Building a System for Advancing Clinico-Genomic Trials on Cancer

Stelios Sfakianakis, Norbert Graf, Alexander Hoppe, Stefan Rüping, Dennis Wegener, and Lefteris Koumakis

Abstract The analysis of clinico-genomic data poses complex computational problems. In the project ACGT, a grid-based software system to support clinicians and bio-statisticians in their daily work is being developed. Starting with a detailed user requirements analysis, and with the continuous integration of usability analysis in the development process, the project strives to develop an architecture that will substantially improve the way clinico-genomic trials are conducted today. In this paper, results of the initial requirements analysis and approaches to address these requirements are presented. We also discuss the importance of appropriate metadata to tailor the system to the needs of the users.

1 Introduction

The goal of the Advancing Clinico-Genomics Trials on Cancer (ACGT[1]) project is to develop an open-source and open access IT infrastructure that provides the biomedical research community with the tools needed to integrate complex clinical information and make a concrete step towards the tailorization of treatment to the patient[4]. The necessity of such an environment is evident today more than ever due to the recent advancements in

Stelios Sfakianakis · Lefteris Koumakis
Institute of Computer Science, FORTH, Greece, e-mail: {ssfak,koumakis}@ics.forth.gr

Norbert Graf · Alexander Hoppe
University Hospital of Saarland, Paediatric Haematology and Oncology, D-66421 Homburg, Germany, e-mail: {norbert.graf,alexander.hoppe}@uniklinikum-saarland.de

Stefan Rüping · Dennis Wegener
Fraunhofer IAIS, Schloss Birlinghoven, 53754 St. Augustin, Germany, e-mail: {stefan.rueping,dennis.wegener}@iais.fraunhofer.de

[1] http://www.eu-acgt.org

high throughput genomics and post-genomics technologies. These technologies yield an enormous pool of data that needs to be managed, analysed, correlated, and comprehended for the treatment of diseases like cancer and for the benefit of the community at large.

In this paper we discuss how the clinical requirements for such an environment can be addressed in a large-scale system and how appropriate meta data can be used to achieve satisfaction of the end user. The rest of the paper is structured as follows: Section 2 gives an overview over the user requirements analysis that was conducted in ACGT and highlight the main challenges. Section 3 introduces the main ACGT architecture, and in Section 4 some new approaches to address the user requirements are presented. Section 5 concludes.

2 The End-User View

Treatment and survival of patients with cancer is increasing steadily for most age groups as shown in Fig. 1 which gives the average annual percentage change over a period of 10 years. One of the most important reasons for this success story is the enrollment of patients in prospective clinical trials. Nevertheless, for most clinical trials in cancer, the number of patients recruited is much lesser than the number of eligible patients. In adults only 5% of cancer patients are participating in such trials. Therefore, higher rates are of utmost importance, especially in those cancers with still a dismal prognosis. To achieve this goal it is necessary to facilitate the building and running of clinical trials and to attract more patients to participate. In addition, the improvement in molecular biology has to be taken into account to create more clinico-genomic trials.

Recent advances in methods and technologies in molecular biology have resulted in an explosion of information and knowledge about cancer and its treatment. As a result, our ability to characterize and understand the various forms of cancer is growing exponentially. Information arising from post-genomics research and combined genetic and clinical trials on one hand, and advances from high-performance computing and informatics on the other, is rapidly providing the medical and scientific community with an enormous opportunity to improve prognosis of patients with cancer by individualizing treatment. To achieve this goal, a unifying platform is needed that has the capacity to process this huge amount of multi-level and heterogeneous data in a standardized way. Multi-level data collection within clinico-genomic trials and interdisciplinary analysis by clinicians, molecular biologists and others involved in life science is mandatory to further improve the outcome of cancer patients. It is essential to merge the research results of biomolecular findings, imaging studies and clinical data of patients and to enable users to easily join, analyze and share even great amounts of data. To provide a functional

