TECHNOLOGY AND INFORMATICS 126

From Genes to Personalized HealthCare: Grid Solutions for the Life Sciences

Proceedings of HealthGrid 2007

Editors: Nicolas Jacq, Henning Müller, Ignacio Blanquer, Yannick Legré, Vincent Breton, Dominique Hausser, Vicente Hernández, Tony Solomonides and Martin Hofmann-Apitius

IOS Press

FROM GENES TO PERSONALIZED HEALTHCARE: GRID SOLUTIONS FOR THE LIFE SCIENCES

Studies in Health Technology and Informatics

This book series was started in 1990 to promote research conducted under the auspices of the EC programmes' Advanced Informatics in Medicine (AIM) and Biomedical and Health Research (BHR) bioengineering branch. A driving aspect of international health informatics is that telecommunication technology, rehabilitative technology, intelligent home technology and many other components are moving together and form one integrated world of information and communication media. The complete series has been accepted in Medline. Volumes from 2005 onwards are available online.

Volume 126

Recently published in this series

ISSN 0926-9630

From Genes to Personalized HealthCare: Grid Solutions for the Life Sciences

Proceedings of HealthGrid 2007

Edited by

Nicolas Jacq

HealthGrid, France

Henning Müller

University and Hospitals of Geneva, Switzerland

Ignacio Blanquer

Universidad Politécnica de Valencia, Spain

Yannick Legré

CNRS-IN2P3, France

Vincent Breton

CNRS-IN2P3, France

Dominique Hausser

Observatoire technologique, Switzerland

Vicente Hernández

Universidad Politécnica de Valencia, Spain

Tony Solomonides

University of the West of England, Bristol, United Kingdom

and

Martin Hofmann-Apitius

Fraunhofer Institut für Algorithmen und Wissenschaftliches Rechnen SCAI, Germany

IOS Press

Amsterdam • Berlin • Oxford • Tokyo • Washington, DC

ISBN 978-1-58603-738-3
Library of Congress Control Number: 2007924025

Publisher
IOS Press
Nieuwe Hemweg 6B
1013 BG Amsterdam
Netherlands
fax: +31 20 687 0019
e-mail: order@iospress.nl

Distributor in the UK and Ireland
Gazelle Books Services Ltd.
White Cross Mills
Hightown
Lancaster LA1 4XS
United Kingdom
fax: +44 1524 63232
e-mail: sales@gazellebooks.co.uk

Distributor in the USA and Canada
IOS Press, Inc.
4502 Rachael Manor Drive
Fairfax, VA 22032
USA
fax: +1 703 323 3668
e-mail: iosbooks@iospress.com

PRINTED IN THE NETHERLANDS

From Genes to Personalized HealthCare: Grid Solutions for the Life Sciences
N. Jacq et al. (Eds.)
IOS Press, 2007

Introduction

HealthGrid 2007 (http://geneva2007.healthgrid.org) is the fifth edition of this open forum for the integration of grid technologies and their applications in the biomedical, medical, and biological domains to pave the path towards an international research area in the HealthGrid field. The main objective of the HealthGrid conference and of the HealthGrid Association is the exchange and discussion of ideas, technologies, solutions and requirements that interest the grid and the life-science communities to foster the integration of grids into health. Participation is encouraged for grid middleware and grid application developers, biomedical and health informatics users, and security and policy makers to participate in a set of multidisciplinary sessions with a common focus on the application in the health domain.

HealthGrid conferences have been organized on an annual basis. The first conference, held in 2003 in Lyon (http://lyon2003.healthgrid.org), reflected the need to involve all actors – physicians, scientists and technologists – who might play a role in the application of grid technology to health, whether healthcare or bio-medical research. The second conference, held in Clermont-Ferrand in January 2004 (http://clermont2004.healthgrid.org) reported research and work in progress from a large number of projects. The third conference in Oxford (http://oxford2005.healthgrid.org) had a major focus on the results and deployment strategies in healthcare. The fourth conference in Valencia (http://valencia2006.healthgrid.org) aimed at consolidating the collaboration among biologists, healthcare professionals and grid technology experts. This fifth conference will focus on the five domains defined by the European Commission as application areas for grids in the biomedical field: molecules, cells, organs, individuals, and populations. For each of these five domains, an invited speaker will give a state of the art address followed by concrete projects. The conference venue in a hospital setting should also help to locate healthgrid research where it undoubtedly belongs, in the biomedical field. Potential users need to be shown that grids have now gone beyond hype and can show concrete applications that demonstrate the success of the technology.

The conference includes a number of high-profile keynote presentations complemented by a set of high quality peer reviewed papers. The number of contributions to this conference has increased from previous occasions, reaching the number of 55 submissions of papers, demonstrations and posters from principal authors coming from 21 countries (ordered by the number of contributions: Italy, Switzerland, United Kingdom, Germany, USA, France, Spain, Poland, Russia, Taiwan, Australia, Belgium, Brazil, Cuba, Cyprus, Czech Republic, Greece, Hungary, Portugal, Romania, and Ukraine). Considering the affiliations of all the authors of the papers, the number of contributing countries is extended to 22 countries with Venezuela. These proceedings have been organized in eight chapters. Five chapters focus on state of the art of the grid research and use at molecule, cell, organ, individual and population levels. Two chapters present security and imaging papers. The last chapter includes the best poster contributions.

From among the themes of the conference, it may be thought that molecules present the most amenable target for a healthgrid approach. Although this may be true, it is far from obvious. Even in applications such as BLAST which treat symbolic 'words' in a molecular alphabet, there are a great many problems to be addressed, both when searching for large query strings and in seeking to maintain an appropriate 'rollback' distributed database, as Trombetti et al. show. Two papers, by Andreas Quandt et al. and by Zosso et al., apply grids in 'tandem mass spectrometry' a highly demanding application for protein identification. It is equally good to see applications to screening in Malaria (Jean Salzemann et al.) and early detection of Alzheimer's (Nabil Abdennadher et al.), while the application to sleep disorders in their full breadth (Sebastian Canisius et al.) demonstrates the considerable breadth of application that is possible on a grid platform. Hernandez et al. provide an overview of the substantial progress that has been made in the field in the wake of large platform projects.

By contrast, applications in the cellular and organ domains appear to present as much difficulty in conceiving possible projects as in devising grid deployment schemes. Sinnott et al. identify and are in part motivated by the intrinsic value of microarray data as a major issue, while Roberta Alfieri et al. consider a mathematical model of the cell cycle – two extremes in the exploration of cell issues. Marienne Hibbert et al. describe a molecular medicine model and Emerson and Rossi discuss a simulation of the human immune system. Among organs, Andrés Gómez et al. demonstrate a radiotherapy tool and Blanquer Espert et al. explore the management of DICOM objects on a grid.

The domain of the individual is highly attractive, since it provides an entry point to the whole HealthGrid project and the possible grid health record. Two papers by the SHARE collaboration provide outline 'road maps' for the creation of viable healthgrids. Jenny Ure et al. took schizophrenia as a case study for a series of workshops to understand what it might take to create appropriate ontologies for data integration. Michal Kosiedowski et al. discuss a grid-based Electronic Medical Library and Shu-Hui Hung et al. consider the merits of the treatment of asthma in a grid system. Stefan Rüping et al. seek to extend workflow management for knowledge discovery from combined clinical and genomic data towards a fuller electronic health record. Giovanni Aloisio et al. treat the paradigm of Service-Oriented Architecture in bioinformatics.

Thus we arrive at the population domain, where Fabricio Silva et al. report on a grid-based epidemic surveillance system in Brazil for essentially slow evolving diseases. Tiberiu Stef-Praun et al. present the Swift system of workflow description and execution.

Apart from these five domains, two topical areas are considered in depth: security and imaging. In security, in two separate papers Jean Herveg and Luigi Lo Iacono explore familiar questions of data protection and of pseudonymization respectively, providing a theoretical perspective against which some practical proposals by Petr Holub et al. on the Access Grid model and Harald Gjermundrød et al. on the EGEE model can be compared. In imaging, we have papers on Globus MEDICUS (Stephan G. Erberich et al.) which federates DICOM devices through a grid architecture and KnowARC (Henning Müller et al.) on facilitating grid networks for the biomedical research community. Finally, Adina Riposan et al. report on the successful use of multimodal workflows in diabetic retinopathy research.

ACKNOWLEDGEMENTS

The editors express their gratitude to the program committee and the reviewers; each paper was read by at least two reviewers, including the editors. The editors want to thank for the remarkable work that the staff of the HealthGrid association has invested in these conference proceedings and on the organisation of the conference, particularly Yannick Legré.

Opinions expressed in these proceedings are those of individual authors and editors, and not necessarily those of their institutions.

Healthgrid 2007 Programme Committee (last name alphabetical order)

Chairman:

Martin Hofmann-Apitius	Fraunhofer Institut für Algorithmen und Wissenschaftliches Rechnen SCAI, Germany

Howard Bilofsky	University of Pennsylvania, USA
Ignacio Blanquer	Universidad Politécnica de Valencia, Spain
Vincent Breton	Centre National de la Recherche Scientifique, France
Anca Bucur	Philips Research, The Netherlands
Jose Maria Carazo	Centro Nacional de Biotecnología, Spain
Bastien Chopard	University of Geneva, Switzerland
Manuel Doblaré	University of Zaragoza, Spain
Alejandro Frangi	Pompeu Fabra University, Spain
Fabrizio Gagliardi	Microsoft, Switzerland
Dominique Hausser	Observatoire technologique, Switzerland
Patricia Hernandez	Swiss Institute of Bioinformatics, Switzerland
Vicente Hernandez-Garcia	Universidad Politécnica de Valencia, Spain
Chun-Hsi Huang	University of Connecticut, USA
Emmanuel Ifeachor	University of Plymouth, United Kingdom
Alfonso Jaramillo	École Polytechnique, France
Nicolas Jacq	HealthGrid, France
Wang Jun	Beijing Genomic Institute, China PRC
Mary Kratz	University of Michigan, USA
Peter Kunszt	Swiss National Supercomputing Center, Switzerland
Jimison Lavindrasana	University of Geneva, Switzerland
Yannick Legré	Centre National de la Recherche Scientifique, France
Richard McClatchey	University of the West of England, Bristol, United Kingdom
Fernando Martin-Sanchez	Institute of Health Carlos III, Madrid, Spain
Carlos Martinez	Generalitat Valenciana, Spain
Luciano Milanesi	Consiglio Nazionale delle Ricerche, Istituto di Tecnologie Biomediche (CNR-ITB), Italy
Steffen Moeller	University of Lübeck, Germany
Johan Montagnat	Institut National de Recherche en Informatique et Automatique, France
Georges de Moor	State University of Ghent, Belgium
Henning Müller	University and Hospitals of Geneva, Switzerland
Luis Núñez de Villavicencio	Universidad de los Andes, Venezuela
Armando Padhila	Faculdade de Engenharia da Universidade do Porto, Portugal
Christian R. Pellegrini	University of Geneva, Switzerland
Michael Podvinec	University of Basel, Switzerland
Simon Robinson	Empirica GmbH, Germany
Andres Santos	Universidad Politécnica de Madrid, Spain
Ferran Sanz	Universitat Pompeu Fabra, Spain
Tony Solomonides	University of the West of England, Bristol, United Kingdom

Antonio Sousa	Instituto de Engenharia Electrónica e Telemática de Aveiro, Portugal
Alfredo Tirado-Ramos	University of Amsterdam, The Netherlands
Petra Wilson	CISCO, Belgium
Zhiwei Xu	Institute of Computing Technology, China PRC

This page intentionally left blank

Contents

I. State of the Art of the Grid Research and Use at Molecule Level

This page intentionally left blank

From Genes to Personalized HealthCare: Grid Solutions for the Life Sciences
N. Jacq et al. (Eds.)
IOS Press, 2007

Tandem Mass Spectrometry Protein Identification on a PC Grid

D. Zosso[a], M. Podvinec[a,1], M. Müller[b], R. Aebersold[b], M.C. Peitsch[c], T. Schwede[a]
[a] *Swiss Institute of Bioinformatics and Biozentrum, University of Basel, Switzerland*
[b] *Institute of Molecular Systems Biology, ETHZ, Switzerland*
[c] *Novartis Institutes of BioMedical Research, Basel, Switzerland*

Abstract. We present a method to grid-enable tandem mass spectrometry protein identification. The implemented parallelization strategy embeds the open-source x!tandem tool in a grid-enabled workflow. This allows rapid analysis of large-scale mass spectrometry experiments on existing heterogeneous hardware. We have explored different data-splitting schemes, considering both splitting spectra datasets and protein databases, and examine the impact of the different schemes on scoring and computation time. While resulting peptide e-values exhibit fluctuation, we show that these variations are small, caused by statistical rather than numerical instability, and are not specific to the grid environment. The correlation coefficient of results obtained on a standalone machine versus the grid environment is found to be better than 0.933 for spectra and 0.984 for protein identification, demonstrating the validity of our approach. Finally, we examine the effect of different splitting schemes of spectra and protein data on CPU time and overall wall clock time, revealing that judicious splitting of both data sets yields best overall performance.

Keywords. Proteomics, Tandem MS, Protein Identification, PC Grid.

Introduction

Proteomics is the systematic, parallel study of ensembles of proteins found in particular cell types or present under particular exogenous or endogenous conditions. These studies promise deep and detailed insights into the control and function of biological systems by comparing samples from different tissues, developmental stages or disease states [1]. Along with other high-volume experimental techniques, proteomics is a cornerstone of systems biology, an emerging discipline characterized by the systematic and quantitative large-scale collection of data, linked to the use of computational biology to model and predict the behavior of complex biological systems [2].

Proteomics and systems biology are increasingly applied in clinically relevant fields of research, such as the biology of cancer, diabetes or other multifactorial diseases [3, 4]. Moreover, they are vital tools in the development of new drugs [5] and the study of biomarkers (i.e. disease-related alterations of the protein composition of accessible body fluids) in the diagnosis of diseases, the adaptation of therapy to inter-individual variation in drug metabolism, response, and toxicity. [6-9].

[1] Corresponding Author: Swiss Institute of Bioinformatics, University of Basel, Klingelbergstr. 50-70, CH-4056 Basel; phone: +41 61 267 15 83, fax: +41 61 267 15 84. michael.podvinec@unibas.ch

To identify proteins present in a particular biological sample, tandem mass spectrometry (MS/MS) combined with bioinformatics analysis is commonly applied. Briefly, the following steps are performed: First, the proteins contained within a biological sample are separated into fractions along one or multiple dimensions, such as size, isoelectric point, or hydrophobicity. Subsequently, fractions are digested by specific proteases, ionized and injected into a mass spectrometer, where peptide parent ions are selected, fragmented, and the mass fingerprint of the fragments is acquired.

Bioinformatics approaches for protein identification are computationally expensive (see [10, 11]). Observed peptide fragment masses are compared against peptide mass fingerprints computed from a database of protein target sequences. To model peptide fingerprints, the protein database must be expanded into a list of expected peptides by chopping sequences at specific proteolytic cleavage sites. This step models protein digestion prior to MS/MS analysis. Further database expansion criteria may include coverage of missed or unanticipated cleavage sites, mutations of single amino acids, and constant or potential amino acid mass modifications. The last can occur during sample preparation or may be post-translational modifications (PTM) of biological relevance. Expansion of peptide mutations and of a list of potential PTMs must be exhaustive, since completeness of annotated mutations and PTMs is not guaranteed in protein sequence databases. As all these expansion criteria are orthogonal, their joint extension rapidly leads to combinatorial explosion of the computational complexity.

A single MS/MS experiment routinely results in tens of thousands of spectra, e.g. in [12], which leads to considerable computational requirements for analysis and forces strict limits on expansion criteria to prevent the search problem from becoming intractable. On the other hand, experiments typically yield a significant proportion of spectra that cannot be assigned to a peptide in spite of good data quality. It therefore makes sense to allow for as many biologically meaningful peptide modifications as possible in an attempt to match previously unidentified spectra. Such searches require the availability of large computational resources, as can be provided by a compute grid. Moreover, protein identification from MS/MS spectra is essentially a data-parallel problem, and is well suited for efficient grid-based execution.

In this paper, we demonstrate how x!tandem [13], a publicly available MS/MS identification tool can be grid-enabled. Our system allows users to submit MS/MS data for analysis via a web front end. Protein identification is performed on a grid of hundreds of desktop PCs. This work extends the *ParallelTandem* parallelization strategy [14] and adapts it to a grid environment. In particular, we have investigated the impact of parallelization on search results in terms of their numerical stability, the stability of the score statistics, detection characteristics (sensitivity and specificity), and runtime.

1. Materials and Methods

1.1. Tandem Mass Spectrometry Protein Identification Tool

X!tandem is an open-source implementation of an algorithm to match a set of peptide tandem mass spectra with a list of protein sequences [13, 15]. The free availability of the executable, its source code and plug-ins make this tool particularly attractive for (academic) grid environments, where licensing costs and models, flexibility, maintainability and portability are important issues. Its pluggable architecture allows the integration of additional scoring schemes.

X!tandem splits the matching process into two sequential steps. First, spectra are assessed against the complete database of proteins with a low level of model complexity, permitting rapid elimination of non-matching sequences and establishing a set of candidate proteins. On these candidates, a second, refined search is carried out, resulting in additional peptide identifications.

At the beginning of both the non-refined and refined search, the protein sequences are expanded into a peptide list. For each peptide, the mass values of its possible fragment ions are calculated, producing an artificial peptide mass fingerprint. This fingerprint is compared to each measured tandem mass spectrum and scored using either the native or a plugged-in scoring scheme.

The native x!tandem scoring scheme, called hyperscore, is based on the dot-product between spectrum (I) and prediction (P) peaks (eq. 1). N_b and N_y are the number of matched b- and y-ions, respectively, i.e. N- and C-terminal peptide fragments.

$$Hyperscore = N_b! N_y! \sum_i I_i P_i \tag{1}$$

In order to assess the statistical significance of a calculated hyperscore, x!tandem computes an expectation value (e-value) as proposed in [16]. For each peptide, a hyperscore histogram of all scored spectra is established. Only the highest-scoring spectrum is supposed to be a valid match and all other spectra are considered as random matches. The p-value of the valid score, i.e. the probability of observing that score at random, can be estimated by log-linear extrapolation of the right-hand tail of this extreme value distribution. Multiplying this value by the number of scored sequences yields the expected number of equal scores, considering the given set of spectra and peptide list.

Once the peptide evidence is established, x!tandem attempts to infer protein identities. Based on the number of peptide hits n of a protein and their respective scores e_i, a protein e-value is calculated according to eq. 2:

$$e_{protein} = \binom{s}{n} \cdot \frac{p^n (1-p)^{s-n}}{s N^{n-1}} \cdot \prod_{i=1}^{n} e_i \tag{2}$$

where s is the number of mass spectra in the dataset, N the number of peptide sequences scored to find the unique peptides, and p is N divided by the total number of peptides in the considered protein expansion. The equation is a Bayesian model for a protein to have obtained the observed number of matches by chance. The first two terms mainly describe the probability of random parent mass matches, a concept also suggested in [17], whereas the product of the underlying peptide e-values takes into account their non-randomness with respect to their fragment mass fingerprints.

1.2. Parallelization Scheme

X!tandem supports multithreading on suitable machine architectures. In a cluster environment, however, parallelization requires inter-node communication tools like PVM or MPI. Such a cluster-based parallelization strategy has been implemented in *ParallelTandem* [14]. The authors introduce the distribution of subsets of the initial mass spectra to different cluster nodes. After a first non-refined step, results are col-

lected, candidate proteins are extracted from all subjobs, and refinement jobs are sent out to the nodes again. A consolidation step calculates the final protein e-values.

A major constraint in grid calculation is data distribution. In most cases, mass spectra files are much smaller than the protein databases they are matched against. The December 2006 release of TrEMBL [18] is about 1.3 GB in size. Distributing the entire file on each target machine can cause data transport and local storage issues. Additionally, [13] demonstrate a log-log-linear reduction of per-spectrum analysis time with increasing size of the set of spectra, whereas the influence of the protein database size on calculation time is proportional. As a consequence, it is opportune to explore the impact of splitting spectrum sets, protein sequence files, or both. The implemented parallelization scheme is shown in Figure 1a. To compensate for the effect of reduced peptide list size on peptide e-values, the output threshold of the non-refined step is lowered proportional to the number of protein database subdivisions, leading to an almost stable number of candidate proteins.

1.3. Grid-Enabled Implementation

End users should not be forced to know all the complexities and peculiarities involved in executing large proteomics analysis jobs in a grid environment. In addition, the use of PVM or MPI is not appropriate for grids due to network latency and node persistence issues. To this end, we have developed a three-layered application service to interface between users and the actual computation (Figure 1b). The service takes care of portal and workflow aspects for proteomics searches and uses existing grid middleware or local resource management systems (LRMS) to submit independent jobs.

The central part of the application service consists of a Perl service daemon, which parallelizes incoming search requests and preprocesses input data. Data sets are deployed to the grid, and jobs are monitored, triggering intermediate or post-processing steps.

The application service interacts with compute resources through a separate layer, which abstracts interaction with the grid middleware or LRMS. This layer has been implemented as a Perl local resource management system interface (LRMSI) module that hides the implementation differences of different grid middleware or batch systems and presents a simple API for handling jobs and job data. This approach has the advantage that changes in middleware versions or even switching from a grid infrastructure to a cluster environment can take place rapidly. The LRMSI is a derivative of the ProtoGRID developed in the context of the SwissBioGrid [19], and already supports a number of common LRMS software.

Requests for analysis are submitted by users through a web portal, and the results of finished analyses can be retrieved and visualized within the portal. We are using a customized version of the SWISS-MODEL Workspace web application [20] as framework for user-based project and data handling.

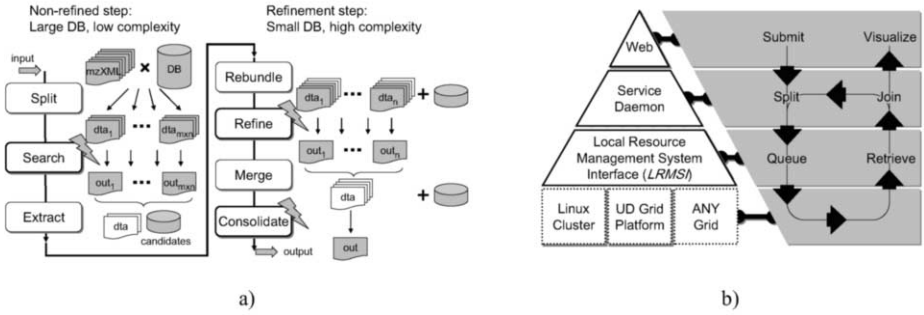


Fig. 1: a) Parallelization scheme of the algorithm. In the non-refined step, work units are formed by the cross-product of *n* sets of spectra and *m* protein database subdivisions. Collected candidate proteins are subjected to a refined search against the original spectra parts (including previously matched spectra) before consolidation. Steps marked with a bolt are executed on the grid. b) Three-layer architecture of the proteomics application service. Requests are received through a web portal. A daemon implements the parallelization scheme and interacts with the underlying compute infrastructure through a LRMS interface layer.

1.4. Validation of Parallelization Scheme

To assess the validity of the proposed grid parallelization scheme, a test case was set up by defining a benchmark consisting of a set of spectra, a standard protein database and common model parameters. Three runs on a single-CPU computer using different analysis protocols (*R1*, *R2*, *R3*) serve as reference to assess the results obtained on the grid. The MS/MS dataset used is a standard set created by analyzing a mixture of 17 proteins from different species on a Micromass MALDI Q-TOF device, provided in mzXML format [21] and obtained from [22]. UniProtKB/Swiss-Prot version 9.1 was used as protein sequence database, containing 241'365 protein sequences (107 MB) [18]. For reference runs and grid analyses, the same x!tandem binary (version 2006-09-15-3) including the k-score plug-in [23] was used running on Microsoft Windows XP.

The peptide expansion model included tryptic digestion with a maximum of 3 missed cleavage sites. Cysteine residue mass was modified by +57 Da to account for cysteine carboxamidomethylation during protein sample preparation. Additionally, the following frequently observed potential modifications were added: methionine oxidation (+16Da), asparagine/glutamine deamidation (+1Da), as well as serine/threonine phosphorylation (+80Da) in refinement.

We define the *R1* reference as the results of a local, single-CPU run of x!tandem. In the *R1* protocol, the whole MS/MS dataset of the mzXML file is matched against the complete native protein sequence database. As the protocol for grid-based analysis comprises more processing steps, several sources of noise can interfere with the final output. In order to get fine-grained information about where score variations may be caused, two modified local runs were carried out that approximate the grid protocol. The *R2* reference was defined as the result of a local standalone run, followed by a consolidation step against the extracted candidate protein list. In this step, previous x!tandem output (including processed spectra and model parameters) is used as input. Here, both spectra set and protein sequences have changed, and the spectra might have undergone minor conversion modifications. A third reference, the *R3* standard, is obtained by converting the MS/MS dataset from mzXML to plain peak lists (DTA format) before executing the R2 protocol. This conversion is a prerequisite for spectra rebundling in the grid environment.

2. Results and Discussion

Peptide and protein e-values both depend on the set of spectra analyzed, the protein database, and on model parameters causing database expansion. Changes in one or more of these factors may strongly influence the resulting e-values both for peptides and proteins. Instability might be increased in protein scores, as in addition to cumulated peptide score shifts and fluctuations, further distortion is injected by the dataset and protein-database related terms in the protein scoring function. To study these effects, we first investigated differences in the results of the three reference protocols, and subsequently compared these to results of the grid-enabled version.

2.1. Reference Performance

The results of both the *R2* and *R3* reference were compared to the output of the single-run *R1* reference analysis. The correlations to the e-values of *R1* are $c_s = 0.94$ for both *R2* and *R3*.

The scatter plot of consensus hits (spectra and proteins that were matched in both analyses) of *R1* and *R2* shows significant fluctuations between corresponding e-values on peptide level (Figure 2). No significant difference was observed between *R2* and *R3*, as spectrum conversion has more influence on the composition of the scored set of spectra than on the e-values of consensus spectra. This means that a considerable amount of *statistical flickering* is injected in the consolidation step, where e-values are based on conditions much different from the original setting. The least-squares linear fit reveals a systematic –0.78 shift of the consolidated log(e-values), corresponding to e-values being almost 6 times smaller in *R2* than in *R1*.

At protein level, e-values show a much higher correlation of $c_p = 0.99$. This indicates that underlying peptide e-value fluctuations are random enough to be compensated in the protein scoring function. However, the peptide e-value shift is cumulated in protein scores and amplified in proteins featuring multiple peptide matches, resulting in increased deviations for highly significant protein identifications.

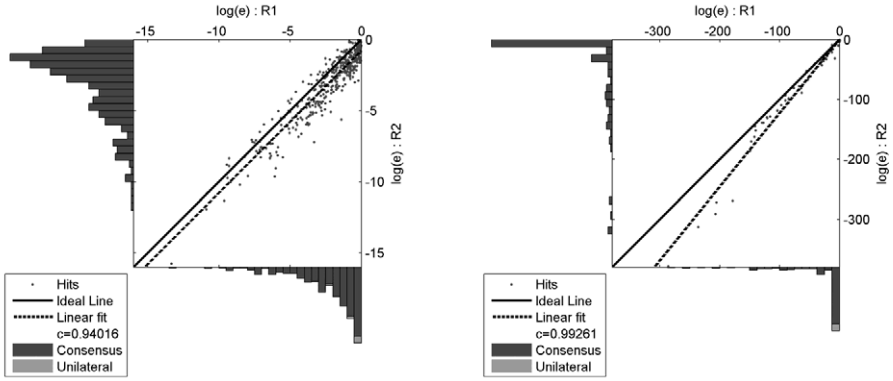
To investigate the impact of the statistical instabilities on the output characteristics, ROC-like sensitivity-selectivity curves of *R1* and *R2* analyses have been plotted in Figure 3. For increasing e-value cutoff levels (decreasing match significance), the number of true positive hits, i.e. matches against peptide sequences present in the known protein mixture, is plotted against the false positives. At spectrum level, no significant difference between the characteristics appears in the first 200 assigned spectra. Beyond, the curves diverge as *R1* yields a few more true positives.

At the level of inferred proteins, there is no significant divergence between the reference schemes. The generally high false positive rate is due to shadow matches by homologous proteins, e.g. from different species.

2.2. Stability of Grid Results

2.2.1. e-Value Correlation

To investigate the stability of the e-value statistics under grid parallelization conditions, the results of grid analyses were compared to the *R1* and *R2* reference e-values. For each grid result, the subset of consensus spectra and proteins were extracted. The distributions of the log(e)-differences (residues) are shown in Figure 4.

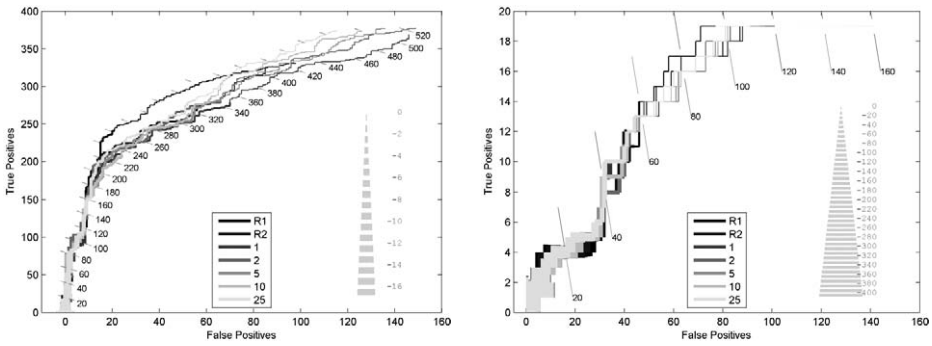


a) b)

Fig. 2: Peptide (a) and protein (b) score correlations of *R2* against the *R1* reference. The histograms show the respective score distributions of consensus spectra/proteins (assigned in both references) (dark) and matches from only one analysis (light).

Varying the number of spectra subdivisions has no significant impact on either spectra or protein score residue distributions. When using different numbers of protein database subdivisions, minor differences in the distributions appear.

Compared to their *R1* counterparts, grid spectra log(e-values) are shifted by approx. –0.8. The negative outliers of protein e-value residues reflect the cumulated shift in high ranking proteins. Compared to *R2*, this shift entirely disappears both for peptides and proteins, confirming the consolidation step as the cause of this shift. Residues spread slightly wider for higher numbers of database subdivisions, as illustrated by both the outliers and the increased inter-quartile distance. This is due to an overcorrection of the adapted output e-value threshold at the non-refined step: about 15% less candidates are selected with 25 protein database subdivisions than when the database is not split. In this way, low-quality matches are removed which score very close to the threshold and reduce residue dispersion.



a) b)

Fig. 3: Spectra (a) and protein (b) ROC-plots of the *R1*/*R2* references and of grid-computed results differing in the number of protein database subdivisions. Line thickness illustrates log(e-value) threshold and the diagonal lines indicate the number of selected matches. The high number of false positives in the protein charts is due to the shadow matches of homologous proteins (e.g. from different species).

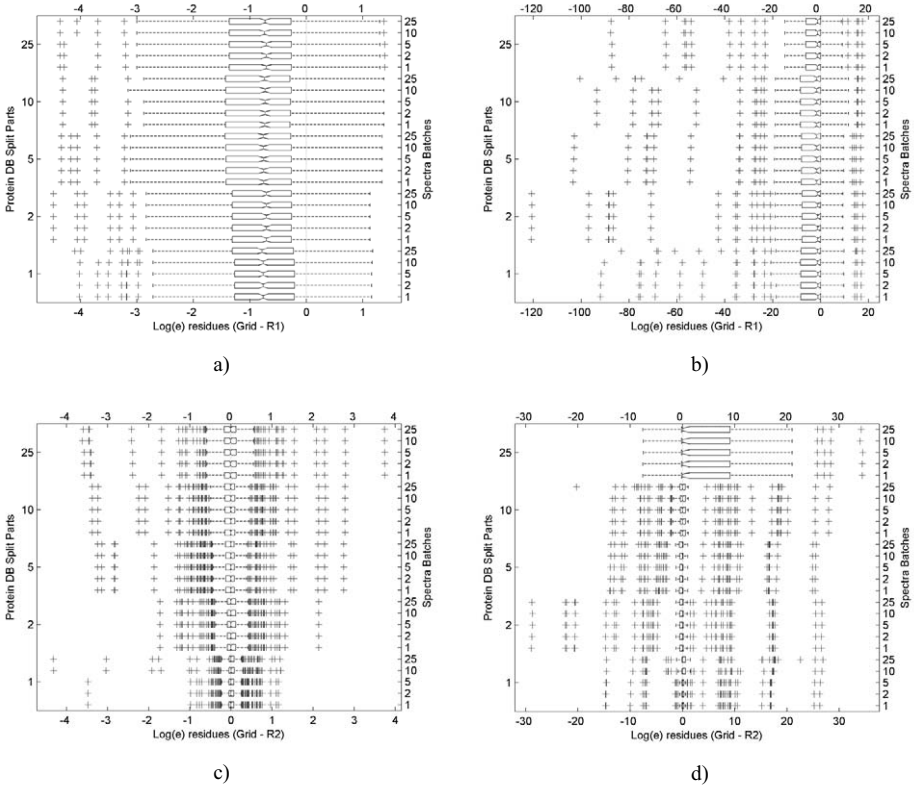


Fig. 4: Residues of spectra and protein log(e-values) between different grid analysis configurations and *R1* (a+b) or *R2* (c+d).

In summary, correlation coefficients of the consensus e-values against *R1* fall in the intervals [0.9327 0.9460] for spectra and [0.9838 0.9880] for proteins. Against *R2*, they improve to [0.9722 0.9972] and [0.9955 0.9974], respectively. Within these bounds, correlations tend to degrade slightly with increasing number of protein database subdivisions.

2.2.2. Detection Characteristics

To compare the analytic power of grid results with local runs, the ROC-curves of different protein database distribution schemes are shown in Figure 3a and b. In this context, the number of sets of spectra plays a minor role and can be neglected.

Grid analyses consistently yielded more spectra hits. However, their true/false-positives ratio is very similar to that of the *R2* reference. Characteristics are almost equal for the best 200 spectra, and differ only in the low-quality part. At protein level, no significant differences appear at all.

These findings, taken together with the e-value correlation data, indicate that the quality of grid results is nearly indistinguishable from searches using the *R2* or the *R1* protocol locally, in particular when considering the top-scoring spectra and proteins, which typically are the most relevant in experiments.

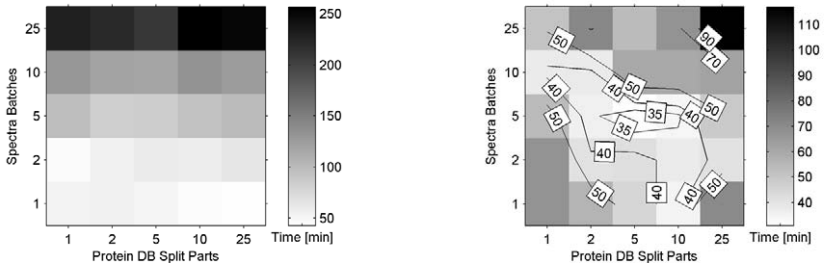


Fig. 5: CPU (a) and project wall clock time (b) for different combinations of spectra set and protein database divisions. The contour lines in the wall time chart emphasize the central region of optimal configuration.

2.3. Numerical Stability and Performance Metrics of Grid-Enabled Executable

CPU time depends on the collective CPU power of the machines a job is running on, a quantity that fluctuates in a heterogeneous, dynamic grid resource. We have measured calculation time for 10 identical distributed x!tandem analyses to quantify the magnitude of these variations. Job splitting parameters were set to a batch size of 500 spectra and 10 protein database subdivisions, giving rise to 50 non-refined and 5 refined work units. The Grid MP middleware estimates the amount of CPU time used by a job based on the share of CPU assigned to a particular work unit, integrated over time and all work units of a job. Despite the wide heterogeneity of PC-grid machines, relative standard deviation between the runs was 5.3%. Moreover, all 10 searches resulted in identical output, demonstrating numerical stability of the grid-enabled application.

Next, CPU times were recorded for different schemes of spectra and protein database splitting. Results are shown in Figure 5a. As expected, splitting generally increases the cumulative CPU time, due to multiplication of the program overhead. This effect is particularly prominent when splitting spectra datasets, as the protein expansion step is computationally expensive.

From the user perspective, the total wall clock time for job execution is perceived as more important, defined as the time between job submission and result retrieval, including queuing and transmission overhead. Corresponding measures have been sampled for the same distribution schemes as above (Figure 5b). We identified a valley of optimal configuration in the wall time charts. Performance gain is most pronounced when splitting the protein database into 5 subdivisions speeds up the non-refined step, and when splitting spectra into 5 subdivisions enhances refinement step performance.

3. Conclusions

In the present work, we have demonstrated how MS/MS protein identification can be transformed into a data-parallel task suitable for efficient grid computing. By developing a multi-layer application service, we manage to abstract the parallelization and grid submission process from the user while maintaining an open architecture supporting various LRMS. We have validated our approach by comparing results obtained on the grid with a number of reference result sets that were calculated on a local single-CPU resource.

Although grid processing introduces fluctuations into peptide and protein scores, the resulting peptide and protein score characteristics do not degrade. However, the fact that match selection ultimately relies on e-value calculations that show extreme dependence on boundary conditions inherently causes statistical instability. Here, we demonstrate how this instability can be minimized, but using the present scoring scheme, it can not fully be eliminated. One future direction in this regard is the inclusion of complementary information and descriptors in the process of peptide and protein inference.

We next investigated the best scheme for job parallelization. While an optimum is expected to exist, the exact position depends mainly on 3 job-specific factors: spectra dataset size, protein database size, and protein expansion model complexity. Ideally, parallel submission to many machines decreases overall time, but the computation/data transfer ratio must be kept high to avoid accruing overhead. In bigger experiments (more spectra, larger databases, more complex model), the optimum is likely to shift towards a higher number of database subdivisions. Grid-based optimization strategies, such as the one presented here, are crucial to address the computational needs of large-scale proteomics studies.

Acknowledgments

We thank Paul Jenö, Suzette Moes and Konstantin Arnold for valuable discussions and technical support. **Author contributions:** MP, MM, RA, MCP and TS initiated the project, MP and TS designed the study. DZ and MP developed the application service. DZ designed the experiments and analyzed the data. MP and DZ wrote the paper.

References

[1] S. D. Patterson and R. H. Aebersold, Nat Genet, vol. 33 Suppl, pp. 311-23, 2003.
[2] J. C. Smith and D. Figeys, Mol Biosyst, vol. 2, pp. 364-70, 2006.
[3] M. R. Flory and R. Aebersold, Prog Cell Cycle Res, vol. 5, pp. 167-71, 2003.
[4] T. Sundsten, M. Eberhardson, M. Goransson, and P. Bergsten, Proteome Sci, vol. 4, pp. 22, 2006.
[5] C. R. Cho, M. Labow, M. Reinhardt, et al., Curr Opin Chem Biol, vol. 10, pp. 294-302, 2006.
[6] S. V. Parikh and J. A. de Lemos, Am J Med Sci, vol. 332, pp. 186-97, 2006.
[7] S. Ciordia, V. de Los Rios, and J. P. Albar, Clin Transl Oncol, vol. 8, pp. 566-80, 2006.
[8] S. Schaub, J. A. Wilkins, D. Rush, and P. Nickerson, Expert Rev Proteomics, vol. 3, pp. 497-509, 2006.
[9] A. Schmidt and R. Aebersold, Genome Biol, vol. 7, pp. 242, 2006.
[10] P. Hernandez, M. Muller, and R. D. Appel, Mass Spectrom Rev, vol. 25, pp. 235-54, 2006.
[11] E. A. Kapp, F. Schutz, L. M. Connolly, et al., Proteomics, vol. 5, pp. 3475-90, 2005.
[12] M. P. Washburn, D. Wolters, and J. R. Yates, 3rd, Nat Biotechnol, vol. 19, pp. 242-7, 2001.
[13] R. Craig and R. C. Beavis, Rapid Commun Mass Spectrom, vol. 17, pp. 2310-6, 2003.
[14] D. T. Duncan, R. Craig, and A. J. Link, J Proteome Res, vol. 4, pp. 1842-7, 2005.
[15] R. Craig and R. C. Beavis, Bioinformatics, vol. 20, pp. 1466-7, 2004.
[16] D. Fenyo and R. C. Beavis, Anal Chem, vol. 75, pp. 768-74, 2003.
[17] J. Eriksson and D. Fenyo, J Proteome Res, vol. 3, pp. 32-6, 2004.
[18] C. H. Wu, R. Apweiler, A. Bairoch, et al., Nucleic Acids Res, vol. 34, pp. D187-91, 2006.
[19] M. Podvinec, S. Maffioletti, P. Kunszt, et al., presented at 2nd IEEE International Conference on e-Science and Grid Computing (e-Science 2006), Amsterdam, The Netherlands, 2006.
[20] K. Arnold, L. Bordoli, J. Kopp, and T. Schwede, Bioinformatics, vol. 22, pp. 195-201, 2006.
[21] P. G. Pedrioli, J. K. Eng, R. Hubley, et al., Nat Biotechnol, vol. 22, pp. 1459-66, 2004.
[22] http://sashimi.sourceforge.net/repository.html, Dec 18, 2006.
[23] B. MacLean, J. K. Eng, R. C. Beavis, and M. McIntosh, Bioinformatics., vol. 22, pp. 2830-2, 2006.

From Genes to Personalized HealthCare: Grid Solutions for the Life Sciences
N. Jacq et al. (Eds.)
IOS Press, 2007

Grid-based Analysis of Tandem Mass Spectrometry Data in Clinical Proteomics

Andreas Quandt [a,b,1], Patricia Hernandez [a,2], Peter Kunzst [c], Cesare Pautasso [d], Marc Tuloup [a], Celine Hernandez [a], and Ron D. Appel [a,b]

[a] *Swiss Institute of Bioinformatics, Geneva, Switzerland*
[b] *University of Geneva, Geneva, Switzerland*
[c] *Swiss National Supercomputing Centre, Manno, Switzerland*
[d] *IBM Research GmbH, Zurich Research Laboratory, Switzerland*

Abstract. Biomarker detection is one of the greatest challenges in Clinical Proteomics. Today, great hopes are placed into tandem mass spectrometry (MS/MS) to discover potential biomarkers. MS/MS is a technique that allows large scale data analysis, including the identification, characterization, and quantification of molecules. Especially the identification process, that implies to compare experimental spectra with theoretical amino acid sequences stored in specialized databases, has been subject for extensive research in bioinformatics since many years. Dozens of identification programs have been developed addressing different aspects of the identification process but in general, clinicians are only using a single tools for their data analysis along with a single set of specific parameters. Hence, a significant proportion of the experimental spectra do not lead to a confident identification score due to inappropriate parameters or scoring schemes of the applied analysis software. The swissPIT (Swiss Protein Identification Toolbox) project was initiated to provide the scientific community with an expandable multi-tool platform for automated and in-depth analysis of mass spectrometry data. The swissPIT uses multiple identification tools to automatic analyze mass spectra. The tools are concatenated as analysis workflows. In order to realize these calculation-intensive workflows we are using the Swiss Bio Grid infrastructure. A first version of the web-based frontend is available (http://www.swisspit.cscs.ch) and can be freely accessed after requesting an account. The source code of the project will be also made available in near future.

Keywords. cyclic workflows, proteomics, tandem mass spectrometry, workflow manager

1. Introduction

Proteomics can be defined as the systematic study of the protein content (called the proteome) of a given cell, tissue or organism, at a given time and under specific conditions [18]. Proteomic research encompasses the identification, characterization and quantification of proteins. It involves various techniques, including single and tandem mass spectrometry.

[1] To whom correspondence should be addressed: andreas.quandt@isb-sib.ch
[2] Contributed to the paper equally as the first author

Studying proteins is a particularly challenging task. Unlike genes, proteins are neither homogenous nor static. Genes are confined in the cell nucleus, while proteins are present in all cell compartments, on the cell surface and even in extra-cellular fluids. Gene patterns are fairly constant while protein patterns mirror even small changes in their environmental changes including pathological and physiological developments. Great hopes are placed in protein separation techniques associated with mass spectrometry in order to discover potential diagnostic markers and therapeutic targets. They allow the rapid detection of proteomic biomarkers by comparing case and control samples such as cancer and non-cancer samples. Examples for biomarkers at the genomic level are single nucleotide polymorphisms (SNP) [10] or DNA methylation [4], and at the proteomic level, protein abundance or protein post-translational modifications. Proteomic biomarkers additionally provide the possibility to monitor the evolution of diseases or drug treatment. This is especially important for the correct treatment of cancer where over-treatment in earlier stages of the disease is a well-known phenomenon. For example, breast cancer patients often display significantly different clinical phenotypes and responses to a specific therapy [12]. Medical doctors can use the additional information provided by the analysis of proteomic biomarkers to identify the best method for treatment and the correct doses to be applied. For example, the treatment with hormonal therapy [16] has fewer side effects than restrain treatments with more noxious effects like radiation [1].

In clinical proteomics, biological samples are analyzed in high-throughput mode, producing thousands of data files each day. The analysis, storage, and distribution of this huge data volume requires the development of new bioinformatics tools capable to use distributed computing resources such as a Grid infrastructure. Grid Computing has been postulated as a modern paradigm to gain access to a large number of computational resources. It enables individual resource owners to share their infrastructure with each other, providing a large overall infrastructure for many different projects. The biggest benefit of a shared infrastructure comes from the optimized usage of resources. Individually, each project would most probably under-utilize dedicated resources or would not get access to enough resources. The Grid takes care of distributing the load on many resources, maximizing the utilization of the resources.

This paper describes the Swiss Protein Identification Toolbox (swissPIT) a new proteomics platform managing several data analysis tools by combining them in user-defined workflows (Figure 1). In order to be able to execute complex analysis workflows, swissPIT uses the infrastructure of the Swiss Bio Grid project. The Swiss Bio Grid (http://www.swissbiogrid.org/) provides a large-enough Grid infrastructure that can be used to perform identification workflows and parameter studies with swissPIT. Currently, five partner institutions are providing resources: the Swiss Institute of Bioinformatics (SIB), the Swiss National Supercomputing Centre (CSCS), the Biozentrum at Basel University, the Friedrich Miescher Institute in Basel and the Functional Genomics Centre Zurich. The infrastructure makes use of the NorduGrid Advanced Resource Connector ARC middleware (http://www.nordugrid.org/) to manage the submission of individual Grid jobs for swissPIT (i.e. the execution of the identification tools with a given parameter set). Two main advantages of grid computing for proteomic experiments are job distribution (e.g. when a great number of users want, at the same time, to analyze their datasets), and job parallelization (e.g. in case of exploration of parameter combinations).

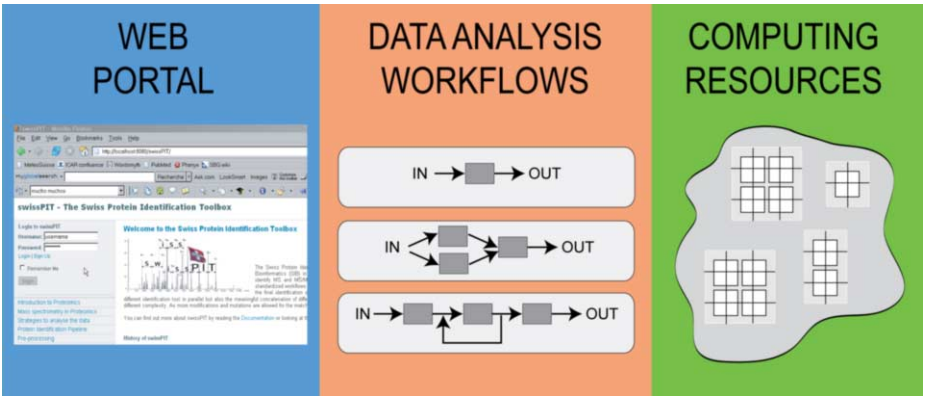


Figure 1. The three aspects of the swissPIT platform: a) the web interface that allows the user to login, choose the programs or workflow to run, upload the data, set the parameters and visualize the results; b) the data analysis workflows; c) the grid environment.

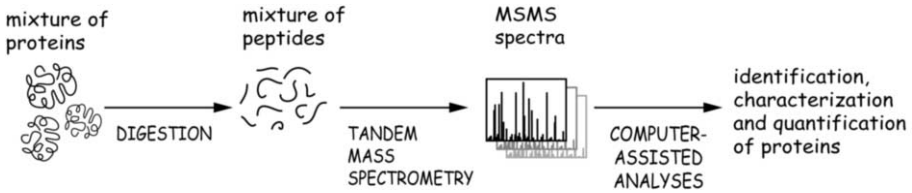


Figure 2. Whole proteins are purified and cleaved into peptides. During the mass spectrometric analysis, peptides are selected based on their m/z values, fragmented into pieces, and the m/z of the obtained fragments are reported in an MS/MS spectrum. Interpretation of the obtained mass spectra is done using specialized software and protein or genomic sequence databases.

1.1. Background

In clinical proteomics, both single and tandem mass spectrometry techniques are used to screen biological samples in high-throughput mode to identify and quantify interesting biomarkers. Unlike single mass spectrometry, tandem mass spectrometry provides information on the protein sequence and is additionally capable to identify proteins in mixtures. Tandem mass spectrometry (MS/MS) requires two mass analyzers in series. The first analyzer separates the peptides according to their mass-to-charge ratio (m/z). Then, selected peptides (called precursors) undergo fragmentation and the m/z ratios of the produced fragments are measured by the second analyzer. This yields MS/MS spectra composed of a precursor peptide mass and of fragment peaks (m/z and intensity). The number of peaks varies from about ten to several hundreds depending on factors like the precursor peptide length, the fragmentation quality, the mass spectrometer type and the peak detection process. The obtained mass spectra can then be used to carry out various analyses. In proteomics studies, the three major topics of interest are the identification, characterization and quantification of the proteins present in a sample. Figure 2 summarizes key steps of a typical MS/MS-based experiment.

To identify peptide patterns within the spectra, bioinformatics tools are applied to screen genomic and protein databases (such as the UniProtKB/swissProt and UniPro-

tKB/TrEMBL [19]) for high scoring matches. Databases are usually composed of tens of thousands to several millions of biological sequences. The most common approach to identify peptides is called Peptide Fragment Fingerprinting (PFF). This approach computes a similarity score between the experimental MS/MS spectrum and theoretical MS/MS spectra constructed from the theoretical digestion of protein sequences to peptides and *in silico* fragmentation of the peptides. The calculation of these similarity scores is often difficult because of the variation between observed peptides and their corresponding database entries. For example, the peptide may carry a post-translational modification that is not documented in the database. The sequence can also be mutated, or sequence rearrangements may have occurred during the digestion process (transpeptidation). All these modifications cause peak shifts in the experimental spectra which then differ from the theoretical spectra in the database. On the other hand, the database may also contain errors, i.e. when the protein sequences are automatically translated from genomic or transcriptomic sequences (e.g. EST contigs). Traditionally, identification tools, such as Sequest [9], Mascot [14] and Phenyx [2], require the user to specify in advance a list of possible modifications taken into account during the matching process. These tools perform what we call a "classical search". Their main advantage is the production of results in a reasonable amount of time, but they cannot identify peptides which carry unexpected modifications or mutations. “Open-modification search” tools, such as Popitam [7], OpenSea [15], GutenTag [6] and InsPecT [17], have been specifically designed to handle unexpected amino acid modifications, but they often need preliminary filtering steps. *De novo* sequencing is another approach, which infers sequence information from the experimental MS/MS spectrum. The identification is then performed by matching the obtained *de novo* sequence with the database peptides. *De novo* sequencing methods require spectra of higher quality with smaller fragment errors and a more or less continuous signal, or at least high-quality signal for several adjacent amino acids. Despite these disadvantages, *de novo* methods may surpass PFF methods, notably when searching genomic databases subjected to sequencing errors, when searching databases composed of homologous sequences or when analyzing a spectrum that originates from a mutated protein or variant [11].

1.2. MS/MS identifications workflows

For many years the combination of different search tools and analysis strategies has been proposed to improve the result quality of the identification process [8]. Hence, the development of a platform for MS/MS identification workflows is one of the most interesting topics in bioinformatics for proteomics. The swissPIT project aims at providing a protein identification platform from MS/MS data. It is being developed on a Grid infrastructure in order to perform complex identification workflows. It is anticipated that the meaningful combination of several analysis tools will allow on the one hand, to identify more interesting biomarkers, and on the other hand to increase the overall reliability of the results. Advantages of such a platform include (1) reduced human intervention by automated execution of software workflows (2) simple data and results sharing between different user groups by user-controlled data storage, (3) and the feasibility of large studies on parameter optimization as well as more complex identification processes by the use of distributed computing resources.

2. Methods

The heart of the swissPIT project is to implement meaningful identification workflows such as to improve the identification quality for given data sets. In order to develop such an individual identification strategy, the same input data has to be analyzed with different parameter settings using the various identification tools. Depending on the number of adjustable parameters, these parameter sweeps may be extensive, i.e. there may be several hundreds or thousands of different parameter sets for which the tools have to be executed and results have to be evaluated. Consequently, we expect thousands of individual executions for each of the identification tools, in order to discover the most adapted parameter combination for a specific data set. So as to be able to perform these calculations within a reasonable amount of time, a high-throughput computing infrastructure has to be used.

2.1. Typical scenario

Due to the advantages and disadvantages of the different search strategies, a typical scenario for an identification process is to apply a classical search followed by an open-modification search. In a classical search, an identification algorithm generally produces a list of accession numbers (ACs) corresponding to peptide sequences matching the identified spectra. In an open-modification search, these ACs can then be used to improve the identification by searching for unexpected modifications. The unexpected modifications can then be used again to perform a new classical search with an updated list of potential modifications.

2.2. Use case workflow

The use case describes an identification process with three searches performed sequentially, where the output of each search is the input of the following search. In swissPIT, the workflow of this use case contains three modules: two classical search modules and one open-modification search module. Modules in swissPIT are defined as workflow parts which have to be executed in a predefined order. For each of the three modules multiple algorithms exist that are executed in parallel to reduce to overall analysis time. To test the first workflows, we decided to implement interfaces to two classical search approaches and two open-modification search tools. Phenyx and X!Tandem [3] are well-known tools for classical search. Phenyx (http://phenyx.vital-it.ch/) is currently the most advanced tool available while X!Tandem being an open-source tool is often used as a reference implementation. Popitam and InsPecT are two more recent tools specialized for detecting unexpected modifications.

Using all four tools, the described use case example can be implemented as follows (see Figure 3): Phenyx and X!Tandem can be started in parallel to perform a first classical search on a specific data set. Both programs will identify a subset of all spectra and produce an AC list. This list is then used by Popitam in the following open-modification search, with Inspect running in parallel. The produced list of possible modifications (modres.xml) can then be used again by a repeated run of Phenyx and X!Tandem in order to identify spectra of the original data set that have not been identified during the first run. Besides the computational needs, a major problem is the handling of the different input and output files. The lack of standards for input and output formats, parameter de-

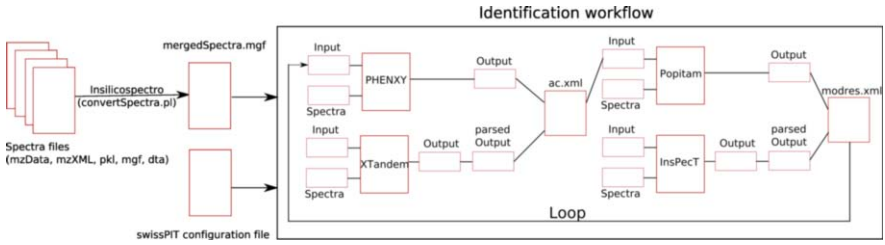


Figure 3. Use case workflow

finitions and presentation of the information content of the results requires many parsing steps in order to use standardized parameters and to merge results for the continued use within the workflow.

2.3. Workflow configuration

We implemented a basic workflow manager that can be configured by a single XML file. Figure 4 shows an example configuration file. It contains three parts necessary to describe a complete workflow for the swissPIT engine: the workflow (A), the modules (B), and the programs (C). In general, a workflow (A) contains different modules (B) with one or more programs (C). Each module and each program has a unique identifier (ID) which is used to execute the workflow with the correct parameter set. Unique ID's are necessary because a workflow can contain modules and programs of the same type but with different parameter sets.

In section A of the configuration file, the modules of the workflow are specified by using the <moduleid> tag. The module ID is unique and points to a specific module in section B of the configuration file. The order of the module ID's is important because the modules are performed in the same sequence as their ID's are specified. If no module matches the id specified in the workflow, swissPIT stops with an error message. The example configuration shown in Figure 4 defines a workflow with two modules (IDs "m01" and "m02"). The IDs are pointing to two modules (B) which are of the type "strictSearch" (m01) and "openModificationSearch" (m02). In addition to the module definition, section B also contains a parameter section. That section specifies parameters which are shared between different module types of a workflow. A module is defined by the parameters specific to this module (shared parameters between the programs of the module) and the programs executed within this module. The program IDs named within the <module> tag are pointing to a specific program definition (C). Depending on the program type, the program definition contains parameters which are not shared with other programs of the workflow. During the execution of the workflow, the various parameters are used to create the individual input files for each program.

2.4. Information flow within a workflow

Figure 3 illustrates the second major problem in realizing identification workflows. In addition to the intensive computational resources needed to execute these workflows, many parsing and converting steps are required to transport information between the tools. The lack of standardized formats has been recognized within the

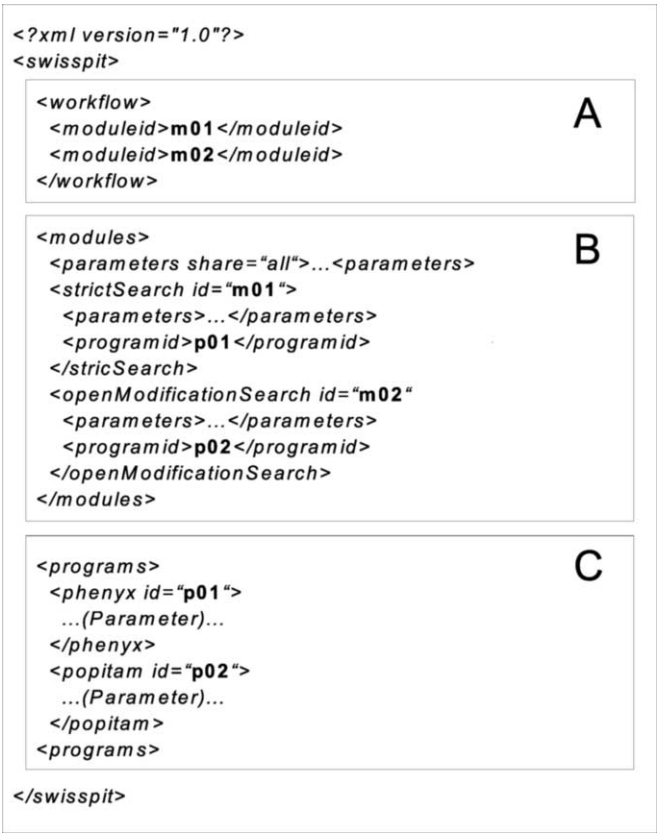


Figure 4. Example scheme of the swissPIT configuration file

scientific community. The Proteomics Standards Initiative (PSI) has been founded in 2002 (http://psidev.sourceforge.net/) to develop, among others, standard formats for raw data/peak lists (mzData) and for analysis results (analysisXML). At the time this article was written, analysisXML was still in development. Therefore, it was necessary to create internal formats for the information transport within the presented workflow. First, we developed the initial configuration file containing parameters which are shared by all programs, shared between programs of a specific search type (strict search (or classical search) and open modification search), and program specific parameters (see Figure 4). These parameters are the common basis to create the individual parameter files for each program. In a second step, we developed parsers to have a common output format for programs of the same module (search strategy). The use of a common output format allows not only the simplification of the information transport between the modules (ac.xml, modres.xml), it is also the first step of the development of a visualization tool displaying the results of different applications in a standardized way, in order to enhance result comparison.

