# Towards an Environment for Data Mining based Analysis Processes in Bioinformatics & Personalized Medicine

Dennis Wegener*, Simona Rossi†, Francesca Buffa‡, Mauro Delorenzi† and Stefan Rüping*
*Fraunhofer IAIS, Schloss Birlinghoven, 53754 Sankt Augustin, Germany
Email: {dennis.wegener, stefan.rueping}@iais.fraunhofer.de
†Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland
Email: {simona.rossi, mauro.delorenzi}@isb-sib.ch
‡Weatherall Institute of Molecular Medicine, University of Oxford
John Radcliffe Hospital, Headington, Oxford OX3 9DS, United Kingdom
Email: francesca.buffa@imm.ox.ac.uk

***Abstract*—Bioinformatics and data mining procedures are collaborating to implement and evaluate tools and procedures for prediction of disease recurrence and progression, response to treatment, as well as new insights into various oncogenic pathways [1], [2], [3], [4] by taking into account the user needs and their heterogeneity. Based on these advances, medicine is undergoing a revolution that is even transforming the nature of health care from reactive to proactive [5]. The p-medicine (www.p-medicine.eu) consortium is creating a biomedical platform to facilitate the translation from current practice to a predictive, personalized, preventive, participatory and psycho-cognitive medicine by integrating VPH models, clinical practice, imaging and omics data. In this paper, we present the challenges for data mining based analysis in bio- and medical informatics and our approach towards a data mining environment addressing these requirements in the p-medicine platform.**

***Keywords*-Data Mining; Scientific Data Analysis; Integration; e-Health; Bioinformatics; Evaluation;**

## I. Introduction

In the last decade, the huge generation and availability of array-based and DNA sequencing technologies data has made possible the generation of knowledge about coding and noncoding-RNAs (nc-RNAs), single-nucleotide polymorphisms (SNPs), and their behaviour in the human and other organisms [6], [7], [8]. The interpretation of this data and the growing interdisciplinary way of performing research enhanced the comprehension of cancer biology. In recent years, using advanced semi-interactive data analysis algorithms such as those from the field of data mining gained more and more importance in life science in general and in particular in bioinformatics, genetics, medicine and biodiversity. Today, there is a trend away from collecting and evaluating data in the context of a specific problem or study only towards extensively collecting data from different sources in repositories potentially useful for subsequent analysis. At the time the data is collected, the type of analysis is not yet known. Content and data format are not focused. Thus, complex process chains are needed for the analysis of the data. Such process chains need to be supported by the environments that are used to set-up analysis solutions. Building specialized software is not a solution, as this effort can only be carried out for huge projects running for several years. Hence, data mining functionality was developed to tool-kits, which provide data mining functionality in form of a collection of different components. Depending on the different research questions of the users, the solutions consist of distinct compositions of these components.

Today, existing solutions for data mining processes comprise different modules that represent different steps in the analysis process. There exist graphical or script-based toolkits for combining such modules. Classical data mining tools, which can serve as modules in analysis processes, are based on single computer environments, local data sources and single users. However, analysis scenarios in medical- and bioinformatics have to deal with multi computer environments, distributed data sources and multiple users that have to cooperate. Users need support for integrating data mining into analysis processes in the context of such scenarios, which lacks today. Typically, analysts working with single computer environments face the problem of large data volumes since tools do not address scalability and access to distributed data sources. Distributed environments provide scalability and access to distributed data sources but the integration of existing modules into such environments is complex. In addition, new modules often cannot be directly developed in distributed environment. Moreover, in scenarios involving multiple computers, multiple distributed data sources and multiple users, reuse of modules and analysis processes becomes more important as more steps and configuration and thus much bigger efforts are needed to develop and set-up a solution.

In this paper, we introduce the field of scientific data analysis in bio- and medical informatics and present today's typical analysis scenarios in bioinformatics, including the roles of the user groups involved, the data sources that are available, and the analysis processes that are set up.

Subsequently, we present the challenges for data analysis processes in today's health information systems in the context of personalized medicine. Based on the challenges and requirements, we present the building blocks that can serve as basis for the development of the data mining environment in a system for personalized medicine.

