ACGT: Advancing Clinico-genomic trials on cancer – Four years of experience

Luis MARTIN[a,1], Alberto ANGUITA[a], Norbert GRAF[b], Manolis TSIKNAKIS[c], Mathias BROCHHAUSEN[d], Stefan RÜPING[e], Anca BUCUR[f], Stelios SFAKIANAKIS[g], Thierry SENGSTAG[h], Francesca BUFFA[i] and Holger STENZHORN[b]

[a] *Biomedical Informatics Group, Universidad Politécnica de Madrid, Spain*
[b] *Department of Paediatric Oncology and Haematology, Saarland University Hospital, Germany*
[c] *Biomedical Informatics Laboratory, FORTH, Greece*
[d] *IFOMIS, Saarland University, Germany*
[e] *Fraunhofer IAIS, Germany*
[f] *Philips Research Europe, The Netherlands*
[g] *Institute of Computer Science, FORTH, Greece*
[h] *RIKEN Yokohama Institute, Japan*
[I] *The Weatherall Institute of Molecular Medicine, University of Oxford, UK*

Abstract. The challenges regarding seamless integration of distributed, heterogeneous and multilevel data arising in the context of contemporary, post-genomic clinical trials cannot be effectively addressed with current methodologies. An urgent need exists to access data in a uniform manner, to share information among different clinical and research centers, and to store data in secure repositories assuring the privacy of patients. Advancing Clinico-Genomic Trials (ACGT) was a European Commission funded Integrated Project that aimed at providing tools and methods to enhance the efficiency of clinical trials in the -omics era. The project, now completed after four years of work, involved the development of both a set of methodological approaches as well as tools and services and its testing in the context of real-world clinico-genomic scenarios. This paper describes the main experiences using the ACGT platform and its tools within one such scenario and highlights the very promising results obtained.

Keywords. Clinical trials, semantic mediation, ontologies, knowledge discovery on databases, workflows

1. Introduction

Advances in research methodologies and technology during the last decade have resulted in a rapid increase of information about cancer in general. Still, heterogeneity of infrastructures and data within clinical and research institutions has limited the ability to extract useful knowledge and to apply it to treatment regimens. Current post-genomic clinical trials often rely on ad-hoc built information systems for handling the

[1] Luis Martín: PhD Student, Group of Biomedical Informatics, Universidad Politécnica de Madrid, Campus de Montegancedo s/n, 28660 Boadilla del Monte, Spain; E-mail: lmartin@infomed.dia.fi.upm.es.

generated data, each based on their own formats and standards. Therefore solutions have to be devised and provided that allow sharing of information gathered in one trial with another, or incorporating external data from disparate sources during the trial if this is required. In addition, guaranteeing the privacy of collected patient data is always an inherently difficult issue. All these tasks further require that some level of syntactic and semantic homogeneity is established for data.

The vision of the Advancing Clinico-Genomic Trials on Cancer (ACGT) project (www.eu-acgt.org) was to tackle the above issues by developing a semantically rich grid infrastructure platform in support of multicentric, postgenomic clinical trials, thus enabling discoveries in the laboratory to be quickly transferred to the clinical management and treatment of patients [1].

2. The ACGT Platform

In order to be able to deal with the complexities of research and management of cancer, it was obvious that a highly elaborate, yet easy to use, technical infrastructure had to be developed. Features such as intuitive access for end-users, coherent content organization and consistence with the way the different user groups carry out their daily work were mandatory. A thorough design and development has led to the construction of a powerful and versatile ontology-driven grid infrastructure named the ACGT Platform (available from http://purl.org/acgt/portal) (Figure 1). This platform comprises a set of tools and services that cover the requirements described above.

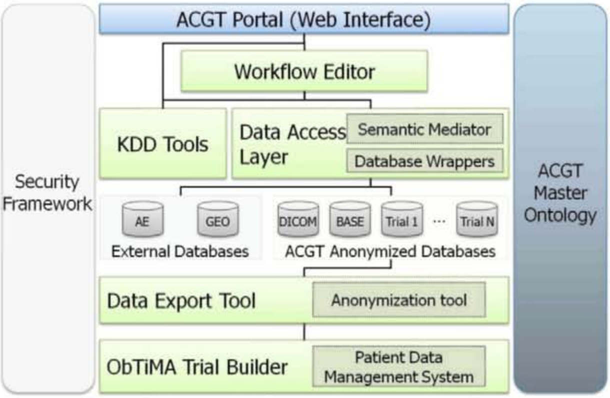

Figure 1. The ACGT Platform. On top, the web interface provides access to the underlying tools (KDD tools, workflow editor). These tools access a set of heterogeneous databases offered by the data access layer. The data of these sources is properly anonymized. The trial builder allows running new clinical trials in the platform.

