MediGRID
Medical Grid Computing

Sax U.[1], Mohammed Y.[1], Viezens F.[1], Lingner Th.[2], Morgenstern B.[2],Hartung M.[3], Rienhoff O.[1]

[1]Department of Medical Informatics, University of Göttingen, [2]Department of Bioinformatics, University of Göttingen, [3]Institute of Informatics, University of Leipzig

Introduction

The project MediGRID 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 as a basis for eScience Services for the community and to help researchers to use these services.

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. Finally the result data can be presented.

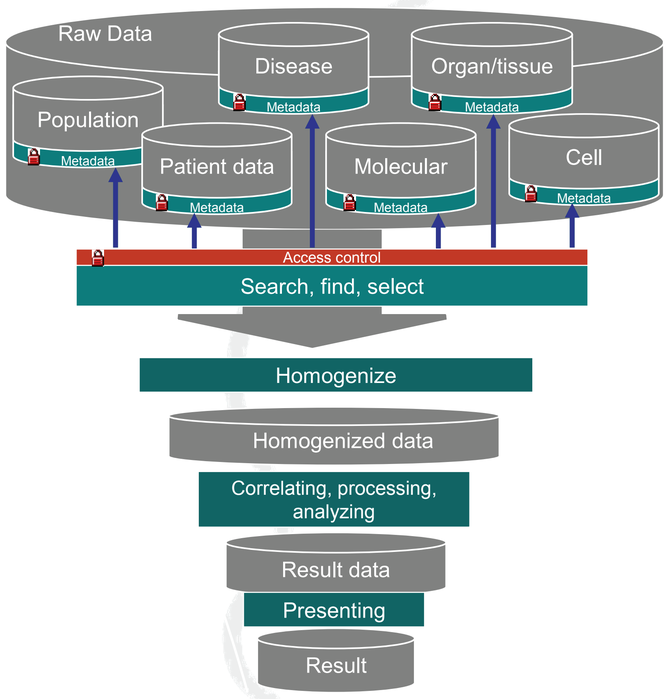

Fig.1: Data Flow in MediGRID

Biomedical data in medical grids are heterogeneous, contain different kinds of information and have different levels of privacy. The data might include information about [1] :

- 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. 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 [2-4].
These are the main reasons for the development of an enhanced security concept for MediGRID (s. Fig. 2).

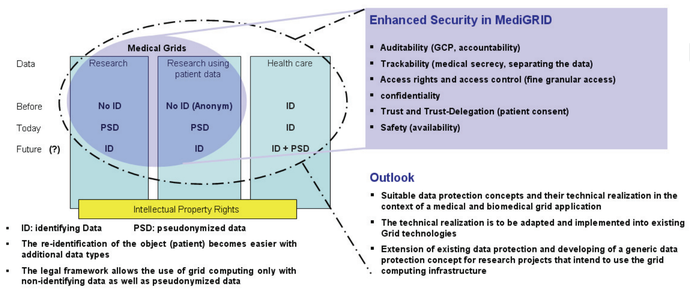

Fig.2: Enhanced Security in MediGRID

As a partner within the German national D-Grid initiative [5], MediGRID consists of seven modules aiming to cover the different aspects in the medical community (s. Fig. 3). Four methodological modules are responsible to construct the suitable infrastructure: ontology, ressource 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.

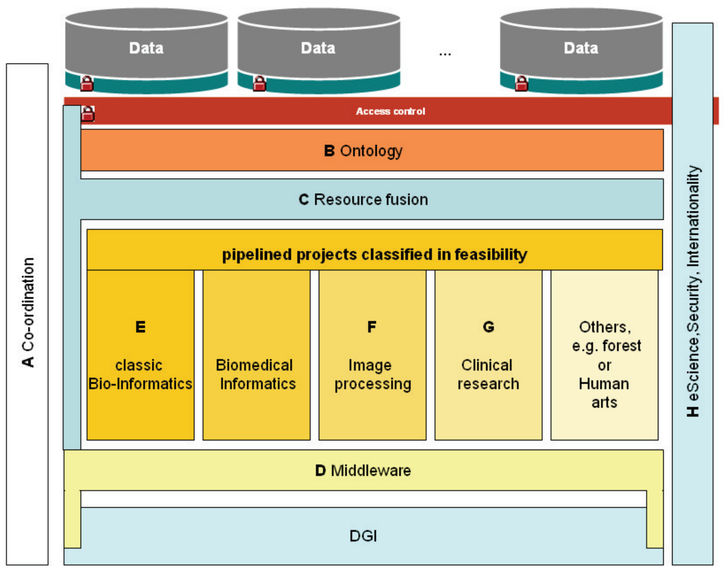

Fig. 3: Structure of the MediGRID project

What's special?

- Data heterogeneity (data from cell to population level)
- Lack of semantic interoperability
- Multidimensional data vs. low case numbers (NP-hard)
- Data Protection and privacy challenges by correlating different data types

First results

Ontology

The successful use of grid services as well as sharing and accessing the data in a medical grid environment needs semantical interoperability. 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 portlet in the MediGRID portal being available for all project partners (s. Fig. 4).