## II. Scientific Data Analysis in bio- & medical informatics

Bioinformatics is conceptualizing biology in terms of macromolecules and applying information technology techniques from applied maths, computer science, and statistics to understand and organize the information associated with these molecules [9]. Typical research questions in bioinformatics are, e.g., finding predictive or prognostic biomarkers, defining subtypes of diseases, classifying samples by using gene signals, annotations, etc. In order to answer such questions, bioinformaticians combine different heterogeneous data sources from private or public repositories, apply different analysis methods to the information extracted from the repositories and interpret the results until they have found good combinations of data sources and analysis methods. This is what we call a scenario.

In the following, we will describe the data sources and repositories, techniques and analysis processes and the user and user groups that are typically involved in bioinformatics scenarios.

### A. Data Sources

Bioinformatics is an area in which analysis scenarios include huge amounts and different types of data. Analyses in bioinformatics predominantly focus on three types of large datasets available in molecular biology: macromolecular structures, genome sequences, and the results of functional genomics experiments such as expression data [9].

Recent advances in technology enable collecting data at more and more detailed levels [10], from organism level, organ level, tissue level up to cellular and even sub-cellular level [9], [11]. In detail, we can distinguish several abstraction levels in multi-cellular organisms:

- **Organism level**: an organism is the biological system in its wholeness, typically including a group of organs. Organism level related data is the clinical data, which usually comes from the hospital database manager.
- **Organ level**: an organ is a group of tissues that together perform a complex function. Organ level related data usually comes from the pathologist.
- **Tissue level**: tissues are groups of similar cells specialized for a single function. Tissue level data usually comes from the pathologist.
- **Cellular level**: in a multi-cellular organism such as a human, different types of cells perform different tasks. Cellular level related data usually is organized by the lab manager;
- **Sub-cellular level**: data at the sub-cellular level is composed by the structures that compose the cells. Usually, a data analyst can retrieve this data by performing ontologies analyses.

Additional information includes the content of scientific papers and "relationship data" from metabolic pathways, taxonomy trees, and protein-protein interaction networks [9]. Comprehensive meta-data describing the semantics of the heterogeneous and complex data are needed to leverage it for further research [12]. To address this issue, efforts exist for describing the data in a comprehensive way by domain specific ontologies [13].

Data from different sources is extensively collected in repositories potentially useful for subsequent analysis [10], [9], [11]. Some of these repositories are publicly available. Common public repositories include, e.g., the following:

- **PubMed**: literature [14]
- **GEO**: for genomics, proteomics and clinical data [15]
- **ArrayExpress** for genomics, proteomics and clinical data [16]
- **miRGen**: integrated database of miRNA targets and positional relationships [17]
- **miRBase**: searchable database of published miRNA sequences and annotation [18]

### B. Techniques and User Groups

Bioinformatics employs a wide range of techniques from maths, computer science and statistics, including sequence alignment, database design, data mining, prediction of protein structure and function, gene finding, expression data clustering, which are applied to heterogeneous data sources [9]. Bioinformatics is a collaborative discipline [10]. Bioinformaticians of today are highly qualified and specialized people from various backgrounds such as data-mining, mathematics, statistics, biology, IT development, etc. and a typical analysis scenario involves multiple users and experts from different departments or organizations. In projects in the bioinformatics area, bioinformaticians are working together with people from IT and with different collaborators:

- IT people usually support bioinformaticians by providing and helping with the needed computational power, network infrastructure and data sharing.
- Clinicians are usually the principal investigators (PIs) or co-PIs of the project and are the key point for patients information access and for experiment design planning.
- Pharmaceuticals Companies are funders and PIs. E.g., they are looking for a biomarker for a specific disease that they possibly could to commercialize a product at the end of the research project.
- Statisticians are specialists that can provide help on correctly analysing the data.
- Biologists are specialists that can provide help on correctly interpret the data. If they work on a wet

laboratory, they usually are key people on managing the samples and the related information.

### C. Analysis Processes

The common procedure for data analysis for scenarios from bioinformatics can be described in an abstract way as follows: There exist various data sources with different types of information. Based on the research question, data of different types is acquired from data repositories. For each data of a certain type there is a pre-processing done. After that, the data is merged. Based on this, the analysis is performed. The analysis step is repeated until a good result is found. Fig. 1 visualizes such a common analysis process from an abstract point of view. Analysis processes involve both manual and automated steps. Results of the analysis processes have to be interpreted to use them, e.g., for support to the clinical decision making process [19], [20].