3. Results

In this article we have described swissPIT, the Swiss Protein Identification Toolbox, a new workflow-based platform for the analysis of mass spectrometry data. By giving access to several analysis tools, swissPIT aims to mix different search strategies to increase the overall identification of spectra. We also argue for the need of high-throughput computing infrastructures like Grids in modern mass spectrometry-based proteomics to perform the in-depth analysis with our identification workflows.

The swissPIT platform has been installed on the CSCS node of the Swiss Bio Grid. Access to this installation is given by request to the authors. This command line version of swissPIT provides the execution of basic identification workflows with parallel program execution and sequential combination of different search strategies. The user can create new workflows by easily modifying the configuration file of swissPIT. He/she also can perform a parameter exploration by specifying different parameter sets which are executed in parallel. A web-based interface allows for the easy control of swissPIT (http://swisspit.cscs.ch). The web interface does not provide the execution of workflows yet, but it is already a useful tool for biologist to use several identification programs with a unified interface. It also allows an improved result comparison for the different tools by providing standardized parameter sets across the programs. The available programs are executed in parallel on the Grid infrastructure for faster analysis of the data. Submissions are handled with a swissPIT server certificate and the user authentication is managed by the login procedure. After the successful submission of a Grid job, the user can monitor its progress through the job monitor. Once all runs are finished, the user can browse the job folder within his/her personal user space to retrieve the individual result files. The results are presented to in their original view and can also be downloaded locally.

Grid Interaction Details

The Swiss Bio Grid aims to create a showcase of a basic infrastructure capable of serving a variety of bioinformatics calculations. There are three projects currently supported on this platform, one of which is the swissPIT proteomics project described in this paper. This is a showcase infrastructure, with the aim of producing real results, initiating the establishment of a more sustainable Swiss national Grid. It is based on two complementary technologies: the NorduGrid ARC middleware (which itself is based on Globus) and the United Devices GridMP desktop Grid. The proteomics project achieves Grid parallelization by parametrization, i.e. the execution of the same programs with different input parameters. Each program is submitted by the swissPIT end-user interface through the NorduGrid ARC submission interface. The GridMP services are not used for this project, as the executables are not all available on the Windows platform. In the current setup, the swissPIT interface acts as a portal to the users, i.e. the actual jobs are all submitted as a single 'swisspit' user to the Grid. The ARC monitoring clients are used to track the progress of each job, and when successful execution has been reported, again the ARC client tools are used to retrieve the output of the job.

4. Outlook

The swissPIT addresses complex problems that have not been tackled on this scale before. The most challenging aspect is to provide high-throughput computing resources to

the user by hiding the complexities of the system behind an intuitive interface. In the future, we have to automate processes such as individual user authentication and authorization. We also have to continue working on solutions for distributed data storage across all nodes of the Grid infrastructure, and to implement mechanisms for data access to allow data sharing between user groups. A further aspect will be the link between these data and shared databases such as Peptide Atlas [5] or Pride [13] which allow the application of statistics on a larger scale. We also have to work on an automatic distribution of the several biological databanks such as Uniprot/Swissprot or Uniprot/TrEMBL in their different versions within the Grid infrastructure, so that results of the distributed parameter sweeps are consistently making use of the same datasets.

In the short term, we are mainly working on improvements of the web interface, as well as on a common visualization interface for the results retrieved from the different search tools. A tool for graphically building and executing workflows is planned, as well as support for an enlarged set of identification algorithms. We also plan to use analysisXML as the internal file format for the information transport to unify the information flow and to simplify the connection to other programs interacting with swissPIT.

5. Acknowledgements

We would like to thank Sergio Maffioletti (CSCS) and Alexandre Masselot (Genebio S.A.) for intensive help with the Swiss Bio Grid infrastructure and the file format parsing.

References

[1] B. S. Bloom, N. de Pouvourville, S. Chhatre, R. Jayadevappa, and D. Weinberg. Breast cancer treatment in clinical practice compared to best evidence and practice guidelines. *Br J Cancer*, 90(1):26–30, Jan 2004.

[2] Jacques Colinge, Alexandre Masselot, Marc Giron, Thierry Dessingy, and Jérôme Magnin. Olav: towards high-throughput tandem mass spectrometry data identification. *Proteomics*, 3(8):1454–1463, Aug 2003.

[3] Robertson Craig and Ronald C Beavis. Tandem: matching proteins with tandem mass spectra. *Bioinformatics*, 20(9):1466–1467, Jun 2004.

[4] Partha M Das and Rakesh Singal. Dna methylation and cancer. *J Clin Oncol*, 22(22):4632–4642, Nov 2004.

[5] Frank Desiere, Eric W Deutsch, Nichole L King, Alexey I Nesvizhskii, Parag Mallick, Jimmy Eng, Sharon Chen, James Eddes, Sandra N Loevenich, and Ruedi Aebersold. The peptideatlas project. *Nucleic Acids Res*, 34(Database issue):D655–D658, Jan 2006.

[6] Ari Frank, Stephen Tanner, Vineet Bafna, and Pavel Pevzner. Peptide sequence tags for fast database search in mass-spectrometry. *J Proteome Res*, 4(4):1287–1295, 2005.

[7] Patricia Hernandez, Robin Gras, Julien Frey, and Ron D Appel. Popitam: towards new heuristic strategies to improve protein identification from tandem mass spectrometry data. *Proteomics*, 3(6):870–878, Jun 2003.

[8] Patricia Hernandez, Markus Müller, and Ron D Appel. Automated protein identification by tandem mass spectrometry: Issues and strategies. *Mass Spectrom Rev*, Nov 2005.

[9] Ashley L. McCormack Jimmy K. Eng and III John R. Yates. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, Volume 5, Issue 11:Pages 976–989, November 1994.

[10] Bao li Chang, Siqun L Zheng, Sarah D Isaacs, Aubrey Turner, Gregory A Hawkins, Kathy E Wiley, Eugene R Bleecker, Patrick C Walsh, Deborah A Meyers, William B Isaacs, and Jianfeng Xu. Polymorphisms in the cyp1a1 gene are associated with prostate cancer risk. *Int J Cancer*, 106(3):375–378, Sep 2003.

[11] Adam J Liska and Andrej Shevchenko. Expanding the organismal scope of proteomics: cross-species protein identification by mass spectrometry and its implications. *Proteomics*, 3(1):19–28, Jan 2003.

[12] Edison T Liu. Classification of cancers by expression profiling. *Curr Opin Genet Dev*, 13(1):97–103, Feb 2003.

[13] Lennart Martens, Alexey I Nesvizhskii, Henning Hermjakob, Marcin Adamski, Gilbert S Omenn, Joël Vandekerckhove, and Kris Gevaert. Do we want our data raw? Including binary mass spectrometry data in public proteomics data repositories. *Proteomics*, 5(13):3501–3505, Aug 2005.

[14] D. N. Perkins, D. J. Pappin, D. M. Creasy, and J. S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18):3551–3567, Dec 1999.

[15] Brian C Searle, Surendra Dasari, Mark Turner, Ashok P Reddy, Dongseok Choi, Phillip A Wilmarth, Ashley L McCormack, Larry L David, and Srinivasa R Nagalla. High-throughput identification of proteins and unanticipated sequence modifications using a mass-based alignment algorithm for ms/ms de novo sequencing results. *Anal Chem*, 76(8):2220–2230, Apr 2004.

[16] Richard I Somiari, Stella Somiari, Stephen Russell, and Craig D Shriver. Proteomics of breast carcinoma. *J Chromatogr B Analyt Technol Biomed Life Sci*, 815(1-2):215–225, Feb 2005.

[17] Stephen Tanner, Hongjun Shu, Ari Frank, Ling-Chi Wang, Ebrahim Zandi, Marc Mumby, Pavel A Pevzner, and Vineet Bafna. Inspect: identification of posttranslationally modified peptides from tandem mass spectra. *Anal Chem*, 77(14):4626–4639, Jul 2005.

[18] M. R. Wilkins, C. Pasquali, R. D. Appel, K. Ou, O. Golaz, J. C. Sanchez, J. X. Yan, A. A. Gooley, G. Hughes, I. Humphery-Smith, K. L. Williams, and D. F. Hochstrasser. From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis. *Biotechnology (N Y)*, 14(1):61–65, Jan 1996.

[19] Cathy H Wu, Rolf Apweiler, Amos Bairoch, Darren A Natale, Winona C Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele Magrane, Maria J Martin, Raja Mazumder, Claire O'Donovan, Nicole Redaschi, and Baris Suzek. The universal protein resource (uniprot): an expanding universe of protein information. *Nucleic Acids Res*, 34(Database issue):D187–D191, Jan 2006.

From Genes to Personalized HealthCare: Grid Solutions for the Life Sciences
N. Jacq et al. (Eds.)
IOS Press, 2007

BGBlast: A BLAST Grid Implementation with Database Self-Updating and Adaptive Replication

Gabriele A. TROMBETTI [a,b], Ivan MERELLI [a], Alessandro ORRO [a,b,1] and Luciano MILANESI [a]

[a] *Institute for Biomedical Technologies – National Research Council, via Fratelli Cervi 93, 20090 Segrate (MI), Italy*

[b] *Consorzio Interuniversitario Lombardo per L'Elaborazione Automatica, via R. Sanzio 4, 20090 Segrate (MI), Italy*

Abstract. BLAST is probably the most used application in bioinformatics teams. BLAST complexity tends to be a concern when the query sequence sets and reference databases are large. Here we present BGBlast: an approach for handling the computational complexity of large BLAST executions by porting BLAST to the Grid platform, leveraging the power of the thousands of CPUs which compose the EGEE infrastructure. BGBlast provides innovative features for efficiently managing BLAST databases in the distributed Grid environment. The system (1) keeps the databases constantly up to date while still allowing the user to regress to earlier versions, (2) stores the older versions of databases on the Grid with a time and space efficient delta encoding and (3) manages the number of replicas for each database over the Grid with an adaptive algorithm, dynamically balancing between execution parallelism and storage costs.

Keywords. Bioinformatics, adaptive database replication.

Introduction

BLAST [1], [2] is a well known and widely used bioinformatics application for comparing (usually unknown) "query" biological sequences, either genomic or amino-acidic, against a set of known "reference" sequences ("Blast Reference Database" or BRD in this paper). BLAST is a variation and approximation of the exhaustive dynamic-programming Smith-Waterman [3] algorithm for local sequence-alignment, resulting in a speed increase of 10-100x, at the expense of some sensitivity [4].

While BLAST sensitivity is generally regarded as still adequate for most circumstances, the speed of BLAST can still be scarce for certain massive computations, which are in fact performed rather commonly by many bioinformatics research groups.

The problem of BLAST speed can be addressed in various ways, the solutions usually belonging to the following groups (a) faster alternatives to BLAST, (b) BLAST execution on clusters and (c) BLAST execution in Grid. Pros and cons for (a) and (b)

[1] Corresponding Author: Alessandro Orro, Institute for Biomedical Technologies – National Research Council, via Fratelli Cervi 93, 20090 Segrate (MI), Italy; E-mail: alessandro.orro@itb.cnr.it

will be mentioned in the next chapter. The solution we present in this paper belongs to (c). As far as (c) is concerned, the main problems usually arising from BLAST execution in Grid are:

1. Defining and enforcing a policy for replication of the BRDs over the Grid. BRDs are large files needed during BLAST execution over the Grid. Due to their size they require allocation on Grid Storage Elements (SEs). Rising the number of replicas for a BRD reduces the Grid queue times for BGBlast runs using that BRD, but also rises the associated storage costs (see chapter 2.2.1 below). Due to the significant size, it is not reasonable to replicate every BRD on a large number of SEs; compromises have to be made.
2. Keeping the replicated BRDs up-to-date.
3. Optionally it might be profitable to store older versions of the BRDs so that BLAST users can reproduce and verify results obtained in the past. The problem in providing this feature is that keeping older versions of BRDs available normally has a very high storage cost.

In the Methods section more details are given regarding the above issues and on how we were able to address them.

1. Related Work

As mentioned in the Introduction, a number of solutions have been developed to address the problem of BLAST speed for large BLAST runs. In this chapter we will report about the main approaches.

1.1. Faster alternatives

Various alternatives to BLAST which are faster and similar in scope are available such as MegaBLAST [5], [6], BLAT [7] and PatternHunter [8]. These alternatives usually are different enough to be not suitable for exactly the same situations as BLAST is, or sometimes can have different drawbacks. As far as the examples are concerned, MegaBLAST and BLAT, albeit much faster, have a lower sensitivity than BLAST. PatternHunter on the other hand claims a similar sensitivity but is a commercial closed source product, and the algorithm is not known exactly. Such drawbacks might or might not be acceptable for the research, depending on the specific circumstances. In addition, researchers aiming at publishing their results might want to use specifically BLAST simply because its reliability is well established and cannot be object of discussion.

1.2. Cluster execution

Various solutions [9], [10], [11] have been developed to parallelize the BLAST algorithm for execution on computing clusters and supercomputers. These solutions have been used for quite some time now and are regarded as reliable. The main drawback of cluster execution for BLAST is the initial cost for purchasing the dedicated cluster, which is high, and might be unreasonably high -relatively speaking- in case the cluster is not going to be used full-time (uneven workloads).

1.3. Grid execution

A number of implementations of BLAST for the Grid environments already exist [12], [13], [14] but in general suffer from the problems already mentioned in the introduction. In this paper we present BGBlast, another Grid implementation for BLAST which we developed evolving the earlier GridBlast project carried out by Merelli and Milanesi [14]. In BGBlast we successfully addressed the issues mentioned in the introduction.

2. Methods

BGBlast (BioinfoGridBlast) has some unique advantages over the existing solutions. BGBlast is an innovative porting of BLAST onto the Grid providing the following capabilities (1) automatic update of the biological databases handled by BGBlast (2) adaptive replication of databases on the Storage Element Grid nodes (3) version regression for the biological databases. BGBlast is the evolution of the earlier project GridBlast [14] on top of which the features (1), (2) and (3) have been added:

1. Automatic Database Updater (ADU): ensures the users always work with the latest version of every Blast Reference Database (BRD), and this without needing human staff to manually monitor the release of newer versions of BRDs or manually performing database updates over the Grid.
2. Adaptive Replication (AR) for the BLAST Reference Databases: ensures that the most used BLAST databases are replicated more times than less used databases. The optimal number of replicas for each BRD is calculated dynamically based on the relative usage of the specific database in recent times. This keeps a constant optimization of Grid queue times vs Grid storage costs.
3. Version Regression for BLAST Reference Databases: allows the user of BGBlast to specify an older version of a certain BRD to be used for the computation. This allows the user to reproduce exactly computations obtained in the past, something that might be needed to confirm results that were obtained. The storage of older version of BRDs is performed with a delta-encoding efficient in both space (storage costs) and time (a short download time and a short time to patch a BRD for regressing it to an earlier version).

2.1. GridBlast core

BGBlast is composed of the following three functional parts: GridBlast core, Database Version Regression (DVR) and Automatic Database Updater (ADU). Here follows a more detailed description. GridBlast [14] is still the core for BGBlast, providing the following capabilities:

1. Factor J parallelization of large BLAST executions. This is done by splitting the user input into J even subset, each taking 1/J of the original time to execute. This is followed by the submission of J smaller BLAST jobs (1/J of query sequences against the target BRD) on the EGEE [17] Grid platform. J is chosen so that jobs of reasonable length are created: neither too small (Grid overhead would be comparatively large) nor too big (insufficient parallelization).

2. A rate limiting feature triggered on very large BLAST executions. This limits the rate at which the jobs are submitted to the Grid so to avoid a sudden massive Grid exploit.
3. Monitoring of every launched launched job and automatic resubmission in case of failure. This is still important nowadays, as the Grid platform is still new and reliability is not excellent.
4. Fetching of the results back after the completion of the Grid jobs. Merging of such results into a single BLAST results file.
5. A recent improvement of the core provides measurements of the queue times and CPU hours consumed by the J Grid jobs for each run of BGBlast. These measurements are passed to the Adaptive Replication Manager and are essential for the correct functioning of the AR functionality (see).

On top of the GridBlast core, the following functionalities have been implemented:

2.2. Adaptive Replication Manager (ARM)

The Adaptive Replication of BLAST Reference Databases is a BGBlast feature for optimizing the number of replicas for each BLAST database dynamically and adaptively.

2.2.1. Motivation for ARM

BLAST Reference Databases (BRDs) are large files, usually in the range 500MB-5GB, and are needed during the run of BLAST on the Grid CEs for each of the J BGBlast-generated jobs. Due to their size, it is not reasonable to download a BRD from a remote location. It is hence necessary to constrain the J jobs to execute on CEs having a replica of the user-requested BRD on a near (local network) SE.

Due to this constraint, the number of CEs to choose from for the BGBlast generated Grid jobs is limited. This impacts the queue times negatively and this is particularly true if the replicas of the requested BRD are few. A massive replication of every BRD on all the SEs of the Grid is not feasible either, because of their size which would make the storage costs unbearable.

Clearly, it is more useful to have additional replicas for BRDs used often, so that the queue times are reasonably small for the most common BGBlast runs, while it is better to have fewer or possibly only one replica for the BRDs used less frequently, in order to reduce the Grid storage costs.

Since the amount of usage of for each of the BRDs cannot be known in advance, we have implemented a dynamic, adaptive replication mechanism to balance between queue times and storage costs.

2.2.2. Methods for ARM

The ARM performs a D days moving average (usually D=10) of the CPU hours and queue times used for each reference database. This statistical measurement is used to compute the optimal number of replicas for each of the BLAST reference databases. This algorithm balances between the additional storage costs incurred in increasing the number of replicas and the benefit of the reduced queue times.

Additionally, when evaluating the addition of a replica the ARM engine also evaluates which of the SEs would be the most advantageous for a replica addition. Similarly, when evaluating the benefit of removing a replica, the ARM engine also evaluates the least advantageous of the currently existing replicas, that is, best for removal. See below for further details on the algorithm.

The measurements of used CPU-hours and queue times experienced for each BGBlast run, and implicitly for each BLAST database, are provided to the ARM by the GridBlast core (see). The dynamic variation of the number of replicas is evaluated, and possibly performed, at each BGBlast run and at the end of each day.

2.2.3. Algorithm details

BGBlast's ARM optimizes the number of replicas for each BRD separately, by minimizing the sum of the storage cost and user wait time cost. The algorithm is an iterative algorithm which converges on the optimal number of replicas and the optimal location for them, simultaneously.

The ARM optimization algorithm at each cycle evaluates the benefit of the *addition* of one replica and the benefit of the *removal* of one replica.

During the evaluation of the *addition* of one replica, the ARM takes into account the specificity of each Grid location suitable for replication (i.e. every SE not yet holding a replica), hence finding the best location for an added replica. The ARM then evaluates whether the addition of a replica in that specific place is profitable or not, using the costs formula.

The best location for adding a replica is ideally a SE having a large amount of free disk space (so as to cause proportionally little impact when adding the replica) near a CE being large in the number of nodes (a larger computing power means that the job queue is generally consumed more quickly).

The costs formula for evaluating variations in replicas numbers considers the Grid queue times to be inversely proportional to the number of nodes useable by BGBlast, i.e. those having a replica nearby (see chapter 2.2.1). The correctness of this assumption can in fact be demonstrated under some simplifying assumptions. The cost of a minute of user wait time is to be specified in the BGBlast configuration file.

The cost of Grid storage is the other cost to be specified in the BGBlast configuration file. The cost of storage is to be expressed in terms of cost per percent of free storage space occupied per day on a SE. This approach was chosen for reflecting the intuitively higher impact on other Grid users that a GB-sized file has when uploaded on a small or already full SE compared to the impact it has when uploaded on a SE with plenty of free space.

The ARM engine hence works by minimizing the sum of the storage cost and user wait time cost, for each BRD separately.

The process for evaluating the benefit of the removal of one replica is analogous. The worst existing replica is chosen using the same kind of analysis as described above. The cost formula is then recomputed while simulating the removal of the "worst" replica, and the result obtained in this way is compared to the cost associated to the current situation. If the cost after the removal of the replica appears lower, the replica is removed.

This algorithm converges quickly.

2.3. Automatic Database Updater (ADU)

BGBlast's ADU engine constantly monitors FTP sites for newer versions of the BRDs registered to be handled by BGBlast. If a newer version of a BRD is detected, the ADU automatically updates all the replicas of such BRD over the Grid. This is not the only action performed by the ADU: the ADU also computes an xdelta patch for regressing the newer version of the BRD to the earlier version of the BRD now being replaced, and uploads the said xdelta patch on a predefined SE. The xdelta patch computed by the ADU, together with the xdelta patches computed during previous database updates, is needed for the DVR functionality (see).

Such xdelta patches are many times smaller than any version of the BRD they refer to, and this makes the storage costs reasonable. In order to further reduce the storage costs, we decided to keep only one replica for the xdelta patches. Also see chapter 2.4 on this topic.

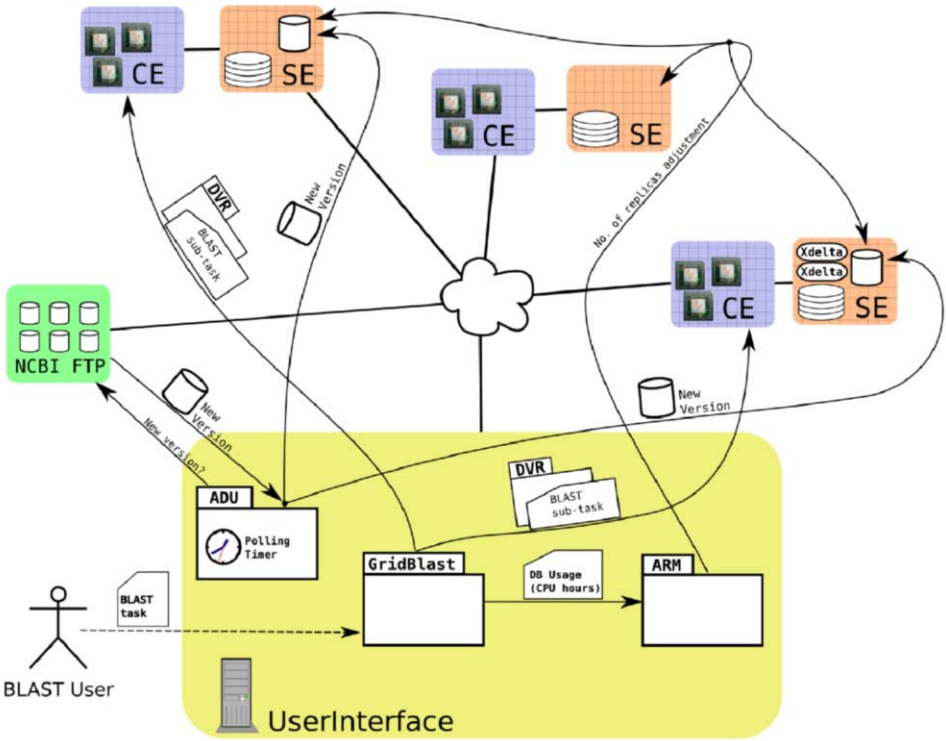


Figure 1. *Interactions between parts of BGBlast and Grid elements.* The user interacts with the *GridBlast* core for launching a BLAST task on the Grid. The BLAST task is split by GridBlast in jobs of equal size ("*sub-tasks*") and sent to Computing Elements (*CEs*) for execution. The code for performing the *DVR* is sent to the CEs together with the Grid jobs. The xdelta patches are stored on a single SE (not replicated) as shown, however, the interaction between the DVR code and the xdelta patches is not shown in the figure. The *ADU* uses a timer-based polling to detect and fetch new versions of a database from the *FTP* site of origin, then it updates the databases located on the Storage Elements (SEs). The *ARM* receives notification of CPU time spent on a database by the GridBlast core, then it adjusts the number of replicas of the database on the SEs.

2.4. Database Version Regression (DVR)

BGBlast provides an option for specifying a *version* (in terms of *date*) of the BRD to be used for the BLAST computation, along with the *name* of the BRD. The requested version of the BRD is obtained from the latest version of the BRD by applying the ADU-generated xdelta patches in sequence, from the newest to the oldest. Each xdelta patch regresses the BRD by one version, and this action is performed until the requested version is reached.

The version regression operation is performed on the Computing Element (CE) after the download of the BRD from the near (local network) SE and prior of starting the computation.

The download of the xdelta patches is generally remote, as the patches are only replicated once on the Grid (see chapter 2.3), and this is in contrast with the download of the BRD (latest version) which is over a local network (see chapter 2.2.1). However, due to the small size of the patches, the patches' download time rarely surpasses that of the BRD (latest version). Since the DVR is also a relatively uncommon request by users, we considered the patches download time an acceptable overhead.

3. Conclusion

We have shown how we were able to improve BLAST performances by distributing the BLAST execution on the EGEE Grid infrastructure, leveraging the power of thousands of CPUs. Additionally we have shown how we could further reduce the queue times while impacting the Grid storage costs as little as possible, by using adaptive replication for BLAST databases, and how we were able to provide version regression for such BLAST databases. Our work can shorten biologists' waiting times for their research, and also acts as a proof of concept showing what can be done to optimize Grid resources and Grid applications.

BGBlast will be available as a service for EGEE in the upcoming BioinfoGRID portal by CNR-ITB. CNR-ITB will be in charge of the maintenance for the Blast biological databases on the Grid.

Acknowledgment

This work was supported by the Italian FIRB-MIUR project “Italian Laboratory for Bioinformatics Technologies – LITBIO” [15] and by the European Specific Support Action BioinfoGRID [16] and EGEE [17] projects.

References

[1] S.F. Altschul, W. Gish, W. Miller, E.W. Myers and D.J. Lipman: **Basic Local Alignment Search Tool**, *J. Mol. Biol.* 1990;215(3):403-10

[2] BLAST - http://www.ncbi.nlm.nih.gov/BLAST/

[3] T.F. Smith and M.S. Waterman: **Identification of Common Molecular Subsequences**, *Journal of Molecular Biology* 1981;147:195-7

[4] http://searchlauncher.bcm.tmc.edu/help/AlignmentScore.html

[5] Z. Zhang, S. Schwartz, L. Wagner, and W. Miller (2000): **A greedy algorithm for aligning DNA sequences**, *J. Comput. Biol.* 2000;7:203–14

[6] http://www.ncbi.nlm.nih.gov/blast/megablast.html

[7] W.J. Kent: **BLAT – the BLAST-like alignment tool**, *Genome Res* 2002;12:656–64

[8] B. Ma, J. Tromp, and M. Li: **PatternHunter: faster and more sensitive homology search**, *Bioinformatics* 2002;18(3):440–5

[9] Y. Qi, F. Lin: **Parallelisation of the blast algorithm**, *Cell Mol Biol Lett.* 2005;10(2):281–5

[10] D.R. Mathog: **Parallel BLAST on split databases**, *Bioinformatics* 2003;19(14):1865–6

[11] A. Darling, L. Carey, and W. Feng: **The Design, Implementation, and Evaluation of mpiBLAST**, *4th International Conference on Linux Clusters* June 2003; San Jose, CA

[12] M. Bayer, A. Campbell, D. Virdee: **A GT3 based BLAST grid service for biomedical research**, *Proceedings of the UK e-Science All Hands Meeting* 2004

[13] F. Konishi, and Y. Shiroto, and R. Umetsu, and A. Konagaya: **Scalable BLAST Service in OBIGrid Environment**, *Genome Informatics* 2003;14:535–6

[14] I. Merelli, L. Milanesi: **High performance implementation of BLAST using GRID technology**, *Proceedings BITS* 2005: p.59

[15] LITBIO – Laboratory for Interdisciplinary Technologies in Bioinformatics – http://www.litbio.org

[16] BioinfoGRID – Bioinformatics Grid Applications for Life Science – http://www.itb.cnr.it/bioinfogrid

[17] EGEE – Enabling Grids for E-SciencE – http://public.eu-egee.org/

From Genes to Personalized HealthCare: Grid Solutions for the Life Sciences
N. Jacq et al. (Eds.)
IOS Press, 2007

Advances in the biomedical applications of the EELA Project

Vicente HERNÁNDEZ[a], Ignacio BLANQUER[a], Gabriel APARICIO[a], Raúl ISEA[b], Juan Luis CHAVES[b], Álvaro HERNÁNDEZ[b], Henry Ricardo MORA[c], Manuel FERNÁNDEZ[c], Alicia ACERO[d], Esther MONTES[d] and Rafael MAYO[d,1]
(On behalf of the EELA Project)
[a] *Universidad Politécnica de Valencia, ITACA – GRyCAP, Spain*
[b] *Universidad de los Andes, Venezuela*
[c] *Cubaenergía, Cuba*
[d] *CIEMAT, Spain*

Abstract. In the last years an increasing demand for Grid Infrastructures has resulted in several international collaborations. This is the case of the EELA Project, which has brought together collaborating groups of Latin America and Europe. One year ago we presented this e-infrastructure used, among others, by the biomedical groups for the studies of oncological analysis, neglected diseases, sequence alignments and computational phylogenetics. After this period, the achieved advances are summarised in this paper.

Keywords. Biomedical applications, E-Science, Grid infrastructure, Information technology

Introduction

Funded by the European Commission, the EELA Project (E-Infrastructure shared between Europe and Latin America) [1] began in January 2006 to build a digital bridge between the existing e-Infrastructure initiatives in Europe and those that were emerging in Latin America, throughout the creation of a collaborative network that shares an interoperable Grid infrastructure to support the development and test of advanced applications.

One of the areas of work, corresponding to one of the four Work Packages of the Project, is the identification and support of Grid enhanced applications. This scientific research covers several fields, but due to the social impact in the Latin American society,

[1] Corresponding Author: Rafael Mayo, CIEMAT, Avda. Complutense, 22, 28040 Madrid, Spain; E-mail: rafael.mayo@ciemat.es

one of the pillars of the Project is Biomedicine and, consequently, the applications that can be ported to the Grid.

Some of them, falling in the three typical categories of bioinformatics applications, computational biochemical processes and biomedical models, were selected [2] and started to be deployed on the pilot EELA infrastructures for both production and dissemination purposes.

This document describes the achieved advances in these biomedical applications in the last year; this is, the half of the lifetime of EELA. For a more detailed explanation of the applications, the reader can consult [2] and [3].

1. GATE

The C++ platform based on the Monte Carlo GEANT4 software [4] GATE (GEANT4 Application for Tomographic Emission) [5] models nuclear medicine applications, such as PET and SPECT within the OpenGATE collaboration [6]. The computational-intensive Monte Carlo simulations prevent hospitals and clinical centres from using them for daily practice in radiotherapy and brachytherapy treatment planning. As a result, the objective of GATE is to use the Grid environment to reduce the computing time of Monte Carlo simulations providing a higher accuracy in a reasonable period of time

Nine Cuban centres are currently testing, as users, the results of the simulation of radiotherapy treatments using realistic models that GATE provides in two main oncological problems:

- Thyroid Cancer (the diseases of thyroids are one of the 5 main causes of Endocrinology treatments) [7]; and,
- Treatment of metastasis with P^{32} isotopes [8] by means of the brachytherapy improving the knowledge on the doses captured from the different tissues by accurate simulation.

The EELA consortium has provided the support for the installation and integration of a small Grid site in Cuba in order to run the simplest jobs concerning Monte Carlo simulations. For those cases where the computer load is greater, a previous preparation of the job is done in Cuba and, after that, the necessary material is brought physically to another EELA sites with a higher bandwidth in order to do the submission process. For this purpose, GATE has been installed in several EELA sites.

Because of all these reasons, the expected impact is very important because many cases are being studied and an important step beyond in the diagnosis of the aforementioned diseases will be reached. Some medical results are expected through 2007.

As a conclusion and added value, Grid will increase the performance to the application, but in this case, it will even be more important, since it is an enabling technology opening the doors to a new range of applications and possibilities. All the centres from Cuba currently testing this application bring to the EELA community around 90 cases per month.

2. WISDOM

The objective of WISDOM (Wide in Silico Docking of Malaria) is the proposition of new inhibitors for a family of proteins produced *by Plasmodium Falciparum* due to this protozoan parasite causes malaria and affects around 300 million people and more than 4 thousand people die daily in the world. The cross-resistance to antimalarials produced by the focus on a limited number of biological targets has derived in a drug resistance for all classes of antimalarials except artemisinins. Because of this the development of new drugs with new targets is necessary, but the process is cost so the economic profit is not clear for the drug manufacturers.

This WISDOM application consists on the deployment of a high throughput virtual screening platform in the perspective of *in silico* drug discovery for neglected diseases.

The interest of the EELA partners centred in three actions:

- The study of new targets for new parasitory diseases.
- The selection of new targets for malaria.
- The contribution with resources for the WISDOM Data Challenge.

The WISDOM platform [9] has performed its second High-Throughput virtual Docking of million of chemical compounds available in the databases of ligands to several targets in the fall of 2006 (see [10] for details of the first one). In this WISDOM Data Challenge-II [11] ULA has proposed two targets in *Plasmodium Vivax,* which is of special interest in the Latin American society.

These targets have been accepted by the consortium and docked within the EELA as was agreed by the WISDOM DC-II decision-makers. This process has been coordinated by UPV and all the EELA sites have acted as donor of computational and storage resources, running effectively more than 40% of the jobs in the Latin American sites. The results will be presented in future works.

Each target requires executing 2422 jobs which take around 1 CPU day each. This large demand of resources is being tackled with the collaboration of resources from institutions of Brazil, Mexico, Venezuela, Spain and Italy. At the end of 2006, half of the experiment has been completed producing 53 GB of results due to 100% of the jobs have been run.

On the one hand, for the EELA infrastructure, WISDOM-DC II has been a crucial test; on the other hand, this high completion in the docked process has debugged some unexpected problems in the original files.

3. BLAST in Grid

The study of the functionality of the different genes and regions is one of the most important efforts on the analysis of the genome. If the queries and the alignments are well designed both functional and evolutionary information can be inferred from sequence alignments since they provide a powerful way to compare novel sequences with previously characterised genes.

The Basic Local Alignment Search Tool (BLAST) [12] finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to databases and calculates the statistical significance of matches. This process of finding homologous of sequences is very computationally-intensive since the searching alignment of a single sequence is not a costly task, but normally, thousands of sequences are searched at the same time.

The biocomputing community usually relies on either local installations or public servers, such as the NCBI [13] or the gPS@ [14], but the limitations on the number of simultaneous queries make this environment inefficient for large tests. Moreover, since the databases are periodically updated, it will be convenient to do the same with the results of previous studies.

BLAST has been ported to the Grid according to different approaches [15] because the number of fragments to be analysed and the periodical updating of the information will be increased. EELA has adopted the availability of an independent Grid-enabled version integrated on the Bioinformatics Portal of the Universidad de los Andes [16] providing registered users with results in a shorter time. Basically the users access the service through this web portal and, after that, to the EELA Grid with a Gate-to-Grid, i. e., an EELA Grid node which provides a WSRF-Based web interface.

The security is provided by means of a MyProxy server which generates manually and temporally certificates that will be retrieved by the UI when required. Besides, some improvements in the return of a selected number of hits and a new system monitoring method have been achieved in the post-processing stage.

BLAST in Grid (BiG) has been used for searching similar sequences and inferring their function in parasite diseases such as the Leishmaniasis (mainly *Mexican Leishmania*), Chagas (mainly *Trypanosoma Cruzi*) and Malaria (mainly *Plasmodium vivax*) producing in this way several scientific results of interest from the first moment. For example, work on the complete genome of the *Plasmodiun Falciparum* for the identification of DHFR

antigenic proteins has already been done. At the same time, it means a technical innovation in the field and the first "own EELA application" running in the Grid.

4. Phylogeny (MrBayes)

A phylogeny is a reconstruction of the evolutionary history of a group of organisms used throughout the life sciences, as it offers a structure around which to organize the knowledge and data accumulated by researchers. Computational phylogenetics has been a rich area for algorithm design over the last 15 years. The inference of phylogenies with computational methods is widely used in medical and biological research and has many important applications, such as gene function prediction, drug discovery and conservation biology [17].

The most commonly used methods to infer phylogenies include cladistics, phenetics, maximum likelihood, and Markov Chain Monte Carlo (MCMC) based Bayesian inference. These last two depend upon a mathematical model describing the evolution of characters observed in the included species, and are usually used for molecular phylogeny where the characters are aligned nucleotide or amino acid sequences.

Due to the nature of Bayesian inference, the simulation can be prone to entrapment in local optima. To overcome local optima and achieve better estimation, the MrBayes program [18] has to run for millions of iterations (generations) which require a large amount of computation time. For multiple sessions with different models or parameters, it will take a very long time before the results can be analyzed and summarized.

The phylogenetic tools are widely demanded by the Latin America bioinformatics community. A Grid service for the parallelised version of MrBayes application is currently being developed and a simple interface will be deployed on the bioinformatics portal of Universidad de los Andes [19] in a similar way to that done for BiG. Nevertheless, some previous works with this parallelization have been successfully submitted and executed in the EELA infrastructure and been demonstrated in the first EELA conference [20].

However, the parallelisation model of MrBayes is different from the approach followed on BiG. Finer-grain parallelism is present and database splitting is not possible. More efficient approaches are being studied to enlarge the availability of resources for this problem, which typically is bounded by memory constraints.

5. Conclusion and new challenges

The EELA e-infrastructure is permitting various collaborative groups in Latin America to use more powerful computational resources than those available on their centres. This

achieves that the lines of investigation stated in the document can be feasible as the computational requirements for them are being met and new results are being offered to the scientific community.

The EELA project currently has four biomedical pilot applications running and the support of the portal developed at ULA, but is also looking for new ones. For this purpose, it is offering its support with initiatives like Grid School [21]. As a result, new biomedical applications such as EMBOSS [22] will be ported to the Grid very soon. With this free Open Source software analysis package specially developed for the needs of the molecular biology user community EELA will keep on working in the biomedical field.

References

[1] http://www.eu-eela.org
[2] EELA Deliverable D3.1.1. http://www.eu-eela.org/eela_deliverables_list.php
[3] M Cárdenas, V Hernández, R Mayo, I Blanquer, J Pérez-Griffo, R Isea, L Nuñez, HR Mora, M Fernández, "*Biomedical Aplications in EELA*", Studies in Health Technology and Informatics 2006;120;397-400.
[4] S. Agostinelli *et al.*, "GEANT4—a simulation toolkit", Nuclear Instruments and Methods in Physics Research A 2003;506;250-303
[5] S. Jan, et al., *"GATE: a simulation toolkit for PET and SPECT"*, Phys. Med. Biol. 2004;49**;**4543-4561
[6] http://opengatecollaboration.healthgrid.org/
[7] D. Navarro *"Epidemiología de las enfermedades del tiroides en Cuba"*, Rev Cubana Endocrinol. 2004;15.
[8] J. Alert, J. Jiménez, *"Tendencias del tratamiento radiante en los tumores del sistema nervioso central"*, Rev Cubana Med 2004; 43(2-3).
[9] N. Jacq, *et al.*, "Virtual Screening on Large Scale Grids", to be published in Parallel Computing, (2007)
[10] N. Jacq, *et al.*, "Grid-enabled Virtual Screening against malaria", to be published in Journal of Grid Computing, (2007).
[11] http://wisdom.healthgrid.org/index.php?id=139
[12] SF Altschul *et al.*, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucl. Acids Res. 1997;25;3389-3402
[13] http://www.ncbi.nlm.nih.gov/
[14] http://gpsa.ibcp.fr/
[15] G Aparicio, S Götz, A Conesa, JD Segrelles, I Blanquer, JM García, V Hernández, "*Blast2GO goes Grid: Developing a Grid-Enabled Prototype for Functional Genomics Analysis*", Studies in Health Technology and Informatics 2006;120:194-204
[16] http://www.cecalc.ula.ve/blast/
[17] K. Lesheng, *"Phylogenetic Inference Using Parallel Version of MrBayes".* http://www.nus.edu.sg/comcen/svu/publications/hpc_nus/sep_2005/ParallelMrBayes.pdf
[18] F. Ronquist, JP Huelsenbeck., "MrBayes 3: Bayesian phylogenetic inference under mixed models." Bioinformatics 2003;19;1572-1574
[19] http://portal-bio.ula.ve/
[20] http://documents.eu-eela.org/getfile.py?recid=273
[21] http://www.eu-eela.org/egris1/index.html
[22] http://emboss.sourceforge.net/what/#Overview

From Genes to Personalized HealthCare: Grid Solutions for the Life Sciences
N. Jacq et al. (Eds.)
IOS Press, 2007

Sleep Medicine as a Scenario for Medical Grid Application

SEBASTIAN CANISIUS [1], THOMAS PLOCH [1], KARL KESPER [1], MATTHEW SMITH [2], BERND FREISLEBEN [2], THOMAS PENZEL [3]

[1] *Philipps-University Marburg, Faculty of Medicine, Sleep Disorders Centre*
[2] *Philipps-University Marburg, Faculty of Mathematics and Informatics*
[3] *Charité Universitätsmedizin Berlin, Interdisziplinäres schlafmedizinisches Zentrum*

Abstract: Sleep medicine is gaining more and more interest and importance both within medical research and clinical routine. The investigation of sleep and associated disorders requires the overnight acquisition of a huge amount of biosignal data derived from various sensors (polysomnographic recording) as well as consecutive time-consuming manual analysis (polysomnographic analysis). Therefore, the development of automatic analysis systems has become a major focus in sleep research in the recent years, resulting in the development of algorithms for the analysis of different biosignals (EEG, ECG, EMG, breathing signals). In this study, an open source algorithm published by Hamilton et al. was used for ECG analysis, whereas the analysis of breathing signals was done using an algorithm published by Clark et al. using also variations of the intra-thoracic pressure for the detection of breathing disorders. The electromyogram (EMG) analysis was done with a self-made algorithm, whereas EEG analyses are currently under development, using both frequency analysis modules and pattern recognition procedures. Although all these algorithms have proved to be quite useful, their validity and reliability still needs to be verified in future studies. Taking into account that during a standard polysomnographic recording data from approximately 8 hours of sleep are collected, it is imaginable that processing this amount of data by the described algorithms very often exceeds the calculating capacity of current standard computers. Using Grid technology, this limitation can be transcended by splitting biosignal data and distributing it to several analysis computers. Therefore, Grid based automatic analysis systems may improve the effectiveness of polysomnographic investigations and thereby diminish the costs for health care providers.

Keywords: sleep medicine, grid computing, polysomnography

Introduction

Sleep loss, excessive fatigue, stress and inattention constitute the social diseases of our century. Within the "24 hour society" people tend more and more to exchange sleep and serenity for gain or pleasure. This gradually leads to an excessive rate of sleep disorders, roughly 20% of the population suffer from one. One major symptom, associated with the occurrence of sleep disorders, is the excessive daytime sleepiness (EDS), showing a prevalence of 5-10% in the young and middle-aged and even 20-30% in the older-aged[1]. Furthermore, EDS and consecutive microsleep have become a major cause for hazardous car accidents over the last decades[2;3]. Taking these facts into consideration it is understandable that sleep medicine and also sleep related research is characterized by a rapidly growing interest.

Among the vast amount of different sleep disorders that are known up to now, some are well investigated with regard to prevalence, causes and treatment. This paper will focus on new investigational approaches for two sleep disorders having a considerable high prevalence and one sleep disorder with dramatic effects for the suffering patient. All three sleep disorders commonly cause excessive daytime sleepiness.

The obstructive sleep apnea syndrome (OSAS) with a prevalence of 4% in men and 2% in women[4] is a well investigated sleep disorder with regard to occurrence, associated symptoms, risk factors as well as overall consequences[5;6]. It is characterized by repetitive cessations of breathing during the night caused by endogenous obstructions of the upper airway followed by awakenings restoring ventilation. These awakenings, however, have impairing effects on the continuation of sleep but also on the cardiovascular system due to repetitive increases in blood pressure and heart rate during the night.

A second, also quite extensively investigated sleep disorder is the Restless Legs Syndrome (RLS), a sensory-motor disorder characterized by dysesthesia and leg restlessness occurring predominantly at night during periods of immobility. The sensations associated with RLS and the urge to move usually interfere with the ability to fall and/or stay asleep. A recent study by Winkelmann et al. revealed a prevalence of daily RLS symptoms of 4.2% in males and 5.4% in females[7;8].

Narcolepsy, the third sleep disorder this paper will focus on, very often has dramatic effects on the quality of life and employment status of the patient. Although the prevalence is quite low (app. 47 of 100.000), this disorder has been extensively investigated due to its dramatic effects for the patient and the immense socioeconomic impact[9].

The investigation of all three sleep disorders requires a very sophisticated overnight examination of patients in a specialized sleep centres, performing a so called polysomnography. During a polysomnography different biological signals of the human body are conducted and digitally recorded. A very important part of each polysomnographic recording is the conduction of the electroencephalogram (electrical activity of the brain) in combination with electrooculogram (eye movements) and electromyogram (electrical muscle activity) using electrodes connected to the patient's body. With the help of these signals, experienced technicians can analyze the overnight sleep of a patient with regard to sleep depth, awakenings and also sequence of sleep depths using standardized rules for the analysis[10]. This results in a sequence of so called sleep stages, allowing a sleep physician to rate a patient's sleep structure and also the effectiveness of sleep for a patient.

In addition to the investigation of brain waves, also ECG waves (electrocardiogram, electrical activity of the heart), respiratory related signals and electrical activity of leg muscles are recorded and analysed in polysomnographic investigations. For example, the number of breathing cessations occurring during the night helps to diagnose the OSAS. For the diagnosis of RLS, periodically occurring leg movements are counted. And one new approach for the investigation of narcolepsy is the determination of short muscle activations, the so called twitches, during the night.

The acquisition of all described biological signals during a whole night (usually app. 8h) produces a huge amount of digital data. For an interpretation of this data with regard to recognition of sleep disorders, usually a time consuming manual analysis of the recorded signals is performed. Furthermore, the quality of the manual analysis depends mainly on the expertise of the scoring technician despite there exist specific rules for most scoring procedures. Furthermore interindividual differences between human scorers make scoring results sometimes barely comparable between different sleep centres.

Automated analysis systems, being independent from human expertise and interindividual differences, have therefore become a major research focus in sleep medicine over the past decades. One major problem has always been the vast amount of data that needs to be processed by automated analysis systems, very often exceeding the calculating capacity of standard computers.

As each biological signal is stored in a single channel of the digital recording, all recorded data can be easily split and analysed separately. Thinking of GRID technology and its primary objective to split and distribute data in small packets for analysis on many different connected computers, digital polysomnographic investigations turn out to be the ideal area for the implementation of GRID based automated analysis systems.

1. Methods

For the automated analysis of polysomnographic recordings to diagnose the above described sleep disorders, our group has developed different algorithms. The first algorithm automatically analyzes the ECG signal and detects the different waves and spikes and calculates the heart rate (time interval between R waves). The second algorithm performs automated analyses of respiratory signals (nasal airflow and esophageal pressure changes) and detects inspiratory airflow limitations. A third algorithm was created for the analysis of the electromyogram signal, determining the muscle tone as well as intermittent activations like periodic leg movements or short muscle activations (twitches).

The analysis of the ECG consists of a fully automated analysis of the ECG wave and automatic detection of QRS complexes (electrical innervation of the heart muscle) within the ECG wave, calculation of R-R-intervals and thereby calculation of heart rate and also heart rate variations during the night. For this purpose, an algorithm developed and published by Hamilton et al. has been used[11]. In Figure 1 the results of the automatic detection of QRS complexes of one example recording are shown. Using this detection algorithm, the width of QRS complexes and the interval between the R-waves can be calculated.

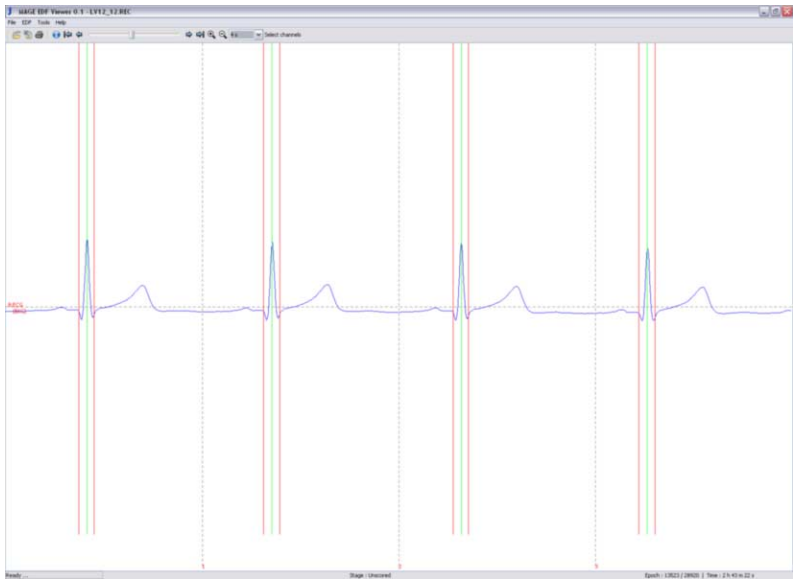


Figure 1: Automatic detection of QRS complex within the ECG signal.

As shown in figure 2, the distribution of R-R interval duration (y-axis) throughout a whole night recording, which is typically around 8h, can be calculated from this algorithm (time on x-axis), thereby compressing the ECG signal data to values ready for further sophisticated analysis.

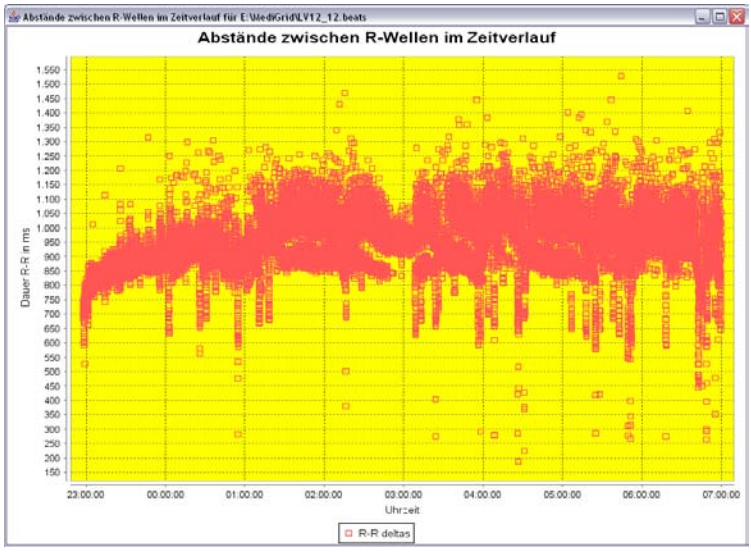


Figure 2: R-R-Intervals throughout the night recording. Although the mean heart rate during the night can be clearly identified, there is a remarkably high variance during some periods of the night. These extreme values may be due to body movements of the patient, very often interfering with the recorded biological signals.

Figure 3 shows the results of a percental distribution of R-R intervals during the whole recording. Thereby the physician can get an overview of the heart rate during the night but also detect frequent variations in heart rate, being a possible indicator for sleep disorders like OSAS or RLS. Furthermore, the analysed data can be used for Fourier transform calculations, calculations of QT-interval (activity phase of the heart).

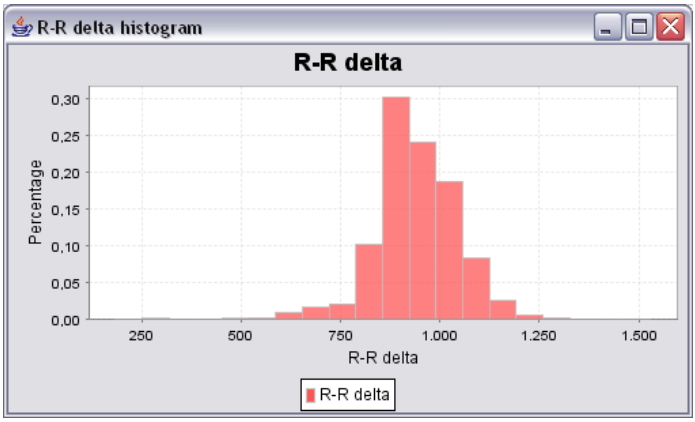


Figure 3: Histogram of the R-R-intervals during the night.

As described in the Introduction, OSAS is accompanied by breathing cessations during the night and consecutive awakening reactions. Usually these breathing cessations (apneas) and also hypoventilations (hypopneas) are sought and counted, using standardized criteria[12;13]. The amount of events during sleep is then one parameter of the treatment decision.

However, not only complete cessations of breathing can cause such awakening reactions and associated increases in nightly blood pressure and heart rate. Even a limited inspiratory airflow can cause such events during the night[14]. The detection of a limited inspiratory airflow, however, represents a difficult task even for experienced technicians. For this reason, the developed algorithm uses not only the nasal airflow signal (amount of air flowing through the nose/min) but also pressure variations in the esophagus for the detection of those inspiratory flow limitations. Usually, this esophageal pressure decreases when we breathe in and increases when we breathe out. Excessive decreases in esophageal with lacking increases in nasal airflow can be a possible sign for inspiratory flow limitations. The algorithm used for this purpose analyzes both signals for the occurrence of the inspiratory flow limitations[15].

Figure 4 shows exemplary results of the developed algorithm. It first detects the inspiratory and expiratory part of the recorded signal. Then esophageal pressure swings and nasal airflow signals are analysed to detect inspiratory flow limitations. The red dot in the lower part of the picture indicates that the marked breath was identified to be flow limited.

Apart from breathing disorders that may have impairing effects on sleep, there also exist neurologic disorders that can affect sleep. The other two sleep disorders described in the Introduction (narcolepsy, restless legs syndrome) represent such neurologic disorders.

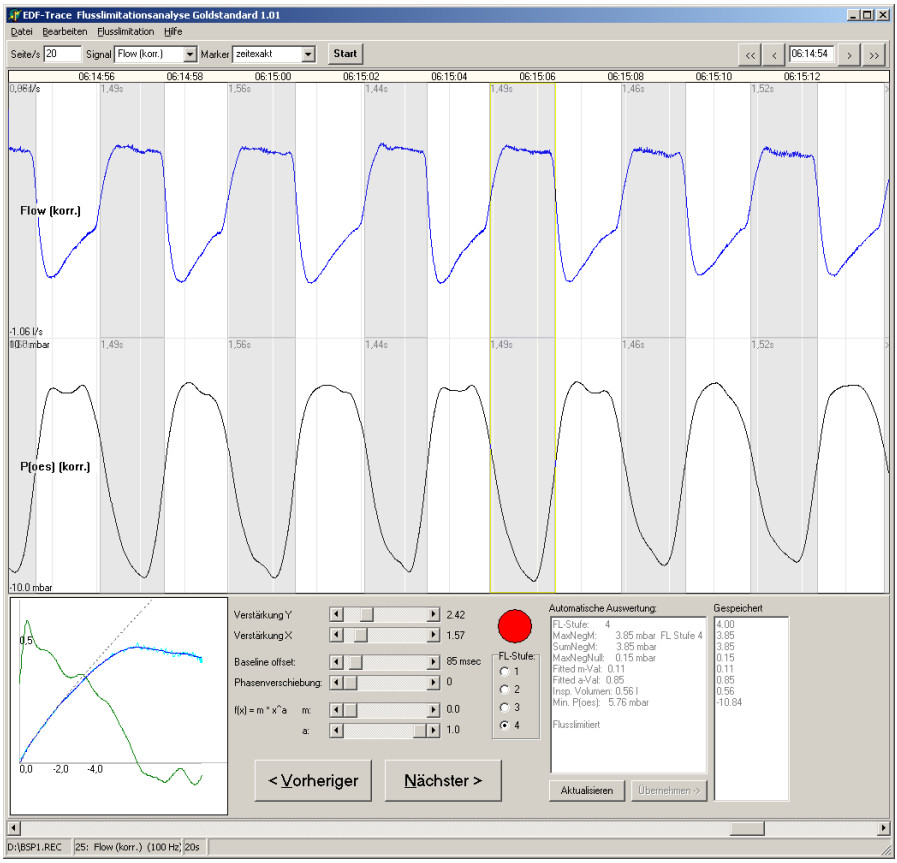


Figure 4: Automatic detection of inspiration (grey background) and expiration (white background), automatic detection of flow limited inspirations

For both disorders, the analysis of muscle activity with the help of the electromyogram (EMG) during sleep can alleviate the finding of the correct diagnosis. The algorithm described here was developed to analyse the overall muscle tone of the recorded EMG signal and search for the occurrence of periodic leg movements during sleep, which can be an indicator for the restless legs syndrome or the periodic leg movement syndrome. Short muscle activations during sleep (so called twitches), which frequently occur in narcolepsy, can also be detected.

Similar to the diagnosis of sleep disordered breathing, periodic leg movements are also scored according to standard criteria and then counted in order to get an estimate of the severity of the disease[16;17].

The analysis is done by calculating upper and lower envelopes for the EMG signal, thereby calculating the amplitude. EMG activations with an interval smaller than 1 second between the amplitude peaks are considered as one activation. Thereby short activations (twitches) can be distinguished from longer activations.

In Figure 5, the results of the automated amplitude calculation using upper and lower envelopes are shown.

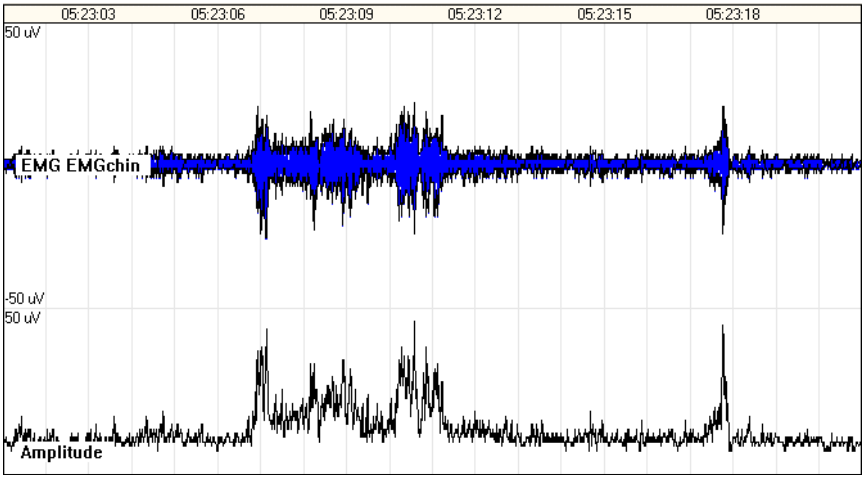


Figure 5: Automated analysis of EMG activations by calculation of an upper and lower envelope and consecutive calculation of the amplitude.