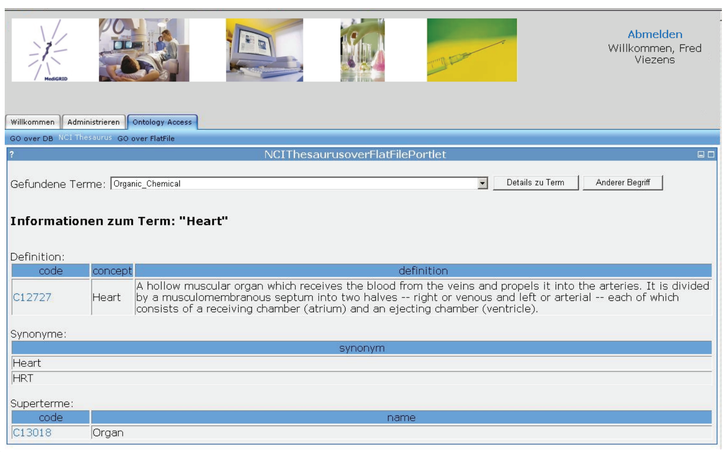

Fig.4: The Ontology Tool integrated as Gridshpere portlet

Bioinformatics

Dialign is a widely used software tool for multiple alignment of nucleic acid and protein sequences (s. Fig. 5). The program is based on local segment-by-segment comparison; this strategy is often superior to more traditional global alignment methods, in particular if distantly related sequences share only local homologies. However, since the program is slower than most global alignment methods, its applicability is limited to data sets of moderate size.

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.
By means of grid portlets the software is easy to use and readily accessible.

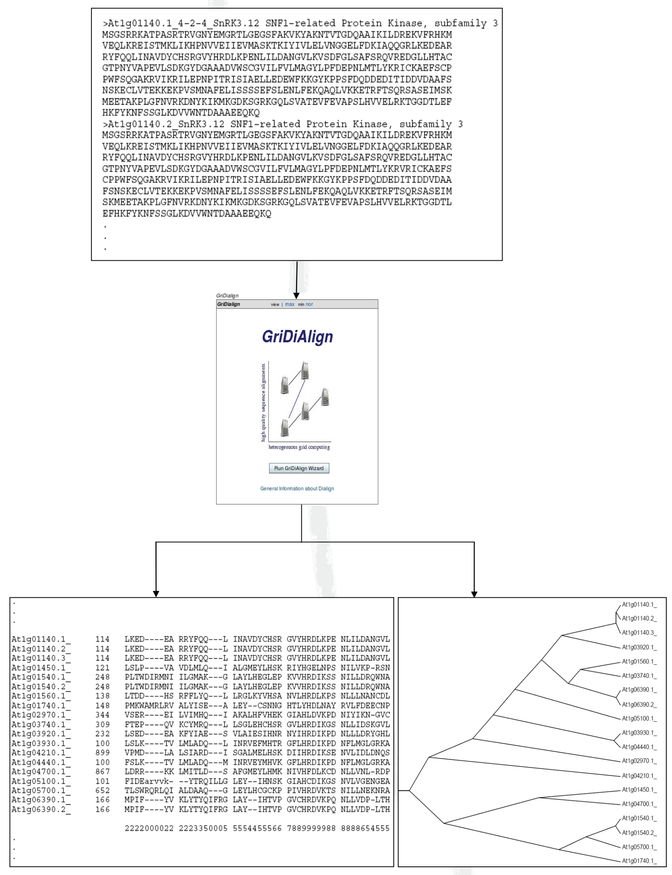

Fig.5: Dialign

AUGUSTUS is a program that predicts gene structures in eukaryotic genomic sequences with high accuracy.
The input DNA sequences are searched ab initio or with hints generated from existing EST (expressed sequence tag) databases. In the latter case a sizeable amount of data will be generated and processed.
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.

Further perspective

As we gained experience with our first medical grid applications, we will be able to "gridify" other communities and applications more easily. The initial incarnation of the infrastructure is set up, the details are still to come. As suggested in figure 1, the big picture includes projects like genome-wide association studies. These studies are becoming more and more interesting for the biomedical community. 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 Martin-Sanchez, F., V. Maojo, and G. Lopez-Campos, Integrating genomics into health information systems. Methods Inf Med, 2002. 41(1): p. 25-30.
2 Butte, A.J. and I.S. Kohane, Creation and implications of a phenome-genome network. Nat Biotechnol, 2006. 24(1): p. 55-62.
3 Lin, Z., A.B. Owen, and R.B. Altman, Genetics. Genomic research and human subject privacy.
Science, 2004. 305(5681): p. 183.
4 Kohane, I.S. and Altman R.B.,
Health-Information Altruists —
A Potentially Critical Resource.
NEJM, 2005. 353 (19): p. 2074-2077
5 www.d-grid.de

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung