in the project for testing and validation purposes.

We will also construct a set of simulated datasets, where variables representing biomarkers will be placed by design. The datasets will serve for testing and benchmarking of the proposed methods.

The project will additionally use other sources of information, which, for the reasons of convenience, will be queried for information specific for the project and will internally be represented as additional data sets. These include, but are not limited to:

DS6 Gene function annotations from Gene Ontology (GO) consortium.[13]

DS7 Various pathways from KEGG, Kyoto Encyclopedia of Genes and Genomes.[14]

DS8 NDEx pathway data base.[15]

DS9 Various marker gene data bases, including CellMarker[16] [**?**] and PanglaoDB [17] [11].

Computational Approaches, Data Mining and Bioinformatics. Computational approaches and development of data mining methods will include:

Data organization (task T1). The project will develop a computational platform with access to the common, server-based databases that will store transcriptome and survival data from third parties and participating SME Genialis. We will use standard software engineering practices to construct this architecture (data server with secure HTTP access, HTTP-based queries, components on clients to access the data). We will locally store the data from other information sources like gene ontologies and pathways to allow for fast computation and utility of such information, and for these reuse existing server and database architecture of Orange [9, 8, 12]. Overall, the software engineering task here is to hide the details of data access and query from the user. Users should be able to access these functions with a single click and focus on data analysis and interpretation.

Gene ranking (T2.1). We will use the standard log-rank test to compare two or more survival curves and hence estimate a selection of gene and its expression threshold to break observed cases to subpopulation. The log-rank test will be used as a baseline. We will compare the inferred ranking of genes to modeling approaches, where we will infer survival models first, and then estimate the information value of its constituents (genes) either directly for linear models (e.g. Cox proportional hazards models) or indirectly. Advanced models will include random survival forests [31] and deep learning [15, 6]. Indirect measurements of contributions of individual features will include game theoretic approaches such as SHAP[18] [21].

Gene set enrichment analysis (part of T2.1). We will use the gene set enrichment analysis [39] to interpret the results of the biomarker ranking and inspect commonalities of the best-ranked genes in terms of functions and pathways.

[13] `https://www.geneontology.org`
[14] `https://www.genome.jp/kegg`
[15] `http://www.ndexbio.org/`
[16] `http://biocc.hrbmu.edu.cn/CellMarker`
[17] `https://panglaodb.se`
[18] `https://github.com/slundberg/shap`

Feature construction and gene set identification (task T2.2). We will use direct and indirect feature selection techniques. The quality of the set will be estimated through the predictive performance of the model. As an alternative approach, we will use model-based feature selection, like $\ell 1$ regularization of Cox and network models.

Gene interaction analysis (T2.3) will examine possible combinations of genes for their combined effect on the survival function. We plan to use model-based evaluation of gene pairs (T2.1) and represent results in interaction maps (T2.6) and networks.

Knowledge-infused biomarker discovery (T2.4) will limit the search for useful combinations of features in task T2.2 to genes with common functional or pathway labels, that is, preferably to those already associated with investigated pathology in the literature. The success of this approach will be measured through gains in algorithm speed and reduction of search space to be considered, while requesting identification of the gene sets of similar quality to those from exhaustive search.

Deep learning (T2.5) will employ variational autoencoders for embedding to represent marker candidates with latent vectors, that is, position them in the latent space where we can easily examine their relatedness and the structure of marker space. We will construct autoencoders from the gathered collection of transcription data and clinical data sets (task T1) and then use transfer learning [12] to adapt the model to a specific target dataset. The embedded representations of potential markers will provide the foundation to construct interpretable visualizations of the search space (e.g., task T2.6).

Gene interactions maps (T2.6) will render information from tasks T2.3 and T2.7 and provide means for graphical explanation and presentation of the search space, exploratory data analysis, and visual interpretation. To render gene maps we will use our own variant of t-SNE [32] called openTSNE[19] [26], which can preserve global structure.

Annotation of point-based visualizations will equip gene maps (task T2.6) and other point-based visualizations of gene and marker search space with functional and pathway labels. We will use this approach to provide assisted interpretation of the search space. To develop this technique, we will extend our approach previously developed for single-cell gene expression analysis (see Fig. 4).

Software Implementation. Over the past two decades, we have been developing a comprehensive data analysis suite called Orange[20] [9, 8, 12]. Orange has been used by thousands of users and is a toolbox of choice in training of data science in hundreds of universities[21]. Orange features a scripting and a visual programming environment. Visual programming offers an intuitive means of combining known analysis and visualization methods into powerful applications. Orange includes a set of visual components for functional genomics (Fig. 3) that enable users who are not programmers to analyse transcriptomic data and to customize their analysis by combining common data analysis tools to fit their needs [12]. Orange framework will be used to implement methods proposed in this project and offer them to the community within an open-source model.