In Figure 6, the detected events are shown in the upper trace. The middle trace shows the described summarization of events. In the lower trace, the original EMG signal is shown.

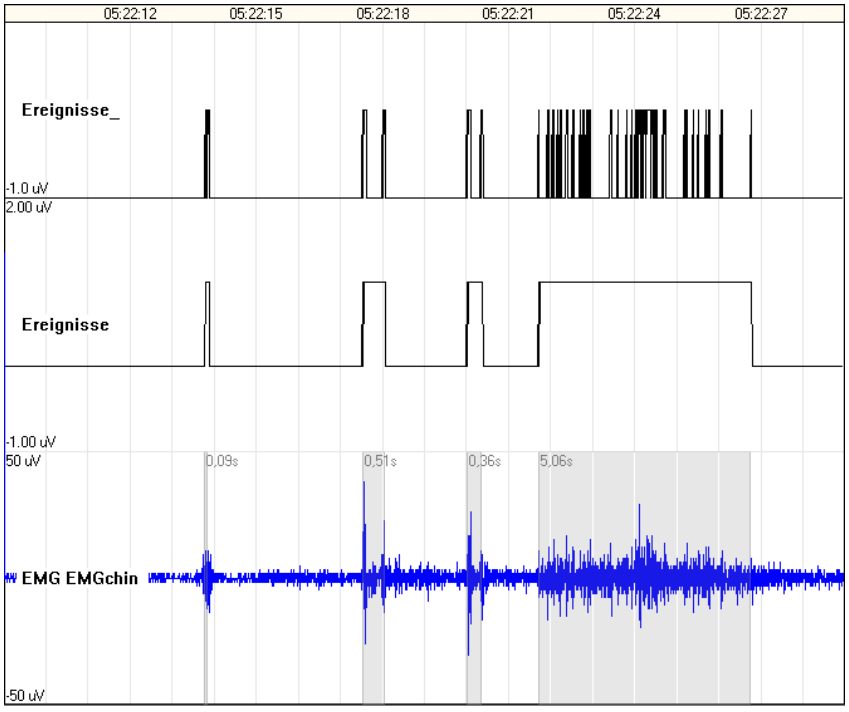


Figure 6: Summarization of coherent EMG activations (interval between activations < 1 sec) to distinguish between short twitches and longer muscle activations.

2. Results and Discussion

First results regarding the automated ECG analysis showed a good detection rate of QRS complexes with an exact calculation of R-R-intervals compared to manual analysis. Using this method, changes in heart rate that may be associated with sleep related disorders seem to be detectable with sufficient precision.

Comparative tests of the algorithm for the analysis of breathing signals showed a precise and thorough detection of inspiration and expiration. Furthermore, the detection algorithm for inspiratory flow limitations in combination with esophageal pressure changes led to a good automatic recognition of flow limited breaths compared to expert opinion. However, further studies investigating the validity of the detection of inspiration and expiration have to be performed, while the validity of the detection of inspiratory flow limitation has been investigated in a study by Clark et al. in a small population of 7 patients [15].

Comparing the manual and automatic scoring of periodic leg movements, the accordance between both methods appeared to be good both for the detection of periodic leg movements and for the detection of short twitches appearing in RBD patients. However, there are no objective scoring criteria for EMG twitches as they exist for periodic leg movements. Further studies are necessary to confirm the reliability and validity of the algorithm. In future publications, the underlying procedures for EMG analysis have to be described in detail.

Limitations with regard to the estimation of usefulness of the presented approaches are lacking information about consumed resources and computational power.

Automatic detection algorithms for sleep stages, as described in the introductory section, are currently in preparation. They will be based on the analysis of frequencies of the EEG waves with regard to alpha, theta, delta and sigma activity. Furthermore, detection of features like sleep spindles and K-complexes in combination with a detection of slow and rapid eye movements from the EOG signal will be integrated in this algorithm. This will allow the correct classification of sleep stages according to the current Rechtschaffen and Kales rules [10].

3. Conclusion

All developed algorithms for the analysis of biosignals from overnight recordings showed a good accordance to expert opinion ratings. Furthermore, all algorithms are written in standard programming languages (Delphi, Java) and can therefore be easily implemented into GRID environments.

All data needed for the analysis can be split into small packets, as polysomnographic data is recorded in separate channels from different sensors. We therefore believe that in short future, Grid based analysis of polysomnographic recordings will be available using the described scenarios. In combinations with different approaches, e.g. to use the ECG changes for the detection of sleep related disorders, it may be possible to have a Grid based polysomnographic analysis. As this

can have improving effects on the effectiveness of the analysis of overnight recordings, costs for polysomnographic investigations that have to be raised by health care providers can possibly be diminished.

Further studies are necessary for the assessment of validity and reliability of the different presented approaches.

Acknowledgements

This study has been supported by the MediGRID project (Bundesministerium für Bildung und Forschung, Förderkennzeichen 01AK803D). More information available at: http://www.medigrid.de

References

[1] Young TB. Epidemiology of daytime sleepiness: definitions, symptomatology, and prevalence. J Clin Psychiatry 2004; 65 Suppl 16:12-16.
[2] Akerstedt T. Consensus statement: fatigue and accidents in transport operations. J Sleep Res 2000; 9(4):395.
[3] Garbarino S, Nobili L, Beelke M, De Carli F, Ferrillo F. The contributing role of sleepiness in highway vehicle accidents. Sleep 2001; 24(2):203-206.
[4] Young T, Palta M, Dempsey J, Skatrud J, Weber S, Badr S. The occurrence of sleep-disordered breathing among middle-aged adults. N Engl J Med 1993; 328(17):1230-1235.
[5] Grote L, Ploch T, Heitmann J, Knaack L, Penzel T, Peter JH. Sleep-related breathing disorder is an independent risk factor for systemic hypertension. Am J Respir Crit Care Med 1999; 160(6):1875-1882.
[6] Hedner J, Grote L. Cardiovascular consequences of obstructive sleep apnea. In: McNicholas W, editor. European Respiratory Monograph: Respiratory disorders during sleep. Sheffield: ERJS Ltd, 1998: 227-265.
[7] Hornyak M, Feige B, Riemann D, Voderholzer U. Periodic leg movements in sleep and periodic limb movement disorder: prevalence, clinical significance and treatment. Sleep Med Rev 2006; 10(3):169-177.
[8] Winkelman JW, Finn L, Young T. Prevalence and correlates of restless legs syndrome symptoms in the Wisconsin Sleep Cohort. Sleep Med 2006; 7(7):545-552.
[9] Ohayon MM, Priest RG, Zulley J, Smirne S, Paiva T. Prevalence of narcolepsy symptomatology and diagnosis in the European general population. Neurology 2002; 58(12):1826-1833.
[10] Rechtschaffen A, Kales A. A manual of standardized terminology: techniques and scoring system for sleep stages of human subjects. Los Angeles: UCLA Brain Information Service/Brain Research Institute., 1968.
[11] Hamilton PS, Tompkins WJ. Quantitative investigation of QRS detection rules using the MIT/BIH arrhythmia database. IEEE Trans Biomed Eng 1986; 33(12):1157-1165.
[12] Practice parameters for the indications for polysomnography and related procedures. Polysomnography Task Force, American Sleep Disorders Association Standards of Practice Committee. Sleep 1997; 20(6):406-422.
[13] Kushida CA, Littner MR, Morgenthaler T, Alessi CA, Bailey D, Coleman J, Jr. et al. Practice parameters for the indications for polysomnography and related procedures: an update for 2005. Sleep 2005; 28(4):499-521.
[14] Calero G, Farre R, Ballester E, Hernandez L, Daniel N, Montserrat Canal JM. Physiological consequences of prolonged periods of flow limitation in patients with sleep apnea hypopnea syndrome. Respir Med 2006; 100(5):813-817.
[15] Clark SA, Wilson CR, Satoh M, Pegelow D, Dempsey JA. Assessment of inspiratory flow limitation invasively and noninvasively during sleep. Am J Respir Crit Care Med 1998; 158(3):713-722.

[16] Chesson AL, Jr., Wise M, Davila D, Johnson S, Littner M, Anderson WM et al. Practice parameters for the treatment of restless legs syndrome and periodic limb movement disorder. An American Academy of Sleep Medicine Report. Standards of Practice Committee of the American Academy of Sleep Medicine. Sleep 1999; 22(7):961-968.

[17] Littner MR, Kushida C, Anderson WM, Bailey D, Berry RB, Hirshkowitz M et al. Practice parameters for the dopaminergic treatment of restless legs syndrome and periodic limb movement disorder. Sleep 2004; 27(3):557-559.

From Genes to Personalized HealthCare: Grid Solutions for the Life Sciences
N. Jacq et al. (Eds.)
IOS Press, 2007

Grid Enabled High Throughput Virtual Screening Against Four Different Targets Implicated in Malaria

Jean SALZEMANN[a], Vinod KASAM[a,b,1], Nicolas JACQ[a], Astrid MAASS[b]
Horst SCHWICHTENBERG[b], Vincent BRETON[a]
[a]*Laboratoire de Physique Corpusculaire, 63177 Aubière cedex, France*
[b]*Fraunhofer Institut for Algorithms and Scientific Computing, Germany*

Abstract. After having deployed a first data challenge on malaria and a second one on avian flu, respectively in summer 2005 and spring 2006, we are demonstrating here again how efficiently the computational grids can be used to produce massive docking data at a high-throughput. During more than 2 months and a half, we have achieved at least 140 million dockings, representing an average throughput of almost 80,000 dockings per hour. This was made possible by the availability of thousands of CPUs through different infrastructures worldwide. Through the acquired experience, the WISDOM production environment is evolving to enable an easy and fault-tolerant deployment of biological tools; in this case it is the FlexX commercial docking software which is used to dock the whole ZINC database against 4 different targets.

Keywords: large scale deployment, Computational grids, Malaria, In silico docking, Virtual Screening, WISDOM.

Introduction

WISDOM stands for World-wide In Silico Docking On Malaria. Malaria together with many other tropical and protozoan diseases is one of the most neglected diseases by the developed countries as well as by the pharmaceutical industries. Plasmodium is the protozoan genus causing malaria. Due to very high costs associated to the drug discovery process as well as due to late stage attrition rates, novel and cost effective strategies are absolutely needed for combating the neglected diseases, especially malaria [1].

In silico screening of chemical compounds against a particular target is termed as Virtual Screening. The costs associated to the virtual screening of chemical compounds are significantly reduced when compared to screening of compounds in experimental laboratory. Beside the costs, virtual screening is fast and reliable [2, 3]. However, it is computationally intensive: docking a single compound within the active site of a given receptor requires about 1 minute CPU. With the development of combinatorial chemistry technology, millions of different chemical compounds are now available in

[1] Corresponding author: Vinod Kasam: Laboratoire de Physique Corpusculaire, Campus des Cézeaux - 24 av des Landais, 63177 Aubière cedex, France; E-mail: kasam@clermont.in2p3.fr.

digital databases [4]. To screen all these compounds and store the results is a real data challenge. To address this problem computational grid infrastructures are used.

WISDOM-I [5] is the first large scale deployment of molecular docking application on EGEE grid infrastructure. It took place from August 2005 to September 2005 and achieved 41 million dockings which is equivalent to 80 CPU years. The docking was performed on Plasmepsins, a aspartic protease involved in haemoglobin degradation. On the biological front three scaffolds were identified, of them one is guanidino scaffold which is likely to be novel as it was not known as a plasmepsin inhibitor before [6].

With the success achieved by the WISDOM-I project both on the computation and biological sides, several scientific groups around the world proposed targets implicated in malaria, which led to the second assault on malaria, WISDOM-II.

1. Materials and methods

Virtual Screening by molecular docking requires a target structure, a chemical compound database and docking software. The targets used in the current project are and Glutatione–S-trasferase (GST, pdbid: 1Q4J) [7], Plasmodium falciparum Dihydrofolate reductase (DHFR) wild type (pdbid: 1J3I), quadrupule mutant (pdbid: 1J3K) [8], Plasmodium vivax Dihydrofolate reductase wild type (pdbid: 2BL9), double mutant (pdbid: 2BLC) [9]. In another experiment the same structures of Plasmodium vivax Dihydrofolate reductase wild type (pdbid: 2BL9), double mutant (pdbid: 2BLC) but after minimization are used. The chemical compound database used is ZINC database [10, 11] and the docking software used is FlexX. To store the results of docking, MySQL databases are used, but it is still in process. FlexX [12, 13] is an extremely fast, robust and highly configurable computer program for predicting protein-ligand interactions. During our experiment, after several control tests, standard parameter settings are used except for two cases: “Place particles” and “Maximum overlap volume”.

2. Procedure

The goal of WISDOM II is two fold, the biological goal is to find the best hits against the targets implicated in malaria and the computational goal is to keep improving the relevance of computational grids in drug discovery applications. Here in this paper we are going to discuss in details the grid architecture and deployment.

2.1. Virtual screening experimental setup

The complete virtual screening experiment is segmented into five different phases.

i. Target preparation
ii. Compound database
iii. Validation of the docking experiment
iv. Screening
v. Result analysis

2.1.1. Target preparation

A standard protocol is used while preparing the target structures. The initial coordinates for all the target structures are obtained from Brookhaven protein database (www.pdb.org). Depending upon the inclusion of the significant residues, cofactors and the binding pocket, active site is defined as 8.0 Å - 10.0 Å around the co-crystallized ligand.

2.1.2. Compound database

The Compound library used for WISDOM was obtained from the ZINC database [14, 15]. The ZINC database is a collection of 4.3 million chemical compounds ready for virtual screening from different vendors. We have chosen to use the ZINC library because ZINC is an open source database and the structures have already been filtered according to the Lipinski rules. Moreover the data are available in different file formats (Sybyl mol2 format, sdf and smiles). A total of 4.3 million compounds were downloaded from the ZINC database and screened against four targets.

2.1.3. Validation of the docking experiment

Re-docking against the co-crystallized compound is performed to check and tune the docking experiment requirements. Re-docking serves as a control for finally selecting the parameters for target structure, before subjecting it to large scale docking. Docking pose is validated at two levels; RMSD value (the lower, the better) and binding pose of ligand (the more similar the docking pose to the co-crystallized ligand, the better).

2.2. Grid infrastructure and Deployment

2.2.1. Grid Infrastructures

The deployments were achieved on several grid infrastructures: Auvergrid [14], EELA [15], EGEE [16], EUChinaGrid [17] and EUMedGrid [18]. All these infrastructures are actually using the same middleware, gLite. EGEE is the main infrastructure offering the largest resources; they are all interconnected with EGEE, in the sense that all of these Grids share some resources with EGEE. In the case of Auvergrid, it is even more evident as all the resources available through the Auvergrid Virtual Organization (VO) are also shared with several EGEE VOs. The EUChinaGrid project for instance made available all the grid sites belonging to its infrastructure; seven Computing Elements in total and two Storage Elements were used to store the databases and result files on the EuChinaGrid.

2.2.2. WISDOM production Environment

WISDOM environment has been used two times in previous large-scale experiments, WISDOM-I in the summer 2005 [19] and a second deployment against avian flu in the spring 2006 [20]. WISDOM environment keeps evolving in order to make it more user friendly and easier to use by non grid expert. The main objective was also to improve the fault-tolerance of the system, in implementing, for instance, a persistent environment, that can be stopped and restarted at any time without risk of loosing significant information, which proved to be also very useful as it enables the whole maintenance of the scripts and code and improve the interactivity with the user, as the

user can also manage jobs finely, for instance force the cancellation and resubmission of a scheduled job. Along with this, we tried to minimize the cost of the environment in terms of disk space and CPU consumption for the user interface. Most of the job files are now generated dynamically: this allows the user as well to modify on the fly the configuration of the resource brokers and the jobs requirements. This way, the user is sure that the next submissions will take these modifications into account. The Figure 1, shows the overall architecture of the environment.

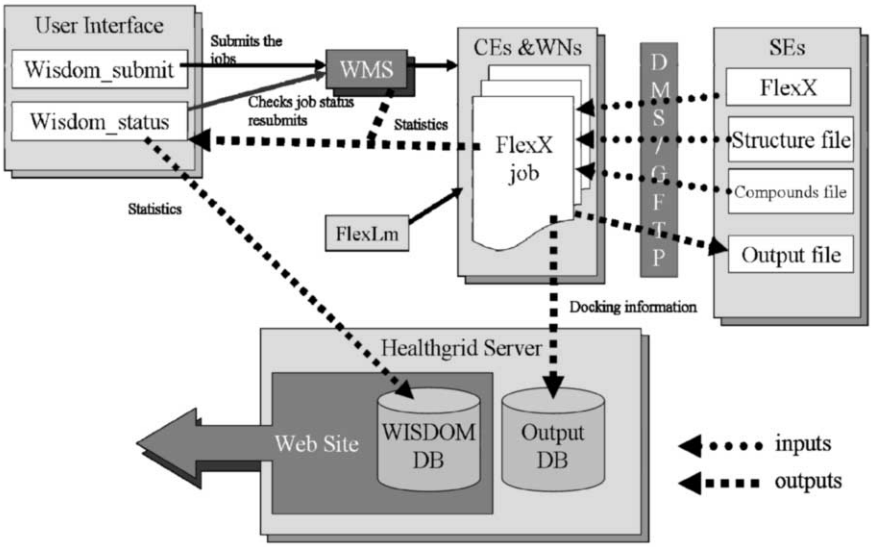


Figure 1. Schema of the WISDOM production environment.

The user is interacting with the system through the two main scripts (widom_submit and wisdom_status) deployed on the user interface. These scripts will take care of job files generation, submission, status follow-up and eventually resubmission automatically. The jobs are submitted directly to the grid Workload Management System (WMS), and are executed on the grid computing elements and worker nodes (CEs and WNs). As soon as it is running, a job transfers all the files stored on the Storage Elements (SEs) via the Data Management System of the grid (DMS) with the gridFTP protocol. Once the job is finished, the outputs are stored back on the grid Storage Elements via the Data Management System and the useful docking results are inserted directly from the grid to a relational database where they can later be more easily queried and analyzed.

2.2.3. Data Challenge Deployment

The deployment was performed on the infrastructures listed under section grid infrastructures, and involved at least one manager to oversee the process on each of them. The three groups of targets (GST, Plasmodium vivax and falciparum DHFR) were docked against the whole ZINC database (4.3 millions of compounds). The database was actually cut into 2,422 chunks of 1,800 compounds each. This splitting

was chosen because we wanted to have an approximated processing time ranging from 20 to 30 hours for each job (one docking process takes from 40s to 1min depending on the CPU power). The subsets were stored on the grid and replicated on several locations whenever possible to improve fault-tolerance. We define a WISDOM instance as one target structure docked against the whole ZINC database, with a given parameter set.

A total number of 32 instances were deployed, corresponding to an overall workload of 77,504 jobs, and up to 140,000,000 docking operations. On these total 32 instances, 29 instances ran on EGEE, and 3 were on Auvergrid, EELA and EuChinaGrid.

As shown in Figure 1, the environment included a FlexLm server that was providing the floating licenses for the FlexX commercial software. The FlexX software was already used during the first WISDOM deployment in 2005, and the license server was identified as a potential bottleneck and point of failure because we had just one server available at this time. For WISDOM-II, up to 3 servers were made available at the SCAI Fraunhofer institute (http://www.scai.fraunhofer.de), with 3,000 licenses available on each server.

As the average duration of a job was around 30 hours, we submitted 1 instance per day, with a delay of 30 seconds between each submission. As one instance was submitted in about 20 hours, the submission process was quite continuous during the first month of deployment. The jobs were submitted to 15 Resource Brokers (the components of the Workload Management System) in a round-robin order. At the end of a job, the results were stored on the grid Storage Elements and directly into a relational database. The job repartition was quite similar to the previous deployments, but here the UKI federation played an even bigger part. For instance, one of the British sites offered for quite a long period of time more than 1,000 free CPUs, which is half of the average used CPUs. Auvergrid, EELA, Euchinagrid and Eumedgrid contributed by running each 3% of the jobs.

3. Results and Statistics

Table 1, shows the overall statistics of the deployment. The number of jobs here are the number of awaited results, but far more jobs were actually submitted on the grid. When a job was done on the grid, the environment checked a status file specifying the final result of the job: a job can be done in the point of view of the worker node, without having produced the result files, in this specific case, the status of the job, which was stored on the grid as well, was labeled as failed, and the environment had to resubmit the job. In some cases, the environment failed at retrieving the status from the grid, and thus considered implicitly the job has failed, even if the job has succeeded. It explains why some jobs ran several times, and why the final completed docking number is bigger than the useful awaited dockings. Anyhow the average docking throughput is coherent with the crunching factor. The crunching factor is the ratio of the total CPU time over the duration of the experiment. It represents the average number of CPUs used simultaneously all along the data challenge and is a metric of the parallelization gain. If we consider 80,000 dockings per hour for 2,000, it means 40 dockings for one CPU per hour, which is coherent with the empiric observation of one docking process lasting approximately 1 minute on a 3.06 Intel Xeon processor.

Table 1. Overall statistics concerning the deployment.

Number of Jobs	77,504
Total Number of completed dockings	156,407,400
Estimated duration on 1 CPU	413 years
Duration of the experiment	76 days
Average throughput	78,400 dockings/hour
Maximum number of loaded licences (concurrent running jobs)	5,000
Number of used computing elements	98
Average duration of a job	41 hours
Average crunching factor	1,986
Volume of output results	1,738 TB
Estimated distribution efficiency	39%
Estimated grid success rate	49%
Estimated success rate after output checking	37%

In the Table 1, the estimated grid success rate is the ratio of successful grid jobs on the total of submitted jobs. The success rate after output checking will consider just the jobs that succeeded in producing the result files, that's why this score is lower. One can notice that these values are very small, but there are several explanations for this. At the beginning of the data challenge, the observed grid success rate was about 80 to 90%, but it decreased constantly because of sites overload. Sometimes the available disk space was decreasing on some Resources Brokers, up to a point where some of the job data could not reach the Computing Element. In other cases, the sites were simply producing a lot of aborted job for an undetermined reason. The Resource Brokers failed again to balance reasonably the jobs on the Computing Elements, and some of them ended up with more than 500 jobs in queue, the site administrator had no other choice than kill all these jobs, producing in a single row more than 500 aborted jobs. Actually, because of the automatic resubmission, this information should not be taken as an overall significant way to evaluate the efficiency of the grid, because the automatic resubmission guaranteed a successful job, and the aborted jobs are not staying a long time on the grid consuming useful resources. The grid is a very dynamic system, and errors can occur at the last minute.

4. Observations and Issues

The scheduling efficiency of the grid is still a major issue. The resource broker is still the main bottleneck, and even if used in high number (>15), is always a source of trouble. Moreover things get worse as load is increasing on the grid. The « sink-hole » effects can result in sites overloading in a very short amount of time, and if not taken care quickly can lead to an impressive overhead caused by the long lasting waiting state of the jobs. Added to that the sometimes unreliable and incomplete information provided by the information system, which does not publish the available slots and VO limitations that would be mandatory to perform an efficient scheduling.

Another issue was that to be able to store and treat the data in a relational database, the machine hosting the database must have good performances or the number of queries coming from the grid may also overload the machine significantly. In this deployment we used a MySQL database and planned to put all the produced result in the same table, but finally we had to spit this database in several ones (one per target), because MySQL would not have been able to withstand the total number of records, It was generating CPU overloads on the machine, which lead to serious slowdowns.

All these elements demonstrate clearly that even if the grid can show very good result in comparison to very simple architecture it is still missing robustness and reliability, and can indeed be improved performance-wise.

5. Conclusions

We have demonstrated the role and significance of computational grids in the life science applications like structure based drug design. Large scale virtual screening on four different targets of malaria was performed in search for potential hits on several grid infrastructures: Auvergrid, EELA, EGEE, EUChinaGrid and EUMedGrid. One of our goals was to further demonstrate the impact of computational grids in life science applications like virtual screening where large amounts of computing power is required; we have achieved it by successfully screening the whole ZINC database for three malaria targets in 76 days instead of 413 CPU years. We have reached during this ten-week period an average docking throughput of 78,400 dockings. MySQL databases are used for the analysis of the docking results, which will ease the final analysis of the virtual screening data. On the biological front 1,738 TB of valuable data has been produced. Analysis of the results (identification of the best hits) is under way: the best hits will be post processed by molecular dynamic simulations and tested in the experimental laboratories.

Acknowledgements

Information on WISDOM collaboration can be found at http://wisdom.healthgrid.org.

The authors acknowledge the contribution of Emmanuel Medernach (AuverGrid), Ignacio Blanquer (EELA), Gabri Aparico (EELA), Enrico Fattibene (EUChinaGrid), Kostas Koumantaros (EUMedGrid) and Domenico Vicinanza (EUMedGrid).

The Enabling Grids for E-sciencE (EGEE) project is co-funded by the European Commission under contract INFSO-RI-031688. Auvergrid is a project funded by the

Conseil Regional d'Auvergne. The BioinfoGRID project is co-funded by the European Commission under contract INFSO-RI-026808. The EMBRACE project is funded by the European Commission within its FP6 Programme, under the thematic area "Life sciences, genomics and biotechnology for health", contract number LHSG-CT-2004-512092.

We gratefully acknowledge the support of BioSolveIT who provided more than 6,000 free licenses for their commercial docking program FlexX.

References

[1] J. G. Breman, M. S. Alilio, A. Mills. Conquering the intolerable burden of malaria: what's new, what's needed: a summary. Am. J. Trop. Med. Hyg. 71S (2004), 1-15.

[2] K. H. Bleicher, H. J. Boehm, K. Mueller, A. I. Alanine. Hit And Lead Generation: Beyond High-Throughput Screening. Nat. Rev. Drug. Discov. , 2 (2003), 369-378.

[3] H. J. Boehm, G. Schneider. Virtual Screening For Bioactive Molecules. Chapter 1, High Throughput Screening and Virtual Screening Entry Points To Drug Discovery. Methods and Principles in Medicinal Chemistry, Volume 10.

[4] R.W. Spencer, High throughput virtual screening of historic collections on the file size, biological targets, and file diversity, Biotechnol. Bioeng 61 (1998), 61-67.

[5] N. Jacq, J. Salzemann, Y. Legré, M. Reichstadt, F. Jacq, E. Medernach, M. Zimmermann, A. Maaß, M. Sridhar, K. Vinod-Kusam, J. Montagnat, H. Schwichtenberg, M. Hofmann, V. Breton, Grid enabled virtual screening against malaria, accepted for publication in Journal of Grid Computing (2007)

[6] V. Kasam, M. Zimmermann, A. Maaß, H. Schwichtenberg, A. Wolf, N. Jacq,V. Breton, M. Hofmann. Design of Plasmepsin Inhibitors: A Virtual High Throughput Screening Approach on the EGEE Grid, submitted to Journal of Chemical Information and Modeling (2006).

[7] M. Perbandt, C. Burmeister, R.D. Walter, C. Betzel, E. Liebau, Native and inhibited structure of a Mu class-related glutathione S-transferase from Plasmodium falciparum, J. Biol. Chem. 279 (2004), 1336-1342

[8] P. Kongsaeree, P. Khongsuk, U. Leartsakulpanich, P. Chitnumsub, B. Tarnchompoo, M.D. Walkinshaw, Y. Yuthavong. Crystal Structure of Dihydrofolate Reductase from Plasmodium Vivax: Pyrimethamine Displacement Linked with Mutation-Induced Resistance, Proc. Natl. Acad. Sci. USA. 2 (2005), 13046-13051.

[9] J. Yuvaniyama, P. Chitnumsub, S. Kamchonwongpaisan, J. Vanichtanankul, W. Sirawaraporn, P. Taylor, M.D. Walkinshaw, Y. Yuthavong. Insights into antifolate resistance from malarial DHFR-TS structures, Nat. Struct. Biol. 10 (2003), 357-365.

[10] ZINC database: J. J. Irwin, B. K. Shoichet. ZINC - A Free Database of Commercially Available Compounds for Virtual Screening. J. Chem. Inf. Model. 45 (2005), 177-182.

[11] A Free database for virtual screening. ZINC is not commercial. http://blaster.docking.org/zinc/ UCSF, University of California, San Francisco.

[12] M. Rarey, B. Kramer, T. Lengauer, G. Klebe, Predicting Receptor-Ligand interactions by an incremental construction algorithm, J. Mol. Biol. 261 (1996), 470-489.

[13] BioSolveIT Homepage: http://www.biosolveit.de

[14] Auvergrid, available at www.auvergrid.fr

[15] EELA, available at www.eu-eela.org

[16] F. Gagliardi, B. Jones, F. Grey, M.E. Begin, M. Heikkurinenn. Building an infrastructure for scientific Grid computing: status and goals of the EGEE project, Philosophical Transactions: Mathematical, Physical and Engineering Sciences, 363 1729-1742 (2005) and http://www.eu-egee.org/

[17] EuChinaGrid, available at www.euchinagrid.org

[18] EuMedGrid, available at www.eumedgrid.org

[19] N. Jacq, V. Breton, H-Y. Chen,L-Y. Ho, M. Hofmann, H-C. Lee, Y. Legré, S. C. Lin, A. Maaß, E. Medernach, I. Merelli, L. Milanesi, G. Rastelli, M. Reichstadt, J. Salzemann, H. Schwichtenberg, M. Sridhar, V. Kasam, Y-T. Wu, M. Zimmermann. Virtual Screening on Large Scale Grids, accepted for publication in Parallel Computing, (2007).

[20] H.-C. Lee, J. Salzemann, N. Jacq, H.-Y. Chen, L.-Y. Ho, I. Merelli, L. Milanesi, V. Breton, S. C. Lin, Y.-T. W. Grid-Enabled High-Throughput In Silico Screening Against Influenza A Neuraminidase, IEEE transactions on nanobioscience, (2006) 288-295.

From Genes to Personalized HealthCare: Grid Solutions for the Life Sciences
N. Jacq et al. (Eds.)
IOS Press, 2007

Porting *PHYLIP* phylogenetic package on the Desktop GRID platform *XtremWeb-CH*

Nabil ABDENNADHER[a], Regis BOESCH[a]
[a]*University of Applied Sciences, Western Switzerland*

Abstract. This paper describes the parallelization (gridification) of the phylogenetic package *PHYLIP* on a desktop GRID platform termed *XtremWeb-CH*.
PHYLIP is a package of programs for inferring phylogenies (evolutionary trees). It is the most widely-distributed phylogeny package. *PHYLIP* has been used to build the largest number of published trees. It's known that some modules of *PHYLIP* are CPU time consuming; their sequential version can not be applied to a large number of sequences.
XtremWeb-CH (XWCH) is a software system that makes it easier for scientists and industrials to deploy and execute their parallel and distributed applications on a public-resource computing infrastructure. Universities, research centres and private companies can create their own *XWCH* platform while anonymous PC owners can participate to these platforms. They can specify how and when their resources could be used. The objective of *XWCH* is to develop a real High Performance Peer-To-Peer platform with a distributed scheduling and communication system. The main idea is to build a completely symmetric model where nodes can be providers and consumers at the same time.
In this paper we describe the porting, deployment, and execution of some *PHYLIP* modules on the *XWCH* platform. The parallelized version of *PHYLIP* is used to generate evolutionary tree related to HIV viruses.

Keywords. Phylogeny, *PHYLIP*, Grid, *XtremWeb-CH*.

Introduction

It is commonly accepted that contemporary genes, genomes, and organisms evolved from ancestors under the influence of natural selection. Consequently, the knowledge of the evolutionary tree behind their origin is crucial for understanding these entities. Knowledge about the relationships within gene families plays an important role in understanding, for example, the origins of biochemical pathways, regulatory mechanisms in cells as well as the development of complex systems. For example, knowing relationships between viruses is central for understanding their ways of infection and pathogenicity.

In a medical context, the generation of a life tree for a family of microbes is particularly useful to trace the changes accumulated in their genomes. These changes are due, inter-alia, to the "reaction" of viral strains to medical treatments.

In this context, computer applications dealing with the reconstruction of evolutionary relationships of organisms, genes, or gene families have become basic

tools in many fields of research [1, 2, 3, 4]. These applications "reconstruct" the pattern of events that have led to "the distribution and diversity of life". These relationships are extracted from comparing Desoxyribo Nucleic Acid (DNA) sequences of species. An evolutionary tree, termed life tree, is then built to show relationship among species. This tree shows the chronological succession of new species (and/or new characters) appearances. The majority of reconstruction methods of evolutionary trees optimize a predefined objective function. Thus, a given tree can easily be evaluated. The "optimal" tree is the one which is supposed to be the most "realistic" one.

The problem of finding an optimal evolutionary tree has been shown to be NP-complete for a quite number of reconstruction methods. In order to reduce the computational burden and to limit the vast number of trees to be examined, heuristics have been suggested: stepwise insertion with local and global optimizations [5], the Quartet Puzzling algorithm [6], star decomposition [7], etc. Recently, Bayesian approaches [8], genetic algorithms [9], and simulated annealing [10] have entered the field. However, approximate and heuristic methods do not solve the problem since their complexity remains polynomial with an order greater than 5: $O(n^m)$ with $m > 5$. Parallelization of these methods could be useful in order to reduce the response time of these applications.

The most widely-distributed phylogeny packages are *PHYLIP* [11] and *PAUP* [12]. These packages have been used to build the largest number of published trees. This paper deals with the parallelization of a sub-set of modules implemented by the *PHYLIP* package. It particularly describes the parallelization of the heuristic reconstruction method Fitch (proposed as a module in the *PHYLIP* tool).

The targeted machine is a network of computers equipped with the *XtremWeb-CH* (www.xtremwebch.net) middleware. The *XtremWeb-CH (XWCH)* project aims to build an effective Peer-To-Peer (P2P) System for CPU time consuming applications. Initially, *XWCH* is an upgraded version of a Global Computing environment called *XtremWeb (XW)* [13]. Major improvements have been brought in order to obtain a reliable and efficient system. The software's architecture was completely re-designed. The communication routines based initially on Remote Procedure Calls (Java RMI) were replaced by socket communications. New modules were added in order to enrich the system by new functionalities.

A typical *XWCH* platform is composed of one coordinator and several workers (remote resources). The coordinator is a three-tier layer allowing "connection" between the users and the workers.

XWCH supports three new features which, from our knowledge, do not exist in similar "prototypes": support of communicating tasks, direct communication between workers and granularity and load balancing management. These features are described in [25, 26] and will not be detailed in this paper.

This document is organized in 5 sections. After the introductory section, section 1 presents the sub-set of the *PHYLIP* modules that was ported on *XWCH*. Section 2 describes the different components of the *XWCH* middleware. Section 3 presents the gridification of *PHYLIP* on *XWCH*. Section 4 presents some experiments carried out in order to evaluate the proposed gridification. Finally, section 5 gives some perspectives of this research.

1. PHYLIP

PHYLIP (the PHYLogeny Inference Package) is a package of programs for inferring phylogenies (evolutionary trees). Developed during 1980s, *PHYLIP* is one of the most widely-distributed phylogeny packages. It has been used to build the largest number of published trees. *PHYLIP* has over 15,000 registered users. The package is available free over the Internet, and written to work on as many different kinds of computer systems as possible. The binary and source code (in C) are distributed. In particular, already-compiled executables are available for Windows, MacOS and Linux systems.

Methods that are available in the package include parsimony, distance matrix, and likelihood methods, including bootstrapping and consensus trees. Data types that can be handled include molecular sequences, gene frequencies, restriction sites and fragments, distance matrices, and discrete characters.

Five modules were ported on *XWCH*: *Seqboot, Dnadist, Fitch-Margoliash, Neighbor-Joining* and *Consensus*. Input data of these modules are nucleotide sequence data (DNA and RNA) coded with an alphabet of the four nucleotides *Adenine, Guanine, Cytosine,* and *Thymine*. Each nucleotide is denoted by its first letters: *A, G, C* and *T*. Every nucleotide sequence belonging to the input data is a leaf node of the evolutionary tree to be constructed.

The evolutionary tree is composed of several branches. Each branch is composed of sub-branches and/or leaf nodes (sequences). Two sequences belonging to the same branch are supposed to have the same ancestors. To construct the tree, the application defines a "distance" between all pairs of sequences. Evolutionary tree is then gradually built by sticking to the same branch, the pairs of sequences having the smallest distance between them. Even if the concept is simple, the algorithm is a CPU time consuming. This complexity is due to two factors:

1. Methods used to group sequences into branches are complex. As an example, the *Fitch* program, one of the most used methods, takes two hours to execute on a Pentium 4 (3 GHz) with 120 sequences.
2. The application constructs not only one tree from the origin data set, but a set of trees generated from a large number of bootstrapped data sets (somewhere between 100 and 1000 is usually adequate). These data are randomly generated from origin data. The final (or consensus) tree is obtained by retaining groups that occur as often as possible. If a group occurs in more than a given fraction of all the input trees, it will definitely appear in the consensus tree.

Seqboot is a general bootstrapping and data set translation tool. It is intended to generate multiple data sets that are re-sampled versions of the input data set. It involves creating a new data set by sampling *N* characters randomly with replacement, so that the resulting data set has the same size as the original, but some characters have been left out and others are duplicated.

Dnadist uses sequences to compute a distance matrix. It generates a table of similarity between the sequences. The distance, for each pair of sequences, estimates the total branch length between the two sequences, it represents the divergence time between those two sequences.

Fitch-Margoliash (Fitch) and *Neighbor-Joining (NJ)*: These two programs generate the evolutionary tree for a given data set. *Fitch* method is a time consuming method. Its sequential version can not be applied to a large number of sequences.

Consensus: This program constructs the consensus tree from the collection of intermediate trees generated from bootstrapped data sets.

The application, as developed, has two parameters (fed by the user): the set of nucleotide sequences from species under investigation and the number of replications. The higher is the replication, the finest is the result.

2. XtremWeb-CH

The majority of Global Computing (GC) projects adopted a centralized structure based on a Master/Slave Architecture: BOINC [14], Entropia [15], United Devices [16], Parabon [17], XtremWeb [13], etc. A natural extension of the GC consists on distributing the "decisional degree" of the master in order to avoid any form of centralization. Thus, architectures such as Clients/Servers and Master/Slaves would be withdrawn. This concept, known as Peer-To-Peer, was successfully used to share and exchange files between computers connected to Internet and broadcast micro-news among internet users. The most known projects are BitTorrent [18], eDonkey [19], Kazaa [20], Gnutella [21], Freenet [22] and FeedTree [23].

XtremWeb-CH (XWCH) is composed of four modules: coordinator, worker, warehouse and broker. The coordinator module is the main component of *XWCH*. It is considered as the master of the *XWCH* system; it has the responsibility of managing communication between the clients (users) and the workers (resource providers).

The worker module is installed on each provider node. It manages execution of tasks and the transfer of data from/to the worker. Workers are considered as the slaves of the *XWCH* system.

A broker module is a “compiler” which transforms the user request (application submission) into a set of tasks compliant to the “format” recognized by *XWCH*. Every family of applications has its own broker. The *XWCH* broker module can be compared to the Globus broker which is responsible of transforming a high level RSL (Request Specification Language) request into a low level RSL request [24].

2.1. The coordinator

It is a three-tier architecture which adds a middle tier between client and workers. The coordinator accepts execution requests coming from clients, assigns the tasks to the workers according to a scheduling policy and the availability of data, transfers binary codes to workers (if necessary), supervises task execution on workers, detects worker crash/disconnection and re-launches tasks on any other available worker. The coordinator is composed of three services: the workers’ manager, the tasks’ manager and the scheduler.

2.1.1. The Workers’ Manager

The workers’ manager maintains a list of connected workers. It receives four types of *common requests/signals* from the workers: *Register Request (RR)*, *Work Request (WR)*, *Life Signal (LS)* and *Work Result Signal (WRS)*. The *Register Request* allows a worker to subscribe nearby the coordinator. When the Workers’ Manager receives a *Work Request*, it searches for the most appropriate task [25] to be assigned to the

concerned worker. During the execution of the task, workers send *Life Signals* to the coordinator to inform about their status. When a worker finishes its execution, it sends a *Work Result Signal* to inform the coordinator about the location of the data it has produced.

2.1.2. The Tasks' Manager

A parallel and distributed application is composed of a set of communicating tasks whose structure is described in [25] and [26]. A task is considered to be "ready" for execution if its input data are available. It is in "blocked" status if its input data are not yet available. Two lists are maintained by the Tasks' Manager: *blocked tasks* and *ready tasks*. When receiving a *Work Result Signal*, the Tasks' Manager checks whether the new available data correspond to input data awaited by one or several blocked tasks; it updates the lists of blocked and ready tasks accordingly.

2.1.3. The scheduler

A Work Request transmits, as input parameter, the performance that can be delivered by the concerned worker. When receiving this request, the coordinator launches a scheduler module which selects the "most appropriate" ready task to be allocated to that worker. The concept of "most appropriate" is detailed in [26].

2.2. The workers

The worker module includes two components: the activity monitor and the execution thread. The activity monitor controls whether some computations are taking place in the hosting machine regarding parameters such as CPU idle time. The execution thread extracts the assigned task, starts computation and waits for the task to complete.

2.3. The warehouses

XWCH supports direct communication between workers executing two communicating tasks. Direct communication can only take place when the workers can "see" each other. Otherwise (one of the two workers is protected by a firewall or by a NAT address), this kind of communication is impossible. In this case, it is necessary to pass by an intermediary: *XWCH* coordinator for example. However, to avoid overloading the coordinator, one possible solution consists of installing "warehouse" nodes which acts as an intermediary. These nodes are used by workers to download input data needed to execute their allocated task and/or upload output data produced by the task. A warehouse node acts as a repository or file server. It must be reachable by all workers contributing to the execution of a given application.
The protocol is the following:

- The list of available warehouses is received by a worker when it registers nearby a coordinator (*Register Request*)
- When a worker finishes the execution of a task it uploads its result in a one of the known warehouses (selected randomly). Thus, the result is stored in the worker and in the warehouse,

- The worker sends a *work result* signal to the coordinator with the two locations (IP address and path) of the result produced by the given task,
- When a worker sends a *Work Request* to execute a new task, it receives as a reply, the binary code of the allocated task and the two locations of its input data.

3. *PHYLIP* Gridification

The "gridification" is the process of parallelizing and/or porting a High Performance application on a Grid platform. The gridification should take into account several constraints linked to the targeted Grid platform: volatility and heterogeneity of nodes, limited bandwidth of the network, etc.

This section describes the gridification of five modules of *PHYLIP*: *seqboot*, *dnadist*, *Fitch*, *NJ* and *consensus* on the *XWCH* middleware. Communications between tasks are based on file transfers.

As stated in section 1, the application, as developed, has two parameters (fed by the user):

1. set of nucleotide sequences from species (or viruses) under investigation. In the reminder of this paper, the number of sequences is noted by *s*.
2. Number of replications (*r*): used to produce multiple data sets from original DNA sequences by bootstrap re-sampling. The higher is this number, the finest is the result.

The structure of the obtained parallel/distributed application is shown in Figure 1.

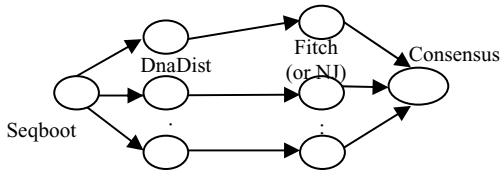


Figure 1. Data flow graph of the modules *SeqBoot, DnaDist, Fitch/NJ* and *Consensus*

The *Seqboot* task generates a multiple data sets. Each of these data is used by a *DnaDist* task to generate one distance matrix. This matrix is then used by a *Fitch* (or *NJ*) task to generate an intermediate evolutionary tree. Finally, the consensus task constructs the evolutionary tree from the intermediate trees. As explained in section 2, the *Fitch* module is time consuming ($O(n^5)$). This is not the case of modules *Seqboot, DnaDist, NJ* and *Consensus* modules.

In order to apply the *Fitch* module to a large number of sequences, a parallel version of this package was designed and ported on *XWCH*. The data flow graph of the parallel implementation of the *Fitch* module is given in Figure 2. Each *Fitch* node in Figure 1 is thus replaced by the graph of Figure 2.

The evolutionary tree is a non-root tree represented by two sets of nodes:

External (or leaf) nodes (square nodes in Figure 2): They represent the sequences under investigation. An external node is always linked to one internal node. When the evolutionary tree is completely constructed, the number of external nodes is equal to *s*.

Internal nodes (circle nodes in Figure 3) are virtuals, they don't represent sequences. Each internal node is linked to exactly three other nodes (internal or

external). When the evolutionary tree is completely constructed, the number of internal nodes is equal to *s-2*.

The evolutionary tree is generated progressively. The *Fitch* algorithm starts by creating a tree with one internal and three external nodes. In each step, the method inserts one sequence (external node) in every possible branch of the already constructed tree, and evaluates an objective function (*Test_Branch* tasks in Figure 2). The selected branch is the one that minimizes a pre-defined criterion *F* (*Best_Topology* tasks in Figure 2). In addition to the external node inserted in each step, an internal node is also created and inserted in the same step. This process is repeated until the insertion of all the sequences. The last step contains 2s – 5 "*Test_Branch*" tasks.

Thus, the number of "*Test_Branch*" tasks for one parallel *Fitch* is $O(s^2)$, *s* being the number of sequences. Since there are a maximum of *r* (*r* = number of replications) *Fitch* tasks, the maximum number of *Test_Branch* tasks is $O(r*s^2)$. The maximum number of parallel *Test_Branch* tasks that could be executed at the same time is equal to: *r*(2s-5)*. The execution time of a "*Test_Branch*" task increases with the size of the evolutionary tree.

4. Experiments

This section presents some performance analysis regarding the gridification of the package *PHYLIP*. Our results demonstrate the performance of the system and highlight promising areas for further research. The objective of these experiments is to validate our approach. They are not carried out to prove that the system delivers a maximum power for a given execution: the project's challenge is to extract, at low cost, a reasonable computing power from a widely distributed platform rather than extracting the maximum power from a local supercomputer or a dedicated GRID platform.

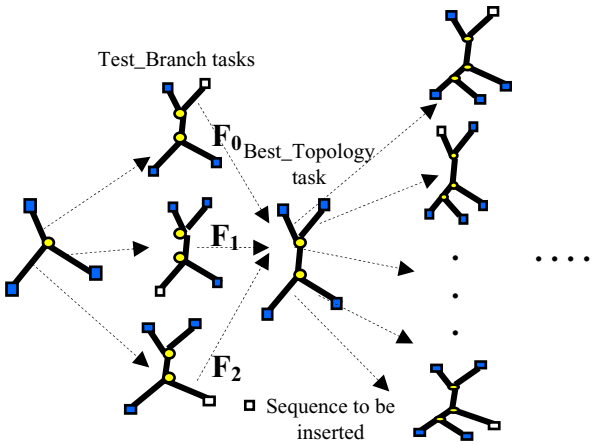


Figure 2. Data flow graph of a parallel *Fitch* task

The parallelized version of *PHYLIP* is used to generate evolutionary tree related to HIV sequences. The application is used by the virology laboratory of Geneva Hospital.

In this context, one needs to keep in mind that the number of sequences *s* can vary from 100 to 300 while the number of replications *r* varies from 100 to 1000.

A specific broker (web service) was developed in order to allow a dynamic configuration of the application regarding the current state (number and performance of the workers) of the platform: number of "*Fitch*" tasks and number of trees generated by each "*Fitch*" task, etc.

The experiments detailed in this section do not implement the parallel version of *Fitch* (Figure 2). They corresponds to the application represented if Figure 1. Executions were carried out on a platform with one coordinator (Linux OS), 250 heterogeneous windows workers ranging from Pentium I to Pentium IV, and 2 warehouse nodes. The workers are geographically located in two different places (Engineering Schools of Geneva and Yverdon). During execution, the 250 workers are used by students; they are often switched off or disconnected.

In order to evaluate the performance of the load balancing strategy implemented by *XWCH*, two versions of *PHYLIP* were deployed on the platform: the first version (*Version 1* in Figure 3) is composed of *r Fitch* tasks. Each task processes one tree. In the second version (*Version 2* in Figure 3), the number of *Fitch* tasks and the number of trees generated by each *Fitch* task are processed depending of the state of the platform (number and performance of workers).

Execution times consumed by the two versions are shown in Figure 3. The difference of execution times in Figure 3 is due to the synchronization between the coordinator and workers.

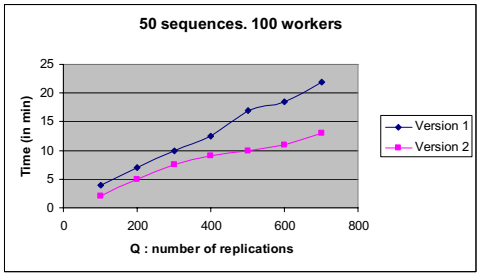
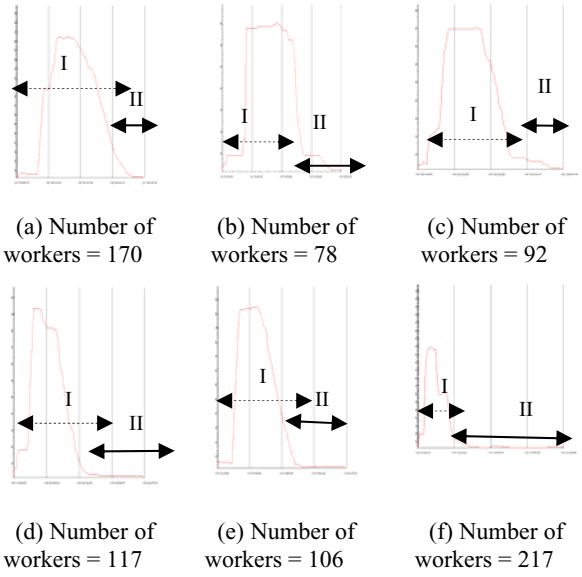


Figure 3. Execution times of *PHYLIP*

Figure 4 illustrates the total number of parallel tasks during the execution of the application. Since the "*Fitchs*" are the most time consuming tasks, this study focuses on the number of these tasks.

Steps I correspond to the execution of the *Fitch* tasks which finish, in general, at the same time. However, some *Fitch* tasks finish their execution later (step II in Figure 4). This is due to at least to one of the following factors:

1. The workers disappear during the execution,
2. As it is implemented today, workers' performance is only represented by the CPU power (CPU frequency). This model is not realistic; the system should take into account other criteria such as main memory, processes, applications and services installed locally on the workers, etc.



(a) Number of workers = 170 (b) Number of workers = 78 (c) Number of workers = 92

(d) Number of workers = 117 (e) Number of workers = 106 (f) Number of workers = 217

Figure 4. X-coordinates: Time, Y-coordinates: Total number of parallel executing *Fitch* tasks.

5. Conclusion

This paper presents the gridification of a sub-set of modules of the phylogeny package PHYLIP on the Large Scale Distributed platform XtremWeb-CH (XWCH). XWCH is a GC environment used for the execution of high performance applications on a highly heterogeneous distributed environment. This middleware can support direct communications between workers, without passing by the coordinator. A scheduling policy is proposed in order to minimize synchronization between coordinator and workers and optimize load balancing of workers.

The porting of *PHYLIP* on *XWCH* has demonstrated the feasibility of our solution. The next step consists of adapting the granularity of the parallel version of Fitch. Two parameters should be fixed according to the state of the targeted platform:

1. Number of parallel Test_Branch tasks executed during the insertion of one sequence.
2. Merging of several Test_Branch and Best_Topology tasks into one task according to the number of sequences.

The current version of *XWCH* allows the decentralization of communications between workers. The next step consists of designing a distributed scheduler. This scheduler shall avoid allocating communicating tasks to workers that can not reach each other and/or not belonging to the same “domain” (Local Area Network for example). This approach offers a strong basis for the development of distributed and dynamic scheduler and could confirm and reinforce the tendency detailed in section 2.

References

[1] http://biowulf.nih.gov/apps/tree-puzzle-doc.html
[2] http://www.tree-puzzle.de/
[3] http://www.dkfz.de/tbi/tree-puzzle/
[4] Heiko A. Schmidt, Phylogenetic Trees from Large Datasets, 'Ph.D.' in Computer Science, Düsseldorf, Germany, 2003.
[5] Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol., 17, 368–376.
[6] Strimmer, K. and von Haeseler, A. Quartet puzzling: A quartet maximum–likelihood method for reconstructing tree topologies. Mol. Biol. Evol., 13, 964–969, 1996.
[7] Adachi, J. and Hasegawa, M. MOLPHY Version 2.3 – Programs for Molecular Phylogenetics Based on Maximum Likelihood, vol. 28 of Computer Science Monographs. Institute of Statistical Mathematics, Minato-ku, Tokyo, 1996.
[8] Huelsenbeck, J. P. and Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics, 17, 754–755, 2001.
[9] P. O. A genetic algorithm for maximum-likelihood phylogeny inference using nucleotide sequence data. Mol. Biol. Evol., 15, 277–283. 1998.
[10] Salter, L. A. and Pearl, D. K. Stochastic search strategy for estimation of maximum likelihood phylogenetic trees. Syst. Biol., 50, 7–17, 2001.
[11] http://www.phylip.com/12. http://paup.csit.fsu.edu/
[12] http://paup.csit.fsu.edu/
[13] Gilles Fedak et al. XtremWeb : A Generic Global Computing System. CCGRID2001, workshop on Global Computing on Personal Devices. Brisbane, Australia, May 2001. http://xtremweb.net
[14] http://boinc.berkeley.edu/
[15] http://www.entropia.com/
[16] http://www.ud.com/home.htm
[17] Parabon Computation, Inc: The Frontier Application. Programming Interface, Version 1.5.2. 2004.
[18] http://www.bittorrent.com/
[19] http://www.edonkey2000.com/
[20] http://www.kazaa.com/us/index.htm
[21] KAN G., Peer-to-Peer: harnessing the power of disruptive technologies, Chapter Gnutella, O'Reilly, Mars 2001.
[22] Ian Clarke. A Distributed Decentralised Information Storage and Retrieval System. Division of Informatics. Univ. of Edinburgh. 1999. http://freenet.sourceforge.net/
[23] http://feedtree.net/
[24] http://www.globus.org/
[25] N. Abdennadher, R. Boesch, A Large Scale Distributed System for High Performance needs. HP-ASIA 2005, Biejing, China, December 2005.
[26] N. Abdennadher, R. Boesch, A Scheduling algorithm for High Performance Peer-To-Peer Platform. CoreGrid workshop, EuroPar 06, Dresden, Germany, 2006

II. State of the Art of the Grid Research and Use at Cell Level

This page intentionally left blank

From Genes to Personalized HealthCare: Grid Solutions for the Life Sciences
N. Jacq et al. (Eds.)
IOS Press, 2007

Security-oriented Data Grids for Microarray Expression Profiles

Richard SINNOTT [a,1], Christopher BAYLISS [a], Jipu JIANG [a]
[a] *National e-Science Centre, University of Glasgow, United Kingdom*

Abstract. Microarray experiments are one of the key ways in which gene activity can be identified and measured thereby shedding light and understanding for example on biological processes. The BBSRC funded Grid enabled Microarray Expression Profile Search (GEMEPS) project has developed an infrastructure which allows post-genomic life science researchers to ask and answer the following questions: who has undertaken microarray experiments that are in some way similar or relevant to mine; and how similar were these relevant experiments? Given that microarray experiments are expensive to undertake and may possess crucial information for future exploitation (both academically and commercially), scientists are wary of allowing unrestricted access to their data by the wider community until fully exploited locally. A key requirement is thus to have fine grained security that is easy to establish and simple (or ideally transparent) to use across inter-institutional virtual organisations. In this paper we present an enhanced security-oriented data Grid infrastructure that supports the definition of these kinds of queries and the analysis and comparison of microarray experiment results.

Keywords. Microarray Expression Profiles, Grid security, Shibboleth

Introduction

The UK Biotechnology and Biological Research Council (BBSRC) Grid Enabled Microarray Expression Profile Search (GEMEPS) project (www.nesc.ac.uk/hub/projects/gemeps) began in March 2006. The fundamental premise upon which GEMEPS is based is that life scientists recognise that it is to their advantage to collaborate, especially with regard to sharing of expensively produced microarray experiments. Academics and researchers will always need to refer to and publish in journals and leading publications in their respective fields, however targeted real time access to research data between collaborators and institutes needs to occur to expedite the knowledge discovery process. Currently this is largely not the case and access to and usage of microarray data sets is limited for a variety of reasons: competitive, ethical, social, political being just a few. To support any form of data sharing models, scientists and their supporting IT staff need technologies that allow them to be fully informed and in control of the security infrastructures by which they make their data sets available and to whom.

The Grid in principle provides an appealing model for access to and usage of distributed and heterogeneous life science data sets. The explosion of data sets across the

[1] r.sinnott@nesc.gla.ac.uk

life science spectrum, and the major compute demands of high through post-genomic research offer direct requirements siuitable for Grid based solutions. However Grid technologies are not a silver bullet or a complete panacea for all of the challenges facing the life science community. The Grid needs agreements and standards on how life science data sets are created, defined and annotated before it can be exploited for data discovery, analysis or linkage. Similarly, understanding of the life science applications and data sets and the specific requirements they impose on computational resources is needed before the Grid can truly solve the compute requirements of this community.

Perhaps the most important aspect to recognise is that technology alone is insufficient to solve the requirements from this domain and must be guided by the wider scientific community needs and experiences. It could be argued that there has primarily been a middleware push as opposed to a scientific pull across the majority of the Grid research communities, and this is especially so in the life sciences. Previous projects such as the DTI funded Biomedical Research Informatics Delivered by Grid Enabled Services (BRIDGES) (www.nesc.ac.uk/hub/projects/bridges) at the National e-Science Centre (NeSC) at the University of Glasgow and funded reports such as [1] have identified that life scientists are especially wary of their data resources being accessed by others without them first exploiting their results, e.g. through journal publication. This cultural issue is especially significant since technologies must be met by scientific willingness to engage and collaborate. Yet the existing Grid security solutions are largely complex and confusing to end users and the supporting IT staff. Thus technologies are needed which simplify as much as possible the access to and usage of a range of data sets and resources more generally. A key and crucial benefit of the Grid is to support site autonomy. Sites should be able to define and enforce their own local policies on access to and usage of data sets. Since large scale post-genomic scientific research is rarely undertaken by a single site, but requires access to a range of data sets and resources including public repositories as well as collaborators private resources, multi-site solutions are needed. The definition of these security policies also needs to be recognised across the multiple institutions involved in collaborative research.

Importantly the scientific community needs to be made aware of what it means to provide controlled access to their research data and the potential ramifications thereof. Biologists tend not to be computer scientists and are unfamiliar with advanced Grid data access or security solutions. As such any solutions that are put forward in this domain have to be intuitive and allay their potential fears on compromises of their research data, or potential exploitation by competitors or third parties. New developments such as gene identification, gene function and development of new targeted drugs offers enormous opportunities for researchers both academically and commerically. As such, they need to be completely satisfied that any new technological solutions will fit into the way in which they wish to work, and importantly protect their research results and data from compromise.