When composing a solution to an analysis problem, bioinformaticians mainly work together with clinicians to provide the best possible solution to the project questions. The solution is typically composed of different modules. Many of them are recycled from previous solutions and often need to be adapted. Other modules will be designed and developed from scratch.

Implementation of bioinformatics scenarios is typically done with tools preferred by the bioinformaticians. Due to the various backgrounds, there is a quite heterogeneous set of tools and languages in use. Thus, analysis processes can be very different, depending of the type of data, the technology used, the tools used, the aim of the study etc. But, in the biomedical field, several steps are common to most of them: quality control, normalization, filtering, visualization, find differentially expressed probe sets, survival analysis.

## III. Challenges & Requirements

Today's data analysis scenarios in bioinformatics face the following challenges:

**Heterogeneous group of users in different locations:** In today's bioinformatics scenarios, users working at different locations have to collaborate. As a proof we can easily check the affiliations of the authors in a PubMed [14] paper. Bioinformaticians of today are from various backgrounds such as data-mining, mathematics, statistics, biology, IT development, etc. Thus, the scenarios involve a heterogeneous and distributed group of users (i.e. see http://www.vital-it.ch/about/team.php, http://bcf.isb-sib.ch/People.html, http://www.ebi.ac.uk/Information/Staff/). Depending on their background, knowledge and type of job, users can interact with an analysis environment in a different way and use different tools. E.g., some bioinformatics people might want to configure and run predefined workflows via simple form-based web pages. Other users might want to design new workflows based on existing components or reuse workflows from colleagues or they might want to develop new components by just writing their analysis algorithms in their own language of choice or use software from colleagues, and might want to integrate them into the system by writing a plug-in module for the code to run within the environment. Advanced users, e.g., might even want to partially modify the structure of the workflow environment. When multiple users work together at different locations and with different background, the set of tools used is also quite heterogeneous. However, the users typically do not have an overview over the full system and no detailed knowledge about all parts of the system. This is especially true if they are involved in huge projects, such as the developers and curators of UCSC [22] or Ensembl [23].

**Large, heterogeneous and distributed data sources:** Today, data is not longer mainly collected and evaluated with focus on a specific problem or study in the bioinformatics and healthcare domain. Instead, data is extensively collected from different sources in repositories potentially useful for subsequent analysis. As the type of analysis is not yet known at the time the data is collected, content and data format are not focused. Moreover, recent advances in technology allow for collecting data on more detailed levels. Thus, the volume of data of a certain type can become very large. In analysis scenarios in the context of bioinformatics lots of different data and data types are involved. People with different responsibilities and analysis questions work with different sets of data sources. The corresponding data sources are distributed by nature. There exist a large number of public data sources and repositories that are accessible via the internet. In addition, private data sources are distributed across several departments of a hospital or institute, or even across different hospitals or institutes. As a result, a huge amount of distributed data is available for usage. For these reasons the scenarios involve an increasing number of data sources and amount of data. Typically, bioinformatics scenarios include the development of a solution based on a certain restricted data repository and the evaluation on public available data or vice-versa. The semantic of the datasets is complex and needs to be described to allow a proper usage. Due to the heterogeneity and complexity of the data, several domain specific ontologies exist for the description of the semantics of the data by comprehensive meta-data.

**Multi computer environments:** Today's analysis scenarios have to deal with distributed and heterogeneous users as well as distributed and heterogeneous data sources. Instead of single-computer environments or environments hosted inside a certain organization, the scenarios involve users working with different tools and distributed data sources managed in different systems spread over the globe. In addition, today's data analysis applications in bioinformatics increase in complexity and in their demand for resources. To address this issue, solutions can be integrated into distributed environments that provide computing resources and allow for scalability, like for example deep sequencing data [24].