Within the project, we will develop a set of components that will be specific to the problem in our project, but also general enough for other survival analysis tasks. We intend to develop components for survival modeling, accuracy estimation, biomarker ranking, identification of groups of markers, gene embedding, and construction of survival-based gene interaction maps. Orange includes components crucial for the proposed project but have already been developed or at least prototyped. These include

[19] `https://github.com/pavlin-policar/openTSNE`
[20] `https://orangedatamining.com`
[21] `https://orangedatamining.com/blog/2021/2021-01-11-orange-in-classroom/`

access to gene annotation libraries, gene set enrichment, GO browser, and visualization tools, including t-SNE and Kaplan-Meier curves.

Experimental Validation. We will use the data sets DS1 to DS4 to experimentally validate both the computational techniques and the visual interfaces we will construct in the project. There are three aspects of validation we will assess:

- **predictive performance**, where we will compare inferred biomarkers to those published in the literature, and compare the results of our marker discovery techniques to those already published;
- **interpretability**, which will be assessed with domain experts working with Genialis, a participating SME. The assessment of interpretability will be both quantitative, in terms of how well can we link identified biomarkers to known functional and pathway labels, and subjective, in terms of agreement of domain experts with the proposed result;
- **usability**, where we will use hands-on workshops with domain experts to assess if they can use the developed tool independently after a three-hour training.

26.4 Available research equipment over 5.000 €

The project will use the computational infrastructure of Bioinformatics Laboratory of University of Ljubljana, which includes CPU cluster (about 500 processors), NFS storage (about 500 TB), and a cluster of GPU processors (about 20 GPUs), collectively valued at about 200.000 EUR. We will require no special computational equipment besides the existing for the proposed project.

26.5 Project management

The project will join two highly compatible R&D teams. The Bioinformatics Laboratory from University of Ljubljana will bring to the project its extensive expertise in data mining, machine learning, bioinformatics, and computational phenotyping. The project will be co-financed by Genialis, a data science and drug discovery company focused on new ways to treat disease. Blending computational biology and AI-based methods, Genialis merges and models data at the intersection of clinical and translational medicine. Genialis is trusted by biopharma and big pharma alike, to validate targets, predict biomarkers and optimally position novel drugs. Together, Genialis and its partners are bringing improved solutions to drug discovery to change people's lives.

The project will be managed by UL (project manager). The management of the project will be organized through regular meetings of the management board, with one appointed representative from each institution, and through regular meetings of the project members. The collaboration platform will be based on GitHub and will be made available in the earliest stage of the project.

We will complete the project in three years. Fig. 6 gives the detailed time line of the projects.