In this paper we outline the solutions developed at the National e-Science Centre (NeSC) at the University of Glasgow to support seamless Grid based access to a range of services that allow discovery, analysis and comparison of microarray experiments.

The rest of the paper is structured as follows. Section 1 outlines the background to microarray data sets and associated standards. Section 2 focuses on the architecture and implementation of the GEMEPS infrastructure focusing in particular on security aspects

and services to ascertain microarray experiment similarity. Finally in section 3 we present our conclusions and plans for the future.

1. Background to GEMEPS

The GEMEPS project aims to develop a Grid infrastructure for discovery, access, integration and analysis of microarray data sets. Through the GEMEPS infrastructure scientists should be able to ask the following kinds of questions and obtain appropriate results based upon their privilege:

- who has run a microarray experiment and generated similar results to mine?
- who has undertaken experiments and produced data relevant to my own interests, e.g. for a particular phenotype, for a particular cell type, for a particular pathogen, on a particular platform or microarray chip set?
- show me the results from a particular collaborator;
- show me the conditions and analysis associated with experimental results similar to mine.

In all of these scenarios, the model we consider is for sites to keep and maintain their own data and define their own security policies on access and usage. This model has a psychological benefit to encourage collaboration, namely that scientists are not simply making their data publicly available for example in one of the existing repositories such as Gene Expression Omnibus (GEO) at NCBI [2], ArrayExpress [3] or CIBEX [4]. Scientists are often reluctant to publish their data in such repositories until they have published results in recognised journals which can, depending on the journal be a long and protracted affair. As a result, these public repositories tend to be populated with older data sets. It is also the case that these data repositiories provide various kinds of services through which the repositories themselves might be searched or mined. These repositories typically require data sets to be MIAME compliant [5]. The stated goal of MIAME is to outline the minimum information required to interpret unambiguously and potentially reproduce and verify an array based gene expression monitoring experiment. Whilst the details of particular experiments may be different, it is the intention of MIAME to define a core that is common to most experiments. MIAME is not a formal specification, but a set of guidelines concentrating on the content of information and various metadata that needs to be captured to facilitate re-use or reproduction of experimental results. Most major journal publications now require data associated with journal papers to be published in combination with the paper itself.

A MIAME description typically describes the design of: array platform - containing the description of the common features of the array and the description of each array design element; gene expression experiment - containing a description of the overall experimental design; the samples used; how extracts were prepared; which hybridisation procedures were followed and ultimately what data was measured and how it was analysed and normalised.

MIAME compliance is not prescriptive in the sense that all or a given subset of the various sections that might be associated with a given experiment must be given. These sections are usually provided in free text format, along with recommendations requiring maximum use of controlled vocabularies or external ontologies. MIAME recognises that

few controlled vocabularies have been fully developed, hence it encourages users to provide their own qualifiers and values identifying the source of the terminology. Of those that are available, the Microarray Gene Expression Data Society (MGED) [6] is one of the more established ontologies for microarray experiment description.

Several data formats have also been defined and applied across different sites with different user communities. These include: Microarray Gene Expression Markup Language (MAGE-ML) [7] is part of the MGED family of standards and is MIAME compliant and XML based. Many major repositories, such as GEO, ArrayExpress and CIBEX support results being deposited in MAGE-ML as well as supplying data in that format. Simple Omnibus Format in Text (SOFTtext) [8] is a simple text based format designed by GEO. Unlike MAGE-ML, SOFTtext is not XML based using instead keywords for describing platform, sample and results. It has fewer fields than MAGE-ML yet is still MIAME compliant. GEO supports submissions in this format and makes results availiable in it as well. Since SOFTtext is based around a simple format it is easy to parse and use. MIAME Notation in Markup Language (MINiML) [9] is an XML based format used by GEO and is equivalent to SOFTtext. The NCBI accepts data deposited in MINiML format and makes records available in this format. MINiML can be considered an XML equivalent to SOFTtext as it provides the same properties, however in XML form. NCBI has made a schema for MINiML avaliable allowing a validating parser to confirm that a MINiML file is well formed. This is a distict advantage over SOFTtext where there is no formal definition of how the files should be formatted. As with the other SOFT formats MINiML is MIAME compliant yet has fewer fields than MAGE-ML. The relative simplicity of MINiML when compared to MAGE-ML has direct advantages for usability and associated learning curve.

SOFTmatrix [10] is a new format developed by NCBI based on MIAME and using Microsoft Excel spreadsheet templates. Excel spreadsheets are one of the most common ways in which scientists keep their own microarray experimental results.

As seen a multitude of on-going efforts in how to describe and annotate the data and metadata associated with microarray experiments and results exist. It is within this context that the GEMEPS project is developing a security oriented Grid infrastructure for discovery and comparison of microarray experiment profiles.

2. Secure Data Grids within GEMEPS

The Open Middleware Infrastructure Institute (OMII - www.omii.ac.uk) and the Open Grid Service Architecture - Data Access and Integration (OGSA-DAI - www.ogsadai.org.uk) technologies were applied to produced an alpha prototype showing how Grid data access middleware could be used for access to and usage of microarray data sets (shown in the striped boxes). This was primarily used for feasibility studies.As stated, GEMEPS recognises and distinguishes between public microarray resources and private resources such as those created and maintained at the Sir Henry Wellcome Functional Genomics Facility (SHWFGF) at the University of Glasgow. The SHWFGF has been the primary source of requirements driving GEMEPS development.

2.1. Security Aspects of GEMEPS

To simplify the end user experience in accessing and using this resource and in providing advanced Grid security, we have provided Shibboleth based single sign-on [11,12] The Shibboleth technologies are developed by the Internet2 community and offer a simple yet secure way in which single sign on to a variety of distributed resources can be supported. The Shibboleth framework provides a mechanism for exchanging attributes across different organisations for the purpose of authentication and authorisation. It enables a user to access a protected resource or service at a remote domain (commonly referred to as a Service Provider (SP) or target) by using the user's own home security domain (commonly referred to as an Identity Provider (IdP) or origin) to perform user authentication. The framework uses X.509 certificates for the underlying secure attribute exchange. An important advantage the framework provides is that the user is not required to possess an X.509 certificate. This is because Shibboleth allows inter-institutional sharing of resources within a trusted federation where it is the responsibility of the home institutions to authenticate their users. Therefore Shibboleth directs the users to their home institution for authentication. The information which is exchanged as attributes helps to determine whether to grant the user access to the resource at the SP. To achieve this Shibboleth uses Security Assertion Mark-up Language (SAML) [13,14], an OASIS standard for exchanging authentication and authorisation statements, between the IdP and the SP. When the user is authenticated, the Shibboleth component at the SP establishes and maintains a session with the user's web browser on behalf of the resource the user is accessing. This session consists of cookies which are passed between the web browser and web server. The cookies are associated with a security context which holds the user's authentication information and a set of attributes describing the user's identity. With this, the user can access the resource (or resources across the federation/virtual organisation) more than once without repeating the Shibboleth authentication process until the cookies expire or are deleted from the user's machine.

The Shibboleth framework enables the creation of a federation to build trust relationships between participating organisations for inter-institutional access of resources. These organisations exchange attributes using the Shibboleth protocols and abide by a common set of policies and practices. Currently within GEMEPS Shibboleth access to the portal is based upon a single IdP (at the University of Glasgow) however extensions to this to allow access by other users from remote institutions are feasible and work is on-going in this area.

The Shibboleth framework devolves responsibility of user authentication to the user's home institution. This avoids the need to create a separate authentication system that is exclusive to the GEMEPS system, e.g. a portal log-in. It is of course possible to set up single usernames and passwords for the GEMEPS portal however this is not an especially scalable solution since users collect many usernames and passwords in the course of their research and often keep the same username/password combination to minimize the amount of information they have to memorize. Furthermore, the Shibboleth framework also provides a scalable and an extensible solution for managing access to resources. By using the Shibboleth framework it is possible to accommodate a growing number of users from different institutions as part of a federated access management arrangement. The Shibboleth framework is also being adopted by a number of higher educational institutes in the UK to develop the next generation access management system

for their users. Indeed the UK federation was established in November 2006, and many other federations are being established internationally.

Once a user has successfully authenticated at their home institution and in turn through Shibboleth they are authenticated to the GEMEPS portal they are presented with a variety of services (portlets). We note that which attributes are released from a given IdP to a particular SP is configurable and dependent upon the attribute release and attribute acceptance policies. Through the NeSC DyVOSE project (www.nesc.ac.uk/hub/projects/dyvose) we have implemented solutions which exploit delegation of authority to allow the dynamic creation or revocation of these attributes across multiple sites [15,16]. Alternatively the roles themselves and how they are assigned to users at different sites can be achieved in a variety of other ways. For example, if their authentication system is based upon LDAP then an LDAP editor can be used to simply add these roles to the particular users by the local system administrator. UK academia more generally has identified and recommended four key attributes based around the eduPerson object class [17]. These will be used to determine that an individual is a member of a recognized UK HE/FE institution. Often this is sufficient information for a service provider to allow/deny access, e.g. e-Journals might only need to know that a user is a member of a given university to allow/deny access (based on whether that institution has a subscription to that particular journal say).

The returned attributes are used to configure and customize (authorize) the resources that are accessible to the user. Thus here the attribute *gemeps_human_genome_researcher* is here used to allow access to human genome related experiments. Typically the management and organization of these roles and their relationship to security policy is achieved through some form of coordinated access control policy. Role Based Access Control (RBAC) technologies such as PERMIS (www.permis.org) can be used to enforce authorization policies and manage hierarchies of roles and their associated privileges. Practical explorations of these technologies are described in detail in [18,19,20].

We note that several portlets are available within this portal for querying the status of the Grid infrastructure, for running large scale bioinformatics BLAST applications on a variety of large scale HPC facilities. The GEMEPS portlet allows for discovery and comparison/similarity checking of microarray experiments. This is achieved firstly through meta-data querying, and then for rank correlation evaluation.

2.2. Meta-data Query Services

The various mark-up languages associated with microarray experiments allow for capturing and annotation of a variety of information about particular microarray experiments. Examples of the kinds of information that are typically captured include the specification of: how the data was processed; the type of molecules extracted from the biological material; the unique ID of the samples; descriptions of the scanning and image acquisition protocols (both hardware and software); the conditions used to grow or maintain organisms or cells prior to extract preparation; the name of the company, laboratory or person that provided the biological material; the protocols used to isolate the extract material; the platform used, e.g. GPL570 for the human genome; the species under study, e.g. homo sapiens, rattus rattus, etc; the biological material and the experimental variable(s) for the sample; a description of the experiment more generally and the protocols used for hybridization, blocking and washing, and any post-processing steps used such as staining.

Scientists in the first instance would like to be able to query across a range of experiments based on any one or more of these kinds of search terms. Through such queries immediate and meaningful experimental results can be returned. Thus scientists are unlikely to be interested comparing experimental results from homo sapiens and barley for example. We note that at the gene name level however, it is often the case that common gene name clashes do exist across species for example. To support this basic metadata querying, the GEMEPS project has implemented a simple user oriented portlet that allows for a variety of these kinds of information to be used for querying over available (subject to authorisation privileges) data sets.

2.3. Similarity Services

Having identified the set of experiments that are relevant to a researcher, finer grained analysis and comparison is needed to understand how relevant these data sets actually are. To support this GEMEPS has explored two different rank correlation cooefficient algorithms based upon Spearman Rank [21] and Kendall Tau [22]. The implementation of these algorithms is facilitated by the building of indexes from microarray experiments. Amongst the challenges associated with these algorithms in this domain are the difficulties in identifying and relating gene names across different experiments. Different experiments might use their own naming schemes, different ontologies such as MGED. Within the course of the project we have explored Life Science Identifiers (LSids) to address these issues. LSids are designed as Uniform Resource Names (URN) written in the form: urn:lsid:<authority>:<database>:<object>:<version> where <authority> is the name of the authority who issued the LSid, <database> is the name of the authority's database the LSid is stored in and <object>:<version> identifies the object within the database and its revision.

LSids are intended to serve as persistent identifiers allowing them to be used without later being reassigned. They allow to map to exactly the same set of bytes permanently. This means that an LSid, once assigned, is permanently attached to a specific encoding of its data which cannot be updated or corrected. An immediate advantage of this is that makes LSids usable as references. The LSid specification suggests using an LSid proxy, e.g. lsid.biopathways.org, to resolve LSids. The biopathways resolver provides LSids for many existing data sets such as the NCBI databases, ArrayExpress and SwissProt for example.

At the time of writing, it is unclear whether LSids will solve the problems arising in uniquely identifying information in the life science domain. For example, the closure of the Interoperable Informatics Infrastructure Consortium (i3c) means the loss of RDF metadata associated with LSids. References to this data still appear in examples and tutorials but the i3c itself website no longer exists. The only implementations of the LSid stack found are from the IBM LSid project on sourceforge. The logs of the source repository reveal little activity with the majority of the code remaining untouched since 2004.

To address this, the project has focused upon developing solutions targeted towards SOFTtext and MINiML. The implementation of these algorithms produces initial results as would be expected. Thus for example, when the results from one experiment are compared with itself a correlation co-efficient of 1 is returned. When the inverse of the results of an experiment, i.e. reversing the gene expression ranking, the algorithm returns

-1 as expected. The implementation itself offers currently just the Spearman ranking co-efficient however the Kendall Tau correlation co-efficient is also under development and will be rolled out in due course.

The interface to this system allows users to either upload their own experimental results for comparison from a file (given as a sequence of gene names, or as a sequence of <gene name, expression value> orderings), or they can cut and paste these, e.g. from an Excel spreadsheet - a common technology used by most life science researchers in managing their microarray results. The final result of these experiments are given as a ranking of most similar experiments. The most similar given with the level of similarity (Spearman rank coefficient) given. With the most similar experiment identified, the end user may then follow the hyperlink to obtain more information about this particular data and how the results were obtained etc.

3. Conclusions

At the time of writing the GEMEPS project has been on-going for 10 months and has a further 2 months remaining. The project has faced many challenges in reaching the current implementation status. Perhaps the greatest of these is in understanding, managing and linking the bioinformatics data resources. The lack of a common naming system and a variety of different mark-up languages and ontologies to describe the results of microarray experiments adds to the overall complexity in development of Grid based systems.

We believe that the adoption and roll out of Shibboleth to simplify the access to and usage of secure bioinformatics data resources is key to the success of this and other solutions. Allowing scientists to remain autonomous and keep their own datasets locally but allow secure and controlled access by remote collaborators offers a much more appealing model to scientific data sharing than centralised public microarray repositories. Trust plays an important role in Shibboleth (or any security based system). Sites trust remote institutions to authenticate their users correctly. To address this the University of Glasgow has rolled out a unified account management system based on Novell nSure technology. With this system, there is a one-one mapping between members of the university - either staff or students - and the privileges that they possess. Thus users do not have different usernames and passwords for different systems. Through this system, when a member of staff or student leaves, all of their privileges are removed. This system is being rolled out as part of the JISC funded Glasgow early adoption of Shibboleth (GLASS) project (www.nesc.ac.uk/hub/projects/glass).

The work on GEMEPS is still on-going and we are now focusing on several areas. These include hardening and scaling the existing system. Thus for example, the Spearman Rank algorithm does not scale well when large sets of genes in one experiment are not available (being expressed) in another. This might at first instance imply that the two experiments are disimilar hence should have a low correlation co-efficient. However it is often the case that biological meaning from microarray experiments should only be determined from the most significantly expressed genes. Thus below a given limit, the expression values and associated gene orderings cannot be guaranteed due to statistical variation when conducting the experiments. To address this we are exploring various cut off scenarios for experiment comparison. Thus a scientists might only be interested

in the top 10, 100, 1000 expressed genes. Alternatively a scientist might only be interested in genes expressed where the expression value itself is above a given cut off. These combinations and their impact of result accuracy offer important ways to increase biological understanding of the accuracy of microarray experiments. A further scenario we are exploring is where a scientist is interested in experiments where a particular gene is expressed or expressed above a particular varlue.

A further enhancement to this system is to support large scale expression profile matching. Thus when many thousands of matching experiments are returned from metadata queries and need to be compared with one or more experiments, then use of large scale HPC facilities is needed. To address this we are exploring user driven access to and usage of large scale HPC facilities such as the UK e-Science National Grid Service (www.ngs.ac.uk) and the ScotGrid system (www.scotgrid.ac.uk) at Glasgow. In achieving this, the Grid will be completely shielded from the user.

4. Acknowledgements

The work described here was suported by a grants from the UK Biotechnology and Biological Sciences Research Council (BBSRC). We gratefully acknowledge their support. We also acknowledge inputs to the science we are trying to support from Dr Pawel Herzyk at the Sir Henry Wellcome Functional Genomics Facility at the University of Glasgow. Dr John Watt is acknowledged for his Shibboleth expertise and support.

References

[1] P. Lord, A. MacDonald, R.O. Sinnott, Large-scale data sharing in the life sciences: Data standards, incentives, barriers and funding models: The "Joint Data Standards Study", prepared for The Biotechnology and Biological Sciences Research Council, The Department of Trade and Industry, The Joint Information Systems Committee for Support for Research, The Medical Research Council, The Natural Environment Research Council and The Wellcome Trust.
[2] Gene Expression Omnibus (GEO), www.ncbi.nlm.nih.gov/geo/
[3] P. Rocca-Serra, A. Brazma, H. Parkinson, U. Sarkans, M. Shojatalab, S. Contrino, J. Vilo, N. Abeygunawardena, G. Mukherjee, E. Holloway, M. Kapushesky, P. Kemmeren, G. Garcia Lara, A. Oezcimen, S.-Assunta Sansone. ArrayExpress: a public database of gene expression data at EBI. C R Biol, 326(10-11):1075–1078, Oct 2003.
[4] K. Ikeo, J. Ishi-i, T. Tamura, T. Gojobori, and Y. Tateno. CIBEX: center for information biology gene expression database. C R Biol, 326(10-11):1079–1082, Oct 2003.
[5] Minimal Information About a Microarray Experiment (MIAME), http://www.mged.org/Workgroups/MIAME
[6] Microarray Gene Expression Data Society (MGED) Ontology Working Group, http://www.mged.org/ontology
[7] MicroArray and Gene Expression Markup Language (MAGE-ML), http://www.mged.org/Workgroups/MAGE/mage-ml.html
[8] Simple Omnibus Format in Text (SOFTtext), http://www.ncbi.nlm.nih.gov/projects/geo/info/soft2.html
[9] MIAME Notation in Markup Language (MINiML), http://www.ncbi.nlm.nih.gov/projects/geo/info/MINiML.html
[10] Simple Omnibus Format in Matrix (SOFTmatrix), http://www.ncbi.nlm.nih.gov/projects/geo/info/soft2.html

[11] Shibboleth Architecture Technical Overview, http://shibboleth.internet2.edu/docs/draft-mace-shibboleth-tech-overview-latest.pdf
[12] Shibboleth Architecture Protocols and Profiles, http://shibboleth.internet2.edu/docs/draft-mace-shibboleth-arch-protocols-latest.pdf
[13] OASIS, Assertions and Protocol for the OASIS Security Assertion Markup Language (SAML) v1.1, 2 September 2003, http://www.oasis-open.org/committees/security/
[14] Security Assertion Markup Language (SAML) version 2.0, March 2005, http://www.oasis-open.org/specs/index.php#samlv2.0
[15] R.O. Sinnott, J. Watt, O. Ajayi, J. Jiang, J. Koetsier, A Shibboleth-Protected Privilege Management Infrastructure for e-Science Education, 6th IEEE International Symposium on Cluster Computing and the Grid, CCGrid2006, May 2006, Singapore.
[16] R.O. Sinnott, J. Watt, O. Ajayi, J. Jiang, Shibboleth-based Access to and Usage of Grid Resources, IEEE International Conference on Grid Computing, Barcelona, Spain, September 2006.
[17] eduPerson Specification, http://www.educause.edu/eduperson/
[18] R.O. Sinnott, Security Focused Federation of Distributed Biomedical Data, Proceedings of UK e-Science All Hands Meeting, 4-6 September 2003, Nottingham, England.
[19] R. O. Sinnott, M. M. Bayer, J. Koetsier, A. J. Stell, Advanced Security on Grid-Enabled Biomedical Services, Proceedings of UK e-Science All Hands Meeting, September 2005, Nottingham, England.
[20] R. O. Sinnott, M. M. Bayer, J. Koetsier, A. J. Stell, Grid Infrastructures for Secure Access to and Use of Bioinformatics Data: Experiences from the BRIDGES Project, 1st International Conference on Availability, Reliability and Security, (ARES'06), Vienna, Austria, April, 2006.
[21] Griffiths, D., A Pragmatic Approach to Spearman's Rank Correlation Coefficient, Teaching Statistics 2, pp. 10-13, 1980.
[22] Kendall, M. (1948) Rank Correlation Methods, Charles Griffin & Company Limited

From Genes to Personalized HealthCare: Grid Solutions for the Life Sciences
N. Jacq et al. (Eds.)
IOS Press, 2007

The Molecular Medicine Informatics Model (MMIM)

Marienne HIBBERT[a,1], Peter GIBBS[b], Terence O'BRIEN[c], Peter COLMAN[d], Robert MERRIEL[e], Naomi RAFAEL[f], Michael GEORGEFF[g]

[a] *Project Director, Melbourne Health, VPAC and University of Melbourne,* [b] *Chief Scientist, Melbourne Health and Ludwig Institute,* [c] *Neurosciences leader, Melbourne Health and University of Melbourne,* [d] *Diabetes leader, Melbourne Health and Walter and Eliza Institute,* [e] *Steering Committee Chairman, Melbourne Health,* [f] *Senior Database Administrator, Melbourne Health,* [g] *Solution Manager, Monash University. MMIM, Melbourne, Australia*

Abstract. In 2005, a major collaboration in Melbourne Australia successfully completed implementing a major medical informatics infrastructure – this is now being used for discovery research and has won significant expansion funding for 2006 - 2009. The convergence of life sciences, healthcare, and information technology is now driving research into the fundamentals of disease causation. Key to enabling this is collating data in sufficient numbers of patients to ensure studies are adequately powered. The Molecular Medicine Informatics Model (MMIM) is a 'virtual' research repository of clinical, laboratory and genetic data sets. Integrated data, physically located within independent hospital and research organisations can be searched and queried seamlessly via a federated data integrator. Researchers must gain authorisation to access data, and inform/obtain permission from the data owners, before the data can be accessed. The legal and ethical issues surrounding the use of this health data have been addressed so data complies with privacy requirements. The MMIM platform has also solved the issue of record linking individual cases and integrating data sources across multiple institutions and multiple clinical specialties. Significant research outcomes already enabled by the MMIM research platform include epilepsy seizure analyses for responders / non responders to therapy; sensitivity of faecal occult blood testing for asymptomatic colorectal cancer and advanced adenomas over a 25-year experience in colorectal cancer screening; subsite-specific colorectal cancer in diabetic and non diabetic patients; and the influence of language spoken on colorectal cancer diagnosis, management and outcomes. Ultimately the infrastructure of MMIM enables discovery research to be accessible via the Web with security, intellectual property and privacy addressed.

Keywords. Molecular Medicine Informatics Model, MMIM, Research, Grid, Record Linkage

Introduction

The convergence of life sciences, healthcare, and information technology is revolutionizing the discovery of new treatments and the optimal use of available

[1] Corresponding Author: Dr Marienne Hibbert, PhD, MMIM Project Director, 6 North, Main Block, Melbourne Health, Grattan St, Parkville, Victoria, 3050, Australia; E-mail: Marienne.Hibbert@mh.org.au

therapies. Researchers now have the capability to analyse human biology at the finest level through genomics and link to clinical outcomes data giving them the potential to understand the fundamental causation of human disease and predict outcomes. This will drive the development of new drugs, new diagnostics, and lead us to the era of personalised medicine. Key to making the required associations between genotype and phenotype is access to detailed clinical data in sufficient numbers of patients to ensure studies are adequately powered. Few institutions alone have sufficient numbers to perform meaningful analyses, particularly where stratification is performed to look at specific disease attributes. Further, clinicians need to look beyond their own specialty into datasets of other disease groupings, analyzing the impact of co-morbidity. To achieve the research objectives promised by this new era in science, sharing of clinical data between research groups and institutions becomes critical.

The impetus for the development of MMIM came from a recognition of the need to maximise collaborative research across Australia and internationally. A cohesive approach between disciplines was needed so that research data collection became a one time only exercise with the data subsequently available to assist in answering multiple research questions across various clinical disciplines and jurisdictions. In addition new and emerging data sets such as genomic data could be linked to more traditional clinical and outcome data. Thus researchers could examine genetic, genomic and proteomic profiles, all factors that may influence treatment outcome, with respect to toxicity and potential benefit.

The MMIM Project enables research from multiple perspectives, including:

- Genetic predisposition;
- Environmental exposures;
- Health Screening activity;
- Genomics, proteomics & epigenetics;
- Co morbidities;
- Treatment strategies;
- Outcomes.

The objective of the project is to maximise collaborative research efforts, both in Australia and internationally through the development of a federated data integration infrastructure that is enabling:

- Linking and testing of multiple hypotheses without collecting / recollecting their own data;
- Identifying patient numbers for clinical trials based on clinical information or genetic profile;
- Researching suitable pre-symptomatic testing and early intervention based on genotype data;
- Analysing summary/statistical information across institutions and from diverse databases.

1. Materials and Methods

1.1. MMIM Background

Phase 1 of the MMIM project was successfully completed in 2005. It was a pilot project funded by the Science, Technology and Innovation Infrastructure Grant

program (STI) of the Victorian Government Department of Innovation, Industry and Regional Development (DIIRD) through Bio21 Australia Limited. This phase delivered the successful integration of data across five hospital sites (The Alfred, Austin Health, The Royal Melbourne Hospital, Peter MacCallum Cancer Centre, and Western Hospital) and two medical research institutes (Ludwig Institute for Cancer Research, and Walter and Eliza Hall Institute). This stage of the project involved three disease types, namely, colorectal cancer, epilepsy and diabetes.

Phase 2 of the MMIM Project is currently funded until 2007 by the Australian Government Department of Education, Science and Training (DEST) through the University of Melbourne. Phase 2 is integrating data across additional Victorian and interstate hospitals including: Box Hill Hospital; Cabrini Health; Flinders Medical Centre; Monash Medical Centre; The Queen Elizabeth Hospital; Royal Adelaide Hospital; Royal Children's Hospital; Royal Hobart Hospital/Menzies Research Institute; Royal Women's Hospital and St Vincent's Health. Additional disease types to be integrated include multiple sclerosis, stroke, Parkinson's disease, cystic fibrosis, asthma, and brain cancer.

Phase 3 of the project is funded by a grant from DIIRD over a three year period until June 2009 to provide support for the creation of an Australian Cancer Grid and:

- The Infrastructure expansion – the data grid;
- The research activity and outcomes.

1.2. Overview of the MMIM Federated Mode - Technologies

The MMIM project is a federated model where each participating site retains ownership and control over their own data sources and data collection systems. The architecture can be seen diagrammatically in Figure 1.

The data sources used for integration were established as clinical research databases written and maintained by specialist clinicians in their own healthcare facilities. These typically have highly detailed clinical information including surveillance and treatment subsets.

Data was uploaded from these database systems nightly or manually loaded (for static datasets) into a 'cache' database, a local DB2 UDB database termed a Local Research Repository (LRR) located at each site. The distributed LRR databases were federated using IBM Websphere Information Integrator running on a single federating server termed the Federated Data Integrator (FDI). Information Integrator makes remote databases appear as local DB2 table views, allowing single SQL queries to be executed against all federated data.

Public domain databases were also federated into the system including a local XML flat file and resources from the National Library of Medicine (Genbank, Medline and Uniprot) via an Internet web service. Both data sources appeared relational to the end user even though they were not.

A unique number was assigned to each patient termed a Unique Subject Index (USI) by transferring certain identifying information to Sun (SeeBeyond) e-Ways and replicated back to the LRRs via the FDI.

The security system included a number of notable features. Each LRR was connected to the FDI via Virtual Private Network (VPN) connections, which ensure data privacy and encryption. Views block all identifying information, allowing end users to see only the clinical data in conjunction with the USI. User access to these views on the FDI is controlled by assigning DB2 database roles which define privileges

to the table/view level. DB2 Query Patroller is used on the FDI to track all queries for audit purposes. Access is controlled by assigning permission at the table level.

SAS Enterprise Guide was used as the interface for researchers to perform queries, statistical analysis and construct reports.

1.3. The MMIM Architecture

Diagrammatically the phase 2 architecture is shown in the diagram below.

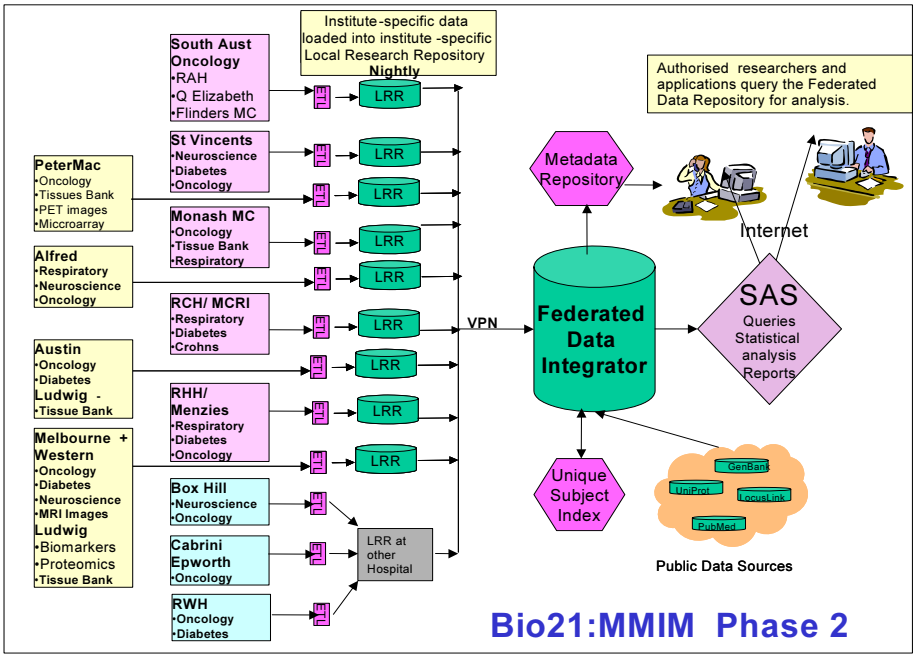


Figure 1. Schematic Architecture of the Bio21:MMIM data grid showing the secure data flow and technologies. Authorized researchers can access the integrated and de-identified data via the Internet.

1.4. MMIM Data Flows-Phase 1 and 2

The key features of the system include:

Connectivity:

Each participating research institution has a local data storage facility (the LRR). All data stores are connected with a secure technology involving double encryption Virtual Private Network and DB2-DB2 encryption. This is the technology commonly used in industries to link various sites of operations for major corporations, such as the banks).

Data loading:

On a nightly or ad-hoc basis, the clinical research data is loaded into the LRR at each individual site. This loading process uses an extraction, transform and load software feature that is installed on each LRR.

Unique Subject Index (USI) data flow:

The USI is a unique number given by the system to each patient; so data accessible to researchers for an individual is linked across databases, but de-identified. In order to de-identify patients, selected identifying data is sent on a nightly basis, from each LRR to a Unique Subject Index (USI) program, where the unique number is generated and stored in an Oracle database. This number is then pushed back to the LRR and stored in encrypted form.

End-user query data flow:

Researchers can only access de-identified data, except where they are performing research on their own databases. The identifying data remains stored on the USI data store with extremely restricted and controlled access. No Health Data is stored externally. Only authorized researchers can log in to MMIM via the Statistical Analysis System (SAS) and perform queries on the de-identified clinical research data. This data flows from the LRRs through the Federated Data Integrator (FDI), the engine that brings together the data from the various institutions (LRRs) to enable analysis. This de-identified data is then put into the researcher's secure folder for statistical analysis (the SAS server). All data queries performed are tracked and logged for security audit purposes.

Security levels:

There are two main levels of security for access to MMIM. Firstly, authorization and usernames and password must be obtained so only authorised researchers who have access to the FDI can query the de-identified clinical data. Secondly, the MMIM System Administrators who are responsible for building and maintaining the system have access to all parts of the system (currently there are only two MMIM project authorised personnel with this level of access).

1.5. MMIM Data Flows Phase 3

As MMIM expands and technology changes, new software called DataStage (IBM) has been added.

This enables the following:

- Improvements in data quality by correcting systematic errors in the data;
- Transformation of data into a simple structure.

As with all the MMIM servers, the DataStage server has only restricted access by system administrators.

The key extra process in this Phase 3 change is data 'passing' through DataStage. Data will be sent from the LRR at each site to the "DataStage server" where it is

transformed and cleansed (its quality improved and validated) in transit, and then sent into the LRR at the site. As with all the data flows, identifying data is always separated from clinical information. All these operations on the data are performed in computer memory and not stored.

1.6. Improving the Quality of Source Data

In MMIM Phase 1 source data was accepted as provided by participant sites with the project being in a pilot phase. With MMIM having passed acceptance testing for the pilot and entering Phase 2 and 3, it was clear that the databases and systems used for clinical data collection vary at local sites and often are not robust enough to ensure high quality and consistent data. As a simple example, the field 'sex' may be stored as 'M/F', 'male/female', '0/1' or the field may contain a mixture of all 3 or other characters such as '?'.

The project disease/tumour team leaders (in particular the colorectal cancer group) have been working with colleagues in other hospitals and disease areas to standardise data fields and collection processes as far as possible.

Further MMIM has been working cooperatively with groups such as the Cancer Council Victoria (CCV) which has through Victorian state funded projects such as the Victorian Cancer Outcomes Network (VCON) been trialing standardised cancer data capture models.

All of these initiatives will in time contribute to standardised and better quality data for cancer and other diseases of relevance to researchers wishing to utilise MMIM.

1.7. Record Linkage - The Unique Subject Index

The Unique Subject Index (USI) is the key element in linking patient records across disparate data sources within and across institutions. It ensures compliance with privacy.

Linking patient/subject records and assigning USI identifiers to data allows patients to be linked across multiple institutions and databases while also observing legal, ethical, privacy and data ownership constraints

The USI is developed based on matching of six key demographic data items:

- Surname
- Given name
- Middle name or Initial
- Date of Birth
- Gender/sex
- Digits 5 to 9 of the Medicare Number

The software checks new records for a match against existing subjects, using probabilistic matching and a score is assigned on the basis of match / non-match for each data item. "Fuzzy logic" is used for transpositions, soundex matches, common "dummy" names (e.g. Babe of, Twin 1). Manual checking of subjects in the "grey area" between thresholds can be undertaken by the data owners.

1.8. The metadata management – business glossary

The MMIM system provides the ability to search for the information in MMIM and discover whether the required data is available - the metadata as opposed to access the data itself – the technical data. MMIM users can search and discover the

- clinical areas covered (diseases & database purpose)
- sites contributing
- types of data collected (pathology, procedures, genetic)
- detail of data elements
- have enough information to request access

MMIM chose IBM's Websphere Business Glossary (WBG) to provide terminology management capabilities. Definitions of standard terms, attaching standard terms to items (databases, tables, or columns), defining the hierarchies of terms, specifying the preferred terms, synonyms and having categories of terms with hierarchies have been implemented. This functionality is important in search and discovery of metadata as users may use non standard or non familiar terminology (or may misspell words) when describing items. For example, "date of birth" may be described as "dob", "gender" may be described as "sex", etc.

The Business Glossary has open access from the Internet and searching for information can occur by browsing, by drilling down the 'trees' or searching using keywords. interpret the data fields, to run queries and to understand the data models so they have the knowledge to join data across databases.

1.9. Addressing the Issue of Privacy

MMIM infrastructure + processes: The project obtained independent legal advice from lawyers and privacy experts at all stages to ensure that measures taken to protect privacy continue to be timely and relevant as the project grows.

Site and project specific Human Research Ethics Committee approval has been obtained for all participating sites, as a prerequisite to proceeding with implementation of MMIM at participating sites. All data outside of the hospital LRR is de-identified. All health data is de-identified. Log-ons and passwords are used, Virtual Private Networks are utilised for transmission of data with secure internet access. Researcher access is provided to specified tables in MMIM only on application with the research/purpose fully described and only after approval of the application by the MMIM steering committee and data is only available to researchers in de-identified form.

1.10. Intellectual Property

A Collaboration agreement that all participating sites must sign to join MMIM explicitly provides for recognition of both Background and Project Intellectual Property.

The project has a set of standard IP management and commercialization processes. Default IP positions are agreed. However, individual research projects are free to negotiate appropriate terms on a case by case basis.

2. Research Results

Examples of research outcomes to date the areas epilepsy and neuropsychiatry, (1–3) evaluating the sensitivity and specificity of FOBT compared to colonoscopy over a 25 year period, (4) and the evaluation of colorectal cancer patients in the areas of biomarkers and therapy. (5–8)

3. Discussion

3.1. Building the Federated Data Integration Infrastructure

Phase One of MMIM successfully built the system infrastructure and federated database integration capability outlined in the methodology section above. This technology allowed the issue of patient privacy, patient record-linkage as well as researcher intellectual property to be protected. Acceptance of the system have meant a further data sets were successfully integrated in five Victorian public sector sites and two medical research institutes (Ludwig Institute for Cancer Research, and Walter and Eliza Hall Institute). This stage of the project involved data sets (clinical, genomic, tissue bank & biomarkers) for three disease types, namely, colorectal cancer, epilepsy and diabetes.

Phase two of MMIM is building on this success to include a further ten public health sites in Victoria and three states and territories. This phase involve a further four medical research institutes (*p*health CSIRO, Murdoch Children's Research Institute, Monash Institute of Medical Research, Neurosciences Victoria) and link more than 35 disease databases.

Phase 3 will expand the technology across the Regional Integrated Cancer Services within Victoria and the Metropolitan Melbourne Hospitals which together with the Victorian and interstate metropolitan public hospital sites implemented in Phase 1 and 2 will create the South eastern Australia part of an Australian Cancer Grid.

3.2. Powering Future Research

The MMIM Project has transformed the way that research can be undertaken giving approved and authorised researchers unprecedented access via the internet to a virtual repository of privacy–protected data not previously available.

From their own work stations researchers can now:

- Link genomic data to clinical / outcome data in Colorectal Cancer and Epilepsy;
- Test multiple hypotheses without collecting / recollecting their own data (with data owner approval);
- Research suitable pre-symptomatic testing and early intervention based on genotype data;
- Research genetic, genomic and proteomic profiles, factors that may influence treatment outcome, with respect to toxicity and potential benefit;
- Analyse summary/statistical information across institutions and from diverse databases.

The following table summarises the traditional approach to research data collection and assembling of databases with that offered by the MIIM Project.

Table 1. Comparative Advantages of Using MMIM

Using Traditional Standalone Research Databases	**Using the MMIM Data Grid**
Static	*Dynamic*
Data at one point in time	Data refreshed & updated
Often one-off 'dump'	Live link to clinical research data
Linked once	Links made on-demand
Often anonymised data	Codified–ethically re-identifiable in exceptional circumstances and privacy protected
Data leaves 'owners' control	Data owners control access
Minimum data sets	Minimum + legacy data
Must specify exact data/query up-front - can only answer one off specific research question	*Discovery* Research analysis / explorative 'Quality' type clinical reports Clinical Data collected at healthcare site Discovery tools and potential for iterative and exploratory research on a theme (to approved data)
Usually population based studies	Clinical and Biomedical Data collected at healthcare site and population data

3.3. Research, Publications and Teaching

The MMIM virtual repository has enabled collaboration between multiple institutions, both within and across disease specialties, and between clinical researchers, bio-informaticians and Information Technology specialists. This in turn has expanded research capacity and productivity as follows:

- Within and across disease groups;
- Between data owners;
- Between data owners and researchers across academic institutions in Australia;
- Between data owners and researchers overseas;
- New research data types –e.g. imaging (MRI)

At the same time MMIM can enable expanded teaching and training resources (data sets and tools) in the health research field (medical informatics, genomics, proteomics) which is being developed including the publication of thesaurus/glossary - metadata schemas and ontologies for medical research.

4. Conclusion

MMIM provides a privacy protected data grid for connecting heterogeneous and dispersed data for medical researchers.

The platform is operational and is growing and developing to become scalable and sustainable. It continues to incorporate new data and provide tools for researchers.

References

[1] Yerra R, Kilpatrick C, Matkovic Z, King B, Rafael N, Hibbert M, Brand C, and O'Brien, First Seizure Clinic Experience: Heterogeneity of patient population and prognosis, *Epilepsia* **46** suppl 8 (2005), 360.

[2] Yerra R, Kilpatrick C, Matkovic Z, Belbin S, Rafael N, Hibbert M, Brand C, and O'Brien T, Syndromic diagnosis of epilepsy in the First Seizure Clinic population, *Journal of Neurosciences*, **238** (2005), S150

[3] Mark Walterfang, Ronald Siu, Dennis Velakoulis. The NUCOG: validity and reliability of a brief cognitive screening tool in neuropsychiatry patients. *Australian and New Zealand Journal of Psychiatry* **40** (2006), 995-1002

[4] Macrae FA, Slattery MA, Brown GJ, O'Dwyer MA, Murphy C, Hibbert ME, St John DJB. Sensitivity of faecal occult blood testing (FOBT) for asymptomatic colorectal cancer and advanced adenomas over a 25 year experience in colorectal cancer screening. Gastroenterological Society of Australia Australian Gastroenterology Week, Brisbane 2005*, Journal of Gastroenterology and Hepatology* **20** (2005), A34.

[5] Lim E, Jones IT, Gibbs P, McLaughlan S, Faragher I, Skinner I, Chao MW, and Johns J. Subsite-specific colorectal cancer in diabetic and non-diabetic patients. *Cancer Epidemiology, Biomarkers & Prevention* **14** (2005), 1579–1582.

[6] Rodrigues J, Lim E, McLaughlin S, Faragher I, Skinner I, Chao M, Croxford M, Chapman M, Johns J, Gibbs P The Influence of Language Spoken on Colorectal Cancer Diagnosis and Management'. *ANZ Journal of Surgery* **76** (2006), 671.

[7] Gibbs P, McLaughlin S, Skinner I, Jones I, Hayes I, Chapman M, Johns J, Lim L, Faragher Re: Completion of therapy by Medicare patients with stage III colon cancer. . *Journal of the National Cancer Institute* **98**(21) (2006), 1582.

[8] Reiger NA, Barnett FS, Moore JWE, Neo E, Badahdah F, Ryan AJ, Ananda SS, Croxford M, Johns J, Gibbs P The quality of pathology reporting impacts on lymph node yield in colon cancer*, Journal Clinical Oncology* **25** (2007), 463.

Acknowledgments. This project was funded by the Victorian Government Department of Innovation and Regional Development via Bio21 Australia Limited. Phase 2 was funded by the Commonwealth Government of Australian, Department of Education, Science and Training, via their Systemic Infrastructure Initiative. Phase 3 is funded by the Victorian Government Department of Innovation and Regional Development via the University of Melbourne.

From Genes to Personalized HealthCare: Grid Solutions for the Life Sciences
N. Jacq et al. (Eds.)
IOS Press, 2007

ImmunoGrid - The Virtual Human Immune System Project

Andrew EMERSON [a,1] and Elda ROSSI [a]
[a] *High Performance Systems Division, CINECA, 40033 Casalecchio di Reno, Italy*

Abstract. ImmunoGrid is a 3 year project funded by the European Union which began in February 2006 and establishes an infrastructure for the simulation of the immune system that integrates processes at molecular, cellular and organ levels. It is designed for applications that support clinical outcomes such as the design of vaccines, immunotherapies and optimization of immunization protocols. The first phase of the project concentrated on improving and extending current models of the immune system. We are now entering the second phase which will design and implement a human immune system simulator. Since the new models are orders of magnitude more complex than the previous ones, grid technologies will be essential in providing the necessary computer infrastructure. The final phase of the project will validate the simulator with pre-clinical trials using mouse models.

Keywords. Immunology, Grid computing

Introduction

The immune system is a complex and adaptive learning system which has evolved to defend the individual against foreign invaders. It has multiple levels (molecular, cellular, organ and tissue, organism, and organism-to-organism) and is also combinatorial in nature with a large number of products; there are typically more than 10^{15} antibodies and 10^{12} immune system cell clones in a single individual. The function of the immune system depends on both the genetic composition and the previous exposure, i.e. the experience of the organism.

Immune intervention, such as vaccination, is the most effective method for the control of disease and the greatest achievements include eradication of smallpox, near-elimination of polio, and savings of some 170 million person-years. Vaccination has been used in the control of over two dozen diseases by the 50 or so successful vaccines which have been developed to date. These vaccines largely protect against infectious diseases, although recent vaccine developments offer great hope for treatment for a broader range of diseases. Large-scale studies of the immune system, also known as immunomics, is the key factor driving the current wave in vaccine development. These include genomics and proteomics, analysis of the diversity of pathogens or complexity of the human immune system, high-throughput screening or immunoinformatic tools for the management

[1] Corresponding Author: A. Emerson, High Performance Systems Division, CINECA, 40033 Casalecchio di Reno, Italy; E-mail: a.emerson@cineca.it

and analysis of vast quantities of data. Computational models are becoming increasingly important in immunomics:

1. Experimental approaches are expensive and it is impossible to perform systematic experimental studies of immune processes in humans.
2. Because of ethical issues, there are stringent limitations as to what experiments can be performed in humans.
3. Computational approaches can compensate for the limitations of allowable studies.

Computational modeling is ideal for complementing experimental research and clinical observations. The usefulness of computational approaches to the study of immune system has been demonstrated, but computational models that encode the natural-size immune system have not been developed because of the past limitations of computational infrastructures. Grid technologies now enable the complexity of the computational models to be matched with that of the natural human immune system. Such computational models have applications not only in theoretical immunology (better understanding of immune function), but also in clinical medicine in diagnostics and the development of vaccines and therapies for cancer and infectious disease.

This paper is intended as a high level view of the ImmunoGrid project which is currently at the half way stage. We start by giving an overview of the computer models employed in our simulations of the immune system and the computational resources required. We then propose a grid infrastructure and discuss the possible technologies which we will be used in its implementation.

Specific applications of the ImmunoGrid Simulator will appear in subsequent publications.

1. Immune system models

Methods for modeling the behaviour of the immune system generally are designed for different length scales, ranging from interactions at the molecular level (e.g. peptide binding to receptors) to system approaches which operate at cellular and organ dimensions. A novel aspect of the ImmunoGrid project is that it integrates both levels, resulting in the most realistic models yet realised.

The simulator at the heart of the ImmunoGrid project originally derives from the ImmSimm program, a cellular automata model of the human immune response [1]. Variants and parametrisations of this program have been successfully employed in the modeling of, for example, the response to generic bacteria and viruses (C-ImmSimm, [2]) and the Triplex cancer immunoprotection vaccine (SimTriplex, [3]). The stochastic cellular automaton represents many different interacting entities (antigens, antibodies, T cells, B cells, peptides, etc) and hence is capable of modeling many aspects of the immune response. Although very successful, the computational requirements impose limits on the sample sizes and complexity that can be handled in current simulations. For example, very often two dimensional grids are used, typically representing only a few mm^3 of tissue. ImmunoGrid has already started to employ much larger three dimensional models which accurately describe the spatial distribution of the entities. When complete the ImmunoGrid simulator will be able to deal with systems which approach the complexities of real immune systems.

2. Computational requirements for the Simulator

Since the first major objective of ImmunoGrid was the research and validation of improved immune system models, the implementation of the Simulator has only just started. However, from an initial analysis of the new models the following requirements have emerged:

1. **Processing power.** The most important requisite is for high computational processing power. Extending the SimTriplex model (used for the simulation of the cancer immune response), for example, to larger samples based on three dimensional lattices requires powerful computing facilities. Fortunately, parallelisation of the SimTriplex program is possible and efficient and the implementation of this program with the MPI message passing library on the supercomputing clusters available to the ImmunoGrid consortium has has already begun. For the C-ImmSimm program (designed for the response to bacteria and viruses), the other main component of the Simulator, a high-throughput approach is more appropriate. Here the simulation involves a genetic algorithm which runs C-ImmSim for every genotype in the simulation space representing the set of therapies possible until convergence. Thus, a very large number (e.g. many tens of thousands) of simulations are launched, with each run requiring typically about 10min CPU time for the current models. The program is therefore "embarrassingly parallel" and is readily implemented on a parallel computer.
2. **Data federation.** Although less critical, access to data resources, either in the form of primary databases or data produced by the simulations themselves, will need to be made available to all the components of the Simulator. Technologies for performing data federation between geographically separated computing resources, such as SRB [4] or OGSA-DAI[5], are being considered.
3. **Visualisation and user portals.** One goal of the project is to create a tool which will allow clinicians or health researchers to design new vaccination protocols based on the simulations performed while another important objective, strongly promoted by the EU, is to provide facilities for educational purposes. Thus, at a later stage it will be possible for non-specialist audiences such as the general public to have access to ImmunoGrid. For these reasons, user portals and visualization facilities are essential. Basic visualization tools, to show for instance the spatial distribution of the interacting entities, have already been designed, while a "ImmunoGrid user portal" forms part of the grid implementation (see next section).

As mentioned above enhanced versions of the basic simulator tools have already been written and are currently being tested. In the next section we will describe how we are integrating these and further tools to create the Simulator.

3. Grid implementation

It should be mentioned at this stage what we mean by "grid" since the term in computational science can have various meanings and interpretations. In our project, the interpretation is primarily that of a "virtual organisation", an important concept in many grid

projects and implying a single point of access for heterogeneous resources and administrative domains. This is not to suggest that the programs themselves only require modest resources but that for the moment we don't anticipate direct communication between the various grid nodes during execution. This is technically feasible by using, for example, grid or web services, but is in practice fairly complicated and also incurs some costly overheads. In addition, these technologies are still not yet mature and standards are continually evolving. It should be emphasised however that we do expect CPU processing requirements to increase rapidly during the project as models of more complex tissues and organs are simulated, but these will run on separate clusters. These models will themselves be scaled up to the whole organism level, i.e. the mouse, to facilitate validation by the pre-clinical trials (based on transgenic mice, see the next section). Towards the end of the project the Simulator will be extended to the human natural scale, which is about 1,000 times bigger than that of the mouse.

The aim of ImmunoGrid is not to develop new grid middleware so we are evaluating the grid-enabling technologies currently available to the consortium members and these are described below.

3.1. EnginFrame portal

EnginFrame[6] is a commercial product based on "agents" which manage computing resources on behalf of users, interacting with the underlying operating system, batch schedulers, grid middleware, user applications and so on. With this software it is possible to create a web-based portal which provides a single point of access to heterogeneous resources, allowing the user to submit and monitor jobs, browse the user's disk space and obtain information regarding the status of the resources. It is also possible to design specific interfaces for commonly used applications, e.g. the different versions of the simulators we are setting up in this project. But in addition to hiding the complexities of the operating system (usually some dialect of UNIX), portals built with EnginFrame have two features which make the environment very attractive for building grid infrastructures:

1. Single sign-on. When combined with a standard password authentication system (e.g based on LDAP, a common UNIX protocol) a user will need to enter username and password only once, i.e. on entry to the portal, in order to access the resources and run jobs. This may require careful mapping of the user accounts between the hosted computing platforms, but this is generally straightforward, at least at a single site.
2. Transparent access to job schedulers. This is a crucial feature because many schedulers and queuing systems are used in supercomputing centres around the world. A non exhaustive list includes PBS, Torque, LSF and LoadLeveler. Support for some schedulers is already built into the product (e.g. LSF) while for others (e.g. LoadLeveler) a plug-in needs to be written and installed.

A computing portal based on EnginFrame is currently under development at the CINECA computing centre where a trial version is being beta-tested by selected users. The CINECA portal is a good test of the technology because this centre provides four different publicly available supercomputing clusters, with different operating systems (Linux and IBM AIX) and two different batch schedulers (LSF and LoadLeveler). The unified access to the batch schedulers is expected to be popular with users who generally

find them difficult to use. In principle, there should be no problem in incorporating other systems external to CINECA into this portal or one built for ImmunoGrid with a custom interface for submitting C-ImmSimm or SimTriplex jobs.

3.2. Application Hosting Environment

The Application Hosting Environment (AHE) [7] is a sophisticated grid infrastructure currently in use at the site of the UK partner of ImmunoGrid. The AHE has been designed to provide the simplest possible service interface to a client for submitting jobs to highly complex grids. The AHE, is a lightweight web service hosting environment able to operate over multiple administrative domains. The AHE stores all the necessary information about how an application should be run on the various computational resources of a grid and provides a uniform interface to the client for running that application across those resources. The AHE interfaces with GridSAM [8], a job submission and monitoring service. GridSAM provides an abstraction of resources, such as Globus Toolkit and Condor on the resource side. Therefore the AHE interacts in a clear transparent way with complex grid architecture and submission systems, such as queuing systems, without interaction from the client. The AHE provides resource selection, application launching, workflow execution, provenance and data.

3.3. Specific applications

One or both of the above technologies will be used to build the framework for the ImmunoGrid simulator, a prototype of which will be available by the end of 2007. But, although the infrastructure and portal will be expected to cope with the day-to-day running of the ImmunoGrid, we anticipate that there will be simulations which will be beyond the computational resources normally available within the consortium. In these situations applications will be made to some large European Grid-infrastructures like DEISA[9], an e-infrastructure connecting the major supercomputing centres in Europe which provides facilities for particular problems which require substantial computing resources or the NGS, the UK National Grid Service[10]. Thanks to the use of the portal, the applications will run on these large infrastructures in a transparent way, hiding the implementation details on the different computing environments. Thus, some nodes on our grid will in fact be other grid infrastructures.

4. Simulator validation by pre-clinical trials

Validation of the ImmunoGrid simulator will be performed by one of the consortium members who has expertise in performing pre-clinical trials with transgenic mice models, such as BALBneuT which carries the HER-2/neu oncogene. The predictions of the simulator will be tested in vivo by the mice models, with healthy and pathological individuals, and will provide feedback for improving the computer models. Later there will be comparison with experimental data for human systems, where available.

5. Comments

The European Commission has identified the Virtual Physiological Human as one of the challenges of their seventh Framework Programme (FP7). ImmunoGrid will clearly have an important role in this effort and collaborations with other projects are being sought.

The project began in February 2006 and has already extended the immune system models and developed more sophisticated parametrisations. Small-scale simulations on single grid nodes are currently being run. By the end of 2007 a demonstration version of the grid-enabled simulator will be available for general use. With the final version of the simulator it will be possible to perform immunological simulations vastly more complex and detailed than those available currently.

Acknowledgments

The ImmunoGrid consortium consists of the following entities: CINECA (Italy), the University of Queensland (Australia), CNR-Rome (Italy), CNRS-Montpellier (France), Birkbeck College London (UK), Dept. of Experimental Pathology (Bologna, Italy) and the University of Catania (Italy). This project is funded as part of the European Union FP6 programme, contract no. FP6-2004-IST-028069.

References

[1] F. Celada, and P.E. Seiden: A computer model of cellular interaction in the immune system. *Immunology Today* (1992), 13:56–62.
[2] F. Castiglione, M. Bernaschi and S. Succi: Simulating the immune response on a distributed parallel computer. *Int. J. Mod. Phys. C* **8** 527–545.
[3] F. Pappalardo, P.-L Lollini, F. Castiglione and S. Motta: Modeling and simulation of cancer immunoprevention vaccine. *Bioinformatics* Vol. 21 no 12 (2005) 2891–2897.
[4] A. Rajasekar, M. Wan and R. Moore: "MySRB & SRB: Components of a Data Grid," hpdc, p. 301, *11th IEEE International Symposium on High Performance Distributed Computing* (HPDC-11 '02), (2002).
[5] The OGSA-DAI Project, http://www.ogsadai.org.uk/.
[6] For more information see http://www.enginframe.com.
[7] P. V. Coveney, S. K. Sadiq, R. Saksena, M. Thyveetil, S. J. Zasada, M. McKeown and S. Pickles, "A Lightweight Application Hosting Environment for Grid Computing", *Proceedings of the UK e-Science All Hands Meeting 2006* (2006), 217–224.
[8] See for example, http://gridsam.sourceforge.net/2.0.0/index.html.
[9] For more information about DEISA see http://www.deisa.org
[10] See, for example, http://www.grid-support.ac.uk/.

From Genes to Personalized HealthCare: Grid Solutions for the Life Sciences
N. Jacq et al. (Eds.)
IOS Press, 2007

Parameter Estimation for Cell Cycle Ordinary Differential Equation (ODE) Models using a Grid Approach

Roberta ALFIERI [a,b,1], Ettore MOSCA [a,c], Ivan MERELLI [a], Luciano MILANESI [a,b]
[a] *Institute for Biomedical Technology, National Research Council, Milan, Italy*
[b] *Consorzio Interuniversitario Lombardo per L'Elaborazione Automatica, Milan, Italy*
[c] *University of Milano-Bicocca, Milan, Italy*

Abstract. Cell cycle is one of the biological processes that has been investigated the most in the recent years, this due to its importance in cancer studies and drug discovery. The complexity of this biological process is revealed every time a mathematical simulation of the processes is carried out. We propose an automated approach that mathematically simulates the cell cycle process with the aim to describe the best estimation of the model. We have implemented a system that starting from a cell cycle model is capable of retrieving from a specific database, called Cell Cycle Database, the necessary mathematical information to perform simulation using a grid approach and identify the best model related to a specific dataset of experimental results from the real biological system. Our system allows the visualization of mathematical expressions, such as the kinetic rate law of a reaction, and the direct simulation of the models with the aim to give the user the possibility to interact with the simulation system. The parameter estimation process usually implies time-consuming computations due to algorithms of linear regression and stochastic methods. In particular, in the case of a stochastic approach based on evolutionary algorithms, the iterative selection process implies many different computations. Therefore, a large number of ODE system simulations are required: the grid infrastructure allows to distribute and obtain the best model that fits the experimental data. The computation of many ODE systems can be distributed on different grid nodes so that the execution time for the estimation of the best model is reduced. This system will be useful for the comparison of models with different initial conditions related to normal and de-regulated cell cycles.

Keywords. Cell cycle models; parameter estimation; grid platform.

Introduction

The cell cycle is a crucial event in the life of every organism. It consists of a series of coordinated and oscillating steps which allow the cell to grow and duplicate correctly. It is an important biological process frequently studied in correlation to tumour disease

[1] Corresponding Author: Roberta Alfieri; E-Mail: roberta.alfieri@itb.cnr.it

and it is considered a valuable target for drug discovery. Thus, the typical systems biology approach can be applied to study this process in order to verify the impact that differently regulated genes can have in normal and cancer cells. The identification of cell cycle models has been frequently reported in the recent literature: in particular cell cycle models for the budding yeast *S. cerevisiae* are more advanced [1] but also detailed models for mammalian cell cycle are forthcoming [2-4].

The ultimate objective of these studies consists in the mathematical simulation of different biological processes which have been described by a set of kinetic equations that define the biochemical reactions, and dynamic equations, structured as Ordinary Differential Equations (ODE) system, that quantify the biological processes. The kinetic equations require initial parameters, such as the rate constants for each reaction, and also the initial concentration of the model species. The mathematical simulation of the system can be performed via many different mathematical software, both license-free such as XPPAUT and Copasi or requiring license, like Mathematica and MatLab. For the simulation of a biological process it is necessary to define the equations describing the system and to set the initial parameters required for the calculation.