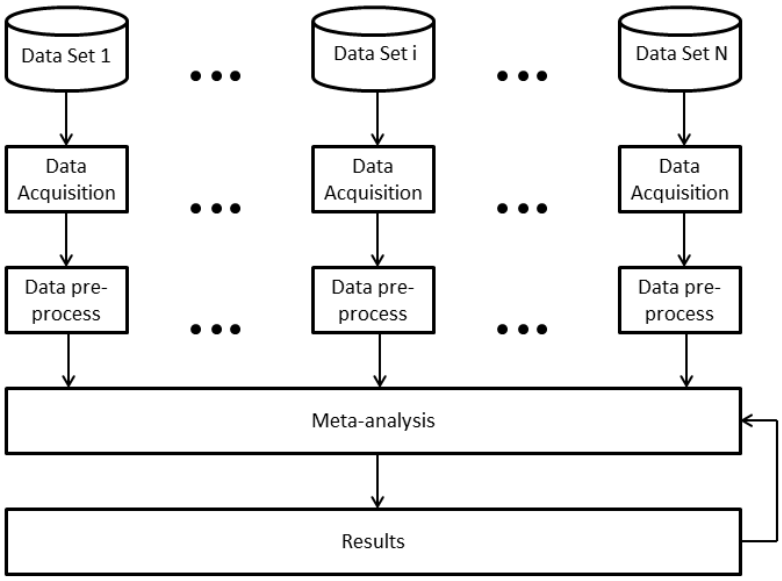


Figure 1. Common procedure in scenarios from bioinformatics.

**Complex process chains:** Content and data format of the data collected in the area of medicine and bioinformatics are not focused on a certain problem or research question. Thus, complex process chains are needed for the analysis of the data. Building specialized software for each analysis problem that is going to be addressed is not a solution, as this effort can only be carried out for huge projects running for several years. Thus, such process chains need to be supported by the environments that are used to set-up analysis solutions.

## IV. P-MEDICINE APPROACH

In this section, we will introduce our approach towards the data mining environment in the p-medicine system that aims at addressing the requirements presented above. First, we will present some lessons learned from prior projects. Second, we will show the building blocks that are foreseen for the data mining environment, Third, we will introduce methods for evaluating our system.

### *A. Lessons learned from prior projects*

In [29] and [28] we presented some lessons learned from building a data mining environment in the ACGT project (http://eu-acgt.org), which has the goal of implementing a secure, semantically enhanced end-to-end system in support of large multi-centric clinico-genomic trials. The various elements of the data mining environment can be integrated into complex analysis pipelines through the ACGT workflow editor and enactor. In the following, we will summarize the lessons learned.

As the construction of a good data mining process requires to encode a significant amount of domain knowledge, it cannot be fully automated. By reusing and adapting existing processes that have proven to be successful, we hope to be able to save much of this manual work in a new application and thereby increase the efficiency of setting up data mining workflows. While a multitude of tools for data mining, bioinformatics, and statistics on clinical data exists, the question of quality control and standardization, as well as ease of use and reusability, remains largely unanswered yet.

With respect to workflow reuse, we made the following experiences in setting up and running an initial version of the ACGT environment:

- The construction of data mining workflows is an inherently complex problem when it is based on input data with complex semantics, as it is the case in clinical and genomic data.
- Because of the complex data dependencies, copy and paste is not an appropriate technique for workflow reuse.
- Standardization and reuse of approaches and algorithms works very well on the level of services, but not on the level of workflows. While it is relatively easy to select the right parameterization of a service, making the right connections and changes to a workflow template is quickly getting quite complex, such that user finds it easier to construct a new workflow from scratch.
- Workflow reuse only occurs when the initial creator of a workflow describes the internal logic of the workflow in many details. However, most work ow creators avoid this effort because they simply want to ”solve the task at hand”.

In order to be able to meaningfully reuse data mining workflows, a formal notation is needed that allows to flexibly express both technical information about the implementation

of workflows, and high-level semantic information about the purpose and pre-requisites of a workflow.

In summary, the situation of having a large repository of workflows to chose the appropriate one from, which is often assumed in existing approaches for workflow recommendation systems, may not be very realistic in practice.

### B. Building blocks for the data mining environment

We identified a set of building blocks that can serve as basis for the p-medicine data mining environment:

- **Reusing available components**: a method for the integration and reuse of data mining components that have been developed in a single computer environment into distributed environments.
- **Developing new components**: a method for interactive development of data mining components in distributed environments.
- **Reusing existing analysis processes**: a method for the integration and reuse of data mining based analysis processes that involve several analysis steps.
- **GUI and system interfaces**: interfaces that address different levels of granularity for users to work with the system or to extend the system.

In the following, we will describe the building block in more details.

*1) Reusing available data mining components:* To supports users in using standard data mining modules and other available modules with small effort there is a need for an approach to integrate data mining modules being developed for a single processor environment into a distributed environment. We assume that there is not yet an existing comprehensive solution for the data mining problem, but that the data mining problem can be solved by using and correctly composing available data mining components.