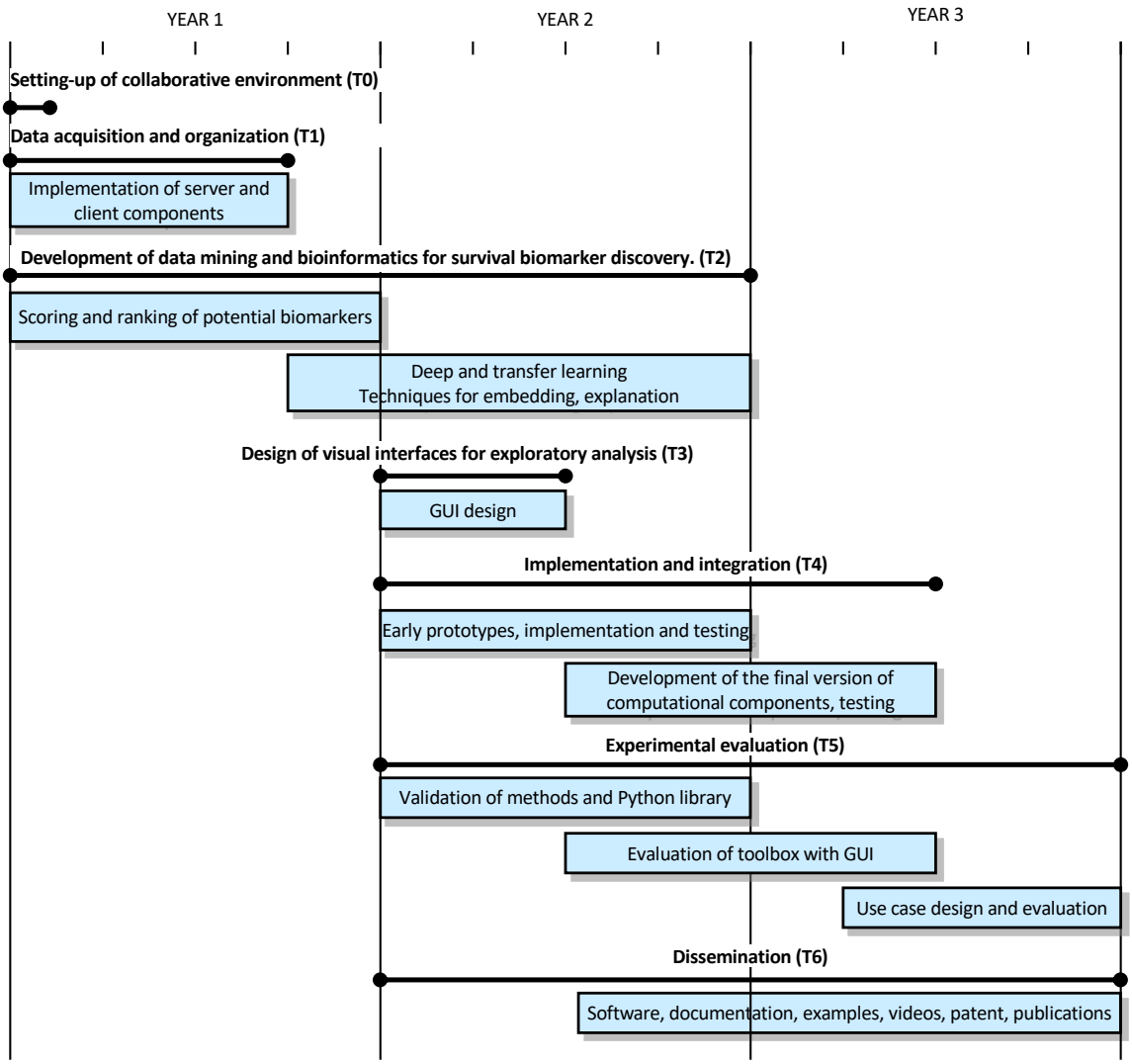

Fig. 6. Project's time line. In a nutshell, we will start with organizing the data sets. Then, we will develop a Python library for marker discovery in survival analysis. We will proceed with the design and implementation of the graphical user interface for new components of Orange data mining library. All components of the developed system will be tested in collaboration with Genialis, our partner SME. A special task is devoted to dissemination, which includes open-source software, documentation, examples, video material, scientific publications, and filing of a patent.

References

[1] D. G. Altman. Categorising continuous variables. *British Journal of Cancer*, 64(5):975–975, Nov 1991.

[2] T. Barrett, S. E. Wilhite, P. Ledoux, et al. NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res*, 41(Database issue):D991–995, Jan 2013.

[3] J. Budczies, F. Klauschen, B. V. Sinn, et al. Cutoff finder: a comprehensive and straightforward web application enabling rapid biomarker cutoff optimization. *PLOS One*, 7(12):e51862, 2012.

[4] H. B. Burke. Predicting clinical outcomes using molecular biomarkers. *Biomarkers in Cancer*, 8:BIC–S33380, 2016.

[5] J. Chen, Z. Wang, W. Wang, et al. SYT16 is a prognostic biomarker and correlated with immune infiltrates in glioma: A study based on TCGA data. *Int Immunopharmacol*, 84:106490, Jul 2020.

[6] T. Ching, X. Zhu, and L. X. Garmire. Cox-nnet: an artificial neural network method for prognosis prediction of high-throughput omics data. *PLOS Computational Biology*, 14(4):e1006076, 2018.

[7] T. Ching, X. Zhu, and L. X. Garmire. Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data. *PLoS Comput Biol*, 14(4):e1006076, 04 2018.

[8] T. Curk, J. Demsar, Q. Xu, et al. Microarray data mining with visual programming. *Bioinformatics*, 21(3):396–398, Feb 2005.

[9] J. Demšar and B. Zupan. Orange: Data mining fruitful and fun - a historical perspective. *Informatica*, 37:55–60, 2013.

[10] C. Doersch. Tutorial on variational autoencoders, 2021.

[11] O. Franzén, L. M. Gan, and J. L. M. Björkegren. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database (Oxford)*, 2019, 01 2019.

[12] P. Godec, M. Pančur, N. Ilenič, et al. Democratized image analytics by visual programming through integration of deep models and small-scale machine learning. *Nat Commun*, 10(1):4551, 10 2019.

[13] J. Hao, Y. Kim, T. Mallavarapu, et al. Interpretable deep neural network for cancer survival analysis by integrating genomic and clinical data. *BMC Medical Genomics*, 12(10):1–13, 2019.

[14] D. Jia, S. Li, D. Li, et al. Mining TCGA database for genes of prognostic value in glioblastoma microenvironment. *Aging (Albany NY)*, 10(4):592–605, 04 2018.

[15] J. L. Katzman, U. Shaham, A. Cloninger, et al. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1):24, 2018.

[16] J. L. Katzman, U. Shaham, A. Cloninger, et al. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med Res Methodol*, 18(1):24, 02 2018.

[17] Y.-W. Kim, D. Koul, S. H. Kim, et al. Identification of prognostic gene signatures of glioblastoma: a study based on tcga data analysis. *Neuro-Oncology*, 15(7):829–839, 2013.

[18] W. Li, X. Kong, T. Huang, et al. Bioinformatic analysis and in vitro validation of a five-microRNA signature as a prognostic biomarker of hepatocellular carcinoma. *Ann Transl Med*, 8(21):1422, Nov 2020.

[19] X. Liao, G. Zhu, R. Huang, et al. Identification of potential prognostic microrna biomarkers for predicting survival in patients with hepatocellular carcinoma. *Cancer Management and Research*, 10:787, 2018.

[20] X. Liao, G. Zhu, R. Huang, et al. Identification of potential prognostic microRNA biomarkers for predicting survival in patients with hepatocellular carcinoma. *Cancer Manag Res*, 10:787–803, 2018.

[21] S. M. Lundberg, G. Erion, H. Chen, et al. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat Mach Intell*, 2(1):56–67, Jan 2020.

[22] E. Martinez-Ledesma, R. G. Verhaak, and V. Treviño. Identification of a multi-cancer gene expression biomarker for cancer clinical outcomes using a network-based algorithm. *Sci Rep*, 5:11966, Jul 2015.

[23] M. Mazumdar and J. R. Glassman. Categorizing a prognostic variable: review of methods, code for easy implementation and applications to decision-making about cancer treatments. *Statistics in Medicine*, 19(1):113–132, 2000.

[24] J. Monforte and S. McPhail. Strategy for gene expression-based biomarker discovery. *Biotechniques*, Suppl:25–29, Apr 2005.

[25] R. Pazdur. Endpoints for assessing drug activity in clinical trials. *Oncologist*, 13(2):19, 2008.

[26] P. G. Poličar, M. Stražar, and B. Zupan. opentsne: a modular python library for t-sne dimensionality reduction and embedding. *bioRxiv*, 2019.

[27] Y. F. Ramos, S. D. Bos, N. Lakenberg, et al. Genes expressed in blood link osteoarthritis with apoptotic pathways. *Ann. Rheum. Dis.*, 73(10):1844–1853, 2014.

[28] R. T. Relator, A. Terada, and J. Sese. Identifying statistically significant combinatorial markers for survival analysis. *BMC Medical Genomics*, 11(2):45–55, 2018.

[29] R. Singh and K. Mukhopadhyay. Survival analysis in clinical trials: Basics and must know areas. *Perspectives in Clinical Research*, 2(4):145, 2011.

[30] A. R. Sonawane, S. T. Weiss, K. Glass, and A. Sharma. Network medicine in the age of biomedical big data. *Frontiers in Genetics*, 10:294, 2019.

[31] J. M. Taylor. Random Survival Forests. *J Thorac Oncol*, 6(12):1974–1975, Dec 2011.

[32] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[33] Z. Wang, G. Chen, Q. Wang, et al. Identification and validation of a prognostic 9-genes expression signature for gastric cancer. *Oncotarget*, 8(43):73826, 2017.

[34] J. N. Weinstein, E. A. Collisson, G. B. Mills, et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet*, 45(10):1113–1120, Oct 2013.

[35] D. M. Witten and R. Tibshirani. Survival analysis with high-dimensional covariates. *Statistical Methods in Medical Research*, 19(1):29–51, 2010.

[36] S. Y. Woo and S. Kim. Determination of cutoff values for biomarkers in clinical studies. *Precision and Future Medicine*, 4(1):2–8, 2020.

[37] G. Yang, Y. Zhang, and J. Yang. A Five-microRNA Signature as Prognostic Biomarker in Colorectal Cancer by Bioinformatics Analysis. *Front Oncol*, 9:1207, 2019.

[38] S. Yousefi, F. Amrollahi, M. Amgad, et al. Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. *Scientific Reports*, 7(1):1–11, 2017.

[39] A. Zacho, J. Nielsen, and C. Cederqvist. Relationship between type of tobacco used and localization of tumour in patients with gastric cancer. *Acta Chir Scand*, 141(7):676–679, 1975.

[40] Y.-Z. Zhang, L.-H. Zhang, Y. Gao, et al. Discovery and validation of prognostic markers in gastric cancer by genome-wide expression profiling. *World Journal of Gastroenterology: WJG*, 17(13):1710, 2011.

[41] M. Zitnik, F. Nguyen, B. Wang, et al. Machine Learning for Integrating Data in Biology and Medicine: Principles, Practice, and Opportunities. *Inf Fusion*, 50:71–91, Oct 2019.

[42] M. Žitnik, E. A. Nam, C. Dinh, et al. Gene Prioritization by Compressive Data Fusion and Chaining. *PLoS Comput Biol*, 11(10):e1004552, Oct 2015.