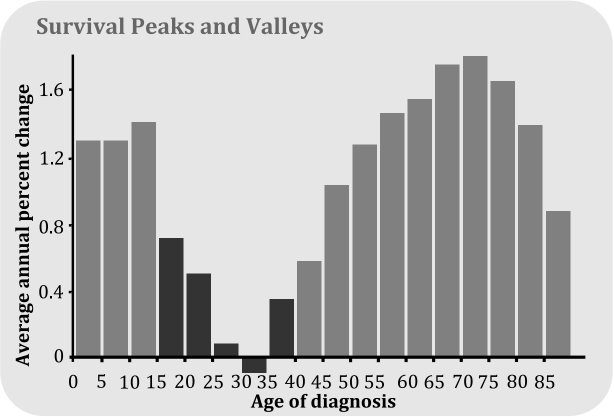

Fig. 1 Average annual change in survival in patients with cancer [5].

and user-friendly platform it is of utmost importance that the development of such a platform is user-driven and evaluated by end users right from the planning and development phase. Tools and software developed within ACGT are based on the user's needs and have to be in accordance with ethical and legal requirements of the European Community.

The project has selected indicative Clinical Trials on Cancer, namely breast cancer, pediatric nephroblastoma and in-silico modeling and simulation of tumor growth and response to treatment, for the initial requirements gathering activity. Since ACGT sees the requirements engineering process as a structured set of activities which will lead to the production of the final system requirements, an iterative requirements engineering process has been adopted, mainly based on scenarios and prototyping. Inputs to the requirements engineering process are information about existing systems, user and stakeholder needs, organizational standards, regulations and other domain information. As clinico-genomic trials are in the center of ACGT a Clinical Trial Management System is of utmost importance, to collect clinical, biomedical, imaging and other trial specific and relevant data. ACGT will provide such a tool, called ObTiMA [7, 8], whose functionality includes administrative and scientific aspects of clinico-genomic trials.

It has to be stressed that such a complex platform as ACGT, dealing with extremely sensitive data (patient data) and used by many different, sometimes multi-role, end-users, having different needs and requirements, a Data Protection Framework for ACGT is mandatory. This is based on the anonymization of patient data, the informed consent from participating patients and the binding of partners/centers by contracts to the ACGT policies and procedures and will ensure compliance with the Data Protection regulations. In addition and from an ethical point of view it is strongly demanded to let patients participate in and have a measure of influence over the processing of their genetic data.

To assure the success and functionality of the ACGT environment end-users are involved in every step of the development process.

2.1 Main Requirements

In summary, the user requirements that have been identified by the clinical experts for the ACGT environment can be divided into the following aspects:

- Appropriateness: the data analysis environment should provide the appropriate tools and services to support users in the state-of-the-art scientific analysis of biomedical data. Section 4.1 introduces the use of the GridR component [9], a "gridified" version of the well-known R statistical software, which is a de-facto standard for many kinds of biomedical data analysis.
- Extensibility and reusability: the platform should be easily extensible to new tasks and existing solutions should be easily reusable and transferable to similar problems. Extensibility is addressed in Section 4.1, while an important aspect of reusability, namely quality control, is described in Section 4.5
- Performance: the system must be performant enough to facilitate large analysis and optimization tasks, which calls for an efficient use of the grid architecture. Challenges exists not only because of the size of the data sets (see Section 4.2), but also from their complexity and heterogeneity (Section 4.3), which is a result of the distributed nature of pan-european clinical trials.
- Security: The system must be secure and protect the privacy of the involved patients. This is discussed in Section 4.4.
- Usability: the system should be easy to use for inexperienced users, but also provide a powerful interface for experts. Usability is best achieved by a continuous process of evaluation and optimization. In Section 4.6, approaches to automatically identify parts of the system that require a high amount of attention are discussed.