In the DataMiningGrid and in the ACGT project, approaches for the modelling of the characteristics of data mining application and infrastructure principles for the integration of the data mining applications into distributed environments based on the modelling have been contributed [30], [28]. [30] presents a meta-data schema definition (Application Description Schema) as solution, which is used to grid-enable existing data mining applications. The ADS is used to manage user interaction with system components in order to grid-enable existing data mining applications, to register and search for available data mining components on the grid, to match analysis jobs with suitable computational resources, and to dynamically create user interfaces. The approach allows for an integration by users without deeper knowledge on the underlying distributed systems and without any intervention on the application side, and thus addresses the needs of the community to support users in using standard data mining tools and available components.

The GridR service [28] allows for reusing R script based data mining components. The underlying method of GridR reduces the complexity of integrating and handling analysis scripts in distributed environments. Instead of registering each single application as separate component in the environment, the method is technically based on a single service with complex inputs and outputs that allows for providing the algorithm as parameter.

*2) Developing new data mining components:* In addition to reusing components from single computer environments, users like bioinformaticians and biostatisticians typically need to interactively develop data mining components and services in the analysis environment interactively to allow for combining information from different data sources and applying different methodologies to the information extracted from these repositories.

In the ACGT project, a method for interactive development based on novel infrastructure principles that allow for profiting from the functionality and support of standardized tools and environments was contributed [28]. The approach supports the development of data mining solutions by the integration of data mining scripts and services into complex analysis systems and their processes. The GridR toolkit [31] is based on the approach for interactively developing data mining components and services in distributed environments. In addition to providing a single service as interface for the execution script based data mining components, the method allows for interactively developing data mining components and services in eScience environments based on extensions to the R environment that interface with the API of middleware components of distributed systems. The approach efficiently supports users when it is necessary to enhance available or develop new data mining components. Users are enabled to interactively develop data mining based data analysis processes directly within an distributed environment.

*3) Reusing existing analysis processes by Data Mining Patterns:* In today's analysis solutions in bioinformatics, complex process chains have to be set-up. The composition of such process chains is a huge effort. Thus, reuse of processes becomes much more important. However, analysis processes often cannot be used directly, as they are customized to a certain analysis question and the information on how the process was set-up and which requirements have to be met for applying the process is often not available. [32] contributes the concept of Data Mining Patterns. Data Mining Process Patterns allow for facilitating the integration and reuse of data mining in analysis processes. The underlying approach is based on encoding requirements and pre-requisites inside the analysis process and a task hierarchy that allows for generalizing and concreting tasks for the creation and application of process patterns. The data mining pattern approach supports users in scenarios that cover different steps of the data mining process or involve several analysis steps. Data mining patterns support the description of data mining processes at different levels of

abstraction between the CRISP model [33] as most general and executable workflows as most concrete representation. Hence, they allow for easy reuse and integration of data mining processes.

*4) GUI and System interfaces:* Today's environments for data mining in the context of bioinformatics scenarios have to support users to work with the system or to extend the system in different levels of granularity One of the reasons why a lot of current tools are not used by bioinformaticians is that they are a black box, i.e. it's not easy to modify the tools to new situations and requirements. There is a need for an open system which can be accessed in layers, depending on the wish of the user.

In conclusion from our experience a system is needed that allows the IT-wise people in the Institutes to modify it, allows the maths and stats people to plug in their models easily, and allows the biologist and clinicians not to see the analysis algorithms but still to understand what they are doing as, e.g., often they will need to justify it in grants. Other fundamental features would be that once a workflow is created, it can be used on data external to the specific repository, and that a web-page can be easily created for each workflow and customized by the bioinformatician. This web-page would contain for example links for the input and links for the output.

Such requirements could be addressed by taking over ideas from business process management systems such as jBPM or YAWL, which provide the functionality of exposing processes via web interfaces automatically and providing web interfaces for certain tasks in the processes which require human interaction.

### C. Evaluation Methods

The evaluation of the processes that involve data mining and data analysis, are implemented accordingly ISO [26] and IEEE [27] standards criteria mainly, and involves both users and developers in the testing process. When the data are made of human samples, like patients, normal tissues from donors or from the patients themselves, enrolled or not in clinical trials, GCP-compliant data management must be observed and satisfied [25].

Usually this kind of analyses have many goals that differ for each category of end-user (clinicians, data miners, bioinformaticians, statisticians, etc.), thus the achievement of the objectives needs an evaluation process based on realistic scenarios tailored on user's needs and requirements.