The simulation of a single set of equation can be performed on a single workstation because the numerical integration of an ODE system is not very time consuming. On the other hand, High Performance Computing techniques, like grid, are extremely useful to perform the parameter estimation that is the evaluation of the best set of parameters which define the model relating to a specific experimental dataset. The parameter values have high impact on the accuracy of the models in representing real biological systems but these values are difficult to estimate experimentally. Generally the estimation of the kinetic parameters *in silico* is performed by fitting the data by computing a number of ODE systems with different parameters and verifying the best solution.

This problem has been recently faced by Zwolak et al. [5] who implemented an algorithm through which the estimation of the best set of parameter fitting the experimental data is possible. The parameter estimation is performed using ODRPACK [6] which finds an estimate for the rate constant by minimizing the weighted orthogonal distance between the experimental data set and the calculated model.

A different way to computationally solve the problem of the parameter estimation has been suggested by Dhar et al. [7]. The technology implemented relies on an Adaptive Swarm Algorithm [8], which is based on simulation of social behaviour in a flock of birds. This algorithm is highly suitable for constrained multi-objective optimization problems. The models are simulated over the grid through GridX meta scheduler and Globus.

1. Methods

In the context of a systems biology study of the cell cycle process, we developed a system for the automatic computation of cell cycle models. Our system relies on the Cell Cycle Database, a resource which integrates the most useful information about genes, proteins and models related to the budding yeast and mammalian cell cycle processes. Our computational system allows users to solve the ODE system which mathematically describes the biological system using grid technology. The aim of this work is the development of a parameter estimation pipeline on the top of this system, in

order to find the best model that fits the experimental data. Furthermore, the system integrates mathematical data for each model, such as kinetic and dynamic equations, the initial parameters and the initial model components concentrations.

1.1 The Cell Cycle Database

The pipeline designed for the model simulation relies on the relational database named Cell Cycle Database [http://www.itb.cnr.it/cellcycle], an integrative resource which collects the main information regarding genes and proteins involved in budding yeast S.cerevisiae and mammalian cell cycle process. A specific section of the database, which is dedicated to store the main information related to the yeast and mammalian cell cycle models published in the recent literature, allows the user to interface with the pipeline for model simulation. The implemented system is able to provide information on the published models, such as the detailed publication data (e.g. authors, PubMed ID, abstract, journal information), the diagram of the model and the related XML file.

1.2 Model Simulation Engine

The pipeline is composed of a series of PHP scripts that allow the user to extract information both from model repositories and from the XML file which describes the whole model. Moreover, these scripts interface users with the model simulation input in order to choose the parameters of the computation and retrieve the model behaviour related to the given conditions.

The simulation software chosen for our system is XPP [10], a computational device frequently used in systems biology numerical calculations. XPPAUT allows the solution of differential equations using many different options for the numerical algorithm. It is widely used for the modelling of different biological pathways [4] and it requires simply formatted input files. XPPAUT is very portable, has a simple input file format and can be run without a GUI: so it is a perfect candidate for solutions using a grid application.

1.3 XML files and model equations

Models stored in the Cell Cycle Database are encoded in Systems Biology Markup Language (SBML) [10], an internationally supported and widely used computer-readable format for representing models of biochemical networks. Some SBML models included in the database are manually generated using the JigCell Model Builder software [11], a model editor which allows the construction of biochemical reaction networks in SBML format.

Mathematical expressions in SBML are represented using Mathematical Markup Language (MathML) [12] an XML-based language especially created to represent mathematical expressions. Some MathML based components, such as algebraic constraints, assignment and rate rules, function definitions, kinetic laws and stochiometric matrix are crucial to describe a biological pathway mathematically.

The models are essentially based on differential equations and they can describe abundances, kinetics and binding affinities of pathway components and their interactions [13]. In this work we consider models based on a system of nonlinear ordinary differential equations (ODE system) in which each X_i state variable (usually species concentrations) can be described by the Eq. (1):

$$\frac{dX_i}{dt} = F_i(X_1, X_2, ..., X_n; p_1, p_2, ..., p_m); i = 1, ..., n \quad (1)$$

where the function F_i is the rate of change or the rate law of the state variable X_i and pi are parameters of the function F_i. The time course of each state variable is obtained by solving the ODE system which requires a set of initial conditions $X_i(t=0)$. One of the biggest problems of the differential equations approach is the experimental estimation of the numerical values of species initial concentrations and parameters. A way to bypass the experimental determination of these values is by parameter estimation, or curve fitting, even if these techniques need many quantitative and qualitative experimental data against which to fit the parameters.

1.4 The grid platform

Using the developed user interface, the computation of the simulation is submitted to a High Performance Computing platform, the grid. This solution plays a key role in the development of a system that must compute a large number of independent ODE system solutions in order to perform parameter estimation on an experimental dataset. In this work we successfully tested the possibility of porting the ODE solver system on the grid, by the creation of an infrastructure able to support users and to perform efficiently the distribution of computation.

The developed solution relies on the wide area grid platform of the European EGEE project, a network of several *Computing Elements*, that are the gateways for the computer clusters on which jobs are performed while the grid core is a set of *Resource Brokers* delegated for controlling the execution of the different jobs. This grid infrastructure is based on the *Globus Toolkit* which represents an ideal communication layer between the different grid components.

The computational resources are connected to a *Resource Broker* that routes each job onto a specific *Computing Element* and takes into account the directives of the submitting script, called JDL. The JDL script, composed using the Job Description Language, specifies the *InputSandBox*, that lists the files to be submitted to grid, and the *OutputSandBox*, which are the output files to be retrieved. The software that gives access to the distributed platform is made of a set of tools by which secure communications can be established between the grid infrastructure and the *User Interface*.

2. Implementation

The core of our technology is a PHP library (in the middle of Fig. 1), that creates a pipeline relying on independent engines (at the bottom in Fig. 1) in order to generate web pages for the interaction with users (at the top in Fig. 1). The visualization of the SBML model requires different components: data retrieval from the Cell Cycle Database, the SBML parser and the MathML to HTML converter. The converter itself is a pipeline which accomplishes the translation of mathematical expressions included

in SBML models from MathML to HTML and so makes the their visualization possible to view on a web browser [14 - 17]. Through the interface which allows the setting of initial conditions and XPPAUT internal options, the PHP level starts a simulation using XPPAUT in the grid platform.

2.1 User Interface

The simulation system has an interactive interface where users can set different values for species concentration, parameters and XPPAUT internal options. This interface allows users to explore model behaviors starting from different initial conditions or setting the solver in different way, by changing, for example, the integration method. Algebraic rules are listed to lead the user to a correct initialization.

In the SBML file model species concentrations and parameters values could be initialized internally by assignment rules: in this case for species or parameters the string "AssignRule" is placed in the input form associated with it. The XPPAUT input file is created by adding users selections to an existent XPPAUT input file, stored in the database, which contains equations and other information for a correct simulation.

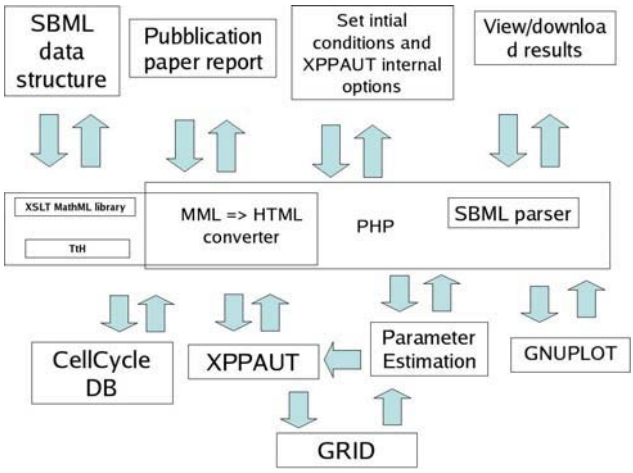


Figure 1: The simulation engine workflow. At the bottom, there is the system used to simulate the models; in the middle, the software pipeline that manages the system; at the top, the user interface.

2.2 The distributed approach

The user interface works on the top of a set of script that are delegated for job submission, the monitoring and the retrieval of the results. Using this solution, the simulation can be coordinated by a single server, on which the grid User Interface software is installed, obtaining a scalable system according to the grid performance.

For each set of ODE system simulation which must be calculated, a grid job is submitted: it means that according to the user parameters selected a JDL script is dynamically generated with the information about the input and the related job requirements. The number of equations which have to be simulated in a specific group is related to the computation time needed for each job and then to the scalability of the system. The jobs are routed by the *Resource Broker* to the best *Computing Element* that

is available at the moment. The ODE solver system is deployed on the grid node at job execution time and the results are retrieved to the User Interface from where the execution of the jobs are automatically monitored and, in case of failure, re-submitted to the grid infrastructure.

When a simulation job is finished the result can be downloaded and quickly viewed on a 2D plot. A graphical interface allows the users to plot one or more of the species involved in the model on the y axis, while on the x axis generally the time is indicated. Thus, concentration versus time or concentration versus concentration plots can be generated. Graphs are images exported in png file format using Gnuplot [18].

3. Results

As test case let us consider a model of G1 to S transition in mammalian cell cycle process [3]. This model is essentially focused on the main key points which characterize the G1 to S transition in mammalian cells: the restriction point R and the progression towards the S-phase. These points involves a small set of proteins, including the transcription factor family of E2F/DP dimers (E2F1-6, DP1 and DP2), the pocket protein family, including the tumor suppressor pRB (retinoblastoma), which are the central regulators of the mammalian cell cycle. In particular, E2F/DP regulate the transcription of a large number of genes which have a crucial role in the G1 to S transition, while pRB has a crucial role as the main inhibitor for the progression if the cell cycle from the G1 phase to the DNA synthesis phase (S-phase).

The model requires to be manually written in SBML format using JigCell Model Builder [11]. By accessing the Cell Cycle Database web interface the user can explore and simulate the model. The models stored in our resource can be analyzed by the user from two different point of view: the first concerns the general information relating to the model, mainly regarding the model publication paper, the second focuses mainly on the information related to the SBML model data structure that makes the simulation of the model possible.

Using the first possibility of exploration the user can retrieve information related to abstract, model wiring diagram and the list of all the proteins involved in the model which are linked to their Cell Cycle Database entries. The abstract of the paper is directly linked to E-Biosci system [19], a literature search engine developed in order to automatically search the scientific literature based on a description as input query. The E-BioSci system relies on performing a conceptual fingerprinting comparison in several literature databases.

On the other hand, browsing the model information from the mathematical point of view, the user can explore the whole SBML data, selecting the different model component to visualize. In case of the G1 to S transition model user can observe 5 unit definitions, 1 compartment, 9 species, 41 parameters, 22 reactions, 2 function definitions and 9 ODEs. An example of the visualization of the equations is shown in Figure 2. The direct simulation of the model is possible through a specific interface, as described before. This interface is composed of four tables which contain species, algebraic rules, parameters and XPPAUT internal options such as total integration time or the integration method. All components are initialized with the default values. Algebraic rules remind the user that some species have constraints on their initial values, while the string “AssignRule” reminds the user that the associated species or

parameter values cannot be initialized because there is an assignment rule defined in the model.

When the simulation has been correctly completed, the output retrieving and the results visualization are possible. User can choose the species to plot on the graphs. The plot of the time course of pRB (labelled pRB_1) and E2F/DP (labelled E2F1_1) is shown in Figure 3 as an example. Even if initial conditions are different (and hence the quantitative solution will be different) we can conclude that the model behaviour calculated with our system and the one calculated in the work previously discussed [3] and used as test case are the same.

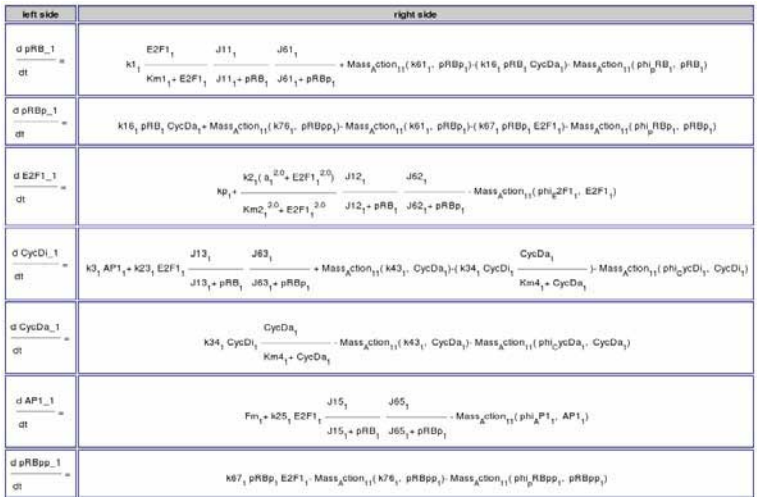


Figure 2: ODE system from the test-case model which is shown in the user interface.

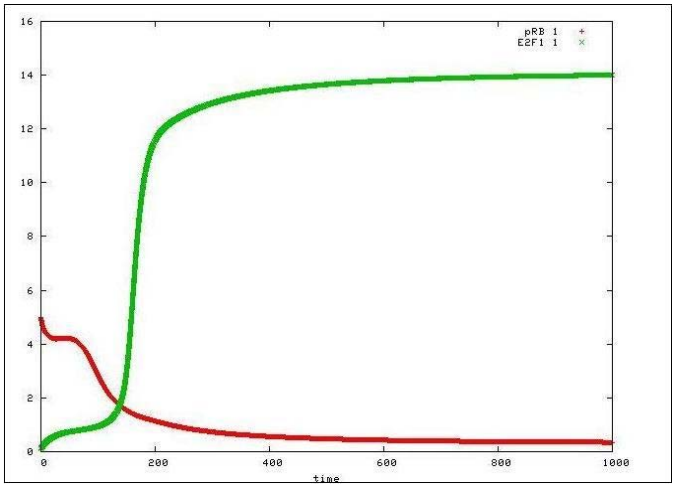


Figure 3: The results plot interface, where pRB (red curve) and E2F (green curve) time course are shown.

4. Discussion

The implementation is now capable of simulating single ODE system which describe a specific cell cycle model stored in the Cell Cycle Database. The model simulation engine is able to solve a single simulation run, but it is only the preliminary step towards the implementation of a wider system able to estimate the model which fits with real biological data the best, through a parameter estimation pipeline. The best model that fits the experimental data in a statistical sense can be found both through stochastic and deterministic mathematical methods.

In the context of stochastic approach to parameter estimation, some methods for global optimization can be considered. In particular, the evolutionary computation are population-based stochastic methods which rely on the idea of biological evolution [20]. The evolutionary computing methods generate solutions close to optimum by iteratively creating new "generations" in numerical form. Those methods are generally classified into three groups: Genetic Algorithms, Evolutionary Programming and Evolutionary Strategies, which is considered the most efficient and robust especially for continuous problems, like ODE systems resolution [21].

In the case of the deterministic approach, the estimation of the best model is possible through linear regression, that is a statistical method of modeling through a linear function the conditional expected value describing the model in function of the parameter. The most frequent linear regression method used in systems biology is the method of least-squares [22], a mathematical optimization technique which attempts to find a function which closely approximates the data in order to find the model that best fits the biological measurements.

We are implementing a system for the parameter estimation in the context of computational biology based on the grid technology. Our approach aims to find the best parameter set by computing many different simulation with the Evolutionary Strategy algorithm using the grid platform. This system essentially differs from the other grid-based parameter estimation approach [5,7] in the type of algorithm used and grid platform on which the computation is performed. Considering the ODE system describing the G1 to S transition model presented as example in the previous results [3], the simulation software XPPAUT needs approximately 4 seconds to achieve the numerical solution for 1000 time units using a Stiff integrator on a Intel Pentium 2.0 Ghz CPU with 1GB RAM. In the case of evolutionary computation to perform the parameter estimation, where a population of 300 individuals for 10000 generations is considered, the total time spent for a single simulation takes about 140 days.

To overcome the complexity of the parameter estimation process, which can be very time-consuming due to the high number of parameter combination values and simulations needed to fit data, the computation can be distributed on several computers using techniques of High Performance Computing, like grid, which makes the parameter estimation possible through the use of thousands CPUs.

Through mathematical modeling the prediction of the system behavior is possible and also unexpected properties of the system may emerge. In particular, the simulation of the cell cycle pathway allows a better understanding of cell cycle control in normal and transformed mammalian cells which is useful to put on a more rational basis the discovery of anticancer drugs.

5. Conclusion

We present a grid-oriented approach to solve ODE systems describing cell cycle models, in order to make the numerical simulations of the biological process easier and more accurate. We choose to perform simulations using a High Performance Computing platform like the grid because our system is designed with the aim to estimate the best model computing many different simulations of each model. To accomplish this task we implement a pipeline useful to visualize the mathematical information related to cell cycle models and a system to simulate the whole process using the grid platform.

Acknowledgement

This work has been supported by the Italian FIRB-MIUR projects "Italian Laboratory for Bioinformatics Technologies - LITBIO" and "ITALBIONET" and by the European "Specific Support Action BioinfoGRID", "EGEE" projects and the INTAS grant .

References

[1] K.C. Chen, L. Calzone, A. Csikasz-Nagy, F.R. Cross, B. Novak, JJ. Tyson, Integrative analysis of cell cycle control in budding yeast, *Mol Biol Cell.* **15(8)** (2004), 3841-62.
[2] Z. Qu, J. N. Weiss & W. R. MacLellan, Regulation of the mammalian cell cycle: a model of the G1-to-S transition, *Am J Physiol Cell Physiol*. **284** (2003), C349-64.
[3] M. Swat, A. Kel, H. Herzel, Bifurcation analysis of the regulatory modules of the mammalian G1/S transition, *Bioinformatics*, **20(10)** (2004), 1506-11.
[4] A. Csikasz-Nagy, D. Battogtokh, K.C. Chen, B. Novak, J.J. Tyson, Analysis of a generic model of eukaryotic cell-cycle regulation, *Biophys J*, **90** (2006), 4361-79.
[5] J.W. Zwolak, J.J. Tyson, L.T. Watson, Parameter estimation for a mathematical model of the cell cycle in frog eggs, *J. Comput Biol.*, **12(1)** (2005), 48-63.
[6] Paul T. Boggs, Richard H. Byrd, Janet Rogers Donaldson and Robert B. Schnabel (1989), ``ODRPACK -- Software for Weighted Orthogonal Distance Regression," *ACM Transactions on Mathematical Software*, Vol. 15, No. 4, pp 348-364.
[7] P.K. Dhar et al., Grid Cellware: the first grid-enabled tool for modelling and simulating cellular processes, *Bioinformatics*, **21(7)** (2005),1284-1287.
[8] T. Ray, P. Saini, Engineering design optimization using a swarm with an intelligent information sharing among individuals*, Eng. Opt*, **33** (2001), 735–748.
[9] B. Ermentrout, *Simulating, Analyzing, and Animating Dynamical Systems: A Guide to XPPAUT for Researchers and Students*, SIAM 2002, Philadelphia, USA.
[10] M. Hucka et al, The Systems Biology Markup Language (SBML): A Medium for Representation and Exchange of Biochemical Network Models, *Bioinformatics*, **19(4)** (2003), 524-531.
[11] M. Vass, N. Allen, C.A. Shaffer, N. Ramakrishnan, L.T. Watson, and J.J. Tyson, The JigCell Model Builder and Run Manager, *Bioinformatics* **20(18)** (2004), 3680-3681.
[12] *Mathematical Markup Language (MathML) Version 2.0 (Second Edition)* [http://www.w3.org/Math/]
[13] T. Ideker, D. Lauffenberger, Building with a scaffold: emerging strategies for high- to low-level cellular modelling, *TRENDS in Biotechnology*, **21(6)** (2003),255-62.
[14] *Extensible Stylesheet Language (XSL) Version 1.1*. Edited by Anders Berglund (IBM). [http://www.w3.org/Style/XSL/].
[15] *XSL Transformations (XSLT) Version 1.0* Edited by James Clark [http://www.w3.org/TR/xslt].
[16] *XSLT MathML Library Version 2.0* Edited by Vasil I. Yaroshevich, [http://xsltml.sourceforge.net/].
[17] I. Hutchinson, *TTH: a "TEX to HTML" translator* [http://hutchinson.belmont.ma.us/tth/].
[18] T. Williams, C. Kelley, GNUPLOT: An Interactive Plotting Program, Version 3.7 organized by: David Denholm.

[19] Grivell L, E-BioSci: Semantic networks of biological information**.** *Information Services & Use* **23** (2003),179–182. [http://www.e-BioSci.org/index.html].
[20] D.B. Fogel, Evolutionary computation: Toward a new philosophy of\machine intelligence. *IEEE Press*, New York, 2000.
[21] N. Saravanan, D.B. Fogel, K.M. Nelson, A comparison of methods for self-adaptation in evolutionary algorithms, *Biosystems*, **36** (1995), 157–166.
[22] J. Cohen, P. Cohen, S.G. West & L.S. Aiken, *Applied multiple regression/correlation analysis for the behavioral sciences. (2nd ed.)* Hillsdale, NJ: Lawrence Erlbaum Associates, 2003.

III. State of the Art of the Grid Research and Use at Organ Level

This page intentionally left blank

From Genes to Personalized HealthCare: Grid Solutions for the Life Sciences
N. Jacq et al. (Eds.)
IOS Press, 2007

Monte Carlo Verification of IMRT treatment plans on Grid

Andrés GÓMEZ[a], Carlos FERNÁNDEZ SÁNCHEZ[a], José Carlos MOURIÑO GALLEGO[a], Javier LÓPEZ CACHEIRO[a], Francisco J. GONZÁLEZ CASTAÑO[b], Daniel RODRÍGUEZ-SILVA[b], Lorena DOMÍNGUEZ CARRERA[b], David GONZÁLEZ MARTÍNEZ[b], Javier PENA GARCÍA[c], Faustino GÓMEZ RODRÍGUEZ[c], Diego GONZÁLEZ CASTAÑO[c], Miguel POMBAR CAMEÁN[d]

[a]*Fundación Centro Tecnológico de Supercomputación de Galicia (CESGA), Santiago de Compostela, Spain*
{agomez,carlosf,jmourino,jlopez}@cesga.es

[b]*Departamento de Ingeniería Telemática, University of Vigo, Spain*
{javier,darguez}@det.uvigo.es

[c]*Departamento de Física de Partículas, University of Santiago de Compostela, Spain*
{javierpg,faustgr}@usc.es

[d]*Hospital Clínico Universitario de Santiago, Santiago de Compostela, Spain*
mrpombar@usc.es

Abstract: The eIMRT project is producing new remote computational tools for helping radiotherapists to plan and deliver treatments. The first available tool will be the IMRT treatment verification using Monte Carlo, which is a computational expensive problem that can be executed remotely on a GRID. In this paper, the current implementation of this process using GRID and SOA technologies is presented, describing the remote execution environment and the client.

Keywords: Radiotherapy, Monte Carlo, treatment plan verification, IMRT, EGEE, GRID, gLite.

1. Introduction

Intensity Modulated Radiation Therapy (IMRT) is a state-of-the-art technique in radiation therapy that allows the delivery of a non-uniform photon fluence for each incident angle of the X-ray beam generated by a medical linear accelerator (linac). It presents clinical advantages over conformal radiation techniques (CRT), which only adjust the shape of the radiation beam to the shape of the tumour. Usually, the calculation of the directions of incidence and the shape of the radiation fields that have to be delivered to build up the desired dose distribution (treatment planning) is done using local software tools called treatment planning systems (TPS) such as Pinnacle,

XiO, Oncentra, Corvus, etc., running on workstations at hospital premises. Specialized personnel (radiotherapists) compute treatment plans either employing their previous knowledge and experience, trial-and-error class-solutions or, for more complex treatment plans, built-in optimization tools. Treatments are tailored to deliver uniform doses to the planned target volumes (PTVs) while keeping the dose to surrounding tissues, especially to the organs at risk (OARs), within the prescribed tolerances. In both cases the goals to be achieved are specified by the radiation oncologists. There are two common methods to deliver an IMRT treatment in computer controlled linacs: step-and-shoot, where the leaves of the multileaf collimator (MLC) are moved in discrete steps between two consecutive irradiations, and dynamic-MLC, where the leaves are moved continuously during irradiation.

The requirements in maximum computation time force TPS tools to perform approximations both in dose calculation engines and optimization algorithms. The most accurate dose calculation techniques included in those codes are based on convolution/superposition methods (C/S) [1], which suffer from certain limitations in high density gradient regions. Due to the complexity of the IMRT plan and the compulsory approximations during dose optimization, each IMRT treatment has to be experimentally verified prior to its actual delivery to the patient, involving the final accelerator and dose measurement units. These in-phantom expensive measurements require large amounts of time and could be avoided or minimized by employing Monte Carlo techniques [2] to simulate the treatment *in silico*.

Nowadays, in developed countries, more than 40,000 people per million inhabitants have been diagnosed a cancer [3] yearly and approximately 50% of them receive radiotherapy. In 1999, more than 56,000 patients were irradiated only in Spain [4]. All treatments follow a planning protocol to ensure the quality and effectiveness of the session, and the treatments should be planned in a short period of time (the mean time between the first visit and the begining of radiotherapy treatment is 18,87 days in Spanish public hospitals [4]). Many radiotherapists have to plan over 600 to 1,200 patients per year, with a mean value of 925 [5]. This situation puts a high pressure on them, raising the need of new optimization tools. There are over 10,000 accelerators worldwide that irradiate around 4 millions patients yearly [3]. Only in Spain, there are over 115 particle accelerators in 70 hospitals with radiotherapy facilities [6], four of them in Galicia including the *Complexo Hospitalario Universitario de Santiago*, which is a partner in the eIMRT [7] project.

As a result of the extremely long CPU time and the large amount of plans to calculate, Monte Carlo verification is a clear best-case of GRID technologies exploitation. This is the aim of the eIMRT project which is devoted to produce new tools based on computational intensive algorithms for helping the radiotherapist to plan and verify the radiotherapy treatments. One of these new tools is the verification of planned treatments using Monte Carlo methods, which is presented in this paper. We hope that this tool has a significant impact due to the high number of hospitals that may benefit from it.

The paper is divided in four sections. First of all, a brief description of eIMRT architecture is given. The next section presents the verification process. The third section describes the computational implementation on GRID and discusses the associated difficulties. The paper ends with a final section dedicated to future work and conclusions.

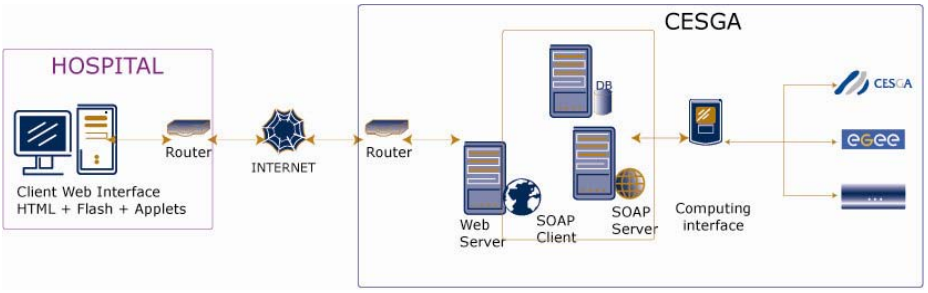


Figure 1. High-level eIMRT architecture

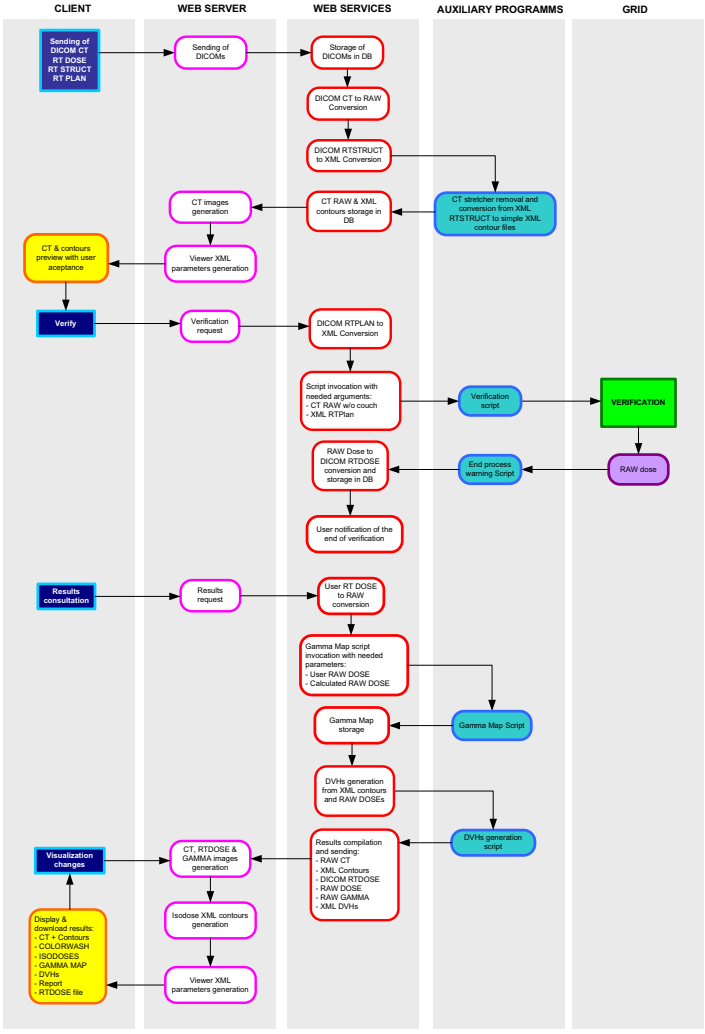


Figure 2: Schematic flow chart of the verification process**.**

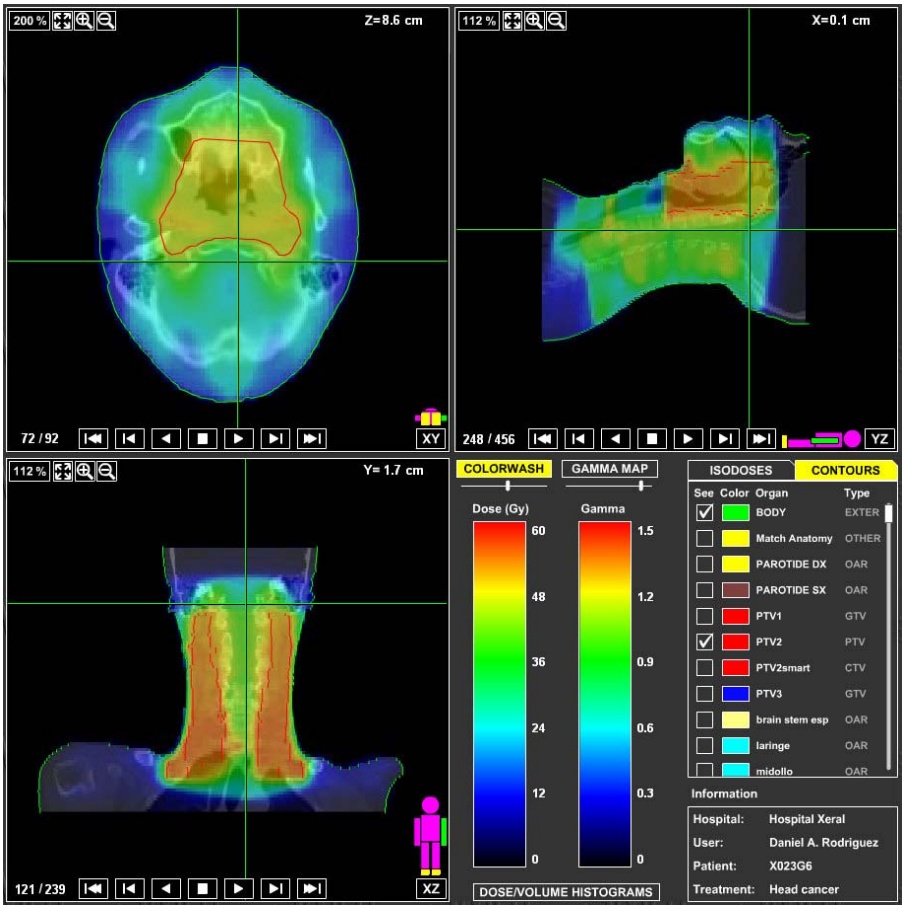


Figure 3. User visualization interface: It shows the slices of the patient in three directions (XY, YZ and XZ), presenting information about the calculated doses, the areas to consider and the comparison between the reference dose (calculated by Monte Carlo) and the treatment dose (calculated by the TPS) using gamma maps.

2. eIMRT Architecture

The eIMRT architecture has been described in a previous paper [7]. Figure 1 shows a general overview. It comprises four layers: client, application server, computing interface/data server and computing elements. The client has been developed for demonstration purposes because all the system is designed following the SOA (Service Oriented Architecture) paradigm. It is divided in two layers: a web server based on Coocon [8], which calls the web services and transforms SOAP messages to HTML, and an Internet navigator. Also, there are two special plugins, one based on Java to upload the DICOM files and anonymize them, and another one based on Flash for visualization of results (see Figure 3 for a view of the visualization interface). Note that the client interface is rather simple. Complexity is completely enclosed at the server side, which accesses high-throughput computing via web services [9]. Due to the open architecture, we can use several computing services. In fact, we have implemented the

computing interface for using the local cluster facilities through a queue system (Sun Grid Engine) or for submitting the jobs to the GRID, in this case using the gLite middleware [10]. Nowadays, the authentication with the GRID infrastructure is done using a single certificate for all the tasks, but in the near future personalized authentication based on certification will be implemented.

Currently, the most important web services are:

- UserManagement, which manages all the information related to user and control sessions.
- FileManagement, which makes all the operations for uploading and controlling the files related to the treatments as DICOM CT, DICOM RTPlan, etc.
- TreatmentManagement, for managing the information and operations related to a treatment.
- Verification, which submits and controls the operations related to the verification of a treatment.
- MapManagement, for generating different maps to compare two dose distributions. This web server provides an open interface for different types of maps. Currently the gamma maps [11] are implemented, but other maps can be supported.
- Monitorization, which allows the monitorization of the status of a computational operation, such as verification, and alerts the final user when it ends.

In the near future, other services will be implemented for the optimization of treatments or the characterization of accelerators (currently an expensive off-line process).

3. Verification of IMRT Treatment Plans

To **validate a treatment**, the end user employs the system to check the dose distribution that has been calculated at the hospital (for instance, with a TPS) against the dose distribution associated to the same treatment resulting from a more accurate dose calculation method (Monte Carlo at the current stage). A schematic representation of the full process is shown in figure 2. The output of both methods is compared, for example using the gamma maps generated by the MapManagement service. The eIMRT project has implemented a Monte Carlo verification process based on the well known and validated BEAMnrc package [12], and comprises five phases (see figure 4 for a schematic view):

- **Phase 1: Accelerator simulation**. It takes the data describing the geometry of the linear accelerator, its radiation source and the treatment to be verified. Basically it takes the information from the input DICOM RTplan of the treatment and produces the input files in the right format for the BEAMnrc Monte Carlo code. There is a single input file for the accelerator head simulation (from the bremsstrahlung target to the bottom of the collimators) for each of the following:
  a) Field in CRT treatments. In Conformal Radiation Therapy the shape of the radiation field is adjusted to fit the profile of the tumour according to the beam's eye view.
  b) Segment in IMRT step-and-shoot treatments.
  c) Control point in IMRT dynamic-MLC treatments.

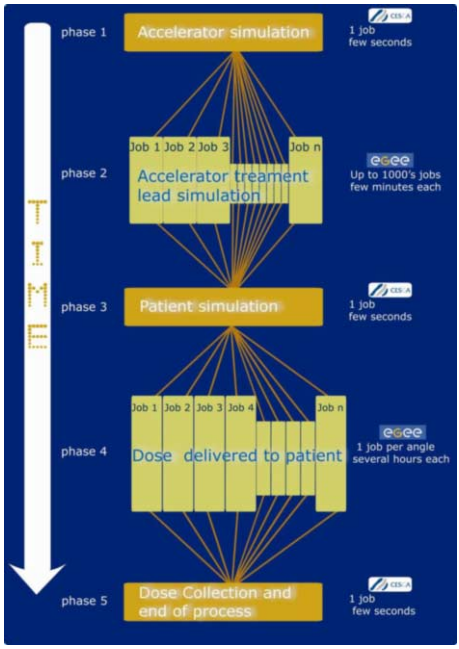


Figure 4. Schematic view of the Monte Carlo verification module. The output of phase 5 can be compared with the input dose distributions using,for example, the gamma maps generated by the MapManagement service.

Due to the reduced CPU time, this step can be executed locally at the server side. It produces from few files for CRT treatments (about one per angle) to several thousands in case of IMRT.

- **Phase 2: Accelerator treatment head simulation**. The whole accelerator treatment head is simulated for each input file, optimizing the variance reduction techniques to maximize the particle production and scoring for that field shape. This step executes the BEAMnrc code for each input file. It needs about 20 hours of CPU (in a Pentium IV at 3.0GHz) for each treatment. It generates one output file for each input.
- **Phase 3: Patient simulation.** The particles associated to each field shape calculated in the previous step are collected and grouped together attending to the energy of the beam and gantry, table and collimator angles. Consistency checks are also performed in order to ensure that no particles were lost during the simulations of the previous step. CT data describing a patient is converted to densities for further calculation of the dose delivered by each of the fields. It takes into account the data of the characterization of the computerized tomography (CT). The output of this phase produces the input files for simulating the dose deposition inside the patient using also Monte Carlo.
- **Phase 4: Dose delivered to the patient.** The dose-inside-the-patient is calculated using the DOSXYZnrc Monte Carlo code [12]. Since this task is highly parallelizable, it can be divided in many different independent jobs. Currently, the

calculation is divided in one job for each incident angle, although divisions with finer granularity are possible. It needs about 35 hours in a single CPU.

- **Phase 5: Dose collection and end of process.** The results are merged into a final, single dose distribution. The dose distribution normalized per unit of primary fluence (i.e. the amount of radiation that reaches the target of the accelerator where the electrons collide to produce photons) is converted into an absolute or relative dose distribution by comparison with TPS results (i.e. the actual dose that the accelerator delivers).

Once the Monte Carlo process has finished, the radiotherapist can manually compare his independently calculated dose with the dose calculated by the Monte Carlo verification. For this task, we have developed a special service that produces gamma maps, taking the dose maps as input. This task only needs a few seconds of CPUs and the result can be graphically displayed on the client (see figure 3). No single value can be produced to assess the quality of the treatment plan, which is a decision to be taken by the radiotherapist.

4. GRID Execution

The previously described Monte Carlo verification process requires a high computational capacity. For each treatment verification, the execution of hundreds of short jobs is necessary in phase 2 and in phase 4, only few jobs (usually less than 10) run, but each one consumes several hours. So, adding all phases, tenths of hours of CPU are consumed in each treatment verification. Also, it may be the case that a treatment has to be verified several times before achieving a good solution. Therefore, a large amount of computational resources is necessary to reduce the time-to-solution. Since not all the treatments have to be verified, the infrastructures of different hospitals or a GRID infrastructure can be shared to fulfil the expectations of the radiotherapists and get a solution within a proper time frame for clinical purposes.

Each one of the described phases of the treatment verification process has different computational necessities. Consequently, they will be executed in different contexts. They will be described in detail from the computational point of view.

Phase 1, which simulates the accelerator, is sequential and requires a short execution time, so can be executed locally. It needs a single input file that specifies the files that are necessary for the execution of the whole process, like DICOM files of the patient, the files that describe the accelerator, as well as other files related to the process itself that are independent from the treatments. All the necessary files for phase 2 are produced as the output, one for each process, as well as a text file with as many rows as Monte Carlo processes to be executed and three columns: one indicating the input file, another one with the EGSnrc executable to be used for this job and a third one that indicates the file with the data of the accelerator. This intermediate command file is produced to make the verification process independent of the final infrastructure. So, it can be used as the input for different types of infrastructures such as GRID, clusters or any new type of middleware. Even better, in the future we can mix several infrastructures for a single treatment, executing the final jobs in the GRID and in a local cluster simultaneously.

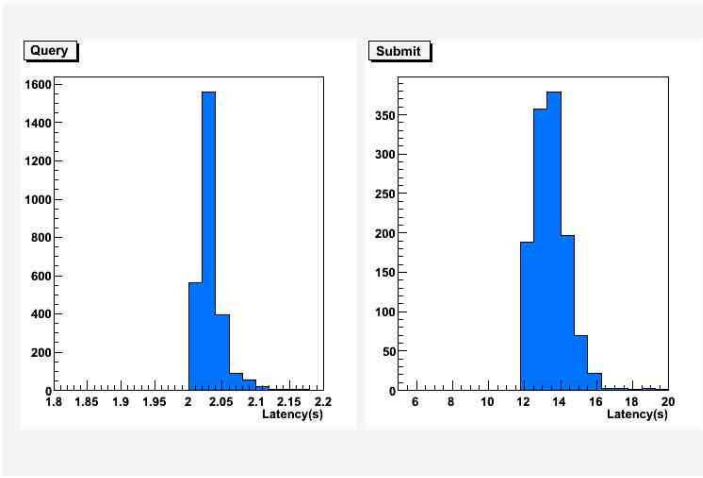


Figure 5: Measured latencies (in seconds) for query (left) and submit (right) operations in EGEE. The mean time for query the status of a job is 2.04 ±0.02 with a minimum latency of 2 seconds and for submitting a single job is 13.46 ± 0.94 seconds.

After this first pre-processing stage, all jobs of phase 2 have to be executed. The number of jobs of this phase depends on the treatment and the accelerator, but it is of the order of several hundreds (more than 1,200 in some cases). Nevertheless, its individual run time is low, in the range of a couple of minutes. Since all of them are independent, these jobs can be sent to the GRID, taking advantage of the many resources available. A JDL file is created for each process, indicating the executable and the necessary input files. The files that depend on the treatment are sent in the InputSandBox, whereas those that are common to all processes (as the executables) are copied from the Storage Element at the beginning of the job. All the outputs are recovered using the OutputSandBox. Due to the current undesirable high latency for sending jobs to the EGEE infrastructure (see figure 5), if compared to the real execution time, the tasks are grouped in a few jobs taking into account this latency and the CPU time, in order to optimize the final elapsed time. Once they are sent to the GRID, the status of all the jobs is continuously monitored to produce useful information for the user (who can ask for the status of the verification process at any time) and to resubmit failed jobs due to infrastructure failures. Again, the high latency of the EGEE infrastructure affects the monitoring process, yielding it infeasible when the jobs are individually submitted. By grouping them, this task becomes feasible because the number of independent jobs is limited.

Once phase 2 has finished satisfactorily, phase 3 gets all the output data of the accelerator simulation and produces the input files for the patient simulation, using the same input file of phase 1 and the results of phase 2. It produces the input files for the next phase and a command file describing the tasks in the next phase. It is a sequential job with reduced elapsed time that can be executed locally in the server.

Phase 4 is similar to the second one. It requires less jobs (initially, one per incident angle, i.e., between 3 and 7 in usual treatment plans, although each job can be parallelized), but each one requires several CPU hours. They are sent to the GRID following the same procedure as in phase 2, but, given the characteristics of the jobs to execute, in this case there is no grouping. Monitoring and fault-tolerant processes are

also executed. This phase is the most computationally expensive, with a long elapsed time. Since each job can be trivially parallelized, we expect to reduce the final elapsed time of this task in the future. This is an on-going work because we want to use the same cluster for each single job in order to reduce the time for uploading and downloading files, and it is a feature only recently available in EGEE.

Finally, in phase 5 all the output files are downloaded from the GRID, post-processed and merged in a single dose file of a few MBs, sending an alerting message to the final user. This is a sequential short task that resides in the server.

5. Conclusions and Future Work

The decoupled eIMRT architecture is a cost-effective solution to speed-up the CPU-greedy processes in advanced radiotherapy planning: accelerator characterization, treatment validation and treatment optimization. As far as we know, there are no similar distributed environments for verification and optimization of radiotherapy treatments, although other desktop tools as DoseLab [13] for dose comparison or CEER [14] are available. In this paper, an implementation of the IMRT verification on the GRID has been presented. The two most computationally demanding steps have been implemented to be executed on a GRID or a distributed environment, because they are well suited for this kind of architectures due to their highly parallel nature. They have been tested on the EGEE infrastructure. We have identified some problems with submission and monitoring due to the high latencies of both tasks in the actual infrastructure. Although the total CPU time is not very high (about 100 hours for each treatment), a short response time is needed to fulfil the expectations of the radiotherapists, who need to check the treatment as soon as possible. Therefore, the problems we have studied represent a good best-case for an interactive GRID environment, as proposed by the Interactive European Grid Project [15].

The ongoing project will produce new tools and improvements in the near future. The next step is the inclusion of the treatment optimization process and the characterization of any accelerator (So far only previously characterized accelerators are allowed). From the point of view of GRID technologies, the full system should improve security, including end-to-end authorization based on certificates, the full support of GRID distributed storage or the access to DICOM files directly from the hospitals without having to upload them. Also, we plan to use the Workload Management Server [10] for submitting jobs, included in the last production version of gLite. It allows bulk submission and shared sandboxes, which will be very helpful for phase 2 and 4.

In the near future, a full Monte Carlo TPS system will be available. It will benefit both from the improvements in Monte Carlo simulation and the increasing multi-core CPU power in the workstations. However, due to the large number of treatments to be planned in each hospital, we believe that a distributed GRID environment based on the SOA paradigm will still be useful, and it will easily provide access to computer power, new functionalities and algorithms from desktop computers.

Acknowledgments

This research has been funded by the PGDIT05SIN00101CT grant (*Xunta de Galicia*, Spain), and partially by FSE. We also appreciate the collaboration of the University of Wisconsin and University of Maryland, mainly from Professor Robert Meyer.

This work makes use of results produced by the Enabling Grids for E-sciencE project, a project co-funded by the European Commission (under contract number INFSO-RI-031688) through the Sixth Framework Programme. EGEE brings together 91 partners in 32 countries to provide a seamless GRID infrastructure available to the European research community 24 hours a day. Full information is available at http://www.eu-egee.org.

References

[1] T.R. Mackie, J.W. Scrimger, and J.J. Batista, "A convolution method of calculating do for 15-MV x rays", Med. Phys. 12, 188-196 (1985)
[2] P. Andreo, "Monte Carlo techniques in medical radiation physics", Phys. Med. Biol. 36, No 7, 861-920 (1991)
[3] U. Amaldi, G. Kraft, "Particle accelerators take up the fight against cancer", CERN Courier, Volume 46, No 10, pp. 17-20 (2006)
[4] G. López-Abente Ortega et.al, "La situación del cáncer en España". Ministerio de Sanidad y Consumo. 2005. ISBN: 84-7670-673-1
[5] A. Iglesias Lago, *Planificadores 3D y simulación virtual del tratamiento. Situación en España. Supervivencia asociada a su aplicación.* Santiago de Compostela: Servicio Galego de Saúde, Axencia de Avaliación de Tecnologías Sanitarias de Galicia, avalia-t; 2003. Serie Avaliación de Tecnoloxías. Investigación Avaliativa: IA2003/01
[6] Catálogo Nacional de Hospitales 2005, Ministerio de Sanidad y Consumo, http://www.msc.es/ciudadanos/prestaciones/centrosServiciosSNS/hospitales/home.htm
[7] A. Gómez, et. al., "Remote Radiotherapy Planning: The eIMRT Project", in Challenges and Opportunities of Health Grids, V. Hernández and others, eds., IOS Press, 2006
[8] Web Services, http://www.w3.org/2002/ws/
[9] The Apache Cocoon Project, http://cocoon.apache.org
[10] gLite, http://glite.web.cern.ch/glite/
[11] D. A. Low, W. B. Harms, S. Mutic, J. A. Purdy, "A technique for the quantitative evaluation of dose distributions", Med. Phys. 25, 656-661 (1998)
[12] D. W. O. Rogers, B. A. Faddegon, G. X. Ding, C.-M. Ma, J. We and T. R. Mackie, "BEAM: a Monte Carlo code to simulate radiotherapy treatment units", Med. Phys. 22, 503-524 (1995).
[13] DoseLab, http://doselab.sourceforge.net/index.html
[14] J. Deasy, A. Blanco and V. Clark, "CEER: a computational environment for radiotherapy research", Med Phys. 30, 979-985 (2004).
[15] http://www.interactive-Grid.eu/

From Genes to Personalized HealthCare: Grid Solutions for the Life Sciences
N. Jacq et al. (Eds.)
IOS Press, 2007

TRENCADIS – Secure Architecture to Share and Manage DICOM Objects in a Ontological Framework Based on OGSA

Ignacio BLANQUER[1], Vicente HERNANDEZ[1], Damià SEGRELLES[1], Erik TORRES[1]
[1]*Instituto de Aplicaciones de las Tecnologías de la Información y de las Comunicaciones Avanzadas, Universidad Politécnica de Valencia, Camino de Vera S/N, 46022 Valencia Spain (phone: +34 96 387 7007 ext.88254; Fax: +34 96 387 7274; e-mails: iblanque@dsic.upv.es; vhernand@dsic.upv.es; dquilis@itaca.upv.es; etorres@itaca.upv.es)*

Abstract. Today most European healthcare centers use the digital format for their databases of images. TRENCADIS is a software architecture comprising a set of services as a solution for interconnecting, managing and sharing selected parts of medical DICOM data for the development of training and decision support tools. The organization of the distributed information in virtual repositories is based on semantic criteria. Different groups of researchers could organize themselves to propose a Virtual Organization (VO). These VOs will be interested in specific target areas, and will share information concerning each area. Although the private part of the information to be shared will be removed, special considerations will be taken into account to avoid the access by non-authorized users. This paper describes the security model implemented as part of TRENCADIS. The paper is organized as follows. First introduces the problem and presents our motivations. Section 1 defines the objectives. Section 2 presents an overview of the existing proposals per objective. Section 3 outlines the overall architecture. Section 4 describes how TRENCADIS is architected to realize the security goals discussed in the previous sections. The different security services and components of the infrastructure are briefly explained, as well as the exposed interfaces. Finally, Section 5 concludes and gives some remarks on our future work.

Keywords. Grid, Security, DICOM

Introduction

The use of digital medical images in hospital environments has changed the way in which radiologists work and cooperate. The generalization of Digital Imaging and Communications in Medicine [1] (DICOM), as a world-wide standard for the transmission and exchange of medical images, has made it possible to share images across a wide set of users and applications. Today most European healthcare centers use the digital format for their databases of images (PACS, Hard Disk Shared Directories, etc…). These databases can be located at different physical sites into a single medical centre or even also in different centers. Therefore, this diversity in the source of images, the distributed location of the storages and managing systems difficulties the transparent sharing of images and the development of secure

collaborative environments, where new applications of Medical Imaging (Advanced Image-based Diagnosis, Non-Affine Registration/Fusion Applications, Context-Based searching of DICOM images etc...) can be built.

On the other hand, DICOM does not only imply images coming from the radiology departments, other type of images and movies (dermatology, endoscopes, etc.) and other type of information such as radiology reports are being coded into DICOM. DICOM Structured Reporting [2] (DICOM-SR) codes and integrates radiology reports with seamless references to findings and Regions of Interests on the associated images. Structuring radiology reports offers a comparable way to code reports enhancing the capability of tools to search and to extract knowledge.

To tackle this challenge, a software architecture, namely TRENCADIS[1] [3] [4] was developed and implemented, comprising also a set of services as a solution for interconnecting, managing and sharing selected parts of medical DICOM data for the development of training and decision support tools. The organization of the distributed information in virtual repositories is based on semantic criteria. Different groups of researchers could organize themselves to propose a Virtual Organization (VO). These VOs will be interested in specific target areas (e.g. pediatric oncology), and will share information (studies and reports) concerning each area. Subsets of those images could be obtained for a specific study (e.g. neuroblastoma). Finally, in each subset, users can make complex queries (e.g. male patients above one year with irregular findings of more than 3 mm). Of course, the sharing of the information must be secure and within the VO. Although the private part of the information to be shared will be removed, special considerations will be taken into account to avoid the access by non-authorized users (even those with administrative privileges at remote sites). It is important to notice that data security is a key requirement for biomedical grid applications, since not only the obvious technical requisite of ensuring integrity and validity of computations must be guaranteed, but also the necessity of being liable to heterogeneous national legal regulations and developing procedures to be accepted by the medical community [5]. Once the architectural design [6], the ontological framework [4] and the management of DICOM SR [3] have been completed, this paper describes the security model implemented as part of the architecture proposed in [3] [4].

1. Objectives

This paper aims to describe and discuss the security model implemented as part of TRENCADIS. The security services of grid are not altogether different from those of other distributed system paradigms. Specifically, an effective security model must ensure a set of security primitives: Identity verification (authentication), authorization, access control, data integrity, data confidentiality and availability.

TRENCADIS architecture combines the security requirements in three main objectives:

- Secure authentication and communication.
- Management of VO security policies.
- Confidentiality and privacy control of medical data.

[1] TRENCADIS: Towards a GRid ENvironment to proCess and shAre DIcom objectS. TRENCADIS means “mosaic” on the language of Valencia Region.

In the next points the TRENCADIS security model is described and discussed. First, a general view of the TRENCADIS architecture is presented in which the security model is applied to accomplish the three security aspects proposed. After that, the security model and the components required in the deployment of the architecture are covered.

2. Related Work

2.1. Secure authentication and communication

Almost all Grid components and Grid Middlewares use the Grid Security Infrastructure (GSI) for authentication [7]. GSI also provides mechanisms for secure communication. These mechanisms deal with the security related aspects of connection establishment as well as message protection. Currently, the message protection could be enforced at two different levels: integrity, guaranteed by the message signature, and privacy, guaranteed by the message encryption and signature.

2.2. Management of VO Security Policies

This objective refers to the need of authorization and access control policy mechanisms at VO level.

The TRENCADIS middleware must include a fine-grain security access control for Grid services and resources. The system must consider the agreements reached between the VO and the different organizations which comprise the VO for the management of the shared computing resources.

Virtual Organization Membership Service (VOMS) has long been proposed as a solution to this problem [8]. VOMS utilizes an extended X.509 certificate specification for defining extra attributes. VOMS defines groups, roles and capabilities. Combinations of the names of these serve as attributes for users [9].

2.3. Confidentiality and Privacy Control of Medical Data

As medical data may be stored for years, long-term storage of encrypted data seems to be a reasonable approach for protecting confidential data from being exposed. Different organizations (virtual or not) need to combine their efforts in a collective purpose of controlling data privacy.

The combination of the data controllers in a scheme for managing decryption keys protects data from disclosure [10].

Different administrative domains agreed to contribute with the collective purpose of protect data from being exposed by users granted with physical or administrator access. The complexity of disclosing protected objects requires compromising the security of a certain number of services deployed by completely different administrative domains.

A previous work have introduced the idea of using VOs as a natural way of define a key sharing schema in a Grid environment, considering each VO as a different administrative domain with the responsibility of guard key shares and object owners being enabled to decide the trusted VOs they will use for key sharing [11].

3. TRENCADIS Architecture

The architecture defined in TRENCADIS project is horizontal and can deal with in different type of objects of the DICOM standard (medical images, structured reports, waveforms, etc…).

In particular, this paper focuses on the Security model applied above the architecture. This section briefly explains the different layers defined in the TRENCADIS architecture and the features associated to it. TRENCADIS is a Service-Oriented Architecture (SOA) in which the usage of resources is represented as Grid services.

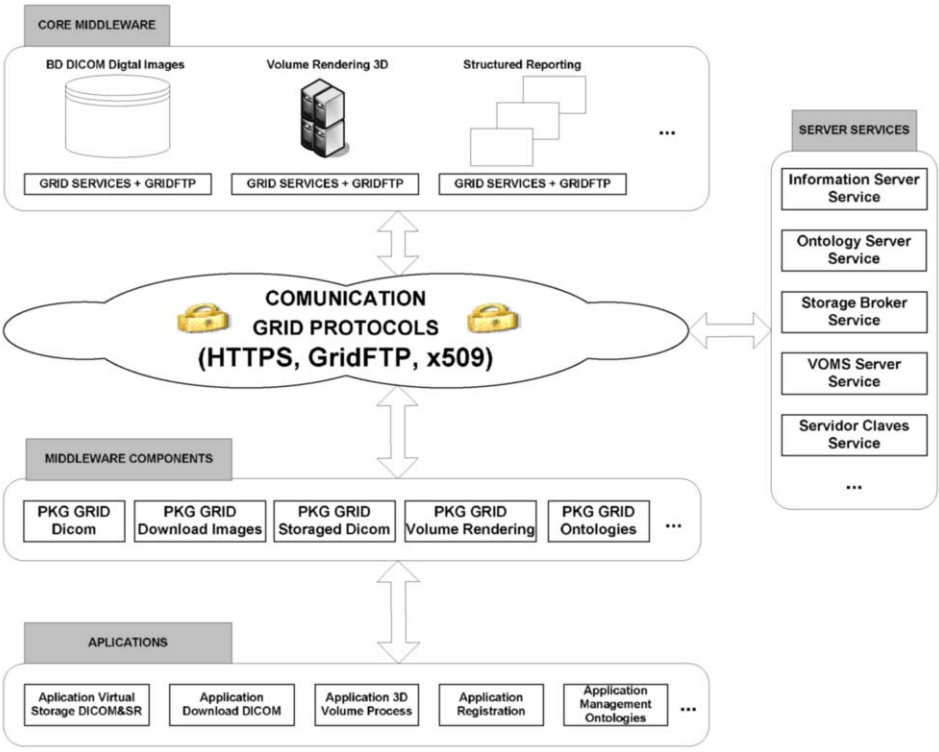


Figure 1. General Scheme of TRENCADIS Architecture.

As described in Figure 1, TRENCADIS comprises five layers:

- *Core Middleware Layer*. This layer provides with the basic resources (Databases, High-Performance Computers, etc…) of the environment, which are offered as services using well defined and standard interfaces in Web Services Definition Language (WSDL), using protocols such as Simple Object Access Protocol (SOAP) or GridFTP and data formats like XML Schemes. It provides upper layers with a unique interface to all resources of the same type. For example, a DICOM storage of structured reports can be implemented using relational databases or a plain directory in a hard disk, but the interaction interface will be the same. The Services of this layer are part of the infrastructure.

- *Server Services Layer*. This layer defines the services that implement server tasks, such as the location of resources. This layer interacts directly with the services from Core Middleware and Server Services Layer and with the upper-level components. The Services of this layer are part of the infrastructure.
- *Communication Layer*. It defines the protocols that will be used by the services that have been implemented in the lowest layers (Core Middleware and Server Services). GridFTP protocol is used for transferring a large amount of data, and SOAP above HTTPS is used for interacting with the services.
- *Components Middleware Layer*. This layer contains the highest components that interact with services of the Core Middleware and the Server Services Layer. These components provide the applications with an object-oriented interface for the development of the applications for managing, processing and sharing DICOM objects in general.
- *Applications Layer*. This layer comprises the applications for managing, processing and sharing DICOM objects.

4. Infrastructure of Security in the TRENCADIS Architecture

This section describes how TRENCADIS is architected to realize the security goals discussed in the previous sections. The different security services and components of the infrastructure are briefly explained, as well as the exposed interfaces.