3 The ACGT Architecture

The complexity and the diversity of user requirements have a strong impact on the design of the ACGT architecture. It is evident that a multidisciplinary and multiparadigm approach is necessary in order to deal with these requirements. For these reasons the ACGT platform is designed according to the

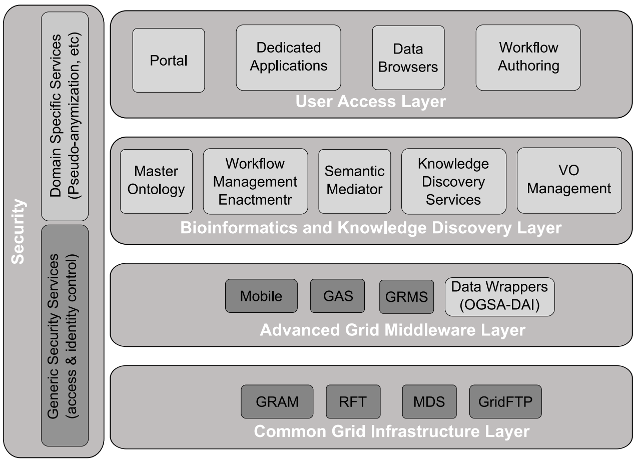
Security
Domain Specific Services (Pseudo-anymization, etc)
Generic Security Services (access & identity control)
Portal
Dedicated Applications
Data Browsers
Workflow Authoring
User Access Layer
Master Ontology
Workflow Management Enactmentr
Semantic Mediator
Knowledge Discovery Services
VO Management
Bioinformatics and Knowledge Discovery Layer
Mobile
GAS
GRMS
Data Wrappers (OGSA-DAI)
Advanced Grid Middleware Layer
GRAM
RFT
MDS
GridFTP
Common Grid Infrastructure Layer

Fig. 2 The ACGT layered architecture

following technologies and standards: Service Oriented Architecture (Web Services), the grid, and the Semantic Web. In essence, the grid provides the computational and data storage infrastructure, the general security framework, the virtual organization abstraction and relevant user management mechanisms etc. The machine to machine communication is performed via XML programmatic interfaces over web transport protocols, which are commonly referred as Web Services interfaces. Finally the Semantic Web adds the knowledge representation mechanisms through the means of OWL ontologies, the implementation-neutral query facilities with the SPARQL "universal" query language and the associated query interfaces.

The adopted architecture for ACGT is shown in Fig. 2. A layered approach has been followed for providing different levels of abstraction and a classification of functionality into groups of homologous software entities. In this approach we consider the security services and components to be pervasive throughout ACGT so as to provide both for the user management, access rights management and enforcement, and trust bindings that are facilitated by the grid and domain specific security requirements like pseudonymization. Apart from the security requirements, the grid infrastructure and other services are located in the first (lowest) two layers: the Common Grid Layer and the Advanced Grid Middleware Layer. The upper layer is where the user access services, such as the portal and the visualization tools, reside. Finally, the Bioinformatics and Knowledge Discovery Services are the "workhorse" of ACGT and the corresponding layer is where the majority of ACGT specific services lie.

4 Addressing the User Requirements in ACGT

In the following, we will try to give a short overview of how to integrate the user requirements of Section 2 into the ACGT grid architecture.

4.1 Extensibility

The requirement for extensibility is very important in the context of grid-enabled data mining [11]. Especially, in order to keep track with new scientific developments, it is crucial to be able to quickly integrate new analysis services or algorithms into a data mining platform. Related to the ACGT environment, extensibility denotes the possibility of extending the environment at the workflow level, at the service level, or at the algorithm level. In order to deal with such requirements we have found that the use of metadata descriptions and the ontology based integration of the ACGT platform components provides a future proof approach to extensibility. In the following we will introduce GridR [9] as an example to demonstrate how the ACGT system can easily be extended by new services and algorithms.

GridR is an analysis tool based on the statistical environment R [3] that allows using the collection of methodologies available as R packages in a grid environment. The aim of GridR is to provide a powerful framework for the analysis of clinico-genomic trials involving large amount of data (e.g. microarray-based clinical trials). The GridR service (see Fig. 3) combines the wide spectrum of methods available in R with an effective distributed grid data management system (DMS) and efficient execution supported by a grid resource management system (GRMS), see [10]. In this fashion, users can make efficient use of distributed, parallel computational resources in their R scripts, while all the technical details are hidden from them. The R code to be executed can be given directly by the user in the form of a script, but in order to increase the possibility of distributing and re-using code, the intended way to execute R code is by storing it in a metadata repository, such that it becomes available to the whole system. Technically, an R function or script f thus becomes an $f-$service. Consequently, users who prefer to work on the workflow level and not edit their own code can make use of available R scripts and even all the single R functions in R libraries in their workflows.

