MediGRID - Medical Grid Computing

Sax U[1], Viezens F[1], Mohammed Y[1], Lingner Th[2], Morgenstern B[2], Vossberg M[3], Krefting, D[3], Rienhoff O[1]

[1]*Department of Medical Informatics, University of Göttingen, Germany*
[2]*Department of Bioinformatics, University of Göttingen, Germany*
[3]*Department of Medical Informatics, Charité, Humboldt University Berlin, Germany*
fred.viezens@med.uni-goettingen.de

Introduction

The project MediGRID [1] combines research institutes from various areas of Medicine, Biomedical Informatics, and other Life Sciences. Numerous associated partners from industry, healthcare and scientific institutions ensure a broad representation of this large community. The main goal of MediGRID is the development of a grid-middleware-platform with Globus Toolkit 4 as a basis for eScience Services for the community and to help researchers to use these services. Here we present the first results of the projects consisting of the infrastructure, first community applications and basic privacy rules.

Materials and Methods

Concerning the data flow in grids, most projects are similar in the lower layers (s. Fig. 1), but the biomedical community has to face particular challenges in the upper layers. The top layer represents the heterogeneous biomedical data sources. Beyond the problem to find the relevant data sets via metadata description, access control to the data is of paramount importance, as the owner of the data are foremost the patients. Due to the heterogeneity of the data we need an additional ontology layer to homogenize the data. Given semantic interoperability the researcher can correlate and analyze the data with biomedical informatics methods. Biomedical data in medical grids are heterogeneous, contain different kinds of information and have different levels of privacy. The data might include information about [2]:

- Population: Epidemiology
- Diseases: Clinical practice, clinical trials
- Patient data: Health record, clinical history, physical exams, lab/imaging studies
- Organ/tissue: pathology
- Cellular: histology
- Molecular: genetic test results and genomic data.

Having these data online with the suitable tools to connect, combine and analyze creates new challenges for data protection and privacy.

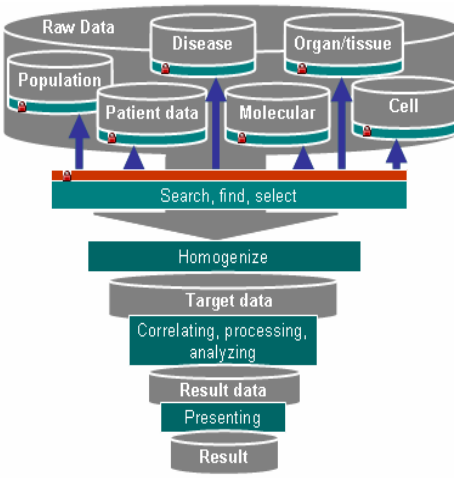

Fig. 1: Data Flow in MediGRID: The top layer contains the raw data, which has to be selected, accessed, homogenized and correlated in order to gain and present a result.

Results

The current privacy concepts do not cover the aspects and abilities of grid computing. Especially the re-identification risk with the combination of different data types has to be assessed. There are severe privacy concerns related to genomic-wide association studies [3-5]. These are the main reasons for the development of an enhanced security concept for MediGRID. Four methodological modules are responsible to construct the suitable infrastructure: ontology, resource fusion, middleware and eScience. On the other hand, three research modules take the initiative to use this national grid infrastructure to assist their work: biomedical Informatics, image processing and clinical research. The MediGRID consortium developed a middleware component as a connector for the medical community to resources of the integration project in D-Grid [6]. The community modules use this middleware to "gridify" their applications in order to show the advantage of grid computing in medicine.

Ontology

Using OGSA-DAI as a standard of Data Access and Integration in grids, the ontology module has successfully developed an ontology tool and implemented it as a first step to be a Gridsphere-Portlet in the MediGRID portal being available for all project partners.

Bioinformatics

Dialign is a widely used software tool for multiple alignments of nucleic acid and protein sequences. Within the MediGRID portal a parallelized version of the software is used to speed up the computationally expensive procedure. In that way distributed computing allows the user to obtain high-quality alignments of bigger databases and longer sequences. *AUGUSTUS* is a program that predicts gene structures in eukaryotic genomic sequences with high accuracy. Since *AUGUSTUS* performs a successive analysis of overlapping sequence sections, it is easy to parallelize. Therefore users benefit from distributed computing with several instances of the program. Grid resources also allow frequent update and centralized storage of huge EST databases.

Image Processing

A *3D image reconstruction for prostate biopsy* will register the different ultrasonic scans helping the physician to have a new viewing of the prostate that was not possible using the traditional 2D methods. The *virtual vascular surgery* helps to calculate and present the animated 3D blood flow field in the brain vessels, which could be used to anticipate the pressure on the walls of the vessels and for example to predict a bleeding risk. The *Brain 4D MRI image processing* application in MediGRID assists the identification of the brain areas research. All three applications are demanding the possibility of massive data volume storage and processing. Because of the constant increase of the data volume coherent with the development in the imaging techniques, a dynamic extendable computing and storage infrastructure is needed, which ideally will be a grid environment.

Privacy

In the first step we are working with non person-related data and set up the first generation of privacy rules. As a preparation for the next phase – dealing with person related data – we defined the additionally necessary advanced security methods like audit and tracking. Furthermore some legal issues concerning virtual organizations and the ownership of data and material have to be addressed.

Further perspective

As we gained experience with our first medical grid applications, we will be able to "gridify" other community applications more easily. The initial incarnation of the infrastructure is set up; some details are still to come. As genotyping constantly gets cheaper, many formerly phenome-related projects around complex diseases consider genotyping within the next couple of years. Beyond the indisputable opportunities of these studies there are quite some challenges to be faced. MediGRID addresses issues like the homogenization of heterogeneous data sources and how do we deal with the well-known privacy problems. Solving those issues will enhance the portability of life science grid services to other projects and other communities.

References

1 www.medigrid.de
2 Martin-Sanchez, F., V. Maojo, and G. Lopez-Campos, Integrating genomics into health information systems. Methods Inf Med, 2002. 41(1): p. 25-30.
3 Butte, A.J. and I.S. Kohane, Creation and implications of a phenome-genome network. Nat Biotechnol, 2006. 24(1): p. 55-62.
4 Lin, Z., A.B. Owen, and R.B. Altman, Genetics. Genomic research and human subject privacy. Science, 2004. 305(5681): p. 183.
5 Kohane, I.S. and Altman R.B., Health-Information Altruists — A Potentially Critical Resource. NEJM, 2005. 353 (19): p. 2074-2077
6 www.d-grid.de

This work was supported by the D-Grid Project MediGRID, funded by the Federal Ministry of Education and Research (BMBF), FKZ 01AK803A-H.