One main focus while designing and developing the ACGT Platform was to ensure data privacy. Data handled in clinical trials are sensible, and the different legislative bodies therefore impose very strict regulations on this aspect. To achieve this objective, all patient's sensible data are initially pseudonymized with dedicated tools ensuring that no patient will be identifiable through the data exposed in the platform. Strong security features such as credential-based data access were added to all platform tools and services thus achieving security in the context of large, distributed data processing.

2.1. The ACGT Master Ontology on Cancer

The ACGT Master Ontology on Cancer (ACGT MO) was developed with the goal of creating a consistent semantic framework to comprehensively describe the domain of post-genomic clinical trials on cancer. This framework is the basis of the semantic interoperability for connecting the different services and data sources in the ACGT Platform. It is written in OWL-DL and contains more than 1600 classes and around 200 properties. The state-of-the-art design principles of the OBO Foundry (http://www.obofoundry.org/) were fundamental in the ontology development. This also includes that well-established ontologies covering parts of the domain were reused as a whole or partly, such as the Foundational Model of Anatomy [2], the OBO Relational Ontology [3] and the Basic Formal Ontology [4]. Other relevant ontologies and terminologies were not directly included since they miss the expected quality criteria but were still used as knowledge source [5].

2.2. Clinical Trial Designer - ObTiMA

ObTiMA is an ontology-based system for creating and conducting clinical trials [6]. It includes a graphical Trial Builder that aids the trial chairman in the design of the Case Report Forms (CRFs) to be used to document each treatment step [7]. The interface allows defining CRF content and layout to capture all relevant patient data during a trial. The resulting descriptions are based on ACGT MO concepts for each CRF item along with metadata, like data type and measurement unit, to setup the trial database.

The second major functionality is the patient data management system. It is automatically set-up based on the items defined in the design phase and guides the user through the treatment of the individual patients according to the defined treatment plans. The MO aids in providing the necessary semantic interoperability so that these data are accessible from other components of the ACGT Platform.

2.3. Data Access Layer

An important challenge in current post-genomic biomedical research is to efficiently manage and retrieve data from heterogeneous sources. In order to provide seamless data access, syntactic and semantic integration needs to take place. The Data Access Layer, comprised by the Database Wrappers (DWs) and the Semantic Mediator (SM) [8] offers this functionality within the ACGT Platform. The DWs deal with the syntactic heterogeneities, offering a uniform interface to the data resources. This includes uniformity of transport protocol, message syntax, data format (RDF), and query language (SPARQL). The SM tackles semantic heterogeneities—i.e. offering a common data model for accessing the data resources exposed by the DWs. The ACGT MO was adopted as the model exposed to clients of the Data Access Layer. Incoming queries in terms of the MO are translated by the SM and redirected to the DWs, with the results being integrated and presented to the client as a single result set.

2.4. KDD Tools

The ACGT Platform comprises a series of knowledge discovery tools for analyzing and extracting useful information from data collected in a clinical trial. With an abundance of such tools available freely, BioMoby [9] and R/Bioconductor [10] being prominent examples, the focus was not set on the development of new tools but rather on seamlessly integrating those existing toolkits in a uniform fashion.

The R language was adopted as the prime tool for carrying out statistical analysis of the data. The GridR tool [11] allows the seamless execution of R jobs in parallel to facilitate the efficient development, execution, and re-use of analytical solutions without the need of knowledge about the underlying architecture on the analyst side.

2.5. The ACGT Workflow Environment

To assist bioinformaticians in creating their complex scientific workflows, a Workflow Editor and Enactment Environment, called WEEE [12], was implemented and made accessible through the ACGT Portal, thus allowing users to combine different web services into complex workflows. An intuitive user interface permits searching registered services—e.g. GridR scripts—and retrieving data through the Data Access Layer. These elements can then be combined and orchestrated to produce workflows that can be subsequently stored in a user's specific area and later retrieved and edited.

Workflows are executed on a remote machine or in clusters in the Grid so there is no burden imposed on the user's local machine. The publication and sharing of workflows is also supported so that the user community can exchange information benefitting from each other's research. WEEE is based on the BPEL workflow standard [13] and supports the BPEL representation of complex bioinformatics workflows.

3. Evaluation: the MCMP Scenario

Validation of the ACGT platform was performed in the context of clinically oriented data analysis scenarios. One such was the MCMP (Multi Center Multi Platform) scenario, with the goal of validating the utility of the platform as an information system to exploit data in the context of clinical trials. The setup consisted of a set of biopsies collected by two institutions using the microrarray platforms Affymetrix and Illumina. The related clinical data were stored in a corresponding clinical trial database. All patient private data were anonymized prior to their inclusion in the ACGT environment.