Along these lines new algorithms can be "gridified" and be seamlessly integrated with the rest of the ACGT grid environment without a need for changing the service's or the R script's implementation.

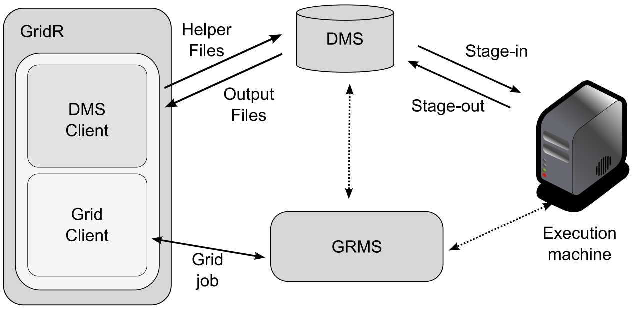

Fig. 3 GridR Architecture

4.2 Large Data Sets

Data is the most valuable asset of ACGT and therefore the platform should be able to manage big data sets in an efficient and secure way. The storage and the transfer of the data is of particular importance and something that should be taken care in a uniform way in the whole ACGT environment.

Data storage requirements are addressed by the grid infrastructure and the ACGT "Data Grid", which controls the sharing and the management of large amounts of distributed data. However, an additional issue has to do with the protocols, infrastructure, and policies for moving these large data sets to the processing nodes where the data analysis is performed. In some cases the grid infrastructure could be employed so that instead of moving the data around, the processing tasks, by the means of grid job submission and scheduling services, are transferred where the data reside. Nevertheless, the majority of services and data processing tools in ACGT are implemented as XML Web Services that are accessible through the network. Being a text format, XML is well known for its unfriendliness for transferring binary data. There are a couple of solutions for this ranging from encoding the binary data in hexadecimal or, most often, in Base64 text format, to using "attachments" in the SOAP messages. Nevertheless these approaches impose additional processing and bandwidth costs and so we opt for another option, which is to transmit *references to data* as part of the Web Services interaction while the data itself can be transferred through "out of band" channels, e.g. by the means of GridFTP. This approach offers the advantage of "quicker" XML interactions, easier and more performant service composition since there is no need to "get" (download) a huge binary data set in order to "give" (upload) it to another service, identity of the data so that they can associated with metadata through their references, etc.

4.3 Complex Data

A particular characteristic of data analysis in clinical trials is that the data used in a statistical analysis can be very heterogeneous and dynamic, meaning that many different tools and approaches may be necessary to analyze the data, but also that intermediate results may become invalid as the trial progresses and more data becomes available. The situation is further exacerbated when improved interactivity can result in changing workflows on the fly and cloning running workflows to explore alternatives in parallel. This involves the risk that the user becomes overwhelmed by the enormous amount of information and choices that are available to him. Hence, approaches to help the user better deal with the possibilities of the system are necessary.

For these reasons, a hyperlinked presentation of information has been proposed as a tool for better supporting the collaboration in scientific communities [6]. In essence, provenance information can be viewed as a graph of services invocations, with edges representing several types of lineage and provenance. For example, relationships such as "produced-by", "part-of", "derived-from", "input-of" etc. can be modeled this way. Each entity (e.g. service, data) is identified by an HTTP URI to provide identification, retrieval, and linking facilities for constructing a web of data and metadata in accordance with the Semantic Web vision [1]. Therefore in ACGT we aim to employ the semantic web technology in order to facilitate the tasks of both the users and the intelligent knowledge extraction services. Users are able to navigate to the information graph formed by the casual and other relationships between and among services and data just by following the hyperlinking paradigm that was popularized by the World Wide Web. On the other hand, semantic web enabled software entities are empowered to take advantage of the semantically rich content and to draw conclusions and knowledge based on the referenced ontologies.

4.4 Security

The sensitivity of the patient data requires a strong security framework to provide enough safety nets in order to maintain privacy, confidentiality, and integrity. The grid middleware already supports much of the necessary infrastructure, in terms of certificate based Grid Security Infrastructure (GSI), the Virtual Organization (VO) abstraction and the user credential management, and the Grid Authorization Services (GAS). In ACGT this "system level" security is complemented by "domain specific" mechanisms like *pseudomymization* that permits the identification of patient specific information without revealing the true person identity. All data is anomymized before their entry in ACGT and even during their analysis all the processing tasks are audited and authorized based on the end users' identity[2].

An interesting assertion about security of services can be made when using tool repositories as described in Section 4.1: with a standard web service, which can be deployed anywhere, it is principally impossible to give technical guarantees about which code is executed, as only the interface of the service is given and standardized. In general this is a desired property of web services, however, when considering data security, this means that external (legal) measures have to be taken to prohibit the service owner from disclosing information about the data. With the use of tool repositories, it can be guaranteed by a central instance, that the code in the repository has been reviewed and is secure to use, because the code that is being executed is directly transported to the execution site from the repository. In addition, this shipment of algorithms allows to analyze the data on a secure site, without needing to transport sensitive data.

4.5 Quality Control

In dynamic, distributed, and heterogeneous environments with multiple actors and complex use cases it is important to have a continuous validation of the different functional components. Therefore an ACGT validation and testing infrastructure is required to constantly monitor the ACGT services and report any malfunctions. This infrastructure for the automatic testing and validation of ACGT workflows and services is useful both for the initial decision making process about the acceptance of a new service and for the monitoring the status of the ACGT services as a whole. The status of ACGT services and workflows is checked with respect to the following criteria:

- Liveness, i.e. that it's "alive" and normally operating
- Correctness, i.e. that it delivers the correct results
- Performance, i.e. that it responds in a timely fashion

A number of tests are developed as scripts for each service according to these criteria. These tests are of course service and workflow specific because different components have different notions of correctness or performance. Nevertheless all of them are given some sample input data and parameters and based on this information they validate the target services according to the services' interface and functionality. The tests are stored centrally and re-evaluated periodically.

The advantage of this testing scenario is that even complex, user-defined workflows can be tested periodically, such that a single user can be notified if a workflow (i.e. a scientific experiment) of hers fails to meet the expected results. In this way, not only software quality, but also the quality of published, clinically relevant findings can be controlled.

4.6 Usability

Much thought in the ACGT project is given to the usability of the final software, including a formal usability analysis and end user integration throughout the runtime of the project to guarantee that the software will meet the requirements of the end users. Usability analysis is an important, but very time consuming process. In this section, we will present some approaches on how to improve the usability of the system using information present in the system's meta data.

The idea is that workflow execution statistics can be gathered together with other meta data and put into relation with the user's content with the system. For example, an analysis of workflows which are often canceled can provide the system's administrator with valuable information on how help users to select a better workflow. A statistic of the execution time of different services can help a developer to choose which services to optimize. A list of often used services can help new users to select good services.

Hyper-linking between meta data, workflow templates and workflow statistics also allow for a more complex reasoning of the users intent. One example could be as follows: the user executes different workflows on a data set, or variants thereof. From the meta data of the data set the system finds out that all the variants of the data set point to the same basic data set (e.g. a trial) and hence can reason that all the workflow executions belong together. It can then search the database of historic workflow executions to see whether a similar groups of workflow have been executed by another user. If this is the case, it is reasonable to assume that both users try to solve a similar problem, and hence the best workflow of the old user can be suggested to the new user. Of course, privacy aspects have to be considered in this kind of scenario.

5 Conclusions

There are a number of projects that aim at developing grid-based infrastructure for post-genomic cancer clinical trials, the most advanced of which are NCI's caBIG[2] in the USA and CancerGrid[3] in the UK. The overall approach in those projects is somewhat different from the one in ACGT. In caBIG, the bottom-up, technology-oriented, approach was chosen, in which the focus was put on the integration of a large number of analysis tools but with weak concern on data privacy issues. CancerGrid on the other hand addresses the very needs of the British clinical community. In contrast, the goal of the ACGT project is develop a pan-european system that is driven by current demands

[2] Cancer Biomedical Informatics Grid, https://cabig.nci.nih.gov/

[3] http://www.cancergrid.org/

from clinical practice. With two on-going international clinical trials actually conducted in the framework of the project, the approach is top-down, with clinicians' and biomedical data analysts' needs at the heart of all technical decisions, considering data privacy issues as central as data analysis needs.

In this user driven endeavor the technical concerns raised by the multiplicity and heterogeneity of user requirements demand state of the art methodologies and technologies. In the ACGT work plan the employment of ontologies and metadata annotations and the realization of intelligent higher level services are the primary implementation targets. Finally, in the realization of this environment, we aspire that the users are also involved. Guided and facilitated by the infrastructure, they can actively participate by creating and sharing information and knowledge. Only this way the ACGT is enriched and improved to become a really useful scientific tool.

Acknowledgments:

The authors wish to thank the ACGT consortium for their contributions and various ideas on which the ACGT project was developed. The ACGT project is funded by the European Commission (FP6/2004/IST-026996).

References

1. Berners-Lee, T.: Linked Data Design Issue. http://www.w3.org/DesignIssues/-LinkedData.html
2. B. Claerhout, N. Forgo, T. Krügel, M. Arning, and G. De Moor (2008): A Data Protection Framework for Transeuropean genetic research projects. Studies in health technology and informatics, vol. 141, 67.
3. R Development Core Team (2008) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org
4. Rüping, S., Sfakianakis, S., Tsiknakis, M. (2007): Extending Workflow Management for Knowledge Discovery in Clinico-Genomic Data. Proc. Healthgrid 2007: From Genes to Personalized HealthCare: Grid Solutions for the Life Sciences 183–193
5. Couzin, J. (2007): Survival in young adults with Cancer. Science **317**
6. Goble, C.,Gocho, O., Alper, P., De Roure, D. (2006): e-Science and the Semantic Web: A Symbiotic Relationship. 9th Int. Conf. on Discovery Science (DS2006), Springer , 1–12.
7. Graf, N., Weiler, G., Brochhausen, M., Scherer, F., Hoppe, A., Tsiknakis, M., Kiefer, S., Aran Lunzer, Yuzuru Tanaka (2007) : The importance of an ontology based clinical data management system (OCDMS) for clinico-genomic trials in ACGT. SIOP 2007, 39th International Society Of Pediatric Oncology Annual Meeting, Mumbai, India
8. Weiler, G., Brochhausen, M., Graf, N., Hoppe, A., Schera, F., Kiefer, S. (2007): Ontology Based Data Management Systems for post-genomic clinical Trials within an European Grid Infrastructure for Cancer Research. SIOP 2007, 39th International Society Of Pediatric Oncology Annual Meeting, Mumbai, India

9. Wegener, D., Sengstag, T., Sfakianakis, S., Rüping, S., Assi, A. (2007): GridR: An R-based grid-enabled tool for data analysis in ACGT clinico-genomic trials. Proc. 3rd Intl. Conf. on e-Science and Grid Computing (eScience 2007), Bangalore, India
10. Pukacki, J., Kosiedowski, M., Mikolajczak, R., Adamski, M., Grabowski, P., Jankowski, M., Kupczyk, M., Mazurek, C., Meyer, N., Nabrzyski, J., Piontek, T., Russell, M., Stroinski, M., Wolski, M. (2006): Programming Grid Applications with Gridge. Computational Methods in Science and Technology **12**.
11. Wegener, D., May, M. (2007): Extensibility of Grid-Enabled Data Mining Platforms: A Case Study. Proc. of the 5th International Workshop on Data Mining Standards, Services and Platforms, San Jose, California, USA, 13–22.