For these reasons a solution is considered good and proper when it successfully meet the evaluation and validation expectation criteria as well the usability ones that are part of the evaluation process.

Several instruments to help in developing a successful solution are available: ISO/IEC 2504n, it is the Quality Evaluation Division of the ISO/IEC 25000:2005, IEEE Std 1063-2001 Standard for software user documentation and IEEE Std 829-1998 Standard for software test documentation.

In general there are several main principles (gold standard) that developed software must satisfy to meet quality assurance criteria:

From a developer point of view:

- The produced software needs to be operational, interoperable with other components (if any) and compliant with the specifications;
- The quality model presented in the standard ISO/IEC 9126-1, classifies software quality in a structured set of characteristics: functionality, reliability, usability, efficiency, maintainability, portability

From the end-user side:

- Scenarios evaluation

When modular projects are developed, the evaluation is an iterative process where scenarios and evaluation procedures evolve as new components get integrated in the environment or as some others are removed; the process is similar to the discovery of a predictive model Fig. 2.

Based on an iterative evaluation process, the quality expectations for software systems are two fold:

- the software must do the right things: software systems must do what they are supposed to do (end-user perspective, validation process)
- the software must do the things right: software systems must perform the tasks correctly (developer perspective, verification process)

The results of a data analysis are usually verified and validated by using the following main procedures:

- Using large amount of data (multiple data sets that belong to different platform) and cross validation methods (mainly: K-fold, repeated random sub-sampling and LOO).
- Verification of benchmarks based on well knowns results, reproducibility of published results and comparison with existing databases
- In house experiments
- Software/tools tested for checking that the correct operation is executed for each planned feature

At the beginning of a project, during the preparation phase, the scenarios and the evaluation criteria can be also used as guidelines for developers to focus their effort towards actual and immediate end-users needs.

## V. Conclusion

In this paper, we presented our approach towards developing a data mining environment for personalized medicine. The approach aims at addressing the needs and requirements for applying data mining techniques to bioinformatics solutions in the context of the p-medicine project.

Challenges for today's bioinformatics scenarios are the heterogeneous set of users in different locations, the large,

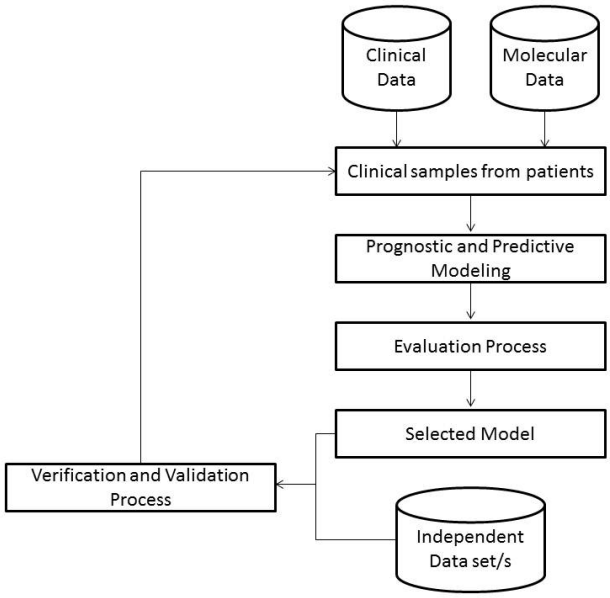


Figure 2. Clinical prognostic and predictive models require an iterative evaluation process.

distributed and heterogeneous data sources and multi-computer environments, and complex process chains for the analysis.

Our approach for addressing these challenges consists of 4 building blocks for the data mining environment:

- a method for reusing data mining components created in single computer environments.
- a method for developing data mining components in distributed environments.
- pattern-based approach for reusing analysis processes including data mining components.
- GUI and system interfaces that allow users to work with the system or to extend the system in different levels of granularity.

In detail, the heterogeneous data will be addressed by extensibility mechanisms and support for ontologies, heterogeneous users will be supported by website-like and expert interfaces as well as by the ability to reuse existing components and processes, and complex process chains will be addressed by the data mining pattern approach.

## ACKNOWLEDGMENT

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement N 270089.

## REFERENCES

[1] Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, Baehner FL, Walker MG, Watson D, Park T, Hiller W, Fisher ER, Wickerham DL, Bryant J, Wolmark N (2004) A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 351 281726 PMID: 15591335 DOI: 10.1056/NEJMoa041588

[2] van t Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415 530536.