All the services presented in this Section were implemented and deployed as part of the *Infrastructure Layer*, a virtual layer that groups the *Core Middleware* and the *Server Services* layers.

4.1. GSI Services

TRENCADIS uses the GSI for enabling secure authentication and communication over an open network. Since the core of GSI is the use of X.509 certificates, an international standard for public key infrastructure (PKI), a set of Grid services is required to manage the issuing and revocation of PKI credentials for services and users.

GSI minimally requires a basic PKI certificate management infrastructure (request, registration and revocation). TRENCADIS deploys the Certificates Management Service in the Server Service Layer, that serve as entry point for new PKI certificates requests managed by one or more Certificate Authority (CA). It is possible to use any of the existent CA implementations (OpenCA, Microsoft Certificate Services, etc).

The Certificates Management Service handles new certificate requests for users and resources which want to be integrated in a specific Grid infrastructure. Registration and revocation decisions must be managed by the authorized operator of the administrative domain of the CA.

The service exposes two different interfaces: a client interface and an administration interface available only for the service administrator.

The client interface exposes the following operations:

- **SendRequest:** It sends the users and resources new certificates requests to the CA.
- **RetrieveCertificate:** It retrieves the certificates issued by the CA.

The administration interface exposes the following operations:

- **SignRequest:** It signs the new certificates requests.
- **RevokeCertificate:** It revokes certificates.

The Communication Layer relies on the Globus Toolkit to provide the client and the server SOAP/HTTPS and GridFTP protocols.

4.2. VOMS Services

Different real domains could be managed as a single administrative domain in TRENCADIS. The architecture supports VO management throughout VOMS.

The VOMS Service is implemented in the Core Middleware and in the Server Service Layers. The components of the service are very similar in both layers. The main difference between the components implemented in each layer is that the Core Middleware Layer requires an additional component to facilitate the interaction with devices.

The VOMS Service manages user memberships, roles and capabilities in one or more VOs. The service is exposed in two different interfaces: a client interface and an administration interface available only for the service administrator. Both interfaces enable a Web client application for the service.

The administration interface exposes the following operations:

- **CreateVO/DeleteVO:** It creates/deletes a VO.
- **CreateUser/DeleteUser:** It creates/deletes a user in the VO.
- **CreateGroup/DeleteGroup:** It creates/deletes a group in the VO.
- **AssignGroup/ResignGroup:** It registers/deregisters a user in a group.
- **CreateRole/DeleteRole:** It creates/deletes a role in the VO.
- **AssignRole/DismissRole:** It assigns/dismisses a role to/from a group.

The client interface exposes the operation RequestProxy that returns a VOMS proxy certificate for a user identified by a user certificate.

The Grid services deployed in the layers Core Middleware and Server Service require a component which reads the attributes list from the user credentials and evaluates a local access control policy returning a decision. The decision could be to permit or to deny the execution of a request submitted to the service. This component is named the Gatekeeper.

In a typical TRENCADIS usage scenario, a subject (e.g. user, service) wants to take some action on a particular virtual storage DICOM. The subject submits its query to the Storage DICOM Service protecting the resource, which examines the request, retrieves the ontologies that are applicable to this request and the subject's VOMS attributes, and contacts a Gatekeeper component that determines whether access should be granted according to the policy rules for evaluating VO groups' access to ontologies.

In this context, technologies for access control and enforcement policies (e.g. SAML, XACML) could be used for exchanging authentication and authorization data between security domains. However, this version of TRENCADIS lacks this functionality for simplicity. Future releases of the architecture were expected to be compliant with some standard access control policy language.

Besides to evaluate decision chains, the Gatekeeper implements a set of operations in the Core Middleware Layer. The Gatekeeper automatically registers authorized users in the grid-mapfile when needed (for example for the GridFTP protocol) and delete it when the requested operation ends.

The Gatekeeper components in the Server Service Layer could require further operations with the purpose of allowing alternative authorization methods.

4.3. Privacy Services

Often, production Grids comprises a few VOs. In real applications, Grids deployed by a single VO are more commonly found that one can expect. In such scenarios, it is not possible to share keys over VOs with a reasonable security level because there are no enough key stores available in different administrative domains.

The VO-based definition of administrative domain limits the generalization of the model. Therefore, considering the experiences acquired by the HealthGrid community throughout the development of a middleware Grid for managing DICOM objects, the concept of the administrative domain is updated in this work, preserving the rest of the architecture.

Parties involved in the deployment of Grid services for storing DICOM objects fits two different categories: organizations feeding the Grid with their own data, and organizations using the data provided by others. Data providers form the group of organizations interested in guaranteeing the confidentiality and privacy of the data. Consequently, administrative domains should be defined as individual organizations (virtual or not) which contribute with their own data to the Grid and which are identified as data controllers with the responsibility of protecting data from unauthorized use.

In practice, different administrative domains can be identified using PKI. Each organization (virtual or not) participating in the sharing scheme must contribute with a CA to the Grid, and must deploy their own key sharing Grid services exposing credentials signed by their own CA. The objects owners' will be enabled to discover administrative domains, properly identified, and to distribute the decryption keys over trusted participants.

The information required to identify the administrative domains that could be combined to retrieve the decryption key is stored in the header of the encrypted object. However, this information is not enough to rebuild the key. Besides, each administrative domain needs to keep track of the object identifiers and its corresponding key share.

The object identifier should provide the ubiquity property to the encrypted object. That is the capability of been physically present in more than one operation or storage element without changing its identity.

A 128-bits Encrypted Object Unique Identifier (EOUID) is randomly generated by a Grid service and assigned to each new encrypted object stored in the grid. The EOUID could be retrieved with a cost-less operation from the header of the encrypted object. All key sharing services references use this identifier.

Privacy Services require the definition of two new components in the infrastructure. The Key Sharing Service and the EOUID Generator Service are implemented in the Server Service Layer.

The Key Sharing Service keeps the decryption key shares and further information related to the encrypted objects as the EOUID and the Message Authentication Code

(MAC) signature for the encrypted object. This service provides to the authorized clients the data needed to decrypt objects in a secure way.

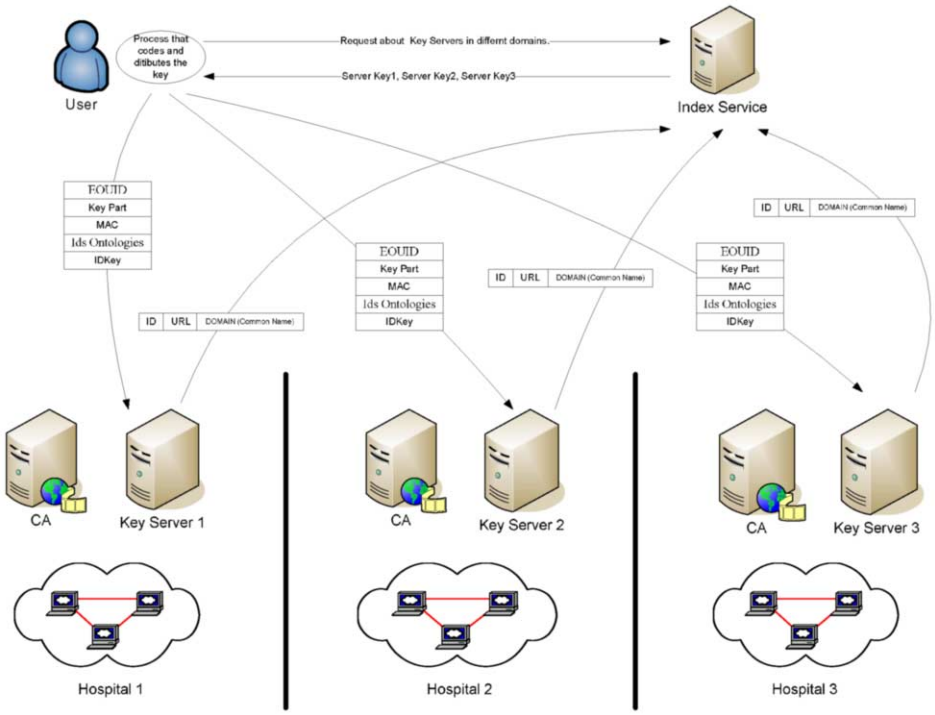


Figure 2. Publication of Key Servers in different administrative domains and distribution of the keys.

Clients need a valid VOMS proxy certificate for executing operations in the service. The service presents the attributes contained in the proxy to the Gatekeeper which enforces the access control policies of the domain. The Gatekeeper could use supplementary information managed by external Grid services to authorize a request.

A work describing Grid middleware that uses an ontological schema to create virtual communities and to define common targets for sharing, transferring and processing DICOM medical images in a distributed environment [6] and a work describing Grid middleware for managing DICOM Structured Reporting Objects [3], have been recently adopted by the medical community. Both projects are based in the TRENCADIS software architecture, and both implement a Gatekeeper component that makes use of the ontology information used for creating virtual storages of DICOM objects to control the access to the virtual storages. Ontologies are managed by an Ontology Server which provides the links between ontologies and VOMS attributes that serve as access control policy.

The Key Sharing Service operations are exposed in a single interface:

- **xmlGetError:** It returns the last error which occurred in the service.
- **xmlSaveSubkey:** It stores a decryption key share in the key server.
- **xmlRetrieveSubkey:** It retrieves a decryption key share from the key server.
- **xmlUpdateKey:** It updates an existent decryption key share.

The EOUID Generator Service produces valid EOUID for a new encrypted object. Although the operations of this service are very simple, it is a good idea to deploy several instances of the service in each administrative domain, because this service could be easily targeted by a malicious client who wants to produce deny of service failures. We propose to deploy one EOUID Generator Service per each Key Sharing Service deployed in the administrative domain.The EOUID Generator Service operations are exposed in a single interface:

- **xmlGetError:** It returns the last error which occurred in the service.
- **xmlGetEOUID:** It returns a new valid EOUID.

5. Conclusions and Future Plans

In this work, a security model for the TRENCADIS architecture has been described. A suitable security level is obtained by defining a set of services and components in the infrastructure (Core Middleware Layer and Server Services Layer). The model aims three main objectives: a) Secure authentication and communication; b) Management of VO security policies; and c) Confidentiality and privacy control of medical data.

The security model is flexible and expansible, and could be integrated with any of the functionalities provided with the architecture through components in the Components Middleware Layer. All general security services have been specified and deployed with the infrastructure, in the same way the additional Gatekeeper component required by new services has been described.

Nowadays, TRENCADIS offers a high-level object-oriented interface with a given functionality [3] [4] (create Virtual Storages, Download DICOM objects, Upload DICOM objects, etc…) through the Components Middleware Layer, and it increases the productivity of code developers for building applications for managing DICOM objects in a secure way. Our focus at this moment is to integrate the security model with the actual functionalities.

A practical deployment of TRENCADIS is the CVIMO architecture. CVIMO (Valencian Cyberinfrastructure for Medical Imaging in Oncology) is a project funded by the regional government of Valencia (Conselleria d'Empresa, Universitat i Ciencia de la Generalitat Valenciana, code GVEMP06/004) involving 5 hospitals of the Land of Valencia interested on sharing Medical Images on the fields of Liver, Lung and Central Nervous System Cancer. This three virtual communities have been deploy and templates for Structured Reports (both for staging and follow-on), following the DICOM standards have been developed. A tool for filling-in the Structured Reports has been developed.

A use case is as follows. When an interesting case is detected, the radiologist sends it from the workstation to the CVIMO local storage of the Hospital through DICOM protocols. Automatically, the system registers this case in the Global Virtual Storage and makes it available as "non-informed" case. Then, the radiologists can fill-in the structured report through a web-browser, and save this report on the local system. Then, both image and structured report are made available for sharing within the VO and indexed through the searching criteria of the ontology in which the structured report was made, considering the information filled-in on the report. Any radiologist with a

valid VOMS certificate belonging to an authorized VO and group can request the downloading of the study to the local machine.

6. Acknowledgements

The authors wish to thanks the financial support received from The Spanish Ministry of Science and Technology to develop the project "Componentes de Nueva Generación para la Explotación Eficiente de Infraestructuras Grid en E-Ciencia", with reference TIN2006-12890.

References

[1] Digital Imaging and Communications in Medicine (DICOM) Part 10: Media Storage and File Format for Media Interchange. National Electrical Manufacturers Association, 1300 N. 17th Street, Rosslyn, Virginia 22209 USA.
[2] "DICOM Structured Reporting". Dr. David A. Clunie. ISBN 0-9701369-0-0. 394 pages softcover.
[3] I. Blanquer, V. Hernandez, and D. Segrelles. "TRENCADIS – a WSRF Grid MiddleWare for Managing DICOM Structured Reporting Objects". Proceeding of HealthGrid 2006. ISBN. Technology and Informatics. Ed. IOPress. ISBN:1-58603-617-3
[4] I. Blanquer, V. Hernández, J. D. Segrelles. "TRENCADIS – A Grid Architecture for Creating Virtual Repositories of DICOM Objects in an OGSA-based Ontological Framework". Lecture Notes in Computer Science (LNCS), Subseries Lecture Notes in Bioinformatics (LNBI). Biological and Medical Data Analysis. ISSN: 0302-9743, ISBN-10: 3-540-68063-2, ISBN-13: 978-3-540-68063-5
[5] I. E. Magnin, J. Montagnat. "The Grid and the Biomedical Community: achievements and open issues" EGEE User Forum, CERN, Geneva, Switzerland, March 1-3, 2006
[6] I. Blanquer, V. Hernández, J. D. Segrelles. "An OGSA Middleware for managing medical images using ontologies". Journal of Clinical Monitoring and Computing. 2005 Oct;19(4-5):295-305
[7] Grid Interoperations Cook Book. https://twiki.cern.ch/twiki/bin/view/EGEE/InteropsCookBook
[8] R. Alfieri, R. Cecchini, V. Ciaschini, L. dell'Agnello, Á . Frohner, A. Gianoli, K. Lörentey, F. Spataro. "VOMS, an authorization system for virtual organizations". In Proceedings of the 1st European Across Grids Conference, 2003
[9] A. Frohner, V. Ciaschini. "VOMS Credential Format", EDG Draft, 2004
[10] L. Seitz, J. M. Pierson, L. Brunie. "Encrypted storage of medical data on a grid". Methods of Information in Medicine. 2005; 44(2):198-201
[11] E. Torres, C. De Alfonso, I. Blanquer, V. Hernandez. "Privacy Protection in HealthGrid: Distributing Encryption Management over the VO". Proceedings of HealthGrid 2006. ISBN. Technology and Informatics. Ed. IOPress. ISBN:1-58603-617-3

IV. State of the Art of the Grid Research and Use at Individual Level

This page intentionally left blank

From Genes to Personalized HealthCare: Grid Solutions for the Life Sciences
N. Jacq et al. (Eds.)
IOS Press, 2007

Grid-supported Medical Digital Library

Michal KOSIEDOWSKI[1], Cezary MAZUREK, Maciej STROINSKI, Jan WEGLARZ
Poznan Supercomputing and Networking Center, Poznan, Poland

Abstract. Secure, flexible and efficient storing and accessing digital medical data is one of the key elements for delivering successful telemedical systems. To this end grid technologies designed and developed over the recent years and grid infrastructures deployed with their use seem to provide an excellent opportunity for the creation of a powerful environment capable of delivering tools and services for medical data storage, access and processing. In this paper we present the early results of our work towards establishing a Medical Digital Library supported by grid technologies and discuss future directions of its development. These works are part of the "Telemedycyna Wielkopolska" project aiming to develop a telemedical system for the support of the regional healthcare.

Keywords. Telemedicine, e-health, medical digital library

Introduction

"Telemedycyna Wielkopolska" is a project run in an interdisciplinary collaboration between Poznan Supercomputing and Networking Center, the Institute of Computing Science of Poznan University of Technology, and the Division of Trauma, Burns and Plastic Surgery of Poznan University of Medical Sciences, which aims to design and develop a system for provisioning telemedical services in the Wielkopolska province [1]. The objective of the project is to propose a set of remote services that will support the regional health care. Although the system and the tools are designed to be used by various medical domains, the pilot use case scenario is based around services for the support of trauma treatment. To this end a system of medical teleconsultations in the area of trauma has been designed and deployed as a pilot for use within several hospitals in Wielkopolska. Further works of the project focus on developing a multimedia Medical Digital Library that will store and enable information and knowledge in the area of trauma.

This digital library makes use of the infrastructure developed in the recent years by the academic community in Wielkopolska and allows to deliver a huge amount of useful information to the regional medical community to support them in their everyday work and to widen their knowledge and thus improve the quality of patient treatment. It also allows to construct value-added services that can support such different activities as medical research or health care system management. In this paper we present the concept and the first results of the development of the Medical Digital Library which was based on the grid infrastructure enabled through the research and deployment performed in the recent years. The grid infrastructure utilized for the needs of the Medical Digital Library is discussed in section 1. Section 2 presents the first

[1] Corresponding Author: Michal Kosiedowski, Poznan Supercomputing and Networking Center, ul. Noskowskiego 10, 61-704 Poznan, Poland

service of the Medical Digital Library that has been deployed for use by the medical community: the Registry of Reference Medical Cases. Other services developed currently and planned for the future are discussed in section 3. We end with conclusions in section 4 where we also draw on some work related to our research.

1. Grid Infrastructure for the Support of the Medical Digital Library

As it was mentioned in the Introduction, the Medical Digital Library is supposed to store and give access to a huge amount of healthcare-related data. This imposes a requirement for delivering an infrastructure capable of efficient storing of this data to the hospitals in the Wielkopolska province. This infrastructure must enable secure sharing of information between the system participants. One should also consider limited investment funds that are available to the healthcare system in Poland. It is the academic community which obviously plays a significant role in supporting the public services.

The Wielkopolska academic community has a long list of success stories connected with the cooperation with public organizations responsible for various areas of public activities such as e-administration [2], hospital informatization [3] or e-safety [4]. The academic community is the owner and operator of powerful infrastructure that can be put to use for the support of the public services such as, for example, health care. This infrastructure includes the broadband optical network named PIONIER [5], and a significant amount of computing and storage resources that were enabled on the grid as a result of several national and international grid research projects conducted in the recent years.

The remote resources enabled on the PIONIER network through advanced grid technologies [6] provide a great opportunity to deliver an infrastructure for the support of the regional Medical Digital Library. More importantly, this infrastructure based on the grid fulfills both the technological and the budget requirements connected with the organization of such a library in Wielkopolska. In this section we discuss several key elements of the grid-based infrastructure that is utilized for the support of a developed distributed digital library.

1.1. Data Storage

One of the most important requirements for the infrastructure supporting the developed Medical Digital Library is the capability of secure and efficient storage of huge amounts of various types of data. This data includes simple textual information, various types of medical images and video files. Moreover, the sources of this data are diverse and range from text and image information entered via a web portal, through DICOM medical images and sequences saved at the point of patient examination, to digital recordings of events such as medical operations. These sources are distributed across individual hospitals and across the region, and often produce sensitive personal data.

The data storage resources owned by the Wielkopolska academic community have been enabled on the grid through the Gridge Data Management System [7]. The system has been designed within the PROGRESS project [8] and further developed within such projects as SGIgrid [9] or ACGT [10]. It will also be used for the construction of the National Data Storage infrastructure [4]. Thanks to the unified access to data storage resources, which is based on the grid standards, it has been possible to enable a

huge virtual storage resource incorporating not only the resources locally available in the Wielkopolska province, but also other resources resulting from the deployment of grid infrastructures within the above-mentioned projects.

The Gridge Data Management System provides mechanisms for file virtualization on the grid. It offers its functionality via GSI-based Web Services that allow to manage the stored files, metadata and file access rights. The physical data is stored within the so called Data Container modules which are also GSI Web Services and enable storage of data on various types of media: generic file systems, tape archivers or relational databases. The transmission of data to and from the system takes place with the use of various data transfer protocols such as, for example, Grid FTP or GASS. Integration of new types of Data Containers and new data transfer protocols is relatively easy thanks to the flexible API system. The Gridge Data Management System is used to enable the above-mentioned distributed data storage resources to be used within the constructed Medical Digital Library. This grid data infrastructure is used to store and share such data as medical images and image sequences, and recordings of various events.

1.2. Data Processing

Putting the data stored within the Medical Digital Library on the grid is also dictated by the need to process the collected data by various applications. These applications will search for knowledge links between individual sets of data and will be used for such scenarios as clinical decision support or medical research. It is envisaged that the developed Medical Digital Library will eventually grow to hold a huge number of data records, thus imposing a significant computing power requirement. Such computing power, just like in the case of data storage resources, is available within the grid infrastructure managed by the academic community that resulted from the national and international grid projects. In addition to the resources resulting from the execution of the already mentioned projects, this infrastructure also includes the GridLab [11] and Crossgrid [12] grid infrastructures.

Apart from delivering a utilizable grid infrastructure, the GridLab project resulted in developing a range of valuable grid technology. This includes the Gridge Resource Management System (GRMS) [13] which allows to manage and access computing resources such as hosts and applications. It offers a GSI Web Services API to submit grid jobs, and hides the complexity of the underlying low-level grid management software. GRMS cooperates with such grid management systems as Globus Toolkit [14], Sun Grid Engine [15] and Unicore [16].

1.3. Security

An important issue of the Medical Digital Library is data security. This includes secure access to the data with the support of a flexible rights management mechanism, secure transmission of data and data integrity. The data and computing grid infrastructure described in the previous two subsections is managed by and accessed via Web Services based tools that expose their functionality in compatibility with the Grid Security Infrastructure [17]. This allows secure communication between the services involved in the realization of a transaction requested by the user and allows single sign-on on distributed resources. Each of the Medical Digital Library users has his/her personal certificate issued by the Polish Grid CA which he/she uses to authenticate into the library.

Further on, the authorization of user operations on resources is maintained by the Gridge Authorization Service (GAS) [18]. GAS is a flexible service that allows to apply different types of authorization policies for different services based on resource centric or role centric models. It can authorize actions for multiple services concurrently. In the Medical Digital Library GAS is used to authorize access to computing grid resources and to library services such as, for example, the Registry of Reference Medical Cases which is described in detail in the next section.

1.4. Data and Service Access

Data and services of the Medical Digital Library are accessed from within various types of user applications. The major user access point is a web portal, but some scenarios involve usage of mobile user interfaces or standalone applications. To enable services such as the already mentioned Registry of Reference Medical Cases or the Medical Teleconsultations Service within various types of user access applications, these services have been realized in line with the Grid Service Provider concept [19, 20]. The Grid Service Provider introduced a flexible architecture of accessing grid services at high-level of granularity to facilitate construction of grid user interfaces. This allows to build user access points on various types of terminals including web browsers, desktop computers and mobile terminals relatively easily.

The "Telemedycyna Wielkopolska" portal, the main user access point, has been deployed within the GridSphere Portal Framework [21] and utilizes its capabilities for building standards-based portals. The telemedical services enabled within this portal also take use of the functionality of Grid Portlets [22], especially in the area of GSI communication with remote services. The mobile terminal is used for the Teleconsultations Service and is planned for delivering clinical decision support at the point of care. The standalone user interface is connected with archiving medical images such as DICOM images and operation recordings.

2. Multimedia Registry of Reference Medical Cases

The Registry of Reference Medical Cases is the first service of the developed Medical Digital Library that has been delivered within the "Telemedycyna Wielkopolska" portal. This aim of this service is to provide functionality for collecting and classifying medical cases in the following two usage scenarios:

- collecting reference medical cases, i.e. such cases that are examples of ideal treatment applied to cure specific injury suffered by a patient;
- collecting all medical cases in the Wielkopolska province that were subject to treatment in the surgery divisions of Wielkopolska hospitals.

Currently the first of these two scenarios has been enabled within the portal. The pilot registry has been deployed to form a library of reference medical cases in the area of trauma.

2.1. The Registry Structure

The Registry of Reference Medical Cases is composed of the following four core elements:

- the medical domain database which holds descriptions of medical domains supported by the particular instance of the registry;
- the UI database which keeps information on UI feel and look for a particular type of end-user terminal in correspondence with the structure of a case belonging to a related medical domain;
- the case database where the actual descriptions of medical cases are stored;
- the data grid resources enabled via the Gridge Data Management System services which store multimedia illustration of the collected cases, such as medical images or video documentation.

The purpose of the medical domain database is to allow easy instantiation of a new registry for the support of a medical domain previously not supported by the Medical Digital Library; the information required to describe a medical case highly depends on the medical case to which it belongs. The domain database allows to use metadata to describe the structure of a medical case connected with a particular domain. The UI database allows to enter information corresponding to the structure of a medical case connected with a particular domain and is later used within the end-user application to construct a user interface. It is very useful to automatically generate the user interface within the portal or a standalone application while enabling on-the-fly modifications to the functionality of the interface when the structure of the medical case changes. Both these databases are stored in a relational database system.

The case database is a place where the descriptions of the actual medical cases are stored. Each medical case description contains basic information on the patient, the results of patient examination, classification of the case according to the standards agreed within the community and the information on the applied treatment. These descriptions are stored in a relational database system. The medical case description is accompanied by the multimedia illustration of the case, for example by medical images or video documentation. These data files are stored on the data grid.

2.2. The Portal Interface

The Registry of Reference Medical Cases is designed in accordance with the Grid Service Provider pattern. It is a Web Service supporting the Grid Security Infrastructure. Its primary access interface is the “Telemedycyna Wielkopolska” portal, but it is also planned to enable some of the functionality on mobile terminals. It will also be possible, in case such a requirement turns up, to enable this service within standalone end-user applications.

The web portal interface of the Registry has been developed as two JSR-168 portlets: one for adding and modifying entries in the Registry and the other for browsing and searching through the Registry. The portlets are compliant with the standards and as such they can be potentially deployed in any JSR-168 compliant portal framework. As it was already mentioned above, we selected GridSphere as the portal engine in which the “Telemedycyna Wielkopolska” portal runs. Selecting GridSphere for the portal framework is connected with both its functionality that fulfills all the requirements for the portal and its full compatibility with the Gridge Toolkit which was used to build the grid infrastructure utilized by the constructed Medical Digital Library.

The Registry is currently used as an education resource for medical doctors, especially surgeons in the Wielkopolska hospitals. The reference cases are provided by the experts in the area of trauma. It is planned to integrate this service within the

Medical Teleconsultations scenario to allow the users to access a simple 'show me similar cases' decision support mechanism and to allow the consultants to point to the Registry entries as a reference within their consultation answers.

2.3. Value-added Services

Introduction of the Registry of the Reference Medical Cases allows us to start the development of some value-added services on top of this core service of the Medical Digital Library. As it was already mentioned, the service can be used not only for collecting the reference medical cases, but it can also be used to collect all the medical cases encountered in the region in the area of the medical domain in question. It is important to note, though, that the organization of this scenario requires a discussion within the medical community so that the collection of data can be possible.

The two scenarios we envisage as those that can be supported by accommodating the developed Registry to collect all medical cases are as follows:

- medical research in which scenario the collected data may be subject to medical research conducted by the researches at the Poznan University of Medical Sciences;
- healthcare system management support in which scenario the collected data may be analyzed to report the current trends in the medical domain in question, thus helping to better estimate needs of the regional health care.

We envisage to employ the computing grid resources to support users working within the above-listed scenarios.

While the deployment of the scenarios in question is relatively easy from the technical point of view assuming the data is already stored in the Registry, one should remember that it also requires developing and implementing protocols for the exchange of information between the hospital information system software used in Wielkopolska and the Registry to facilitate data collection .

3. Other services

The Registry of Reference Medical Cases discussed in the previous section is the first service of the Medical Digital Library that was successfully delivered to the end-users. We are currently working on adding new services to the library. The services that we plan for addition to the Library include the medical data archiving on the data grid and knowledge discovery for clinical decision support. We discuss these two services in this section.

3.1. Medical Data Archiving on the Grid

The archiving of medical data is connected with the requirement received from the medical community to support the hospitals with a cost-effective solution for a PACS system [23, 24]. A natural solution for this type of a system in the "Telemedycyna Wielkopolska" environment is to organize the PACS system on the base of the data grid infrastructure; a similar solution has been proposed by B. J. Liu, M. Z. Zhou, and J. Documet in [25]. To this end we began investigation towards establishing an easy-to-use solution that would seamlessly integrate with the Gridge Data Management

System on the one hand, and with the telemedical services provided within the "Telemedycyna Wielkopolska" portal on the other hand.

The medical data that is subject to storing in the data grid includes the following:

- the image documentation of patient examinations conducted with the use of various types of medical modalities such as, for example, computed tomography, radiography, magnetic resonance imaging and others;
- the digital recordings of events taking place at the hospitals such as, for example, operations performed in the operation rooms.

The long-term purpose of storing these types of data is as follows:

- storage of the DICOM documentation scenario will allow to organize a region-wide archive of radiology documentation which will be accessible to all medical doctors in the province, with a special focus given to the general practitioners;
- the archiving of the course of events during operations will have two applications: the first application is to use the selected recordings as education materials, the second one is to support solving disputes over medical errors.

Both these scenarios will improve safety of the patients and the overall quality of healthcare in the region.

To enable storage and access to the above-listed data types on the data grid we are planning to integrate a DICOM image and a video streaming servers with the Gridge Data Management System as two new protocols for data file transfer. This will allow to easily and immediately serve the data in question embedded within web pages once they are accessed.

The medical data archiving on the grid scenario is currently under design. We expect the first results at the end of February when pilot deployments of the scenario take place.

3.2. Knowledge Driven Decision Support

The data collected with the Medical Digital Library may be a subject of deep analysis aiming at discovering knowledge. A simple example of data analysis is search for medical cases similar to the one currently viewed or processed by a user. In this scenario knowledge discovery is relatively easy to design and implement. However, searching for similar information will go well beyond simple matching pieces of information with well-defined structure such as, for example, results of examinations or classification of a medical case found in the Registry of Reference Medical Cases. It will also apply domain knowledge to evaluate similarity between cases and to search for similarity within the multimedia medical documentation, with a special focus given to image analysis .

Building the domain knowledge will require constant processing of new data to update the knowledge base. This processing of data is envisaged to take place on the grid, which can especially help with the analysis of images as experience of other projects such as, for example, Mammogrid shows [26]. The knowledge collected and constantly updated during the data processing will help to create another service of the Medical Digital Library. This service will be responsible for supporting clinical decision making, including suggesting clinical decisions at the point of care in the 'anytime, anywhere' scheme with the use of mobile terminals. The applications in the area of the clinical decision support are a subject of research by our colleagues from the

Institute of Computing Science of Poznan University of Technology, who have previous experience with constructing such services [27].

4. Conclusions and Related Work

The Medical Digital Library that we discussed in this paper is at early stage of construction. It is envisaged to support the health care in the Wielkopolska province thus improving the quality of treatment received by the patients in the region. The Library is a part of the "Telemedycyna Wielkopolska" system which aims to provide a set of remote services built on top of the optical network and grid infrastructures. The early results of the project show that utilization of the grid infrastructure and grid technology has a big potential providing opportunities to deliver huge data and computing resources that can be cost-effectively used by the healthcare organizations such as hospitals.

Summing up the discussion it is important to draw on other medical digital libraries found in the world. Some examples include the "TeleMedical.com" digital medical library [28] where one can find all sorts of digital medical information ranging from publications in the area of medicine to symptoms, diseases and medications databases. Interesting work has been conducted by Papadakis et al. [29] who designed an architecture for a secure web-based digital library. A medical digital library has been constructed also by Chu at al. [30]. One should note that the term 'medical digital library' is also often used to denote digital libraries with a collection of medical publications like, for example, MeDLib@NIC [31], SUMS Medical Digital Library [32] or a digital library at Hurvey Cushing/John Hay Whitney Medical Library [33]. This type of a digital collection of medical publications is built within the "Telemedycyna Wielkopolska" environment with the use of the dLibra digital library framework [34].

The Medical Digital Library that we are working on is an example of a health scenario supported by the grid technology. When looking at the health scenarios on the grid one can distinguish four major classes of medical applications: clinical, simulation, pharmaceutical and bio-informatic. "Telemedycyna Wielkopolska" is an example of a system providing clinical telemedicine. In this class important research problems include management and integration of distributed medical data [35, 36], deploying medical image analysis on the grid [37, 38, 39], collaborative visualization of medical data [40] and enabling clinical decision support on the grid [41]. A problem close to our work on storing medical images in the grid-based Data Management System is discussed in [42]. Simulation applications are deployed on GEMSS Grid [43] and GAMA [44]. An example of a pharmaceutical grid scenario can be found in [45]. Bio-informatic grid applications are widely designed around the world, for example within the Biogrid project [46].

References

[1] J. Blaszczynski, M. Kosiedowski, C. Mazurek, R. Slowinski, K. Slowinski, M. Stroinski, Sz. Wilk, Telemedical Portal 'Telemedycyna Wielkopolska', *MIT 2006 Medical Informatics & Technology,* Ed. by Ewa Pietka, Jacek Leski, Stanislaw Franiel, Wisla, 2006, 230-235.

[2] P. Gruszczynski, B. Lange, M. Maciejewski C. Mazurek, K. Nowak, S. Osinski, M. Stroinski, A. Swedrzynski, Building a large-scale information system for the education sector: a project experience,

7th International Conference on Enterprise Information Systems, Conference proceedings, Vol. 4, Miami, 2005, 145-150.

[3] *Eskulap - kompleksowy system informatyczny dla jednostki służby zdrowia*, http://www.alma.biz.pl/index.php?id=9.

[4] M. Brzezniak, N. Meyer, R. Mikolajczak, M. Stroinski, KMD – National Data Storage, *iGrid 2005*, San Diego, 2005.

[5] J. Rychlewski, J. Weglarz, S. Starzak, M. Stroinski, M. Nakonieczny, PIONIER: Polish Optical Internet, *ISThmus 2000 Research and Development for the Information Society conference*, Poznan, 2000, 19-28.

[6] J. Pukacki, M. Kosiedowski, R. Mikolajczak, M. Adamski, P. Grabowski, M. Jankowski, M. Kupczyk, C. Mazurek, N. Meyer, J. Nabrzyski, T. Piontek, M. Russell, M. Stroinski, M. Wolski, Programming Grid Applications with Gridge, *Computational Methods in Science and Technology* **12** (2006), Osrodek Wydawnictw Naukowych, Poznan.

[7] M. Kosiedowski, M. Malecki, C. Mazurek, P. Spychala, M. Wolski, Integration of the Biological Databases into Grid-Portal Environments, *Workshop on Database Issues in Bioological Databases DBiBD*, Edinburgh, 2005.

[8] M. Kosiedowski, C. Mazurek, M. Stroinski, PROGRESS – Access Environment to Computational Services Performed by Cluster of Sun Systems, *2nd Cracow Grid Workshop*. Cracow, 2002, 45-56.

[9] M. Lawenda, N. Meyer, T. Rajtar, M. Okon, D. Stoklosa, M. Stroinski, L. Popenda, Z. Gdaniec, R.W. Adamiak, General Conception of the Virtual Laboratory, *Lecture Notes in Computer Science* **3038** (2004), Springer-Verlag, Berlin Heidelberg New York, 1013-1016.

[10] *ACGT - Advancing Clinico Genomic Trials*, http://eu-acgt.org/.

[11] G E. Seidel, G. Allen, A. Merzky, J. Nabrzyski, GridLab – a grid application toolkit and testbed, *Future Generation Computer Systems* **18** (2002), Elsevier Science Publishers B. V, Amsterdam, 1143-1153.

[12] J. Marco, H. Marten, N. Meyer, M. Noga, P. A. M. Sloot, M. Turala, CrossGrid – development of Grid environment for interactive applications, *Pionier 2002: Polski Internet Optyczny: technologie, usługi i aplikacje*, Poznan, 2002, 97-112.

[13] K. Kurowski, B. Ludwiczak, J. Nabrzyski, A. Oleksiak, J. Pukacki, Improving Grid Level Throughput Using Job Migration and Rescheduling Techniques in GRMS, *Scientific Programming* **12** (2004), IOS Press, Amsterdam, 263-273.

[14] I. Foster, Globus Tookit Version 4: Software for Service Oriented Systems, *Lecture Notes in Computer Science* **3779** (2005), Springer-Verlag, Berlin Heidelberg New York, 2-13.

[15] *GridEngine: project home*, http://gridengine.sunsource.net/

[16] D. Erwin, UNICORE – A Grid Computing Environment, *Concurrency and Computation: Practice and Experience* **14** (2002), John Wilesy and Sons, Chichester, 1395-1410.

[17] V. Welch, F. Siebenlist, I. Foster, J. Bresnahan, K. Czajkowski, J. Gawor, C. Kesselman, S. Meder, L. Pearlman, S. Tuecke, Security for Grid Services, *12th IEEE International Symposium on High Performance Distributed Computing HPDC-12*, Seattle, 2003, 48-57.

[18] M. Adamski, P. Grabowski, K. Kurowski, B. Lewandowski, B. Ludwiczak, J. Nabrzyski, A. Oleksiak, T. Piontek, J. Pukacki, R. Strugalski, GridLab Middleware Services for Managing Dynamic Computational and Collaborative Infrastructures (VOs), *13th Annual Mardi Gras Conference*, Baton Rouge, 2005, 35-40.

[19] M. Bogdanski, M. Kosiedowski, C. Mazurek, M. Wolniewicz, GRID SERVICE PROVIDER: How to improve flexibility of grid user interfaces?, *Lecture Notes in Computer Science* **2657** (2003), Springer-Verlag, Berlin Heidelberg New York, 255-263.

[20] P. Grzybowski, M. Kosiedowski, C. Mazurek, Web Services Communication within the PROGRESS Grid-Portal Environment, *International Conference on Web Services ICWS 2003,* Las Vegas, 2003, 340-345.

[21] J. Novotny, M. Russell, O. Wehrens, GridSphere: A Portal Framework For Building Collaborations, *Concurrency and Computation: Practice and Experience* **16** (2004), John Wiley & Sons, Ltd., Chichester, 2004, 503-513.

[22] M. Russell, J. Novotny, O. Wehrens, The Grid Portlets Web Application: A Grid Portal Framework, *Lecture Notes in Computer Science* **3911** (2005), Springer-Verlag, Berlin Heidelberg New York, 691-698

[23] H. Huang, N. Mankovich, R. Taira et al., Picture Archiving and communication Systems (PACS) for Radiological Images: State of the Art, *CRC Critical Reviews in Diagnostic Imaging* **28** (1988), 383-428.

[24] O. Ratib, Y. Ligier, C. Girard, R. Perrier, M. Logean, and J.-F. Vurlod, A picture archiving and communication system based on an open and distributed architecture, *14th Annual International Conference of IEEE Engineering in Medicine and Biology Society*, Paris, 1992, 1197.

[25] B. J. Liu, M. Z. Zhou, and J. Documet, Utilizing data grid architecture for the backup and recovery of clinical image data, *Computerized Medical Imaging and Graphics* 29 (2005), 95-102.

[26] S. R. Amendolia, M. Brady, R. McClatchey, M. Mulet-Parada, M. Odeh, T. Solomonides, MammoGrid: Large-Scale Distributed Mammogram Analysis, *Studies in Health Technology and Informatics* **95** (2003), IOS Press, Amsterdam, 194-199.
[27] W. Michalowski, R. Slowinski, Sz. Wilk, K. Farion, J. Pike, S. Rubin, Design and development of a mobile system for supporting emergency triage, *Methods of Information in Medicine* **44** (2005), 14-24.
[28] *Digital Medical Library*, http://telemedical.com/library.html.
[29] I. Papadakis, V. Chrissikopoulos, D. Polemi, Secure Medical Digital Libraries, *International Journal of Medical Informatics* **64** (2001), Elsevier Science Publishers B. V., Amsterdam, 417-428.
[30] W. W. Chu, D. B. Johnson, and H. Kangarloo, A Medical Digital Library to Support Scenario and User-Tailored Information Retrieval, *IEEE Transactions on Information Technology in Biomedicine* **4** (2000).
[31] S. Singh, S. K. Gaba, and N. Pandita, Architecture and building of medical digital library at National Informatics Centre: what exists and what is required for MeDLib@NIC?, *International Conference on Digital Libraries*, New Delhi, 2004.
[32] *SUMS Medical Digital Library*, http://www.iingroups.com/sums/.
[33] *Digital Library Collections, Harvey Cushing/John Hay Whitney Medical Library*, http://www.med.yale.edu/library/subjects/digital.html.
[34] C. Mazurek, aM. Werla, Distributed Services Architecture in dLibra Digital Library Framework, *8th International Workshop of the DELOS Network of Excellence on Digital Libraries on Future Digital Library Management Systems*, Schloss Dagstuhl, 2005.
[35] D. Budgen, M. Turner, I. Kotsiopoulos, F. Zhu, M. Russell, M. Rigby, K. Bennett, P. Brereton, J. Keane, P. Layzell, Managing healthcare information: the role of the broker, *Studies in Health Technology and Informatics* **112** (2005), IOS Press, Amsterdam, 3-16.
[36] I. C. Oliveira, J. L. Oliveira, F. Martin-Sanchez, V. Maojo, A. Sousa Pereira, Biomedical information integration for health application with Grid: a requirements perspective, *Methods of Information in Medicine* **44** (2005), 161-167.
[37] S. R. Amendolia, F. Estrella, C. Del Frate, J. Galvez, W. Hassan, T. Hauer, D. Manset, R. McClatchey, M. Odeh, D. Rogulin, T. Solomonides, R. Warren, Deployment of a Grid-based Medical Imaging Application, *Studies in Health Technology and Informatics* **112** (2005), IOS Press, Amsterdam, 59-69.
[38] I. De Mitri, The MAGIC-5 Project: Medical Applications on a Grid Infrastructure Connection, *Studies in Health Technology and Informatics* **112** (2005), IOS Press, Amsterdam, 157-166.
[39] J. Montagnat, F. Bellet, H. Benoit-Cattin, V. Breton, L. Brunie, H. Duque, Y. Legré, I.E. Magnin, L. Maigne, S. Miguet, J.-M. Pierson, L. Seitz, T. Tweed, Medical images simulation, storage and processing on the European DataGrid testbed, *Journal of Grid Computing* **16** (2005), 387-400.
[40] J. C. Silverstein, F. Dech, J. Binns, D. Jones, M. E. Papka, R. Stevens, Distributed Collaborative Radiological Visualization using Access Grid, *Studies in Health Technology and Informatics* **111** (2005), IOS Press, Amsterdam, 477-481.
[41] I. Blanquer, V. Hernandez, D. Segrelles, M. Robles, J. M. Garcia J. V. Robledo, Clinical Decision Support Systems (CDSS) in GRID Environments, *Studies in Health Technology and Informatics* **112** (2005), IOS Press, Amsterdam, 80-89.
[42] L. Seitz, J.-M. Pierson, and L. Brunie, Encrypted storage of medical data on a grid, *Methods of Information in Medicine* **44** (2005), 198-201.
[43] S. Benkner, G. Berti, G. Engelbrecht, J. Fingberg, G. Kohlring, S. E. Middleton, R. Schmidt, GEMSS: grid infrastructure for medical service provision, *Methods of Information in Medicine* **44** (2005), 177-181.
[44] A. Bucur, R. Kootstra, R. G. Belleman, A Grid Architecture for Medical Applications, *Studies in Health Technology and Informatics* **112** (2005), IOS Press, Amsterdam, 127-137.
[45] T. Kosaka, Y. Tohsato, S. Date, H. Matsuda, S. Shimojo, An OGSA-Based Integration of Life Scientific Resources toward Drug Discovery, *Methods of Information in Medicine* **44** (2005), 257-261.
[46] K. Fujikawa, W. Jin, S-J Park, T. Furuta, S. Takada, H. Arikawa, S. Date, S. Shimojo, Applying a Grid Technology to Protein Structure Predictor 'ROKKY', *Studies in Health Technology and Informatics* **112** (2005), IOS Press, Amsterdam, 27-36.

From Genes to Personalized HealthCare: Grid Solutions for the Life Sciences
N. Jacq et al. (Eds.)
IOS Press, 2007

Care for Asthma via Mobile Phone (CAMP)

Shu-Hui HUNG, Huan Chin TSENG, Wen Ho TSAI, Hsin Hung LIN
Jen Hsien CHENG, Yi-Ming CHANG
National Center for High-Performance Computing, Taiwan

Abstract. The primary goal of the Care for Asthma via Mobile Phone (CAMP) service is to provide an effective method by which Taiwan's asthma patients can easily monitor their asthma symptoms using a common mobile phone. With the CAMP service, the patient uses his own cellular phone to submit his daily peak expiratory flow rate (PEFR) and answer a simple questionnaire regarding to daily activities. The CAMP service participant then receives an asthma symptom assessment and care suggestion message immediately after imputing his information. This assessment, which is in accordance with the World Health Organization's (WHO) **Global Initiative for Asthma** (GINA) standard, includes weather conditions that might adversely affect the asthma patient (e.g. temperature, pollen count, etc.). This information is, in turn, used to advise the asthma patient how to avoid a severe asthmatic attack.

Keywords. Long-distance care, long-term care, asthma care, chronic disease care, wireless-care system, mobile phone service, e-health, health informatics, tel-health, home care.

Introduction

Taiwan is a small island spreading out geographically along the Tropic of Cancer (N.23.5 latitude). Due to its wide latitudinal range, the temperate often varies by two or three degrees centigrade from North to South.

Taiwan is unique in that more than 50 out of 2,300 people are officially registered with Taiwan's Department of Health as patient's suffering from asthma. Most of Taiwan's asthma patients are classified as "severe" and make trips to the hospital emergency room on a regular basis. According to Taiwan's Department of Health report, more than 1,000 asthma patients die each year in Taiwan. Even though asthma cannot be cured, it does not need to cause death especially in that it can be easily brought under control.

It is very important that asthma patients take the correct dosage of medication. According to physicians, patients who fail to take their medicine correctly often cause their asthma condition to become unstable and sometimes even severe.

In the past, asthma patients had to record their daily peak expiratory flow rate on paper several times throughout the day. Many patients simply become lazy or forgot to record the PEFR. With the CAMP platform, asthma patients no longer need to make or keep a record of their PEFR, rather, they simply input the data directly into their cellular phone. Additionally, the patient receives a health assessment message on their cell phone after imputing their PEFR data.

The asthma care platform has been in service in Taiwan since 2003. Based on input from physicians, the platform has gone through various stages of development including an internet-based phase, a home phone-based phase, and a wireless phase. The platform's data is centrally stored at the National Center for High-Performance Computing and the program results can be viewed on the website. Also, important program notices are sent to patients and physicians from time to time by email or SMS messaging.

1. Methodology

CAMP Objectives are:

- To develop an easy access platform for use by patients and physicians
- To provide asthma attack awareness including changes in the patient's local weather
- To provide long-term asthma care without location or time limitations
- To provide an efficient real-time monitoring service supported by expert medical assistance
- To promote asthma patient care self-management
- To promote more effective diagnosis and treatment of asthma patients
- To provide real-time and historic PEFR relevant statistical flowcharts and support physician's diagnoses
- To illustrate the correlation between asthma symptoms and changes in the weather

2. Primary Platform Service

2.1. Web Service

2.1.1. Patient Participation

Asthma patients can easily access the website to get their personalized information or view their historical and/or real-time data. Additional information such as the latest asthma-related news updates (e.g. a simplified procedure for the measurement of asthma symptoms), medication information, and asthma-related videos can be found on the website as well (Figure 1). Patients are allowed to update their personnel information and view their individual historical flowcharts. Patients are also able to review their asthma care suggestion content messages (Figure 2).

2.1.2. Physician Participation

Using the asthma care website, physicians are able to view their individual patient's raw historical data and the latest asthma-related information/condition. Additionally, the website displays the asthma patient's current condition and weather-related information. This information helps physicians better access their individual patient's conditions and asthma-related needs (Figures 3 and 4).

Figure 1. The CAMP Web Portal

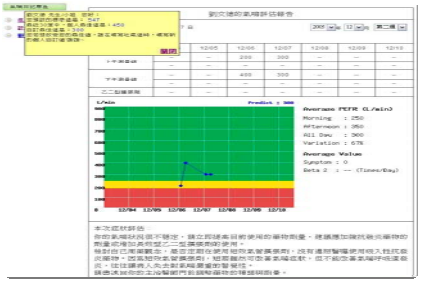


Figure 2. CAMP Historical Data

Figure 3. Monitoring Map

Figure 4. Regional Weather Information

2.2. Mobile Platform Service

2.2.1. Patient Participation

Upon login, patients are required to key in their personal information in order to obtain GINA's standard level of information. According to the patient's input, the CAMP server will calculate the patient's daily PEFR (figures 5 and 6) and send the information to patient's mobile phone automatically. This information is sent to the patient's mobile phone in chart format. It contains all of the patient's asthma care suggestion history and a "condition" color code (e.g. Green = Low Threat, Yellow = Moderate Threat, Red = Severe Threat) (Figure 7) as well.

2.2.2. Location-Based System

As part of the CAMP program, each patient also agrees to submit his physical location data when submitting his PEFR data. The location information is then compared with local real-time weather data and the patient's overall condition. Physicians can also use this information to determine their patient's asthma condition and compare individual patients to one another (figure 8).

Figure 5. User Data Setup

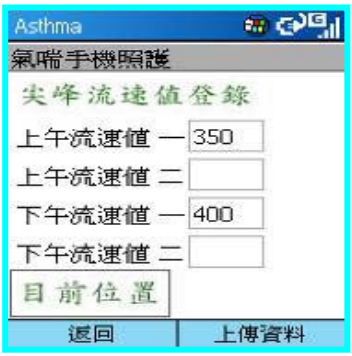


Figure 6. PEFR Input

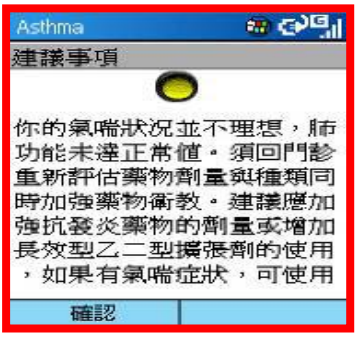


Figure 7. Care Assessment

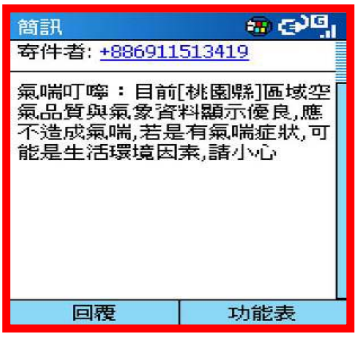


Figure 8. Weather Alert

3. Conceptual Framework

The system consists of two parts that are automatically calculated based on WHO's GINA guidelines. The first part utilizes the patient's age, height and weight to fill in the standard WHO requirement form. The second part uses the physician's expertise to generate data according to GINA guidelines.

The system is made up of more than 18 different disciplines to provide patients appropriate assessment and suggestions. The Asthma Care project is based on more than 15 years of clinical experience and expertise. The care messages the patients receive are based on their daily activities, physical location, current weather conditions, and PEFR (Figure 9).

Figure 9. CAMP Flowchart

4. System Description

The CAMP website is designed to give asthma care through both wired and wireless data transmission. The "wired" part of the system is made up of the asthma care educational platform website which includes an asthma medication introductory video and an asthma symptoms survey questionnaire. Anyone interested in learning more about asthma can access the website using an ordinary Internet connection.

The "wireless" part of the system consists of the CAMP asthma patient participant being able to upload his data to the CAMP web server and receiving a personalized assessment message immediately thereafter. The assessment message takes into consideration of the patient's current asthma status as well as his physical location and any changes in his local weather that might adversely affect his asthma condition.

The data is centrally stored at the National Center for High-Performance Computing (NCHC) where it is analyzed and then archived. The system also utilizes more than 40 weather stations across Taiwan to collect weather-related data. This data is updated to the CAMP server every 30~50 seconds. The CAMP system utilizes the local weather information data and the asthma patient's general condition to produce information that physicians can use to determine the relevance between their patient's symptoms and changes in the local weather (figure 10).

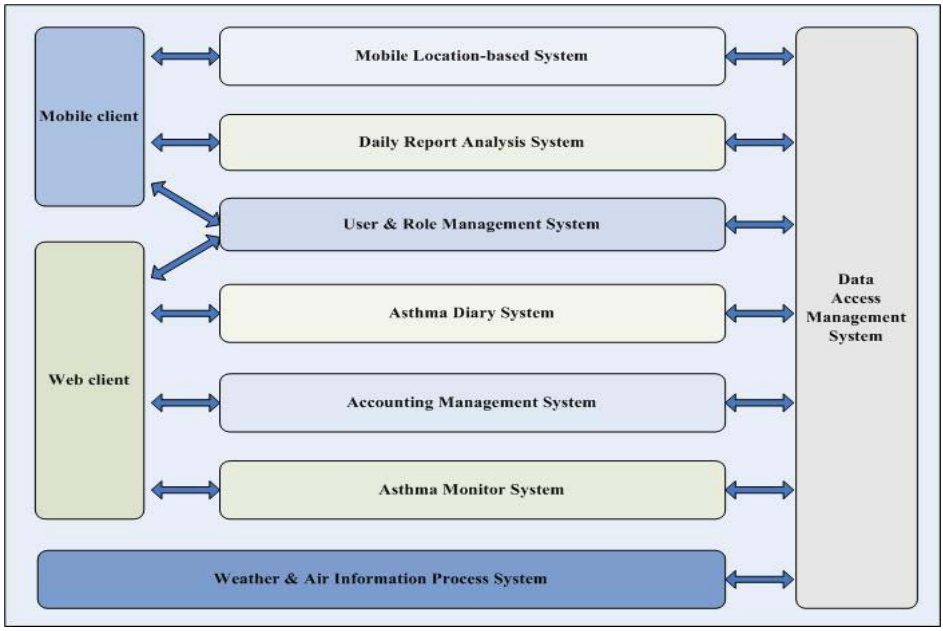


Figure 10. CAMP System Platform

5. Conclusion

The Care for Asthma service has been in use in Taiwan since 2003. Currently, the service provides specialized care to more than 13,000 asthma patients and 400 physicians. Recently, Taiwan's Department Of Health (North Health Bureau) announced that, from 2005 to 2006, the Care for Asthma service program resulted in a 60% reduction in the cost of providing healthcare to Taiwan's asthma sufferers. Figure 11 illustrates the number of Emergency Room visits from 2003 through 2005 for non-CAMP project participants (orange) vs. CAMP project participants (blue).

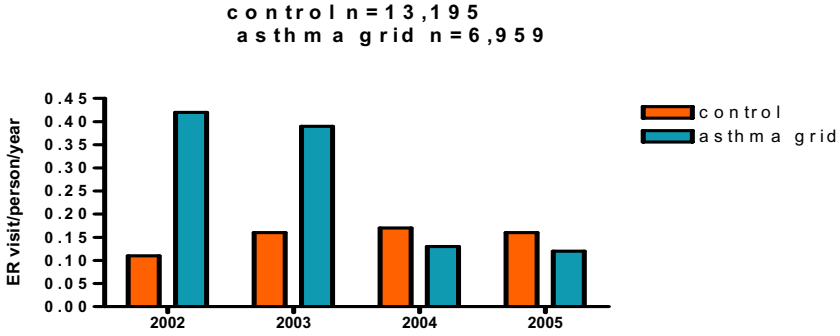


Figure 11. Emergency Room Visits 2003~2005 CAMP participants vs. non-CAMP participants

References

[1] Wagner EH, Austin BT, Davis C, Hindmarcsh M, Schaefer J, Bonomi A. Improving chronic illness care: translating evidence into action. Health Aff (Millwood). 2001;20:64-78
[2] Kilo CM. Improving care through collaboration. Pediatrics. 1999; ;103(Suppl):384-393.
[3] Marks GB, Dunn SM, Woolcock AJ. A scale for the measurement of quality of life in adults with asthma. J Clin Epidemiol. 1992;45:461- 472
[4] Gibson PG, Coughlan J, Wilson AJ, et al. The effects of limited (information-only) asthma education on health outcomes of adults with asthma. Cochrane Database Syst Rev. 2002;(2):CD001005.
[5] Lara M, Nicholas W, Morton S, Vaiana ME, Genovese B, Rachelefsky G. Improving Childhood Asthma Outcomes in the United States. Santa Monica, Calif: RAND; 2001
[6] deOliveira MA, Faresin SM, Bruno VF, de Bittencourt AR, Fernandes ALG. Evaluation of an education program for socially deprived asthma patients. Eur Respir J. 1999;14:908-914.
[7] Lahdensuo A, Haahtela T, Herrala J, et al. Randomised comparison of cost effectiveness of guided self management and traditional treatment of asthma in Finland. BMJ. 1998;316:1138-1139
[8] Mangione-Smith R, Schonlau M, Rosen M, Chan K, Louis T, Keeler E. Measuring the effectiveness of a collaborative for quality improvement in pediatric asthma care: does implementing the chronic care model improve processes and outcomes of care? Ambul Pediatr. 2005;5:75-82
[9] Tsai AC, Morton SC, Mangione CM, Keeler E. A meta-analysis of interventions to improve chronic illness care. Paper presented at Academy Health Annual Research Meeting, San Diego, Calif, June 2003

From Genes to Personalized HealthCare: Grid Solutions for the Life Sciences
N. Jacq et al. (Eds.)
IOS Press, 2007

Data Integration in eHealth: A Domain/Disease Specific Roadmap

Jenny URE[a,b,1], Rob PROCTER[l], Maryann MARTONE[d], David PORTEOUS[a], Sharon LLOYD[b], Stephen LAWRIE[b], Dominic JOB[b], Richard BALDOCK[g], Alistair PHILP[a], Dave LIEWALD[a], Frank RAKEBRANDT[b], Alan BLAIKIE[a], Clare McKAY[b], Stuart ANDERSON, John AINSWORTH[c], Jano van HEMERT[k], Ignacio BLANQUER[e], Richard SINNOTT[a], Christian BARILLOT, Frank Bernard GIBAUD, Alan WILLIAMS[c], Mark HARTSWOOD, Paul WATSON[m], Leslie SMITH[m], Albert BURGER, Jessie KENNEDY[k], Horacio GONZALEZ-VELEZ[n], Robert STEVENS[o], Oscar CORCHO[i], Robin MORTON[a], Pamela LINKSTED[a], Mylene DESCHENES[h], Mark McGILCHRIST[a], Paul JOHNSON[a], Alex VOSS[l], Renate GERTZ[a], Joanna WARDLAW[b]

[a] *Generation Scotland,* [b] *NeuroGrid,* [c] *PsyGrid,* [d] *BIRN,* [e] *HealthGrid Share,* [f] *NeuroBase,* [g] *MouseBIRN,* [h] *P3G Consortium,* [i] *OntoGrid,* [j] *CARO,* [k] *NESC,* [l] *NCESS,* [m] *CARMEN,* [n] *HealthAgents,* [o] *Sealife*

Abstract. The paper documents a series of data integration workshops held in 2006 at the UK National e-Science Centre, summarizing a range of the problem/solution scenarios in multi-site and multi-scale data integration with six HealthGrid projects using schizophrenia as a domain-specific test case. It outlines emerging strategies, recommendations and objectives for collaboration on shared ontology-building and harmonization of data for multi-site trials in this domain.