The process began by associating database concepts to concepts from the ACGT MO—i.e. appropriate semantic mappings were set-up. This allowed retrieving integrated information from the data sources in a homogeneous manner. After that, we constructed and executed the bioinformatics workflow. This workflow, which implemented a methodology linking microarrays and classical clinical data for biomarker discovery, illustrated the capacity of the ACGT platform to repeat complex analysis on an evolving population of patients. This included data retrieval and integration, normalization, analysis and results presentation.

4. Conclusions and Future Work

When launched back in 2006, the ACGT project aimed at providing clinical researchers with an infrastructure to support the requirements of modern clinical trials. From data collection and integration, to workflow design and result analysis, initial studies in the project detected some major points of interest for the area. There were specific needs to cover to alleviate end-users from the most resource-consuming tasks in their daily work.

The combination of thorough analysis of scenarios, research on previously proposed solutions and an extensive tool and service development led, after four years of work, to the completion of the ACGT Platform. Intensive testing within real-world scenarios provided highly promising results. The ontology-driven data integration approach, combined with a focus on user-friendliness, proved to be a key factor in the successful deployment of the infrastructure. Future research will focus on facilitating the integration of external services and its utilization in clinical trial environments. Exploitation, maintenance and sustainability of the infrastructure are the current focal areas in the context of follow-up research and development projects.

Acknowledgements. This research has been supported by the European Commission funded projects ACGT (IST-2005-026996), p-medicine (FP7-ICT-2009-270089) and INTEGRATE (FP7-ICT-2009-270253).

References

[1] Tsiknakis M, Brochhausen M, Nabrzyski J, Pucacki J, Sfakianakis SG, Potamias G, et al. A Semantic Grid Infrastructure Enabling Integrated Access and Analysis of Multilevel Biomedical Data in Support of Postgenomic Clinical Trials on Cancer. *IEEE transactions on information technology in biomedicine*: a publication of the IEEE Engineering in Medicine and Biology Society. 2008 Mar;12(2):205-217.

[2] Rosse C, Mejino JL. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. *Journal of biomedical informatics*. 2003 Dec;36(6):478-500.

[3] Smith B, Ceusters W, Klagges B, Köhler J, Kumar A, Lomax J, et al. Relations in biomedical ontologies. *Genome Biology*. 2005;6(5):R46+.

[4] Smith B, Brochhausen M. Putting biomedical ontologies to work. Methods of information in medicine. 2010 Mar;49(2):135-140.

[5] Brochhausen M, Spear AD, Cocos C, Weiler G, Martín L, Anguita A, et al. The ACGT Master Ontology and its applications - Towards an ontology-driven cancer research and management system. Journal of biomedical informatics. 2011 Feb;44(1):8-25.

[6] Weiler G, Brochhausen M, Graf N, Schera F, Hoppe A, Kiefer S. Ontology based data management systems for post-genomic clinical trials within a European Grid Infrastructure for Cancer Research. Conference proceedings : Annual International Conference of the IEEE Engineering in Medicine and Biology Society IEEE Eng. in Medicine and Biology Society Conference. 2007;2007:6435-6438.

[7] Stenzhorn H, Weiler G, Brochhausen M, Schera F, Kritsotakis V, Tsiknakis M, Kiefer S, Graf N. The ObTiMA System – Ontology-based Managing of Clinical Trials. Stud Health Technol Inform. 2010;160(Pt 2):1090-4.

[8] Martín L, Anguita A, de la Calle G, García-Remesal M, Crespo J, Tsiknakis M, Maojo V. Semantic data integration in the European ACGT project. AMIA Annu Symp Proc. 2007 Oct 11:1042.

[9] Wilkinson MD, Links M. BioMOBY: an open source biological web services proposal. Briefings in bioinformatics. 2002 Dec;3(4):331-341.

[10] Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome biology. 2004;5(10):R80.

[11] Wegener D, Sengstag T, Sfakianakis S, Rueping S, Assi A. GridR: An R-based tool for scientific data analysis in grid environments. Future Generation Computer Systems. 2009 Apr;25(4):481-488.

[12] Sfakianakis S, Koumakis L, Zacharioudakis G, Tsiknakis M. Web-based Authoring and Secure Enactment of Bioinformatics Workflows. 4th International Workshop on Workflow Management. 2009 May;2009:88-95.

[13] Web Service Business Process Execution Language Version 2.0 Specification, OASIS Standard; cited: 29 April 2011. Available from: http://docs.oasis-open.org/wsbpel/2.0/OS/wsbpel-v2.0-OS.html.