[3] Farmer P, Bonnefoi H, Anderle P, Cameron D, Wirapati P, Wirapati P, Becette V, Andr S, Piccart M, Campone M, Brain E, Macgrogan G, Petit T, Jassem J, Bibeau F, Blot E, Bogaerts J, Aguet M, Bergh J, Iggo R, Delorenzi M (2009) A stroma-related gene signature predicts resistance to neoadjuvant chemotherapy in breast cancer. Nat Med 15 6874 PMID: 19122658 DOI: 10.1038/nm.1908

[4] Wistuba II, Gelovani JG, Jacoby JJ, Davis SE, Herbst RS (2011) Methodological and practical challenges for personalized can- cer therapies. Nat Rev Clin Oncol 8 13541 PMID: 21364686 DOI: 10.1038/nrclinonc.2011.2

[5] ood L, Friend SH (2011) Predictive, personalized, preventive, participatory (P4) cancer medicine. Nat Rev Clin Oncol 8 1847 PMID: 21364692 DOI: 10.1038/nrclinonc.2010.227

[6] Tursz T, Andre F, Lazar V, Lacroix L, Soria JC (2011) Implications of personalized medicineperspective from a cancer center. Nat Rev Clin Oncol 8 17783 PMID: 21364691 DOI: 10.1038/nrclinonc.2010.222

[7] Volinia S, Galasso M, Costinean S, Tagliavini L, Gamberoni G, Drusco A, Marchesini J, Mascellani N, Sana ME, Abu Jarour R, Desponts C, Teitell M, Baffa R, Aqeilan R, Iorio MV, Taccioli C, Garzon R, Di Leva G, Fabbri M, Catozzi M, Previati M, Ambs S, Palumbo T, Garofalo M, Veronese A, Bottoni A, Gasparini P, Harris CC, Visone R, Pekarsky Y, de la Chapelle A, Bloomston M, Dillhoff M, Rassenti LZ, Kipps TJ, Huebner K, Pichiorri F, Lenze D, Cairo S, Buendia MA, Pineau P, Dejean A, Zanesi N, Rossi S, Calin GA, Liu CG, Palatini J, Negrini M, Vecchione A, Rosenberg A, Croce CM (2010) Reprogramming of miRNA networks in cancer

and leukemia. Genome Res 20 58999 PMID: 20439436 DOI: 10.1101/gr.098046.109

[8] Nicoloso MS, Sun H, Spizzo R, Kim H, Wickramasinghe P, Shimizu M, Wojcik SE, Ferdin J, Kunej T, Xiao L, Manoukian S, Secreto G, Ravagnani F, Wang X, Radice P, Croce CM, Davuluri RV, Calin GA (2010) Single-nucleotide polymorphisms inside microRNA target sites influence tumor susceptibility. Cancer Res 70 278998 PMID: 20332227 DOI: 10.1158/0008-5472.CAN-09-3541

[9] Luscombe NM, Greenbaum D, Gerstein M.: What is bioinformatics? A proposed definition and overview of the field. Methods Inf Med 40 (2001): 346-58.

[10] Roos, D. S.: Bioinformatics: Trying to Swim in a Sea of Data. Science 291 (16 February 2001): 12601261. http://www.sciencemag.org/content/291/5507/1260.full

[11] Soinov, L.: Bioinformatics and Pattern Recognition Come Together. Journal of Pattern Recognition Research (JPRR), Vol 1 (1) 2006 p. 3741

[12] Weiler G, Brochhausen M, Graf N, Hoppe A, Schera F, Kiefer S: Ontology Based Data Management Systems for post-genomic clinical Trials within an European Grid Infrastructure for Cancer Research. In Proceedings of the 29th Annual International Conference of the IEEE EMBS, Lyon, France, August 23-26, 2007, pp. 6434-6437.