Keywords. Data integration, e-Health, psychosis, multi-scale imaging, ontology integration

1. Introduction

Grid technology has a key role in enabling the development of a European Research Area [1, 2] with the potential to allow querying across heterogeneous and distributed data sets if these can be integrated and represented in ways which are valid, usable and ethically and legally acceptable [3, 4]. In areas such as brain imaging, the opportunities and the challenges of integration have been particularly evident, requiring integration in multi-centre clinical studies of patients in early stages of psychiatric disorders, standardization of scanners and image processing techniques across mental health research networks as well as scalable integration of voxel-based image data at different levels of integration [5, 6], and the development of shared ontologies and spatial

[1] Corresponding Author: Jenny Ure, School of Informatics, University of Edinburgh, Jenny.Ure@ed.ac.uk

frameworks for reporting brain-related data, both for comparison across sites and to build up integrated views of the brain. The paper summarises a range of the problem/solution scenarios in multi-site data integration and multi-scale datasets from a series of eHealth workshops held in 2006 at the UK National e-Science Centre. The first workshop[2] looked at generic issues, and the second[3] brought together six HealthGrid projects[4] in the same disease domain to road map the issues in a single disease domain, using schizophrenia as a testbed. These include discussion of ontology. integration issues and a roadmap of short and medium term objectives for joint working This was co-hosted by the National e-Science Centre and the Generation Scotland national population genomics project[5] (a nationally funded collaboration of the Scottish University medical schools, the NHS and key research institutes), [7] together with the NeuroGrid[6] project [8] funded by the MRC to develop a Grid platform and toolkit for sharing imaging data for research on dementia, stroke and psychosis. This also builds on the ongoing road-mapping process to support collaboration in data sharing in particular disease domains initiated in the HealthGrid Share project[7]. One of the outcomes of the workshop is a wiki-based collaboration[8] between six HealthGrids from the UK, EU and US to develop shared measures of symptom in terms of imaging, genetic, clinical datasets, shared metadata and shared ontological representations of these.

2. Multi-site and multi-scale integration

The eHealth vision of large-scale, seamless data-sharing for research has to be tempered by acknowledgement of the very real barriers standing in the way of its realisation. The data 'supply chain' underpinning the concept of eHealth and translational medicine is a gradual conversion process where many types of error or bias can arise at different stages from sampling, collection, coding, aggregation, analysis or use, sometimes referred to as the 'social life of information' [9].

2.1. Differences in populations

Many of the datasets were drawn from regional or national studies on very different populations, raising the risk of confounding artifact with effect, whether across sites, or across projects. Data sets may be aggregated from very different populations, where there are physiological differences in brain shape and size that reflect ethnic differences rather than disease effects. In neuroscience, for example, in controlling for global brain volumes in high-risk subjects with and without psychotic symptoms, some measures (e.g. using height as a proxy for head size) and use of paternal social class as a proxy for environment may have variable effects across ethnic and cultural groups [5]. Indicating ethnicity as part of the core metadata would go some way towards ameliorating this, however the concept itself is hard to define, and obtaining this

[2] http://www.nesc.ac.uk/action/esi/contribution.cfm?Title=709
[3] http://www.nesc.ac.uk/action/esi/contribution.cfm?Title=684
[4] http://wikis.nesc.ac.uk/mod/Main_Page
[5] www.generationscotland.org
[6] www.neurogrid.ac.uk
[7] http://www.eu-share.org/
[8] http://wikis.nesc.ac.uk/mod/Main_Page

information is fraught with sensitivities, to the extent that some studies do not attempt to document this.

Participants from the consortium of national population genomics projects[9] also pointed out that recruitment, despite best efforts, is rarely truly representative. Participants are essentially volunteers, more often women than men, and many do so as a result of experience of particular illnesses in families. Recruitment itself is also necessarily opportunistic, and reflects the city and often the hospital setting where the clinical tests are carried out. This is also an issue with control groups in other studies, with some evidence of bias effects in small studies [10]. Epidemiologists here highlighted the importance of separating out differential errors that are likely to cancel out with large samples and non differential errors which are accentuated, inducing bias or confounding effects, and may not be easily identified.

Scanned images are assessed by statistical analysis of brain volumes or densities after registration and normalization, using normal scans as a benchmark. HealthGrids using imaging increasingly use sets of human 'phantoms' scanned at all the sites to provide a standard benchmark (although there are also variations in scans between the same individual at different times). In multi-site studies across more than one Grid project this would require further harmonization.

2.2. Differences in Collection, Coding and Collation

Known error rates of 30% were not uncommon when matching test data against patient data. Many of the problems related to the quality of the data originally filled in:

- Missing data
- Incomplete data
- Incorrect data, e.g., the patient's name being entered as "brain".
- Incorrectly formatted data, e.g., a patient name being specified so that the surname is "SmithJohn".
- Data in the wrong field
- Data in the wrong sequence
- Inconsistent data within a single file, e.g. If the patient's age is inconsistent with image date minus birth date.
- Inconsistent data for same patient on different visits e.g. are patients with the same issuer and patient ID but different names really the same patient?

Cleaning and error trapping software can only capture particular types of error and many anomalies would only be recognized by those with local knowledge of the population, the context and the method of data collection. For example, protocols were interpreted in different ways, or local events impacted on the administration of the test. In aggregated data, without this information, anomalies are hard to identify. Strategies for addressing this included wireless notepads or pens used at the data upload stage, so that data incorrectly loaded was automatically validated against the main database as it was stored. As indicated earlier, error trapping software was also used, together with metrics using probabilistic linkage. It was also seen as important to keep links to raw data (or data owners) where possible. Metadata on provenance was seen as very

[9] The P3G consortium www.p3gconsortium.org is working with over 11 Biobanks on harmonisation

important here [11], however some anomalies in data were often only evident to those familiar with the context, and prepared to follow these up on the ground. One study cited measures of heart rate captured in one site which were consistently higher in one site, and where this was consonant with known population differences. Data quality harmonization work between sites in the study highlighted the fact that participants in one site had to climb six flights of stairs before the ('resting') heart rate test while the lift in the hospital was out of order. This would not be evident from inspection of the data alone but required local input. Guidelines, checklists and toolkits were also used for enhancing data quality. The same communities who produce data were seen by many as particularly well placed to evaluate, quality assure and enhance it. In this context, one study proposed to have a panel responsible for ethics, linkage and data quality issues mediating requests for data as well as submission of data, building on their ability to interpret and also act on local processes for data collection [13], and in the extended healthcare team in primary care [14]. Other strategies included data audit tools, and process management tools to make the stages of the data collection and analysis process transparent to allow researchers and users to compare methods and see the strengths and weaknesses of the data [12].

2.3. Differences in Tests and Tools

In the context of schizophrenia, there were a range of tools, techniques and formats for capturing aspects of the same structure or process, making it hard to differentiate between real differences and artefacts. For example, structural magnetic resonance imaging (sMRI) was used by a majority of the participants researching reduced volumes of the medial temporal lobe and other limbic and paralimbic structures in schizophrenia [5]. There was a then a need for harmonization between different types of scanner, differences in the processes of registration or normalization, differences in the settings (even after servicing) etc. Similarly, in genetic analyses, results can be obtained by a range of different methods, such as microarray, in situ hybridization, and immunocytochemistry, thus raising the possibility that differences between datasets may be a function of the testing and analysis process itself, and this must be included in provenance metadata.

A range of strategies were evident here, including ongoing harmonisation/data quality testing between sites, use of shared 'phantoms' as controls and use of common tools. The BIRN project has been active in using all these approaches across US sites, and with collaborators in the UK such as PsyGrid and NeuroGrid, as well as with NeuroBase in France. In a number of other projects, early prototyping provided a vehicle for community engagement [8], and a number of harmonisation studies were underway within and between participating Grid projects. The use of shared tools as freeware is increasingly a strategy, with BIRN again providing a range of these to support (initially) their own work across multiple sites in the use, but increasingly also with collaborating nodes at eScience centres in the UK and EU. For many of the projects, the range of preferred local software and tests was not only an issue in mapping differences, but the fact that some of these are licensed, commercial or IP protected software, making their provision as tools for other unlicensed users particularly complex.

2.4. Differences in Requirements for Confidentiality of Patient Data

In the case of scans of patients at risk of early-onset psychosis in one study, direct access to the imaging data was regarded as too sensitive and the solution agreed was to provide access to derived statistical data on which algorithms could be run. This added some complexity to the workflows and the design as a whole, but aligned the competing requirements of the different stake-holding groups in a way which could be replicated elsewhere. Ethical and legal issues were seen as unresolved problem issues, reflecting a patchwork of disjoint technical ethical legal and administrative domains. The break-out session on ethical and legal issues highlighted the fact that current legal frameworks cannot provide clear answers for emerging new scenarios, and project teams were increasingly aware of the risks of legal challenge, and of delays in recruitment or use of data for ethical reasons or as a function of public perception. Collaborative stakeholder negotiation was increasingly a basis for agreeing a reasoned, and enforceable position within applicable legal and ethical frameworks, and some national initiatives have include the development of tool kits to support coherent approaches across extended health care communities.

2.5. Differences in Semantic, Ontological and Spatial Integration

Neuroscientists now have access to a vast array of large, heterogeneous and multi-dimensional data from multiple sources, and across multiple scales. As observers increasingly point out, the challenges are now more about integrating data and information, making sense of it (in machine and human terms) and representing it in ways that relate to (and evolve through) the aims and frames of reference of different user groups [3, 15, 6]. Integrating heterogeneous and distributed datasets is therefore a challenge for the e-Health and the e-Science vision, and a priority area for regional, national and international bodies supporting research in e-Health and e-Science.

Data at molecular level on synaptic proteins involved in human mental illness, such as schizophrenia, bipolar disorder and mental retardation [16] is even more valuable when integrated with scanning data, and genetic data yet this requires coordination in a spatial/anatomical frame of reference, using a shared data model, and ideally a human and machine readable format. Much as existing pieces in a jigsaw can support new insights about the structure as a whole, and the missing parts of it, the aggregation of disparate information within a shared model can support both interoperability and understanding if there is sufficient opportunity and motivation for joint working [17].

The workshop grew in part from the awareness that six Grid projects were developing different OWL[10] based ontologies to facilitate cross searching and knowledge discovery across multiple data sources, at different anatomical scales. Ontology integration approaches handle multiple different ontologies by identifying mappings between heterogeneous ontologies or by merging them into a single ontology [18], however this is both complex and variable in the results it achieves, given the semantic heterogeneity, and the variable perceptions of ontological relations across groups, and over time.

The approach adopted in the BIRN project has been to facilitate collective development of the underlying semantics, and conceptual relationship from which project ontologies

[10] The Semantic Web comprises the standards and tools of XML, XML Schema, RDF, RDF Schema and OWL. The OWL Web Ontology Language Overview describes the function and relationship of each of these

are then constructed using open source tools such as the BirnLex[11] and Firework Concept Browser tools. (The BIRNLex is a controlled vocabulary including common terms for neuroanatomy, molecular species, subject information, behavioral and cognitive processes, experimental practice and design, and the associated elements of primary data provenance required for large-scale data integration across disparate experimental studies. It also provides a core for the re-use and integration of existing community ontologies - e.g. OBI, CARO, BFO, and GO and some division of work between groups.). A variety of techniques producing data at different scales can now be superimposed onto the neuronal networks to create new models of the human brain [6, 17] in ways which are human and machine readable, and represented in ways which are visually intuitive. As De Roure [19] points out sense-making is increasingly the issue. Sharing anatomical correlates provides a basis for scalable spatial mapping and integration[12] that facilitates both human interpretation and also more extensive data mining and knowledge discovery.

This Open Source approach, sharing tools and resources, has been increasingly seen as a means of adding value, cutting costs, benchmarking approaches and sharing risk towards common ends in the development of sociotechnical systems [20,21]. The aim here has been to develop a dynamic knowledge infrastructure to support integration and analysis, and to identify and assess existing ontologies and terminologies for summarizing, comparing, merging, and mining datasets that include clinical assessments, assays, demographics, cognitive task descriptions, neuroanatomy, imaging parameters/data provenance in general, and derived magnetic resonance imaging data.

Although less of a short term problem, there was also a perceived tension between the benefits of a fixed frame of reference as implied by an onotology, and the fluid nature of knowledge emerging from ongoing research. Ontologies are a tangible model of domain knowledge involving fairly persistent logical and conceptual relations between classes. This is in tension with the accepted model of knowledge implicit in research on an evolving, hypothetical set of relations which will change, and may involve many parallel disputed interpretations at any one time. In business systems, [22, 23] the approach to this problem has been to separate out the core areas that can most easily be standardized, and to allow a range of approaches to evolve 'at the edge' in an evolutionary manner.

2.6. Differences in Diagnosis and Treatment

Another significant difference between projects was in the diagnosis of complex diseases such as the psychoses. Not only are there different measures of symptoms, in relation to the data sets held, the same symptoms can be associated with different formal diagnoses and treatment recommendations. A key task recommended by the collaborating group from the workshop was to achieve agreement on measures of symptom severity.

[11] http://www.nbirn.net/birnlex/index.htm

[12] Similar principles are used in Google Earth, and in web-based self-organising maps (WebSOMS) facilitating sense-making by building on the behaviour of visual and cognitive systems – scanning, zooming, and organising by semantic or physical similarity for example).

3. Recommendations and Outcomes

The workshop provided an opportunity for a collective focus on shared problems and possible strategies with regard to data-sharing. There were a number of general recommendations and specific recommendations for collaborative working on a shared task.

3.1. Collaborative ontology development – building on work of the BIRN Human Brain project

The six HealthGrid projects have a number of common aims in integrating a wide range of datasets in the domain, for particular purposes and the potential therefore to benefit from collaborative working. This was identified as a task that could be jointly pursued through joint Access Grid meetings and a project wiki in the first instance, with secure access to the BIRN ontology effort, developing including common terms for neuroanatomy, molecular species, subject information, behavioral and cognitive processes, experimental practice and design, and the associated elements of primary data provenance required for large-scale data integration across disparate experimental studies. Given small local sample sizes, and a collective interest in large-scale multi-site clinical trials, this was seen as the best approach for a range of reasons.

3.2. Agree on measures of symptoms in relation to the data sets held

Another recommendation was to work towards shared measures of symptoms and symptom severity in datasets. A wiki has been set up for this on http://wikis.nesc.ac.uk/mod/Main_Page , and the Access Grid nodes at the eScience centre will also allow for collaboration in joint development meetings. This complements existing UK and EU collaborations with BIRN supporting the use and evaluation of BIRN applications in the UK eScience community, and the development of ontologies.[13]

3.3. Share and re-use tools and strategies

There were seen to be advantages in re-use of existing work where possible and finding synergies with the work of other groups towards common ends. In harmonisation of the scanning process for example, strategies for accommodating or harmonising inter site differences are important in disentangling disease effects from other factors. Individual projects had invested time and effort in a range of measures that were often duplicated. Sharing techniques and software development effort for general adoption was therefore perceived as offering multiple benefits. Re-use was also seen as possible from other distributed networked systems in e-business and e-learning where similar problem-solution scenarios have been addressed in transferable ways, in the business supply chain for example [22]. The wiki and the Access Grid provide a medium for developing this, as well as opportunities provided by future workshops.

[13] http://www.nesc.ac.uk/action/esi/contribution.cfm?Title=706

3.4. Coordinated Approach to the Development of Spatio-temporal Ontologies

A shared interest for many of the participants was the potential to explore approaches to the development of structural mappings of objects and relations in space and also over time. The diagnosis development and prognosis of disease, and the evaluation of treatment regimes were perceived as potentially crucial in generating benefits for patients. Work from the EMAP/EMAGE[14]and DGEMap[15] projects demonstrated at the workshops and at a prior UK-BIRN workshop[16] were perceived as areas of real interest for future development.

3.5. Data Quality and Situated Local Action

Although not a recommendation, there was a perception from the discussions in both workshops that distributed communities had a particular role in enhancing the quality of integrated resources, and that this is likely to be a key factor in usability [24] and also sustainability. The pattern in more established distributed business contexts has been to move towards models that allow greater leverage of the situated knowledge and agency of local communities to greater advantage [23] and this may also be the case in the context of eHealth.

3.6. Building Technology Around the Cognitive and the Social Process

Much of the discussion has been on the alignment and harmonization of distributed data, but an emerging area of interest has also been the representation of distributed data in spatial and temporal contexts that leverage existing intuitive cognitive and visual architectures to the analysis of complex data (semantic clustering of similar information, layering etc). This has now become a major factor in the usability of the semantic web on the scale implied by HealthGrids. Large-scale initiatives such as BIRN, GIS systems (think Google Earth[17]), and the work of Berriman et al in astronomy [25] highlight the potential of approaches that actively leverage cognitive and also social processes in the design of usable eScience systems.

4. Conclusions

The multiple Grid projects participating at the event[18] were all at different stages of developing data sets of the same type. Collaborating development on this scale can create de facto standards, share cost and risk, and support proposals for changes in the frameworks adopted at a national and international level [26]. Grid-based eHealth projects imply data-sharing across stake-holding groups and governance frameworks. This is a process which requires formal and informal opportunities for collaboration to take place, and perceived rewards for so doing. In the UK the e-Science and e-Social Science centres have been instrumental in providing support for this through the Data

[14] http://genex.hgu.mrc.ac.uk
[15] http://www.dgemap.com/
[16] http://www.nesc.ac.uk/action/esi/contribution.cfm?Title=706
[17] http://earth.google.com/#utm_campaign=en
[18] Appendix of details on the collaborating Grid projects is on http://wikis.nesc.ac.uk/_mod/Main_Page

Integration Theme[19] and agenda setting workshops [27] as well as supporting wiki-based collaboration such as this and joint Access grid meetings. There is, however, a balance to be struck between the benefits of shared standards and ontological mappings on the one hand, and the risks of limiting the diversity of knowledge deriving from multiple models and local variants. e-Business approaches to this same problem increasingly separate core and local in ways which can provide useful synergies, while new approaches in informatics such as agent based decision support systems [28] may offer other less constraining models for data integration in some contexts.

5. Acknowledgements

We would like to acknowledge the support of Generation Scotland, NeuroGrid and the UK eScience Centre for funding the workshop and NeuroGrid and Generation Scotland for funding the research. Generation Scotland is funded by Scottish Funding Council Grant No. SRDG03006 NeuroGrid is funded through MRC Grant Ref no: GO600623 ID number 77729.

References

[1] Breton, V., Dean, K. and Solmonides, T. (2005). The HealthGrid White paper, In *Solomonides, T., McClatchy R., Breton V., Legre, Y. & Norager, S. (eds.) From Grid to HealthGrid*. IOS Press ISSN 0926-9630.

[2] Blanquer I., Hernandez V. and Mas F., 2005 A Peer-to-Peer Environment to Share Medical Images and Diagnoses Providing Context-Based Searching, *Proceedings of the 13th Euromicro Conference on Parallel, Distributed and Network-Based Processing (PDP'05)* - Volume 00 pp 42-48

[3] Goble C.A., Corcho O., Alper P., De Roure D. 2006) e-Science and the Semantic Web: a Symbiotic Relationship, Discovery Science 2006, Barcelona, Spain, Springer-Verlag, Lecture Notes in Artificial Intelligence (LNAI) 4265

[4] Sinnott R.,(2006) Virtual Organisations for Trials and Epidemiological Studies (VOTES) – Experiences & Prototypes after 1 year, In *Re-Use or Reinvention Workshop*, Nat. eScience Centre, Edinburgh

[5] Lawrie SM, Johnstsone EC, Weinberger DR. (editors) (2004) Schizophrenia: From neuroimaging to neuroscience. Oxford University Press, Oxford.

[6] Martone M., (2004), e-Neuroscience: challenges and triumphs in integrating distributed data from molecules to brains, In *Nature Neuroscience Vol. 7, No.5,* May 2004.

[7] Smith, B. Campbell, H., Blackwood D., Connell J., Connor M., Deary I., Dominiczak A., Fitzpatrick B., Ford I, Jackson C., Haddow G., Kerr S, Lindsay R., Mcgilchrist M., Morton R., Murray G., Palmer C., Pell J., Ralston S., St Clair D., Sullivan F., Watt G., Wolf R., Wright A, Porteous D., Morris D., (2006). Generation Scotland: the Scottish Family Health Study; a new resource for researching genes and heritability. BMC Med Genet, pp.74.

[8] Geddes et al (2006) NeuroGrid: Using Grid Technology to Advance Neuroscience, In *18th IEEE Symposium on Computer-Based Medical Systems (CBMS'05)*

[9] Duguid, P. & Brown, J.S. (2000). *The Social Life of Information*, Boston, MA: Harvard Business School Press.

[10] Sterne J.A., Gavahan D., and Egger M., (2000), Publication and related bias in meta-analysis. Power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology, Vol.53, Issue 11, Nov. 2000, pp. 1119-1129*

[11] Zhao J. Goble C., Stevens R., Jun Zhao, Carole Goble (2006), An identity crisis in the life sciences. In *Proc. of the 3rd International Provnenace and Annotation Workshop*, Chicago, USA, May 2006. LNCS. extended paper.

[19] Workshop presentations (1) http://www.nesc.ac.uk/action/esi/contribution.cfm?Title=684
(2) http://www.nesc.ac.uk/esi/events/709/

[12] Van Vlymen J., De Lusignan S., Hague N. Chan T., Dzregah B. (2005), Ensuring the Quality of Aggregated General Practice Data: Lessons from the Primary Care Data Quality Programme (PCDQ).
[13] McGilchrist (forthcoming)
[14] Smith B. H., Watt G.C.M., Campbell H., and Sheikh A. Genetic epidemiology and primary care, *Brit. Journal of General Practice*, March 2006.
[15] Rector A.L. Rogers J.E. (2006)Ontological issues in using a description logic to represent medical concepts: Experience from GALEN , In *Methods of Information in Medicine* , Springer Verlag, Jan. 2000
[16] Armstrong JD, Pocklington AJ, Cumiskey MA, Grant SGN. (2006) Reconstructing protein complexes: from proteomics to systems biology. *Proteomics* 6, 4724 - 4731.
[17] Nonaka, I. and Nishiguchi, T. (eds.) (2001). Knowledge Emergence: Social technical and Evolutionary Dimensions of Knowledge Creation, Oxford University Press, Oxford.
[18] Abels S., and Haak L. and Hahn A., (2005) Identification of Common Methods use for Ontology Integration Task, In Interoperability Of Heterogeneous Information Systems, *Proceedings of the first international workshop on Interoperability of heterogeneous information systems*, (2005) Bremen.
[19] De Roure D., Jennings N., and Shadbolt N., (2003) *Research Agenda for the Semantic Grid, In Grid Computing: Making the Global Infrastructure a Reality*, (Eds) Berman F., Hey AA.J.G. and Fox, G., John Wiley and Sons pp 437-470 2001 to 2003.
[20] Shneiderman B.(1998) Codex, memex, genex: The pursuit of transformational technologies. CHI98, Plenary Address. *Conference on Human Factors in Computing Systems*, Los Angeles, CA, April 1998.
[21] Kling R., McKim G., and King A. (2003) A Bit More To IT: Scholarly Communication Forums as Socio-Technical Interaction Networks. In *Journal of the American Society for Information Science and Technology 54(1)*, 47-67.
[22] Ure J. & Jaegersberg G., (2005) Invisible Architecture,: the benfits of aligning people, processes and technology- case studies for system designers an managers. BCS, ISBN 1-90250-5-59-X
[23] Sawhney, M. & Parikh, D. (2001). Where Value Lives in a Networked World, *Harvard Business Review*, January, pp175-198.
[24] Procter R., Borgmann C., Bowker G., Jirotka M., Olson G., Pancake C., Rodden T., and Schraefel M.C., (2006) Usability research challenges for cyberinfrastructure and tools, In *Conference on Human Factors in Computing Systems CHI '06 extended abstracts on Human factors in computing systems* pp: 1675 - 1678 ISBN:1-59593-298-4
[25] Berriman G.B., Deelman E., Good J., Jacob J. C. Katz D.S. Laity A.C., Prince T.A. Singh G., Su M. (2007) Generating Complex Astronomy Workflows. In Taylor, I.J., Deelman, E., Gannon, D.B. and Shields, M. (eds.) *Workflows for e-Science: Scientific Workflows for Grids.* Springer, 2007. pp. 19-38
[26] Wilson P., and Lessens V., (2006) Rising to the Challenge of e-Health Across Europe's Regions. *In eHealth2006*, May, 2006, Malaga
[27] Lin Y. Procter R., Randall D., Rooksby J., Sharrock W., Ure J., Voss A. (2007) Issues in Ontology Development and Use. (forthcoming)
[28] Arús, C., Celda, B., Dasmahaptra, S., Dupplaw, D., González-Vélez, H., van Huffel, S., Lewis, P., Lluch i Ariet, M., Mier, M., Peet, A., and Robles, M. On the design of a web-based decision support system for brain tumour diagnosis using distributed agents. In *WI-IATW'06: 2006 IEEE/WIC/ACM Int. Conf on Web Intelligence & Intelligent Agent Technology (Hong Kong, Dec. 2006), IEEE*, pp. 208-211

From Genes to Personalized HealthCare: Grid Solutions for the Life Sciences
N. Jacq et al. (Eds.)
IOS Press, 2007

Roadmap for a European Healthgrid

Vincent BRETON[a,1], Ignacio BLANQUER[b], Vicente HERNANDEZ[b], Nicolas JACQ[c], Yannick LEGRE[a], Mark OLIVE[d] et Tony SOLOMONIDES[d]
[a] *Corpuscular Physics Laboratory of Clermont-Ferrand, CNRS-IN2P3, France*
[b] *Universidad Politecnica de Valencia, Spain*
[c] *HealthGrid, France*
[d] *University of the West of England, United Kingdom*

Abstract. This paper proposes a 10-year roadmap to achieve the goal to offer to healthcare professionals an environment created through the sharing of resources, in which heterogeneous and dispersed health data as well as applications can be accessed by all users as a tailored information providing system according to their authorisation and without loss of information. The paper identifies milestones and presents short term objectives on the road to this healthgrid.

Keywords. Grids, medical research, healthcare, medical informatics, medical imaging

Introduction

The concept of grids for health was born in Europe in 2002 and has been carried forward through the HealthGrid initiative [1]. This European initiative has edited, in collaboration with CISCO, a white paper setting out for senior decision makers the concept, benefits and opportunities offered by applying newly emerging Grid technologies in a number of different applications in healthcare [2]. Starting from the conclusions of the White Paper, the EU funded SHARE project aims [3] at identifying the important milestones to achieve the wide deployment and adoption of healthgrids in Europe. The project will devise a strategy to address the issues identified in the action plan for a European e-Health [4] and set up a roadmap for technological developments needed for successful take up of healthgrids in the next 10 years.

In a previous paper [5], we presented an analysis of the adoption of grids for biomedical sciences and healthcare in Europe, identifying bottlenecks and proposing actions. These actions have been further assessed within the framework of the SHARE European. The present paper proposes a technical roadmap for the adoption of healthgrids for medical research in Europe.

1. The Goal

Our goal is to offer to healthcare professionals an environment created through the sharing of resources, in which heterogeneous and dispersed health data as well as applications can be accessed by all users as a tailored information providing system

[1] Correspondence: breton@clermont.in2p3.fr

according to their authorisation and without loss of information. The environment should allow multiple usages depending on the user community and offer to any groups involved in medical and life sciences research means to share information and access resources on demand. Persistent infrastructures are needed for the users who are looking for fully functional services for medical research and aim at scientific production. For the users who are designing distributed applications and projects, toolkits are also needed to enable use of the Grid in a secure, interoperable and flexible manner. These toolkits should allow creating virtual organizations, administering it, manipulating biological and medical data, and interfacing with the infrastructure resources and services whenever needed. The environment should be open and evolving: for those who wish to integrate resources, tools and content, standard specifications should be documented.

The infrastructure itself should give access through standard interfaces to public and private, free and non free computing and storage resources; data stored in public and private databases. The data accessible on the environment include molecular data (ex. genomics, proteomics), cellular data (ex. pathways), tissue data (ex. cancer types), personal data (ex. EHR), population (ex. epidemiology); services for medical research such as analysis tools (workflow, data mining); services for on line patient care (telemedicine, visualization), For the majority of the healthcare professionals, the technical aspects should be completely hidden behind a friendly user interface. As stated above, our goal is to offer services to the healthcare professionals but legal issues may very significantly delay the deployment of healthgrids for healthcare. As a consequence, we are not sure how relevant a purely technical roadmap like the one presented in this paper is for healthcare. As a consequence, we expect the technology to get adopted primarily for medical research. Our roadmap will therefore refer specifically to medical research.

2. The starting point

The state of the art of healthgrids has been described in several documents [2,5,6,7]. We summarize here the main conclusions reached by their authors. Large scale grid infrastructures are available for scientific production in Europe; these infrastructures offer unprecedented opportunities for intensive computing. Several toolkits are now available which start to offer grid services in a secure, interoperable and flexible manner; these toolkits have still to be tested at a large scale on biomedical applications. Europe has witnessed in the last few years the emergence of e-science environments such as ^{my}Grid [8] or VL-e [9]; these environments are science portals for distributed analysis which offer scientists the possibility to carry out their experiments in a familiar environment while using the most recent web service technology and developments. Successful deployment of CPU intensive biomedical applications has been convincingly achieved on grid infrastructures world wide and several projects in Europe and world wide are successfully deploying biomedical grid applications; most of the success stories involve groups which have been active in grid projects for more than 5 years, which shows the present difficulty to use these environments for scientific production. The successful deployments in Europe are almost strictly limited to CPU intensive applications while very few applications involving manipulation of distributed medical data have been demonstrated so far; those were done at a prototype level by recognized team of computer experts in relation to clinicians.

The documents also highlighted several issues which had to be addressed in order to enable the healthgrid vision.

First issues are grid technology: web services are identified as the most promising technology to enable the HealthGrid vision despite their present limitations. There are today many different grid middleware (Globus, gLite, SRB, Unicore, GRIA, etc.) but none of them fulfil the requirements for a healthgrid. The ones which have demonstrated their scalability (gLite, Unicore) have limited functionalities particularly in the area of data management. Some which offer powerful and demonstrated data management functionalities (SRB) do not provide job management services. Moreover, these middleware are not so far built on web services and therefore do not offer standard interfaces. More recent grid middleware based on web services have not yet demonstrated their robustness and scalability.

Second issues are grid deployment: the deployment of grid nodes in healthcare centres such as hospitals or medical research laboratories is still extremely limited. This is due mainly to the present limitations of the existing European grid middleware which do not offer the necessary functionalities for secure manipulation of medical data. Other present limitations include the absence of an easy to install middleware distribution and the lack of friendly user interfaces to the grid for non experts. Just like Windows allows users to use their PC without any knowledge of the operating system, there is a need for user friendly interfaces to the grid.

Third issues are standardization: the definition and adoption of international standards and interoperability mechanisms is required for storing medical information on the grid. This includes for instance recording and ensuring consent, anonymization and pseudonymization.

Fourth issues are communication: there is an evident lack of information on the grid technology in the biomedical community. Communication on grids has been mostly focussed on the particle physics and computer sciences academic communities. Raising the awareness of the biomedical community is a major challenge for the coming years.

3. The road to grid infrastructures for medical research

In the previous section, we have listed a number of issues which need to be addressed in order to build the environment we have set up as our goal. In the following, we focus this study on medical research where adoption of grids is less depending on the evolution of EC legal texts. On the road to healthgrids, we have identified milestones which correspond to important steps forward in the services offered to the medical research community. We need to stress at this point that we are strongly convinced that a bottom-up approach has to be adopted on the road to healthgrids. Starting from the services made available on the existing grid infrastructures, a persistent distributed environment for medical research has to be built. This environment will progressively be enriched with new functionalities as technology progresses. In the United States, the Biomedical Informatics Research Network [1] is a very successful example of this bottom-up approach. In parallel to the building of this persistent environment, more volatile projects using grid toolkits or e-science environments to manage distributed data and knowledge for specific medical applications are very important to develop high level data integration services and spread the grid culture in the medical community. These services will later be made available on the grid infrastructures through standardized interfaces.

3.1. Milestones

Grid deployment is still very limited in European healthcare centres. Several factors contribute to this situation. One is the present human cost for deploying grid elements. Another important factor is the present situation in hospitals where none of the resources is accessible from the outside world. Our first deployment milestone called **MD1** is the **"successful permanent deployment of computing grid nodes inside European medical research centres"**. The goal of this first deployment is to create an environment where groups active in medical research can find resources for large scale simulation and modelling. This requires the development of friendly user interfaces. As a consequence of the present organization, a grid node in a hospital would have to be located outside the hospital firewall. Only anonymized or pseudonymized data could be transferred from inside the hospital firewall to the grid node. Technical solutions exist but they have not been deployed yet. These issues must be addressed and successful deployment of grid nodes inside European healthcare centres must be achieved. The services made available on these nodes are dependent on the technology. As documented in [6], the presently available grid infrastructures are mostly offering services for large scale computing.

As documented in [6], some data management services are emerging which are relevant to manipulate medical images. These services have to reach their full maturity for the second milestone called **MD2**, "**Successful permanent deployment of data grid nodes inside European medical research centres**". The challenge here is to store medical data on the grid and this will only be achieved once the grid middleware will allow doing it in a secure way.

Once medical data are securely stored on the grid, the next issue is to deploy the services to query these data, to build relationships between and to provide appropriate representation to the researchers and to healthcare professionals. Building relationships between these data requires agreeing on standards for representing and storing them, in order to develop knowledge management services to manipulate the data. Once the standards are defined and widely adopted, the next grid deployment milestone **MD3** can be achieved, namely a **"Successful permanent deployment of knowledge grid nodes inside European medical research centres**". The difference here between the data grid and the knowledge grid is that data stored on the data grid are just exploited through simple queries while a knowledge grid offer services to manipulate concepts while ignoring the underlying data model and the grid storage architecture.

3.2. Reaching the milestones

For each of the milestones described above, we need the right technology. We are going to discuss now the technological issues as well as the standardization issues which have to be addressed on the road to healthgrids.

3.2.1. Grid technology issues

We need a grid operating system that allows handling distributed computing for milestone MD1, distributed data storage and query for milestone MD2, data integration and knowledge management for MD3. There are definitely progresses being made in this direction and several software stacks are now available which provide relevant services based on OGSA (GT4 [10], OGSA-DAI [11], GRIA [12], etc). Beside the fact

that these toolkits are not yet offering all the functionalities required by a grid environment for healthcare, a major issue is that most of them have not yet been tested on medical applications and/or at a significant scale. On the other hand, several academic consortia are developing grid middleware which are deployed on the large scale infrastructures in Europe. These consortia are aware of the requirements of the medical community but these middleware are not yet based on web services because their development effort started before the migration of grid standards to web services. These two trends need to converge to deliver a "**tested, robust and scalable grid middleware based on web services that allow job and data management and which complies with EC countries laws on manipulation of personal medical data".** Such a middleware is not for tomorrow and we rather foresee a progressive evolution of the grid operating systems similar to the one observed for personal computers.

Once a grid operating system is developed, a crucial issue is to make it easy to install in a healthcare centre. The availability of a public distribution of the middleware allowing a quick installation and configuration of the grid elements is required. This distribution requires significant resources because a middleware is very complex software with multiple modules and dependencies. Moreover, the technology is still evolving quickly as its standards are still under definition and only few implementations of these standards are now available. As a consequence, the distribution will have to be regularly updated to adapt to this rapid evolution path and be kept backward compatible so that sites using old versions are still able to belong to the same virtual organization as sites configured with newer versions of the distribution. In summary, a major challenge for the development of healthgrids is the **"availability of a free, easy to install and configure, robust and documented distribution of the grid middleware, accompanied by a significant user support"**.

3.2.2. Standardization issues

As already mentioned above, standards are absolutely necessary to the deployment of services which integrate data in bioinformatics and medical informatics, data coming from different medical disciplines and data coming from different countries in Europe. The adoption of standards for the exchange of biological and medical information is still limited to a few specific fields. These standards are needed to build data models, to produce ontologies and to develop knowledge management services. Moreover, they need to be compatible with grid standards so as to allow their implementation on the healthgrids. The largest initiatives in the medical informatics field such as DICOM and HL7 are just starting to study the interface between their standards and web services technology. We suggest focussing the effort on two topics, medical imaging and Electronic Health Records, in a first stage. In view of the importance of these standardization efforts, we introduce standardization milestones **MS1** corresponding to the **"Production of a standard for the exchange of medical images on the grid based on DICOM**", and **MS2** corresponding to the **"Production of a standard for the exchange of Electronic Healthcare Records on the grid**" compatible with HL7.

3.2.3. Communication issues

The lack of information on grids is frequently identified as one of the key reasons why they have raised so far very little interest in the field of medical research. On the other hand, the services offered by the grids have been too limited to really make them a serious alternative to the existing computing models. A convincing communication

relies on success stories demonstrating the impact of grids for medical research. So we identify the need for **a demonstration environment offering a very easy access to the grid for non experts and providing some convincing services for medical research.** The evolution of this demonstration testbed should follow the evolution of the healthgrid through the different deployment milestones. On this dissemination environment, dedicated efforts to promote the technology can be developed.

3.2.4. Security issues

In this document, we are not dissociating the security issues from the other technological issues because they have to be addressed from the lowest middleware layers. Deployment of a data grid for medical research discussed above will only be possible when the middleware will be able to provide all the necessary guarantees in terms of management of personal data. Here is what we perceived as the specific technical challenges related to the handling of medical data on the grid: manipulation of personal data on the grid must strictly obey regulations; these regulations change from country to country in Europe. Services for anonymization and pseudonymization of medical data must be provided. Medical data belong to the patient; a mechanism must be set up to allow any given individual to access to all his/her data or the grid. In the perspective of the usage of grids for healthcare, authentication of health professionals on the grid can not be handled by requesting all of them to get a grid certificate; a mechanism must be set up so that professional cards can be used to provide authentication on the grid.

3.2.5. Summary

We have identified 5 milestones on the road: MD1, called “Computing grid”, corresponding to the successful permanent deployment of computing grid nodes inside European medical research centres; MD2, called “Data grid”, corresponding to the successful permanent deployment of data grid nodes inside European medical research centres; MD3, called “Research K-Grid”, corresponding to the successful permanent deployment of knowledge grid nodes inside European medical research centres; MS1, called “Grid DICOM”, corresponding to the production of a standard for the exchange of medical images on the grid; MS2, called “Grid EHR”, corresponding to the production of a standard for the exchange of Electronic Healthcare Records on the grid.

The milestones MD1, MD2, MD3 correspond to the deployment of infrastructures while MS1 and MS2 are related to the availability of standards. The figure 1 illustrates how the different milestones follow each other on the road and how the progress on the road will depend on the availability of the grid operating system as well as its distribution. For each of the three deployment milestones, computing grid, data grid and knowledge grid, the figure illustrates how two environments are needed, one for scientific production and one for demonstration.

We estimate to about 10 years the time necessary to reach the goal of deploying an healthgrid after successfully achieving the different milestones discussed previously. The technology available today allows reaching the first milestone MD1 today. We estimate about 2 years are needed to achieve MD2 while at least 5 years will be needed for MS1, MS2 and MD3. Finally, our experience shows that about three years will be needed from the day a first knowledge grid is deployed to the day it is robust enough to become an environment for scientific production.

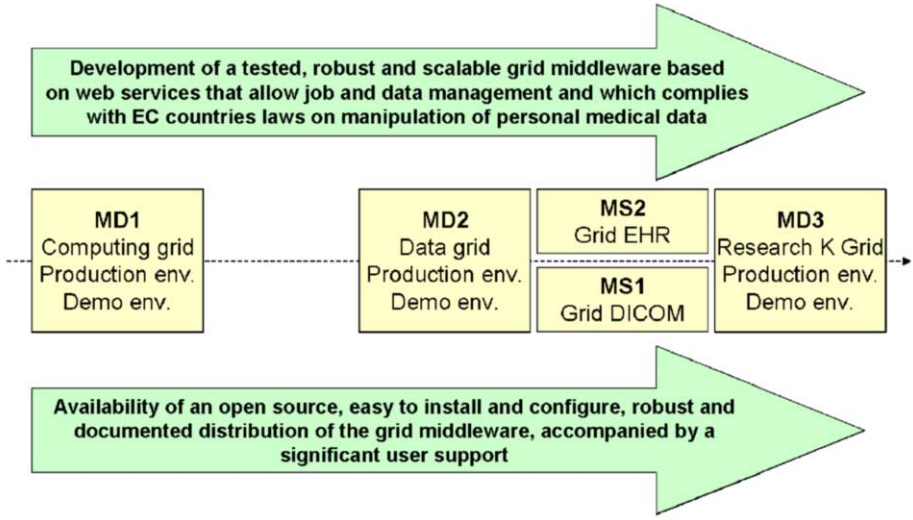


Figure 1. Illustration of the key challenges and the milestones on the road to healthgrids

3.3. Risk analysis

There are three keys to the successful progress on the technical road to healthgrids: the evolution of the technology, including development and distribution of the needed grid middleware; the deployment of stable infrastructures with guaranteed level of services; the deployment of medical applications on these infrastructures. As a consequence, we identify three key technical risks to achieve the vision: the absence of concentration of the critical mass of expertise to develop the grid middleware and its distribution; the absence of agreed standards to share medical images and Electronic Health Records on the grid; the non-adoption of the healthgrid infrastructures by the research community so that they are not used by medical applications. To these technical risks, we add a fourth major risk which is the absence of evolution of the legal texts to allow sharing of medical data: this could completely prevent the deployment of an healthgrid.

4. Short term roadmap (2-3 years)

Successful achievement of the 10-year target depends on immediate actions. We recommend developing R&D activities along three lines:

1. Develop healthgrid infrastructures
2. Deploy biomedical grid applications on the existing infrastructures;
3. Deploy biomedical grid applications using OGSA compliant grid toolkits and e-science environments.

Two out of the three lines (2 and 3) are now actively pursued around the world as biomedical applications are deployed on almost all grid infrastructures and many projects are now under development using web services toolkits or e-science environments. The three lines are needed because there is no dedicated grid environment for medical research and there are no OGSA compliant grid toolkits and e-science environments available on the existing infrastructures. However a convergence of these research axes should be achieved in about 2 to 3 years when the middleware deployed on the healthgrid and on the grid infrastructures will offer web

service interfaces to their grid services so that the grid toolkits and e-science environments will be available to all healthgrid users. In parallel to these R&D activities, we recommend to pursue actively the definition of standards for the sharing of medical images and electronic health records on the grid. Rather than developing new standards, grid experts should as much as possible get involved in the already existing medical informatics standardization bodies.

4.1. Development of healthgrid infrastructures

Having a dedicated infrastructure for medical research is a key to the adoption of grids for medical research. It also solves the problem of handling priorities on a multidisciplinary grid where the services made available have been chosen through a consensus process and specific requirements for the medical community like security issues have low priority for other research communities. This dedicated infrastructure should deploy the most up to date services relevant to biomedical research as soon as they have demonstrated their scalability and robustness. The infrastructure will initially mostly focus on computing services, but should aim at deploying very rapidly secure data management services and later knowledge management tools. In order to offer immediate service to the community, the healthgrid should be built using technologies which are interoperable with the existing infrastructures while keeping the perspective to offer as soon as possible web services interfaces to the grid services.

4.2. Deployment of biomedical grid applications on the existing infrastructures

One can wonder why it is interesting to keep deploying biomedical applications on the existing multidisciplinary infrastructures once an healthgrid is running. The first reason is that one should not reinvent the wheel and take advantage of the services already offered by infrastructures like EGEE [13] and DEISA [14]. The second reason is that the amount of resources and storage available on the healthgrid will hardly ever compare to what EGEE and DEISA provide. The healthgrid should provide the interface and the high level services needed for medical research but whenever there is a need for heavy CPU or storage resources, mechanisms should be available to use EGEE or DEISA to address these needs. As a consequence, through the collaboration between the developers of both healthgrid and infrastructure projects, tools should be designed to achieve this submission in the most transparent way. This means also that deployment of biomedical grid applications will increase on infrastructures like DEISA and EGEE once the healthgrid is deployed. These infrastructures should offer the largest possible palette of services relevant to medical research and exposed through web services interfaces. As a consequence, a continuous effort must be maintained to develop the synergy between the projects and to improve the environment for the deployment of biomedical grid applications on these infrastructures.

4.3. Deployment of biomedical grid applications using OGSA compliant grid toolkits and e-science environments

The services offered for medical research evolve with the grid operating system technology and we expect therefore to witness the evolution of the applications deployed with the progress of the middleware. On the other hand, most of the present European biomedical grid projects require already a high level of data integration

which is beyond the capacities of the existing grid infrastructures. Nowadays, we observe a gap between the needs of these projects and the services made available on the infrastructures. How to address this gap is an important issue. The roadmap exposed above presents the bottom-up approach where infrastructures will progressively offer new services with the evolution of the technology. Toolkits such as OGSA-DAI or GRIA are very relevant to address the specific needs of medical grid projects and their use should be strongly encouraged. As well, high level environments such as [my]Grid or VL-e allow now biologists and healthcare professionals to start manipulating concepts they are familiar with while accessing potentially distributed data. The next step is to achieve the deployment of these environments on grid infrastructures and evolve them towards improved usability by end users.

4.4. Development of standards

The R&D activities described above are going to require the manipulation of medical data on the grid. As a consequence, standards for the exchange of medical data will be more and more required. At all costs, grid experts should avoid developing new standards but rather should as much as possible get involved in the already existing medical informatics standardization bodies. There, they can disseminate the concept of grids and work on extending the existing standards to distributed environments. The Open Grid Forum is a potentially very interesting place where grid experts involved in the different medical informatics standardization bodies could meet and investigate the interface between these standards and the existing grid standards. However, this would require a deep reorganization of the present Life Sciences Research group at OGF which is not working properly. The Healthgrid initiative provides the right framework to coordinate the development of the different standards in collaboration with the OGF and the different medical informatics standardization bodies.

5. Conclusion

In this document, we are proposing a 10-year roadmap to achieve the goal to offer to healthcare professionals an environment created through the sharing of resources, in which heterogeneous and dispersed health data as well as applications can be accessed by all users as a tailored information providing system according to their authorisation and without loss of information. Starting from the state of the art on healthgrids described in several documents [2,5,6,7], we have described a way toward the environment. We have identified 5 milestones on the road:

- MD1, called “Computing grid”, corresponding to the successful permanent deployment of computing grid nodes inside European medical research centres,
- MD2, called “Data grid”, corresponding to the successful permanent deployment of data grid nodes inside European medical research centres,
- MD3, called “Research K-Grid”, corresponding to the successful permanent deployment of knowledge grid nodes inside European medical research centres,
- MS1, called “Grid DICOM”, corresponding to the production of a standard for the exchange of medical images on the grid,
- MS2, called “Grid EHR”, corresponding to the production of a standard for the exchange of Electronic Healthcare Records on the grid.

Achieving these different milestones require the availability of a grid operating system providing all the needed functionalities as well as an easy-to-install distribution of this middleware.

We have identified major technical risks which can prevent the vision to happen:

- the absence of concentration of the critical mass of expertise to develop the grid middleware and its distribution,
- the absence of agreed standards to share medical images and Electronic Health Records on the grid,
- the non-adoption of the healthgrid infrastructures by the research community so that they are not used by medical applications,

To these technical risks, we have added a fourth major risk which is the absence of evolution of the legal texts to allow sharing of medical data which can prevent the deployment of an healthgrid.

Finally, we have proposed in the next 2 to 3 years to develop R&D activities along three lines:

- Develop healthgrid infrastructures,
- Deploy biomedical grid applications on the existing infrastructures,
- Deploy biomedical grid applications using OGSA compliant grid toolkits and e-science environments.

In parallel to these R&D activities, we recommend to pursue actively the definition of standards for the sharing of medical images and electronic health records on the grid in the already existing medical informatics standardization bodies. We consider that the Healthgrid initiative provides the right framework to coordinate the development of the different standards in collaboration with the OGF and the different medical informatics standardization bodies.

6. Acknowledgements

The SHARE project is co-funded by the European Commission under contract n°FP6-2005-IST-027694.

7. References

[1] Information on Healthgrid initiative available at http://www.healthgrid.org.
[2] V. Breton, et al., editors on behalf of the Healthgrid White Paper collaboration, The Healthgrid White Paper, Proceedings of Healthgrid conference, IOS Press, Vol 112, 2005.
[3] Information on Share project available at http://www.eu-share.org.
[4] Action plan for a European e-Health Area, COM(2004) 356, European Commission, http://europa.eu.int/information_society/doc/qualif/health/COM_2004_0356_F_EN_ACTE.pdf.
[5] V. Breton, et al., Proposing a roadmap for Healthgrids, Stud Health Technol Inform. 120:319-29., 2006.
[6] Healthgrid technology baseline report, Share deliverable D3.2 available from http://www.eu-share.org.
[7] V. Breton, et al., HealthGrid, a new approach to eHealth, proceedings of the eHealth 2006 conference, Malaga, 2006.
[8] R.D. Stevens et al., MyGrid, personalised bioinformatics on the information grid, Bioinformatics 191(1) i302-i304, 2003.
[9] H. Rauwerda, et al, The Promise of a virtual lab, Drug Discov. Today. 11(5-6):228-36, 2006.
[10] I. Foster, Globus Toolkit Version 4: Software for Service-Oriented Systems, IFIP International Conference on Network and Parallel Computing, Springer-Verlag LNCS 3779, 2-13, 2005.
[11] Information on OGSA-DAI middleware available at http://www.ogsadai.org.uk.
[12] Information on GRIA middleware available at http://www.gria.org.
[13] F. Gagliardi, et al., Building an infrastructure for scientific Grid computing: status and goals of the EGEE project, Philosophical Transactions: Mathematical, Physical and Engineering Sciences, 363, 1729-1742, 2005 and http://public.eu-egee.org.
[14] Information on DEISA available at http://www.deisa.org.

From Genes to Personalized HealthCare: Grid Solutions for the Life Sciences
N. Jacq et al. (Eds.)
IOS Press, 2007

SHARE Roadmap 1: Towards a debate

Mark OLIVE [a], Hanene RAHMOUNI [a], Tony SOLOMONIDES [a,1], Vincent BRETON [b], Yannick LEGRÉ [b], Ignacio BLANQUER [c], Vicente HERNANDEZ [c], Isabelle ANDOULSI [d], Jean HERVEG [d] and Petra WILSON [e]

[a] *CCCS / CEMS Faculty / UWE, Bristol / Coldharbour Lane / Bristol BS16 1QY / UK*
[b] *LPC / CNRS-IN2P3 / Campus de Cézeaux / 63177 Aubière Cedex / France*
[c] *Universidad Politécnica de Valencia / Camino de Vera s/n / E-46022 Valencia / Spain*
[d] *Centre de Recherches Informatiques et Droit / NAMUR / Belgique*
[e] *European Health Management Association (now at CISCO)*

Abstract. We present the 'HealthGrid' initiative and review work carried out in various European projects. Since the European Commission's Information Society Technologies programme funded the first grid-based health and medical projects, the HealthGrid movement has flourished in Europe. Many projects have now been completed and 'Healthgrid' consulted a number of experts to compile and publish a 'White Paper' which establishes the foundations, potential scope and prospects of an approach to health informatics based on a grid infrastructure. The White Paper demonstrates the ways in which the healthgrid approach supports many modern trends in medicine and healthcare, such as evidence-based practice, integration across levels, from molecules and cells, through tissues and organs to the whole person and community, and the promise of individualized health care. A second generation of projects have now been funded, and the EC has commissioned a study to define a research roadmap for a 'healthgrid for Europe', seen as the preferred infrastructure for medical and health care projects in the European Research Area.

Keywords. healthgrid, e-health, grid applications

Introduction

Grid technology has been identified as a key technology to enable and support the 'European Research Area'. The impact of this concept is expected to reach far beyond eScience, to eBusiness, eGovernment, and eHealth. However, a major challenge is to take the technology out of the laboratory to the citizen. A *healthgrid* is an environment in which data of medical interest can be stored and made easily available to different actors in the healthcare system, such as physicians, healthcare centres, patients and citizens in general. Such an environment has to offer all appropriate guarantees in terms of data protection, respect for ethics and observance of regulations; it has to support the notion of 'duty of care' and may have to deal with 'freedom of information issues'. Working across member states, it may have to support negotiation and policy bridging.

Early grid projects, while encompassing potential applications to the life sciences, did not address the specificities of an e-infrastructure for health, such as the

[1] Corresponding Author: Tony Solomonides, Tony.Solomonides@uwe.ac.uk. The work reported here has been carried out in collaboration with many colleagues in the HealthGrid community Thanks are due to these and numerous other colleagues in Europe, the US and Asia for helpful discussions.

deployment of grid nodes in clinical centres and in healthcare administrations, the connection of individual physicians to the grid and the strict regulations ruling the access to personal data. However, a community of researchers did emerge with an awareness of these issues and an interest in tackling them.

1. The Healthgrid Initiative

Pioneering projects in the application of grid technologies to the health area have recently been completed, and the development of technology to address high level requirements in a grid environment has been making good progress. Because these projects had a finite lifetime and the healthgrid vision required a sustained effort over a much longer period, and because there was a need for these projects to cross-fertilise, the HealthGrid association (http://www.healthgrid.org), was established to bring about the necessary long-term continuity. Its goal is to encourage and support collaboration between autonomous projects in such a way as to ensure that requirements really are met and that the wheel is not re-invented at the expense of other necessary work.

Writing about the healthgrid initiative very soon after its inception, this community identified a number of objectives: identification of potential business models for medical grid applications; feedback to the grid development community on the requirements of the pilot applications deployed by European projects; development of a systematic picture of the broad and specific requirements of physicians and other health workers when interacting with grid applications; dialogue with clinicians and those involved in medical research and grid development to determine potential pilots; interaction with clinicians and researchers to gain feedback from the pilots; interaction with all relevant parties concerning legal and ethical issues identified by the pilots; dissemination to the wider biomedical community on the outcome of the pilots; interaction and exchange of results with similar groups worldwide; and the formulation and specification of potential new applications in conjunction with the end user communities. (Cf [1] for a brief history and [2] for a panoramic vision statement.)

A healthgrid may in theory be deployed to support the full range of healthcare activities, from screening through diagnosis, treatment planning to epidemiology and public health. Thus HealthGrid has been actively involved in the definition of requirements relevant to the development and deployment of grids for health and was among the first to identify the need for a specialist middleware layer, between the generic grid infrastructure and middleware and the medical or health applications.

2. The White Paper: From Grid to Healthgrid

The White Paper [3] defines the concept of a healthgrid more precisely than before: The ultimate goal for eHealth in Europe would be the creation of a single healthgrid, i.e. a grid comprising all eHealth resources, incorporating a ‘principle of subsidiarity’ of independent nodes of the healthgrid as a means of implementing all the legal, ethical, regulatory and negotiation requirements. We may anticipate, however, the development path to proceed through specific healthgrids with perhaps rudimentary inter-grid interaction/interoperational capabilities. Thus, we may identify a need to map future research and advice on research policy, so as to bring diverse initiatives to the point of convergence.

Healthgrid applications address both individualised healthcare and epidemiology with a view to public health. Individualised healthcare is improved by the efficient and secure combination of immediate availability of personal clinical information and widespread availability of advanced services for diagnosis and therapy. Epidemiology healthgrids combine information from a wide population to extract knowledge that can lead to the discovery of new correlations between symptoms, diseases, genetic features and other clinical data. With this broad range of application in mind, the issues below are identified as key features of our analysis.

- Business case, trust and continuity issues: Healthgrids are data- and collaboration grids, but large-scale deployment requires 'security' to be scaled up to a very high level of confidence. Federation of databases introduces additional complexity.
- Biomedical issues: Distributed databases and data mining capabilities, expert system services able to interrogate these, biocomputing, biomodelling and simulation have a strong need for resources that can be provided through the grid. Compliance with medical information standards is necessary.
- Security issues: Security in grid infrastructures is currently adequate for research platforms, but not for real healthcare applications.
- Management issues: The central concept of a 'virtual organisation' (VO) at the heart of eScience, which gave rise to grids, is very apt for healthgrid, but additional flexibility is needed to structure and to control VOs in the large, including, for example, the meta-level of a VO of VOs.

3. The SHARE Project: From White Paper to Road Map

In the White Paper, the HealthGrid community expressed its commitment to modern trends in medical practice, especially 'evidence-based medicine' as an integrative principle, to be applied across the dimensions of individual through to public health, diagnosis through treatment to prevention, from molecules through cells, tissues and organs to individuals and populations. In order to do this, it had to address the question how to collect, organise, and distribute the 'evidence'; this might be 'gold standard' evidence, i.e. peer reviewed knowledge from published research, or it might be more tentative, yet to be confirmed knowledge from practice, and, in addition, would entail knowledge of the individual patient as a whole person. The community also had to address the issues of law, regulation and ethics, and issues about crossing legal and cultural boundaries, expressing these in technological terms – security, trust, encryption, and pseudonymisation. Then it had to consider how healthgrid middleware services would satisfy these requirements; and, if it was to succeed in the real world, how to make the business case for healthgrid to hard-pressed health services across Europe while they are struggling with their own modernisation programmes.

The vision of health that informs the White Paper and the work of Healthgrid since has been defined in the 'Action Plan for a European e-Health Area' [4] as follows:

"… the application of information and communications technologies across the whole range of functions that affect the health sector. e-Health tools or 'solutions' include products, systems and services that go beyond simply Internet-based applications. They include tools for both health authorities and professionals as well as personalised health systems for patients and citizens. Examples include health

information networks, electronic health records (EHR), telemedicine services, personal wearable and portable communicable systems, health portals, and many other information and communication technology-based tools assisting prevention, diagnosis, treatment, health monitoring, and lifestyle management."

The 'vertical integration' implicit in this visionary statement can be translated into more concrete terms by mapping it to its human subjects, their pathologies and the implicit disciplines. The relationships between the different ontological and epistemological levels and the various modalities of data have been captured by Fernando Martin-Sánchez [2] in the schematic diagram below (see fig 1).

In the light of the White Paper and its impact, the EC has funded SHARE, a 'specific support action' project to explore how to realise the vision of the White Paper: the two objectives of the project are:

- a roadmap for research and technology to allow a wide deployment and adoption of healthgrids both in the shorter term (3-5 years) and in the longer term (up to 10 years); and
- a complementary and integrated roadmap for e-Health research and technology development (RTD) policy relating to grid deployment, as a basis for improving coordination amongst funding bodies, health policy makers and leaders of grid initiatives, avoiding legislative barriers and other foreseeable obstacles.