[13] Mathias Brochhausen, Andrew D. Spear, Cristian Cocos, Gabriele Weiler, Luis Martn, Alberto Anguita, Holger Stenzhorn, Evangelia Daskalaki, Fatima Schera, Ulf Schwarz, Stelios Sfakianakis, Stephan Kiefer, Martin Doerr, Norbert M. Graf, Manolis Tsiknakis: The ACGT Master Ontology and its applications - Towards an ontology-driven cancer research and management system. Journal of Biomedical Informatics 44(1): 8-25 (2011)

[14] pubmed: http://www.ncbi.nlm.nih.gov/pubmed/

[15] GEO: http://www.ncbi.nlm.nih.gov/geo/

[16] ArrayExpress: http://www.ebi.ac.uk/arrayexpress/

[17] M. Megraw, P. Sethupathy, B. Corda, and A.G. Hatzigeorgiou (2006). miRGen: A database for the study of animal microRNA genomic organization and function. Nucleic Acids Research, 35: D149-D155. http://www.diana.pcbi.upenn.edu/miRGen.html

[18] miRBase: Kozomara A, Griffiths-Jones S.: Integrating microRNA annotation and deep-sequencing data. NAR 2011 39(Database Issue): D152-D157 http://www.mirbase.org/

[19] Rossi S, Shimizu M, Barbarotto E, Nicoloso MS, Dimitri F, Sampath D, Fabbri M, Lerner S, Barron LL, Rassenti LZ, Jiang L, Xiao L, Hu J, Secchiero P, Zauli G, Volinia S, Negrini M, Wierda W, Kipps TJ, Plunkett W, Coombes KR, Abruzzo LV, Keating MJ, Calin GA. (2010) microRNA fingerprinting of CLL patients with chromosome 17p deletion identify a miR-21 score that stratifies early survival. Blood. 2010 Aug 12;116(6):945-52. PMID: 20393129

[20] Planche A, Bacac M, Provero P, Fusco C, Delorenzi M, Stehle JC, Stamenkovic I. (2011) Identification of prognostic molecular features in the reactive stroma of human breast and prostate cancer. PLoS One. 2011;6(5):e18640. PMID: 21611158

[21] Popovici V, Budinsk E, Delorenzi M. (2011) Rgtsp: a generalized top scoring pairs package for class prediction. Bioinformatics. 2011 Jun 15;27(12):1729-30. PMID: 21505033

[22] http://genome.ucsc.edu/

[23] http://www.ensembl.org/index.html

[24] Hawkins RD, Hon GC, Ren B. (2010) Next-generation genomics: an integrative approach. Nat Rev Genet. 2010 Jul;11(7):476-86.

[25] Ohmann C, Kuchinke W, Canham S, Lauritsen J, Salas N, Schade-Brittinger C, Wittenberg M, McPherson G, McCourt J, Gueyffier F, Lorimer A, Torres F; ECRIN Working Group on Data Centres. (2011) Standard requirements for GCP-compliant data management in multinational clinical trials. Trials. 2011 Mar 22;12:85

[26] http://www.iso.org/iso/home.html

[27] http://www.ieee.org/index.html

[28] Anca Bucur, Stefan Rüping, Thierry Sengstag, Stelios Sfakianakis, Manolis Tsiknakis, Dennis Wegener, The ACGT project in retrospect: Lessons learned and future outlook, Procedia Computer Science, Volume 4, Proceedings of the International Conference on Computational Science, ICCS 2011, 2011, Pages 1119-1128, ISSN 1877-0509.

[29] Rüping, Stefan and Wegener, Dennis and Bremer, Philipp. Reusing Data Mining Workflows. In: Proceedings of the ECML PKDD 2010 Workshop on Third-Generation Data Mining: Towards Service-Oriented Knowledge Discovery (SoKD'10), Barcelona, Spain, 2010, pp. 25-30.

[30] Stankovski, Vlado and Swain, Martin and Kravtsov, Valentin and Niessen, Thomas and Wegener, Dennis and Kindermann, Jrg and Dubitzky, Werner. Grid-enabling data mining applications with DataMiningGrid: An architectural perspective. Future Generation Computer Systems 24 (4), 2008, pp. 259-279.

[31] Wegener, Dennis and Sengstag, Thierry and Sfakianakis, Stelios and Rping, Stefan and Assi, Anthony. GridR: An R-based tool for scientific data analysis in grid environments. Future Generation Computer Systems 25 (4), 2009, pp. 481-488.

[32] Dennis Wegener and Stefan Rping. Integration and reuse of data mining in business processes a pattern-based approach. Int. J. Business Process Integration and Management, Vol. 5 (3), pp. 218-228 (2011)

[33] Shearer, C.: The CRISP-DM model: the new blueprint for data mining. Journal of Data Warehousing, Vol. 5 , Nr. 4, pp. 13–22 (2000)