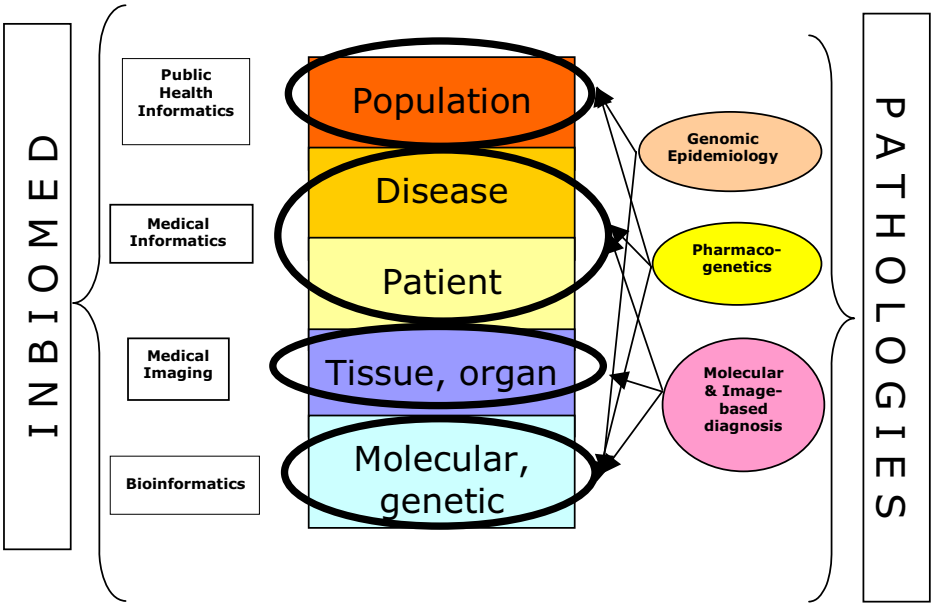


Fig 1 Disciplines, levels of being and pathology diagnostics (F. Martin-Sánchez)

The project must address the questions, *what* research and development needs to be done now? —and *what* are the right initiatives in eHealth RTD policy relating to grid deployment? —with all that implies in terms of coordination of strategy,

programme funding and support for innovation. In summary, therefore, the project will define a comprehensive European research and development roadmap, covering both technology and policy aspects, to guide EU-wide uptake of healthgrid technologies, and their applications into health research and into health care service provision.

4. Technical Road Map: Step One

At a first step SHARE has identified five key deployment and standardization challenges for healthgrids and proposes a series of interlaced milestones to address each of these. The milestones are as follow:

- MD1– deployment of computing grid nodes in medical research centres;
- MD2 – deployment of data grid nodes in medical research centres;
- MD3 – deployment of knowledge grid nodes in medical research centres;
- MS1 – production of a standard for exchange of medical images on the grid;
- MS2 – production of a standard for the exchange of EHR on the grid.

MD1, MD2, MD3 correspond to the deployment of infrastructures while MS1 and MS2 are related to the availability of standards.

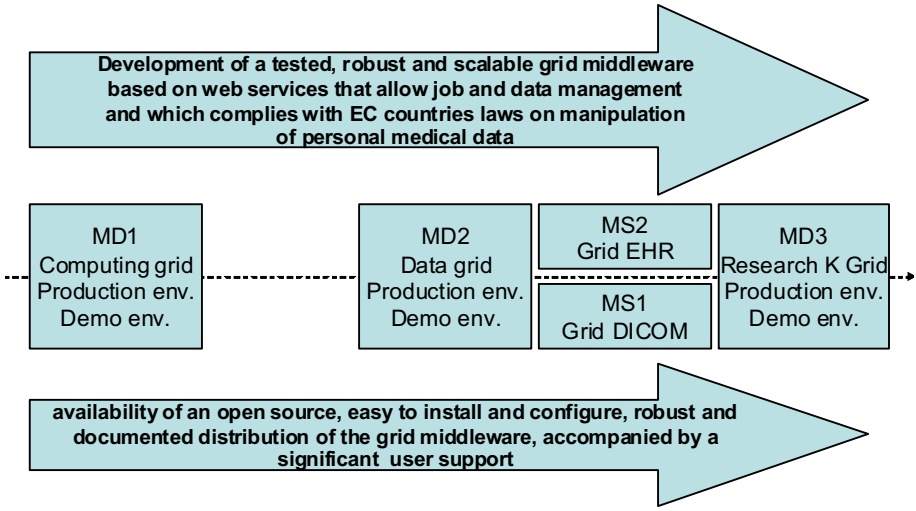


Figure 2.

5. Technical Road Map: Step Two

At a second step we have introduced some overlap to represent concurrency and a cyclical relationship between the standardisation milestones and the second deployment milestone of a data grid production environment. Further discussions have also led to considering “the development of a tested, robust and scalable grid middleware based on Web Services”, and “the availability of a robust, user friendly open source distribution of grid middleware with appropriate user support” as separate milestones.

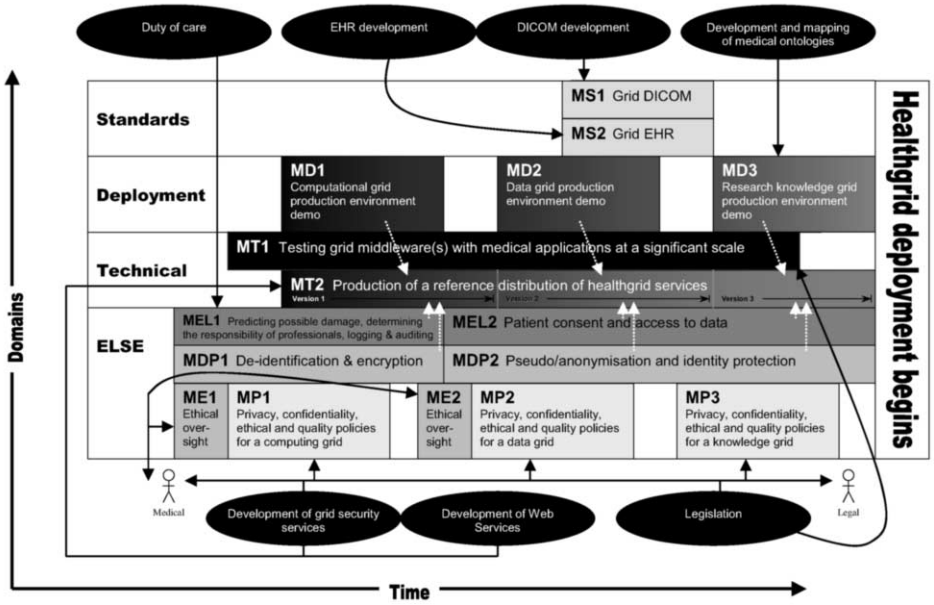


Figure 3.

6. Technology Milestones

MT1 – Testing healthgrids with mid-to-large medical applications
Testing of grid middleware(s) for scalability and robustness should begin at an early stage. It is anticipated that this will be an ongoing activity, with different generations of grid operating systems offering newer, faster and more stable capabilities.

MT2 – Production of a reference distribution of healthgrid services
Shortly after testing has begun, work should commence on the production of a reference distribution. Standards for web services and particularly grid security services are still evolving, and therefore new versions of the distribution should be made available as standards are adopted and implemented.

7. Standardization Milestones

Two standardisation milestones have been created for sharing medical images and records. As examples: MS1 – Grid DICOM and MS2 – Grid EHR respectively, but standards in other areas of bio- and medical informatics will be required. HealthGrid community must liaise with standards developers to define any necessary extensions.

8. Deployment Milestones

MD1 – A computational grid production environment
This initial deployment step would at first seem to require little effort as other projects have attempted it before. However, there are a number of issues associated with the installation of grid nodes in hospitals and medical research centres, including economic factors, hiding the complexity of grid mechanisms from users, and security concerns.

MD2 – A data grid production environment
Although several prototype data grids for medical research have been demonstrated by healthgrid projects, developing and maintaining a production quality data grid will require a number of issues relating to the distributed storage of medical data to be resolved.

MD3-- A knowledge grid production environment
The next task will be to deploy services that can build relationships between data items, and will provide appropriate representation to medical researchers. The development of medical ontologies and the mapping between ontologies will be particularly important for the successful deployment of knowledge grids.

9. Legal and Liability Considerations

The effort to develop healthgrids should not only be directed at the technical and deployment issues. Other challenges also arise which are linked to the legitimacy and ethics of using such a technology and its social impact on the health work-space.

Medical Data Processing between ban and permission
In the healthcare sector, the use of the grid technology implies automated processing of health personal data, especially for therapeutic purposes. The key European principles relevant to the processing of personal data were first established by the Council of Europe and further developed in Directive 95/46/CE of the European Union[2]. The purpose of the Directive is to allow the free flow of personal data between member states of the European Union, to facilitate the internal market and to protect the fundamental rights and freedoms of natural persons, in particular their right to privacy with respect to the processing of their personal data.

The Directive bans the processing of sensitive or medical data, but the principle of this ban is not absolute; some texts in the Directives tend to grant permissions to the processing of personal data under exceptional conditions. For example, the explicit and valid consent of the data subject constitutes the most important source of legitimacy in the processing of medical data although it is subject to very strict conditions for its validity and the data subject may revoke this consent to the processing of his or her medical data at any time and without justification. Processing is allowed if

- it is necessary for the purposes of carrying out the obligations and specific rights of the controller in the field of employment law in so far as it is authorized by national law providing for adequate safeguards;

[2] For references to statutes, directives and other legal references, see SHARE Deliverable 4.2 at [5].

- it is necessary to protect the vital interests of the data subject or of another person when the data subject is physically or legally incapable of giving his or her consent;
- it is carried out in the course of its legitimate activities with appropriate guarantees by a foundation, association or any other non-profit-seeking body with a political, philosophical, religious or trade-union aim and on condition that the processing relates solely to the members of the body or to persons who have regular contact with it in connection with its purposes and that the data are not disclosed to a third party without the consent of the data subjects;
- it relates to data which are manifestly made public by the data subject or is necessary for the establishment, exercise or defence of legal claims;
- it is required for the purposes of preventive medicine, medical diagnosis, the provision of care or treatment or the management of health-care services, and when those data are processed by a health professional subject to the obligation of professional secrecy or by another person also subject to an equivalent obligation of secrecy.

Finally, the European Directive offers the opportunity to the Member States to add exemptions to those listed above, for reasons of substantial public interest. These exemptions should be subject to suitable safeguards. For instance, under Article 8, 4 of the Directive, national exemptions might be adopted for scientific research. For example exemptions could be added to the processing of medical data as long as the following conditions are verified by the data controller.

- The data processing must be legitimate. In other word, the data processing corresponds to one of the social justifications laid down by the European Directive in its Article 7. For example, the data was processed after an unambiguous consent was expressed by the data subject.
- The data controller must ensure the good quality of the data to be processed. For example, the data must be and remain accurate and if necessary up to date.
- The data controller must also respect the rights of the data subject and ensure that these are observed at all stages of the processing.
- The data controller must ensure that security and confidentiality requirements are met.
- The controller is required, prior to carrying out the processing, to provide the relevant national supervisory authority with certain items of information regarding the planned processing. The information recorded will then normally be accessible to data subjects or to third parties.

Personal Data transfer

For EU Member States, the transfer of personal data between two or several controllers established on the territory of one member states or on the territories of several member states involves the problem of communication of personal data to third parties. Indeed, the transfer or disclosure of personal data to third parties is considered to be a processing operation and, as such, is subject to the processing legal requirements discussed in the previous section. However, the transfer of medical data is subject to more specific requirements. Medical data may not be communicated unless the conditions listed below are fulfilled:

- the medical data to be communicated are relevant for the communication purpose;

- the recipient of the communication is subject to confidentiality rules equivalent to those incumbent on healthcare professionals, and the communication is legally authorised and is realised for public health reasons or for another important public interest.

The national legislations of the different member states of the European Union are by and large harmonized by now, and the transfers of personal data between these member states should not create any problem. However, the controller must refrain from transferring personal data to a recipient located in non-EEA countries, if the country involved does not ensure an adequate level of protection, unless the controller adduces adequate safeguards as regards the protection of the privacy, fundamental rights and freedoms of individuals and the exercise of the corresponding rights.

Liability Issues

This section addresses questions of liability that may impact healthgrid participants. In particular, developers need to be aware of these so as to prevent any damage to patient health, rights and freedoms. Healthgrids are complex systems. They involve different actors such as doctors, specialists, hospitals, pharmaceutical companies, data controllers and processors, technicians, etc., located in different countries. Due to this complexity, the establishment of the person to be held responsible for a specific damage can be problematic.

The products and services offered by a healthgrid could turn out to be a cause of death of the patient or of delay in diagnosis and treatment. This may, for example, be due to the lack of regular testing and monitoring of products and services. The question here is who would be liable for such damage? The basic principle has long been established that if a product does not conform to the offer made or causes damage, the consumer (or another person representing him or her) may claim for compensation. Any liability issue will thus normally depend on the general rules of law applicable in the different EU member states.

The situation is also not simple from the services perspective. Services supplied through healthgrids may cause damage to patients. These services could either be services that constitute the grid system or services provided by the internet within the grid domain, including health related websites. For example, currently there is a lack of data and knowledge management services. A citizen might thus be seriously harmed or even die if the information transmitted to the general practitioner treating him or her is not accurate or false, or if it is not supplied on time. As long as the related service is part of the grid infrastructure, Directive 2000/31 (the so-called the 'Directive on Electronic Commerce') might apply.

10. The ELSI Roadmap

Addressing the issues listed above is a challenge as the advice of medical and legal bodies is crucial. A well planned and structured set of actions is required. The ELSI roadmap could be the answer but it won't be enough unless it is harmonised with the technical roadmap milestones.

Ethical and Legal Milestones (MEL1, MEL2)

MEL1 The primary concerns for MEL1 will be liability and determining what the responsibilities of the healthgrid actors are. Possible damage that could happen to the patient could be outlined along with some preventative measures. Logging and auditing

must be addressed early to monitor whether enough testing was done to healthgrid services and products.
MEL2 Patient consent is crucial to the legitimacy of medical data processing and transfer, therefore verifying that the patient has expressed his or her consent should take place prior to any data manipulation. Moreover, appropriate and user friendly ways of allowing patient access to data is also recommended.

Data Protection Milestones
The milestone MDP1is concerned with patient privacy and how it could be best protected within the HealthGrid environment. Patients' identification issues should be discussed, starting with de-identification of medical images. For MDP2 researchers need to make sure robust anonymisation, pseudo-anonymisation and other identity protection techniques are developed and deployed in the grid infrastructure.

Ethical Control Milestones (ME1, ME2)
ME1 will focus on the requirements and tools to facilitate oversight, with automation being explored. ME2 will satisfy the arrangements for automated ethical control for a data grid which will be more complex with long-term data storage.

Policy Milestones (MP1, MP2, MP3)
These policies will cover data processing and transfer issues such as legitimacy, accessing the minimum data required, the ethical transfer of data, compliance with confidentiality rules and limiting the period of data storage.

11. Conclusion

Issues of patient ownership of her/his data and the tension between hospitals' IT policies and the requirements of grids will prove problematic unless addressed with political will; likewise the drag on technology transfer between EC projects. In total, SHARE predicts that the journey from a sustainable computing grid to a generalized knowledge grid should take from seven to fifteen years. The transition to data grid may prove harder than suggested and the transition to knowledge grid will be breaking new ground. It is therefore possible that this timescale will be multiplied by a factor of up to 3 and that a more realistic timeframe might be twenty to fifty years.

References

[1] V. Breton, A.E. Solomonides and R.H. McClatchey A Perspective on the Healthgrid Initiative Second International Workshop on Biomedical Computations on the Grid held at the Fourth IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGrid 2004). Chicago USA, April 2004
[2] Fernando J. Martin-Sanchez The BIOINFOMED Study: Towards the synergy between research in Medical Informatics and Bioinformatics HealthGrid 2003, Lyon (January 16-17th, 2003)
[3] V. Breton, K. Dean and T. Solomonides (Eds) The Healthgrid White Paper pp 249-321 of From Grid to Healthgrid: Proceedings of Healthgrid 2005, IOS Press Studies in Health Technology and Informatics Vol 112, 2005.
[4] European Commission e-Health – making healthcare for European citizens: An action plan for a European e-Health Area COM(2004) 356 final This 'Communication from the Commission' is at: http://europa.eu.int/information_society/doc/qualif/health/COM_2004_0356_F_EN_ACTE.pdf
[5] SHARE Project IST EU Contract FP6-2005-IST-027694 http://www.eu-share.org/

From Genes to Personalized HealthCare: Grid Solutions for the Life Sciences
N. Jacq et al. (Eds.)
IOS Press, 2007

A Services Oriented System for Bioinformatics Applications on the Grid

Giovanni ALOISIO, Massimo CAFARO, Italo EPICOCO, Sandro FIORE and Maria MIRTO [1]

CACT-NNL/CNR-INFM, CMCC, University of Salento, Lecce, Italy

Abstract. This paper describes the evolution of the main services of the ProGenGrid (Proteomics & Genomics Grid) system, a distributed and ubiquitous grid environment ("virtual laboratory"), based on Workflow and supporting the design, execution and monitoring of "in silico" experiments in bioinformatics.

ProGenGrid is a Grid-based Problem Solving Environment that allows the composition of data sources and bioinformatics programs wrapped as Web Services (WS). The use of WS provides ease of use and fosters re-use. The resulting workflow of WS is then scheduled on the Grid, leveraging Grid-middleware services. In particular, ProGenGrid offers a modular bag of services and currently is focused on the biological simulation of two important bioinformatics problems: prediction of the secondary structure of proteins, and sequence alignment of proteins. Both services are based on an enhanced data access service.

Keywords: Grid Problem Solving Environments, Biological Data Banks, Web Services, Grid Computing.

Introduction

Biological experiments are relatively expensive, forcing the scientists to focus on advanced design for experimentation as part of a whole-chain research approach. Furthermore, the data generally contain information outside the scope of the original experiment. Hence, to maximize the scope of experiments, biological data needs to be reusable, shareable and suitable for "in silico" experiments. An in "silico" experiment is a procedure that uses computer-based information repositories and computational analysis to test a hypothesis, derive a summary, search for patterns, or demonstrate a known fact.

In Bioinformatics, an experiment is generally characterized by a search in huge biological databases and by the execution of tools that need to access such data intensively. Moreover, the majority of the experiments are run repeatedly, orchestrating many resources to produce sets of data to be analyzed and validated. Such experiments produce a great deal of fragmented data, each coming from a different resource.

[1]Correspondence to: Maria Mirto, Center for Advanced Computational Technologies of National Nanotechnology Laboratory (NNL/CNR-INFM), University of Salento, Lecce, Italy, Euro-Mediterranean Centre for Climate Change, Lecce, Italy Tel.: +39 0832 298123; Fax: +39 0832 298173; E-mail: maria.mirto@unile.it.

Executing bioinformatics applications involves: i) accessing different biological data banks; ii) accessing experimental data; iii) executing different bioinformatics tools, possibly combining intermediate results, and finally iv) providing results to the users.

There are many open problems arising when implementing this kind of pipeline: i) the biological data banks often use different formats for the same biological entity and are accessible through web interfaces mainly manually; ii) the volume of data grows exponentially: technological advances in biology have made it possible for laboratories to generate un unprecedented amount of data; iii) the data are updated frequently: many data banks are updated daily (e.g. GenBank [1]), whereas others are updated regularly (e.g. PDB [2] is updated on a weekly basis); iv) the data are stored sparsely: each area of molecular biology generates its own databases such as the GenBank DNA database, the protein UniProt database [3], proprietary databases for storing structural information about compounds, databases for storing physical properties and activities of chemical entities, etc.; v) the data are accessed intensively: the public availability of data enables free and easy access for scientists. These data are very often exchanged by researchers as well as among databases.

Therefore, a complete and integrated software environment to execute biological applications is needed, in order to assist scientists and researchers during management and coordination of all of the tasks of an "silico" experiment, while providing large computational power.

Grid Problem Solving Environments (GPSEs) [4] can offer a solution for handling and analyzing so much disparate data connecting many computers within and among institutions through middleware software. Interaction between bioinformatics tools and biological databases should be simplified and each component (i.e. biological data banks, experimental data, bioinformatics tools) should be seen as an atomic service (Web Service) [5] that can be easily integrated in different systems, through standard interfaces and protocols. Moreover, to implement high throughput experiments, the bioinformatics tools need to be Grid-accessible and, to increase performance, they should be parallelized.

This paper describes the design and development of a web based GPSE, named ProGenGrid (Proteomics and Genomics Grid), which allows the design and execution of biological "in silico" experiments on the Grid. Distributed bioinformatics applications are described through a Workflow [6] of Grid services that wrap biological data sources and bioinformatics tools. Main features of ProGenGrid are: i) an integrated environment, where e-scientists can discover available tools and compose them in a graph (Workflow editor). The system guides the user in the selection of these tools, interacting with her, if the graph is not correct; ii) a system for running the experiments on a computational grid and managing the results (Workflow enactment service); iii) enhanced services to access the data, for grid-based sequence alignment and grid-based secondary structure prediction; in the latter case, the predictor is automatically and periodically trained again in order to improve the prediction accuracy. A fundamental feature of this system is that when the basic legacy software packages are updated by third parties, the new versions can simply replace the older ones, since these tools are embedded in high level services and exposed through a WS interface. Moreover, existing alignment services at the NCBI such as BLAST [7], PSI-BLAST [8], etc., although efficient, allow submitting only a sequence at a time, whereas ProGenGrid allows carrying out many submissions in one

shot, specifying more sequences as input. Finally, the implemented Grid-based BLAST and PSI-BLAST services exploit parallelism on clusters.

These important results have been obtained thanks to the use of several Grid Middleware services developed at the CACT/NNL of the University of Salento, Lecce and offered by the Grid Resource Broker (GRB) [9] and Grid Relational Catalog (GRelC) [10] projects, which will be briefly described in the following sections.

The first version of ProGenGrid [11] has been developed as a client-server application. Recently, a GRB based Grid Portal has been developed leveraging GRB middleware services. Both data banks and software tools are installed on geographically distributed resources, in a production grid named SPACI (Southern Partnership for Advanced Computational Infrastructures), and are deployed through the Grid Portal, as in a "virtual laboratory".

The remainder of the paper is organized as follows. In Section 1, the ProGenGrid architecture and its grid services are presented. Section 2 describes the implementation of the system and a case study is presented, while Section 3 recalls related works. Finally in Section 4 we draw our conclusions and discuss future work.

1. The Service-based Architecture of ProGenGrid

The ProGenGrid architecture includes the following logical layers, each one offering specific services (Fig. 1): *Core Services* and *Application Services*.

Application Services include biological tools and the Workflow editor. Such services are based on Grid Middleware services, named Core Grid Services. They offer Data Management services to access distributed biological data banks (both flat files and relational databases), Resource Management services to discover existing analysis tools made available as Web Services and to schedule, submit and monitor these jobs. We describe these services in the next sections.

1.1. Core Grid Services

This section describes the core services of ProGenGrid, respectively the data management and resource management services.

1.1.1. Data Management

In order to support high throughput applications such as alignment tools, we have implemented an enhanced data access service, named Split&Merge [12], to retrieve biological data from both flat files and relational databases. The Split&Merge Data Access Service is a data grid system that allows retrieving biological data and splitting them in order to distribute the computation on several grid nodes.

The Split&Merge architecture is based on the following services: GRelC Data Access (GDA) [10], GRelC Data Storage (GDS) [10], and Splitter/Merger. This service, described in [12], has been extended with the following components, for translating biological data banks into a metadata DBMS: Wrapper (WRP) and GRelC Translator (GTrs).

The main feature of the system is a query protocol that manages the query, distributing the entire record set among several computational grid nodes and processing the subsets locally.

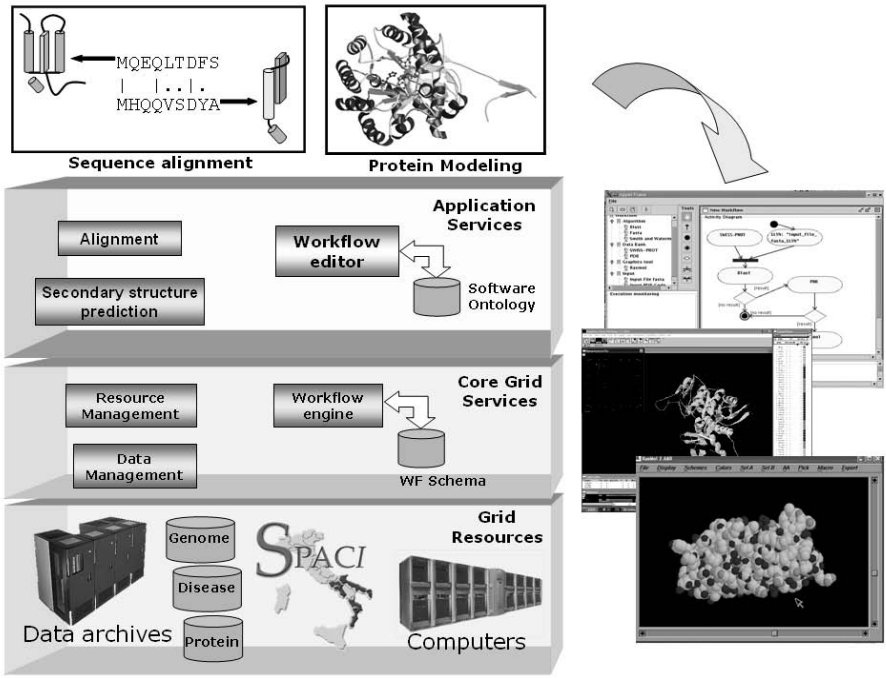


Figure 1. ProGenGrid Architecture

Detailing each service, the GDA acts as a standard front-end for each data bank. This service provides the basic primitives to transparently access and interact with different data sources, concealing the back-end heterogeneity and other low level details (it is built on top of the Globus Toolkit [13], libxml2 and proprietary libraries of several DBMS). The GDA is developed as a Grid Security Infrastructure (GSI) enabled Web Service. To date, the GDA provides (by means of wrappers written using proprietary libraries) dynamic binding to several physical RDBMS such as Oracle or PostgreSQL and ad hoc wrappers for the UniProtKB [3], UTRdb [14], PDBaa [15], Nr [16] and PDB [2] databases.

GDS provides a large set of services to manage group of files on a storage resource (i.e. a Disk Resource Manager, DRM) and to share/make them available among/to a community of users.

Splitter/Merger performs a two phase (split & merge) data management activity (interacting with the GDS). In the former phase the record set is split into several fragments (input data) and uploaded to the GDS, whereas in the latter this component downloads the partial results (output data) from the GDS and merges them composing the final user response.

1.1.2. Resource Management

Resource Management services allow brokering, scheduling and monitoring of jobs on a Computational Grid. ProGenGrid relies on the GRB Scheduler for these functions. It acts as a meta scheduler for the available grid resources and has been designed to be fully compliant with respect to the Job Submission Description Language (JSDL)

specification. The JSDL language has been however extended in order to provide better support for:

- batch jobs: these jobs involve staging of the input and output. MPI/OpenMP jobs are also batch jobs;
- interactive jobs: jobs in which the user can follow the job progress watching the job standard output/error from the client browser;
- parameter sweep: the same application executed several times with different inputs. The GRB scheduler balances the workload on the grid resources;
- workflow jobs: the user composes some jobs in a graph. The GRB scheduler submits all of the sub-jobs of the workflow job without violating the precedence constraints.

A simple job for sequence alignment can be a batch job if a single sequence is run, otherwise it can be considered as a parameter sweep job. Moreover, if adjustment of the input requires several operations, the alignment is a workflow job. Other GRB features are: i) it retrieves grid information using the iGrid or MDS2 (MDS4 planned) information systems; ii) it manages data using GridFTP to automatically stage the executable input and output file(s); iii) it supports credential delegation and single sign-on to the Grid; iv) it configures grid resources and services.

1.2. Application Services

The main application services developed in ProGenGrid are a toolkit for parallel alignment of sequence and a system for the prediction of proteins secondary structure. We now describes these services.

1.2.1. A Grid-based Tool for the Alignment of Sequences on the Grid

One of the key features of this service is based on a technological solution to improve the execution of several alignments tools involved in biological "in silico" experiments. There are two approaches to improve the execution of alignment tools: fragmenting the input and/or the databank on which the search is carried out. We have chosen to use both database and query segmentation for BLAST and query segmentation for PSI-BLAST to achieve this goal. We use this optmized version of PSI-Blast in the Secondary Structure Predictor described in the next section. In the former case (BLAST), if more sequences are considered, the performances are better w.r.t. the sequential case [17] whereas in the latter (PSI-BLAST), we have chosen query fragmentation because many queries are involved in the computation and it seems that database fragmentation can not be beneficial. However, in a future work we would like to exploit an hybrid approach, using both database and query segmentation, depending on the problem, the alignment tool, and the available resources.

The proposed system, named BioGAT, exploits a Grid Platform based on the Globus Toolkit 4.0.3 and in which modules are developed in C/C++ and deployed as Web Services (WS), exploiting the gSOAP Toolkit [18] (with GSI support available as a gSOAP plug-in). BioGAT allows splitting the computation on a biological dataset among several computational nodes. It does not matter how the data is partitioned or what subset is computed by a client, because there are not dependencies among data stored in the same or in different subsets (in the BLAST tool case only an adjustment of several statistical

parameters must be made at the end of the computation). Our aim is to allocate at runtime these subsets on the idlest machines taking into account some dynamic parameters, such as, for instance, the CPU availability of the grid nodes.

There is a difference related to the involved data banks when BLAST or PSI-BLAST computations are considered. In the former case, computation is made on chunks of databases, so each fragment must be indexed. In the latter case, the fragmentation is on the query so the involved data banks must be located on the same machine on which the job is run and indexing is made just once.

1.2.2. Grid Secondary Structure Predictor

Another important service of ProGenGrid is the prediction of the secondary and tertiary structure of proteins starting by their amino acid sequence. To date, we cover secondary structure prediction but in the future the full deployment of the system will also allow tertiary structure prediction.

The protein structure prediction problem is the problem of predicting the 3D structure of the protein (its native conformation) when the list of amino acids is known. Native conformation determines the biological function of the protein [19]. Native conformations are largely built from Secondary Structure Elements (SSEs), which are local motifs consisting of short, consecutive parts of the amino acid sequence having a very regular conformation. Some of them are: $\alpha - helix$, $\beta - sheet$ and *random coil*.

Computational methods available for secondary structure prediction are based on different approaches. The most recent (defined as third generation) are based on neural networks (NNs) and allow carrying out the prediction with an accuracy between 72% and 80%, on specific proteins classes [20,21]. The aim of these techniques is to assign secondary structure elements to segments of the amino acids sequence of unknown structure, starting with the knowledge of a sufficient large number of proteins, whose sequence and three-dimensional structure are known, used as training set.

Proteins that are similar in sequence are likely to have similar structure and function. Hence similarities can give hints about the evolutionary history of certain sequences. By the analysis of different methodologies, it emerges that the quality of prediction has been improved both by the increasing growth of the protein databases and through the application of evolutionary information related to a single protein. Indeed, secondary structure prediction accuracy has improved to over 70% by incorporating the evolutionary information found in multiple sequence alignment (MSA).

An integrated Grid system for the prediction of secondary structure, based on the frequent regular updating of the training protein dataset, has been implemented. The predictor model is based on a feed forward Multi Layer Perceptron (MLP) neural network which is trained with the back-propagation algorithm, and it has been designed using both existing legacy software and novel components. The predictor is based on the evolutionary information found in MSA and calculated by using the PSI-BLAST tool [8] on a set of proteins of known structure (training set), provided by the University of California's research group. The testing set is R126 [20], which has been used by Rost and Sander in their earlier studies on protein secondary structure prediction. Low correlation among these two protein sets yields a reliable way of evaluating predictor results. A comparison of the system with other predictors has been carried out on the same test set and the proposed system is more accurate of an analyzed predictor, JPred [21], by about 2%.

The system exploits advanced mechanisms for management of biological data and allows the automation of the computational process in order to reduce the network learning time and to guarantee periodic updating of the network. Using Grid technology, the training time has been remarkably reduced and the procedure is completely automatic, whereas the implemented predictor has an accuracy percentage of about 76% on a well known testing set available in literature.

1.2.3. Workflow Editor

In order to support complex bioinformatics experiments, a workflow management system composed of an editor and an enactment service has been deployed.

The first prototype was developed as a client-server architecture; the current implementation provides a grid based portal to discover, compose, execute and monitor jobs. The editor, developed in Java, exploits a graphical notation based on UML. The data flow is described as activity diagrams and an opportune workflow language specification is used. This is translated into JSDL and sent to the GRB scheduler. Graph nodes are classified as:

- **execution node**: these nodes represent a job to be executed remotely on a brokered grid resource. This job can be a single batch job or recursively it can represent either a parameter sweep job or a workflow job;
- **condition node**: this node denotes a condition to be evaluated in order to control at runtime the execution flow of the graph. The condition is based on the execution of a test job and on further evaluation of the condition expressed on the outputs. The expression to be evaluated can be constructed using either a specified input file, a specified output file or standard output/error produced by the test job;
- **storage node**: these nodes represent single stage-in or stage-out resources.

Moreover, a relational schema describing the involved software (logical name, path, environment variables, etc.) has been developed using MySQL.

The enactment service allows executing workflows described by a directed graph and supports graphs with cycles and conditions.

2. ProGenGrid Portal Implementation

The GRB grid portal has been customized to expose the services of ProGenGrid. The resulting Grid portal provides the following services: Grid Configuration, Resource Status, File Transfer, Alignment Jobs, Secondary Structure Prediction.

The Grid Configuration section allows setting up grid user's profiles, retrieving the user's credentials, adding a new Virtual Organization (VO), configuring the grid with the possibility to associate grid machines to the users and finally, the Information Service monitors the grid status. It is possible to configure biological data banks, application tools, and to associate several databases to a software. This option allows (through the Alignment web interface) multiple data banks to be used. Finally, job status can be visualized.

Resource Status allows querying several information services to retrieve a resource status, whereas File Transfer allows using GridFTP to transfer files.

Alignment jobs allows submitting sequences for BLAST and PSI-BLAST, whereas for the secondary structure prediction it is possible to use an interface to submit the sequences whose structures need to be predicted. Finally, a workflow interface allows composing several applications in a graph, using drag-and-drop. The portal is available at the following url: https://sara.unile.it/cgi-bin/bioinfo/enter.

As a case study, an experiment regarding multiple sequence alignment (MSA) among each of the human proteins available in the UniProtKB database (about 70845 sequences, retrieved by the *uniprot_sprot_human.dat* and *uniprot_trembl_human.dat* flat files) and those stored in the UniProt NREF data bank has been carried out. Homologue sequences are hence matched to identify functional domains. PSI-BLAST is used for multiple alignment. Moreover, to reduce storage issues and taking into account that a single run of PSI-BLAST involves a set of iterations in order to converge to a solution (we used two iterations), an adjustment of the result based on results retrieved in the last iteration has been carried out. The experiment has been modelled using the ProGenGrid workflow. After running one experiment we produced about 70 thousands alignments files, carrying out an adjustment of the result (i.e., considering only the result of the last iteration). The experiment took about 4 days, using 30 1.4 GHz Itanium 2 processors.

3. Related Work

The Grid is proposed as the next generation infrastructure necessary to support and enable the collaboration of people and resources through highly capable computation and data management systems (Foster and Kesselman, 1998). A number of BioGrid projects are underway, including myGrid [22], GeneGrid [23] and BioSimGrid (Biomolecular Simulation Grid) [24]. These primarily focus on the sharing of computational resources, large-scale data movement and replication for simulations, remote instrumentation steering or high-throughput sequence analysis.

myGrid [22] is a project targeted at developing open source high-level middleware to support personalised in silico experiments in biology on a Grid. myGrid is building services for integration such as resource discovery, workflow enactment and distributed query processing. Additional services are needed to support the scientific method and best practice found at the bench but often neglected at the workstation, notably provenance management, change notification and personalisation. The target users of myGrid are tool and service providers who build applications for a community of biologists.

GeneGrid [23] is a collaborative industrial grid computing R&D project initiated by the Belfast e-Science Centre (BeSC) under the UK e-Science programme and supported by the Department of Trade &Industry. The project aims at providing a platform for scientists to access their collective resources, skills, experiences and results in a secure, reliable and scalable manner by creating a "Virtual Bioinformatics Laboratory". GeneGrid provides seamless integration of a myriad of heterogeneous applications and datasets that span multiple administrative domains and locations across the globe, and presents these to scientist through a simple user-friendly interface. GeneGrid meets this challenge through the GeneGrid Portal - the user interface for the project.

Finally, the aim of the BioSimGrid project [24] is to make the results of large-scale computer simulations of biomolecules more accessible to the biological community. Such simulations of the motions of proteins are a key component in understanding how the structure of a protein is related to its dynamic function.

4. Conclusions and Future work

This paper presented a system, ProGenGrid, for bioinformatics. Advantages of ProGenGrid are: i) the possibility to design and execute a bioinformatics application on a Computational Grid; ii) the access to distributed data through a service that hides their heterogeneity, providing a set of useful libraries; iii) the use of common bioinformatics tools, through a unique interface; iv) the submission of many sequences at a glance to alignment tools; v) the possibility to carry out the training of the neural network using grid resources and the prediction of many sequences at a glance.

Future work will regard the possibility to experiment more complex case studies and full deployment of a service for the protein tertiary structure prediction.

Acknowledgments

This work was supported by the MIUR (Italian Ministry for Education, University and Research) in the LIBI (International Laboratory of BioInformatics) project, F.I.R.B. 2003, under grant RBLA039M7M.

References

[1] D. Benson, I. Karsch-Mizrachi, D. Lipman, and D. W. J. Ostell, "GenBank," *Nucleic Acids Res*, vol. 1, no. 33, pp. 34–38, 2005.

[2] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, and et al., "The Protein Data Bank," *Nucleic Acids Res.*, no. 28, pp. 235–242, 2000, www.rcsb.org/pdb/.

[3] A. Bairoch, R. Apweiler, C. Wu, W. Barker, B. Boeckmann, S. Ferro, and et al., "The Universal Protein Resource (UniProt)," *Nucleic Acids Res.*, no. 33, pp. 154–159, 2005, http://www.uniprot.org.

[4] G. V. Laszewski, I. Foster, J. Gawor, P. Lane, N. Rehn, and M. Russell, "Designing Grid-based Problem Solving Environments and Portals," in *Proceedings of International Conference on System Sciences (HICSS-34)*, 2001.

[5] F. Curbera, M. Duftler, R. Khalaf, N. Mukhi, W. Nagy, and S. Weerawarana, "Unraveling the Web Services Web - An Introduction to SOAP, WSDL, and UDDI," *IEEE Internet Computing*, vol. 6, no. 2, pp. 86–93, 2002.

[6] W. management coalition reference model, http://www.wfmc.org/.

[7] S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman, "Basic local alignment search tool," *Nucleic Acids Res.*, vol. 3, no. 215, pp. 403–410, 1990, http://www.ncbi.nlm.nih.gov/BLAST/.

[8] S. Altschul, T. Madden, A. Schäffer, J.Zhang, Z. Zhang, W. Miller, and D. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Res.*, vol. 17, no. 25, p. 3389Ű3402, 1997, http://www.ncbi.nlm.nih.gov/cgi-bin/BLAST/nph-psi.

[9] G. Aloisio, M. Cafaro, G. Carteni, I. Epicoco, S. Fiore, D. Lezzi, M. Mirto, and S. Mocavero, "The Grid Resource Broker Portal," *to appear in Concurrency and Computation: Practice and Experience, Special Issue on Grid Computing Environments*, 2006.

[10] G. Aloisio, M. Cafaro, S. Fiore, and M. Mirto, "The Grid Relational Catalog Project," in *Advances in Parallel Computing, "Grid Computing: The New Frontiers of High Performance Computing"*, L. Grandinetti, Ed., 2005.

[11] ——, "ProGenGrid: a Grid-enabled platform for Bioinformatics," in *From Grid to HealthGrid, Proceedings of HealthGrid 2005, Tony Solomonides et al. (Eds), IOS Press*, vol. 112, 2005, pp. 113–126.

[12] ——, "A Split&Merge Data Management Architecture in a Grid Environment," in *Proceedings of the 19th IEEE International Symposium on Computer-Based Medical Systems (IEEE CBMS 2006), Salt Lake City Utah (USA)*, 2006, pp. 739–744.

[13] I. Foster and C. Kesselman, "Globus Toolkit Version 4: Software for Service-Oriented Systems," in *IFIP International Conference on Network and Parallel Computing*, S.-V. L. 3779, Ed., 2005, pp. 2–13.

[14] G. Pesole, S. Liuni, and et al., "UTRdb: a specialized database of 5' and 3' untranslated regions of eukaryotic mRNAs," *Nucleic Acids Res.*, no. 27, pp. 188–191, 1999, http://www.ba.itb.cnr.it/UTR/.
[15] G. Wang and R. Dunbrack, "PISCES: improvements to a protein sequence culling server," *Nucleic Acids Res. (in press)*, http://dunbrack.fccc.edu/Guoli/pisces_download.php.
[16] N. B. databases, ftp://ftp.ncbi.nlm.nih.gov/blast/db/.
[17] G. Aloisio, M. Cafaro, I. Epicoco, S. Fiore, and M. Mirto, "BioGAT: a Grid Toolkit for Bioinformatics Sequence Alignment," in *Proceedings of the Grid-Enabling Legacy Applications and Supporting End Users Workshop (GELA), France*, 2006, pp. 77–85.
[18] G. Aloisio, M. Cafaro, D. Lezzi, and R. V. Engelen, "Secure Web Services with Globus GSI and gSOAP," in *Proceedings of Euro-Par 2003, Lecture Notes in Computer Science, Springer-Verlag*, no. 2790, 2003, pp. 421–426.
[19] J. N. Onuchic and P. Wolynes, "Theory of protein folding," *Current Opinion in Structural Biology*, vol. 14, pp. 70–75, 2004.
[20] B. Rost and C. Sander, "Prediction of protein secondary structure at better than 70% accuracy," *Journal of Molecular Biology*, vol. 2, no. 232, pp. 584–599, 1993.
[21] J. Cuff, M. Clamp, A. Siddiqui, M. Finlay, and G. Barton, "JPred: a consensus secondary structure prediction server," *Bioinformatics*, no. 14, pp. 892–893, 1998.
[22] R. Stevens, A. Robinson, and C. Goble, "myGrid: Personalised Bioinformatics on the Information Grid," *Bioinformatics*, vol. 19, no. 1, pp. 302–302, 2004, http://www.mygrid.org.uk/.
[23] N. Kelly, P. Jithesh, and S. W. et al., "The GeneGrid Portal: A User Interface for a Virtual Bioinformatics Laboratory," in *UK e-Science All Hands Meeting (AHM 2005)*, 2005, http://193.61.123.75:8080/GeneGrid/Portal.
[24] B. S. G. Project, www.biosimgrid.org/.

From Genes to Personalized HealthCare: Grid Solutions for the Life Sciences
N. Jacq et al. (Eds.)
IOS Press, 2007

Extending Workflow Management for Knowledge Discovery in Clinico-Genomic Data

Stefan RÜPING [a], Stelios SFAKIANAKIS [b] and Manolis TSIKNAKIS [b]
[a] *Knowledge Discovery Department, Fraunhofer Institute for Intelligent Analysis and Information Systems, Sankt Augustin, Germany*
[b] *Biomedical Informatics Laboratory, Institute of Computer Science, FORTH, Greece*

Abstract. Recent advances in research methods and technologies have resulted in an explosion of information and knowledge about cancers and their treatment. Knowledge Discovery (KD) is a key technique for dealing with this massive amount of data and the challenges of managing the steadily growing amount of available knowledge. In this paper, we present the ACGT integrated project, which is to contribute to the resolution of these problems by developing semantic grid services in support of multi-centric, post-genomic clinical trials. In particular, we describe the challenges of KD in clinico-genomic data in a collaborative Grid framework, and present our approach to overcome these difficulties by improving workflow management, construction and managing workflow results and provenance information. Our approach combines several techniques into a framework that is suitable to address the problems of interactivity and multiple dependencies between workflows, services, and data.

Keywords. Biomedical Semantic Grid, Grid Workflows, Knowledge Discovery

Introduction

Recent advances in research methods and technologies have resulted in an explosion of information and knowledge about cancers and their treatment. The ability to characterize and understand cancer is growing exponentially based on information from genetic and protein studies, clinical trials, and other research endeavours. The breadth and depth of information already available in the research community at large, present an enormous opportunity for improving our ability to reduce mortality from cancer, improve therapies and meet the demanding individualization of care needs [1,2]. Because our knowledge of this domain is still largely rudimentary, investigations are now moving from being *hypothesis-driven* to being "data-driven" with analysis based on a search for biologically relevant patterns.

While these opportunities exist, the lack of a common infrastructure has prevented clinical research institutions from being able to mine and analyze disparate data sources. Moreover, clinical researchers or molecular biologists often find it hard to exploit each other's expertise due to the absence of a cooperative environment which enables the sharing of data, resources or tools for comparing results and experiments, and a uniform plat-

form supporting the seamless integration and analysis of disease-related data at all levels [2]. Similarly, the lack of a unifying architecture is proving to be a major roadblock to a researcher's ability to mine different databases. The vision of the Advancing Clinico-Genomic Trials on Cancer (ACGT) integrated project is to contribute to the resolution of these problems by developing a semantic grid services in support of multi-centric, post-genomic clinical trials, and thus enabling discoveries in the laboratory to be quickly transferred to the clinical management and treatment of patients. In achieving such an objective, extensive use of ontologies and metadata are required. This will be provided by developing a master ontology on cancer and implementing suitable metadata registries, which provide semantically rich information about available data and computational services.

In this paper, we present our approach on how to improve the support for KD in a grid environment, in particular with respect to the requirements put up by the domain of post genomic clinical trials. While we are focusing on the data miner's view on data analysis, we always keep in mind the cross-disciplinarity of the research field (indeed, we will see that a thorough domain understanding and interaction with the end-user is a crucial aspect of data mining). The remainder of the paper is organized as follows: Section 1 gives an overview over the field of KD and lists several requirements that are not satisfactorily supported by current Grid environments. Section 2 gives an overview of the ACGT architecture and an assortment of techniques to improve the support for KD within existing Grid architecture are presented. Section 3 gives an overview of the related work and Section 4 concludes.

1. Knowledge Discovery

Knowledge Discovery has been described as "the non- trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" [12]. A model for a typical data mining process has been proposed by the CRISP project [11]. An interactive, iterative nature of the process is generally characterized as a main property of data mining, see e.g. [13]. The reason is that data mining has an experimental nature, meaning that often several approaches are tried out in parallel before the best one is selected.

1.1. Knowledge Discovery in Clinico-Genomic Data

Data mining methodology and technology has been developed to date for classical business, finance, and customer-oriented application domains. Such domains are characterized by the availability of large quantities of data in an attribute-value based representation, high ratio of examples over attributes in the data sets, and weak background knowledge about the underlying entities and processes.

For biomedical data these conditions do not hold. With high-throughput technologies like microarrays or mass-spectrometry, masses of genomic/proteomic data are produced – about 10^4 gene, or mass/charge (m/Z) features per patient. In addition, legal, ethical and practical limitations make it cumbersome to acquire a high number of patients in a clinical trial – a typical (preclinical) cohort may contain only about 100-200 patients. Under these conditions, standard statistical methods are likely to over-fit the structures

in the data, such that a high amount of domain knowledge is needed to guide the analysis and guarantee the validity of the extracted knowledge [3].

It follows that the challenges of KD in bio-medical data differs significantly from the original problems of data analysis that prompted the development of Grid technologies: instead of the analysis of huge data sets, the problem here lies in the analysis of many small data sets with a plethora of possible analysis workflows. The central factor here is to make effective use of the distributed knowledge of the involved research communities in order to compensate the low statistical significance which results from small sample sizes. Valuable kinds of knowledge include:

- *Knowledge about the semantics of the data*: it is well known that in data mining finding an optimal representation of the data is central for obtaining good results. That is, great care must be taken in the step of pre-processing, e.g. feature selection and construction [8].
- *Knowledge about the plausibility of results*: when there is not enough statistical information about the validity of a hypothesis, one can look for external evidence for or against this hypothesis in the scientific literature, where usually much more knowledge is available than what is encoded in the specific data set. To make use of this knowledge, the interpretability of the models must be ensured [14], and text mining technologies [16] must be incorporated.
- *Knowledge about analysis workflows*,in particular about which workflow is optimal for a given problem. This problem is an instance of the field of workflow mining [9] and related approaches [10].

Therefore, the main challenge for KD in clinico-genomic data is the sharing of knowledge, either in the form of the integration of existing knowledge to design and select appropriate analysis tools, or to manage the discovered knowledge in order to make it available to the research communities. In particular, the following five requirements for a workflow-based environment have to be met:

- *Data is central*: workflow execution gives a central position to the services that process the data and not to the data itself. The reason is that usually these functions remain constant, while the data they are executed on varies from execution to execution. In KD, however, the fixed point is the data set which is to be analysed. Several functions can be executed on the data set to solve the research question, and it is not a-priori clear which approach (workflow) will be optimal. The analyst will typically execute a large variety of functions on the same data set, and readily discard any approach that fails to answer his question. Hence, data and results should play a much more prominent role in workflow construction for KD than in usual approaches.
- *The need to support interactivity*: as stated before, KD is inherently an interactive process. Workflow enactment must support this nature of the process in order to match the trial-and-error way of working in data analysis. For example, a typical step in KD is to check the intermediary results of a workflow and then decide whether to continue the workflow or to modify it and re-start the analysis.
- *Managing dependencies between discovered knowledge and experiments*: as a result from the data- and hypothesis-driven working style, several data sets, workflows, workflow executions, and results will be generated during a single KD process. Traditionally, it is the responsibility of the analyst to keep track of all

these instances and their relationships, e.g. which workflow generated which result, what methods have been tried out on a data set, etc. With the possibilities of collaborative and massively parallel data analysis on the grid, keeping track of all these dependencies becomes infeasible if it is not automatically supported by the system.
- *Mining-specific provenance information and meta data needs to be supported.* Important information that relates to data mining are for example the mapping of services to stages in the CRISP process, a description of data in terms of its role in the mining process (e.g. input data, transformed data, model), the relation between different results (A is better than B), and information about the applicability of services to relevant data types (e.g. nominal, or numeric values, text data, images).
- *Support of building blocks*: it can often be found that a certain sequence of operators that is part of a workflow is meaningful on its own, for example because it corresponds to a phase in the CRISP model, or because a certain operator cannot be applied without a fixed sequence of pre-processing steps. In these cases, a method for the composition of a sequence of services into a single service that can be appropriately marked up and retrieved by a semantic discovery method is very useful.

In the following chapter, we will propose approaches to meet these requirements based on existing Grid architectures.

2. Architecture

The current state of the art architectural style for building distributed applications is the so called Service Oriented Architecture (SOA) where the building block is the notion of Services. A service is usually defined as "a software system designed to support interoperable machine-to-machine interaction over a network" [1] typically using SOAP as the communication protocol for exchanging XML messages complying to an agreed contract expressed in WSDL. The SOA methodology has been also adopted by the Grid computing community and we currently see a convergence of Web Services and Grid Services in the context of Open Grid Service Architecture [27]. An architecture based on services and their interactions has therefore being defined for the development of the ACGT infrastructure. An overview of this layered architecture is depicted in Figure 1.

In the general ACGT environment the workflow authoring and management tasks play a central role to support the KD process. Therefore, a more detailed architecture has been defined to support the publication, discovery, invocation, and management of workflows and it will be further elaborated during the course of the project. Based on the discussion in Section1 we have come up with the following approaches:

2.1. Workflows as services

The Workflow Management Coalition (WFMC) defines a workflow as "The automation of a business process, in whole or part, during which documents, information or tasks are passed from one participant to another for action, according to a set of procedural

[1] http://www.w3.org/TR/ws-arch/

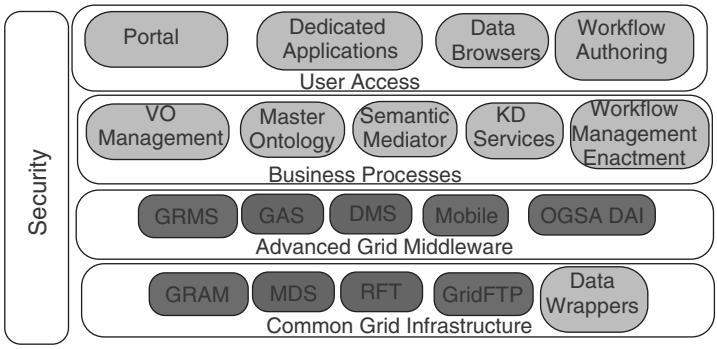


Figure 1. An overview of the ACGT architecture

rules". In other words a workflow consists of all the steps and the orchestration of a set of activities that should be executed in order to deliver an output or achieve a larger or more sophisticated goal. In essence a workflow can be abstracted as a composite service, i.e. a service that is composed by other services that are orchestrated in order to perform some higher level functionality. This idea will be further extended and become concrete during the ACGT project.

On the practical side we are planning to support the implementation of a service that accepts a workflow description in some workflow description language and automatically generates a WSDL that describes this workflow. This will make the workflows easily accessible and reused from the rest of the ACGT environment using familiar web services technologies.

2.2. Ontologies and Semantic Descriptions

The importance of annotation of services and workflows with semantic descriptions cannot be overemphasized. An adequate and convenient way of annotating workflows with metadata is of paramount importance for their posterior discovery among repositories of useful workflows based on their purpose, internals and field of application characteristics.

In trying to respond to this need in ACGT we are developing a domain ontology for cancer research that will enable the semantic integration of data and services. This ACGT Master Ontology will serve as a common reference for concepts related to cancer and their relationships that is both, human-understandable and machine-understandable. In addition to the Master Ontology another ontology specific for the description of the services details is needed. This Service Ontology is necessary for the description of the data formats, the service functionalities, provenance information, and so on. These issues have already been tackled in the myGrid project [30] and we will base our implementation on their work.

2.3. Data as services

In a scientific workflow data is the information that flows between services and triggers their execution and also the final outcome of a workflow execution. In addition to workflows being services we also promote the modeling of data, be it intermediate or final

results of a workflow, as services. We consider this approach to be feasible, since each data source that is somehow accessible can be considered as a service that when invoked returns the data that it encapsulates. This is a common phenomenon in the World Wide Web where accessing a HTTP URI does not usually disclose if the retrieved content is dynamically generated by some web service or not.

It is also desirable, since the handling of data as basic services will enable a unified approach for the integration of data and services into workflows. This means that we can have the same infrastructure for annotating, searching, matching, and composing data, services, and workflows. It also means that data are "first class citizens" in the ACGT environment and not just the channels of communication between the services.

2.4. Breakpoints

In order to improve the interactivity of Grid workflows, we propose the implementation of a so-called Breakpoint Service. The task of this service when it is executed is to store the state of the workflow execution (e.g. the intermediate results, the execution's control point) as a continuation [39] to a user-specific data repository. It will then construct a new *continuation workflow* that starts with extracting this input data from the data repository and continue with the workflow that the Breakpoint Service is a part of. This new workflow will be entered into the workflow repository, with appropriate information linking it to the workflow data and the original workflow it was derived from, and converted into a service to allow an easy access. Finally, the Breakpoint Service will suspend the running workflow.

When the user is notified that his initial workflow has stopped (which will be done by the usual Grid management services), he can then look for the new workflow-representing service (using the usual semantic service discovery and notification services) and decide to inspect the intermediate results of the workflow from his data repository, continue the workflow, i.e. execute the newly constructed continuation workflow, load the continuation workflow into the workflow editor to make appropriate changes, or load the original workflow into the workflow editor to make appropriate changes and restart.

In this way, a significant extension of the interactivity and flexibility of workflows can be achieved within existing architectures.

2.5. Hyperlinked Provenance Information

Provenance information is the key to organizing information in a collaborative, Grid-based system. This becomes an even greater necessity when improved interactivity can result in changing workflows on the fly and cloning running workflows to explore alternatives in parallel. A hyperlinked presentation of information has been proposed as a tool for better supporting the collaboration in scientific communities [15]. In combination with the idea of representing data as services and workflows as services, provenance information can be viewed as a graph of services, with edges representing several types of provenance. For example, edges can represent the relations “produced-by”, “part-of”, “derived-from”, “input-of” etc.

The user can navigate in such a graph just like surfing the web, where each service corresponds to a web page, which could be available in different representations to adapt

to the needs of different user groups, and edges correspond to links. Each entity (e.g. service, data) is identified by a HTTP URI to provide identification, retrieval, and linking facilities for constructing a web of data and metadata in accordance with the Semantic Web vision [37].

The employment of this prototypical web of provenance information on the ubiquitous infrastructure of World Wide Web will have the advantage of user familiarity while offering a basis for more advanced features such as a bridge to the Semantic Web through GRDDL [34] transformations for example.

3. Related Work

A practical approach to provide standard data mining services across a Grid infrastructure is the integration of standard statistical and data mining toolkits, such as Weka [4] or R [5]. This approach was followed by the DataMiningGrid project [2], which focused on facilitating large-scale, distributed data mining operations, but less to support interconnected data mining tasks as they are prominent in the analysis of biomedical data. Hence, this project needs to be complemented with user-oriented services to structure and manage data mining results.

The Smart Tea project [18] and its successor myTea [19], have introduced the concept of an electronic lab book. The idea is to bridge the gap between workflow execution on the grid and work done in the lab in order to enable meaningful interrogation of the process and controlled sharing of the results. Other authors have proposed approaches for linking and browsing lab information as well, most notable the approach for provenance information in the myGrid project [21,20].

IDEA [17] is an Intelligent Discovery Assistant, which provide users with systematic enumerations of valid KD processes, and effective rankings of these valid processes by different criteria, to facilitate the choice of KD processes to execute. It is based on an ontology of KD operators, which guides the combination of services into valid workflows. In contrast to the ACGT approach, IDEA does not make use of a refined ontology on the data, because it is meant to be a general-purpose tool and it cannot be assumed that such an ontology exists.

The Mining Mart system [10] focusses on setting up and re-using best-practice cases of preprocessing data stored in very large databases. A meta-data model named M4 is used to declaratively define and document both, all steps of such a pre-processing chain and all the data involved. The idea of Mining Mart is to publish best-practice cases on the internet. However, the system only contains information about the workflows, not about models, because the system was developed in a business context, where models are usually private information

There is a lot of research in the area of workflows and service composition. There are plenty of available tools for authoring scientific workflows in the Grid [28,26,29] with different characteristics supported by each of them [35].

In myGrid's Taverna Workbench [3] services can be annotated with semantic descriptions based on ontologies and later discovered based on these descriptions. The myGRID ontology [30] will be used also in our work but in linkage with our specific domain ontol-

[2] http://www.datamininggrid.org/

[3] http://sourceforge.net/projects/taverna

ogy for cancer. The information about the workflows, their enactment, and provenance is kept in the myGrid Information Repository (mIR) which is a UDDI compliant registry. The provenance logs are also annotated with concepts from the domain ontology and a web of provenance documents, services, and workflows can be built [21]. The sharing and reuse of workflows in myGrid has recently been enabled by the design of myExperiment[4]. In our case the hyperlinked provenance information is based on the semantic web approach where each identified entity is given a URI that can be "dereferenced" and easily retrieved by both humans and software. Nevertheless, the general concepts are the same and we are investigating methods for reusing parts of the myGrid infrastructure .

Continuations have been proposed in the past as a means for maintaining conversational state in web interactions [38]. The applicability of continuations in workflow enactment has also been described in the past [40] but to our knowledge there are few workflow tools to support them (e.g. BpmScript[5]). An interaction capability has been recently added to the Taverna workbench through the Interaction Service. An interaction service when invoked sends an e-mail to the user with the information needed to stop or continue the workflow by the means of HTTP links.

For the semantic annotation of services we consider OWL-S[6] to be highly relevant, since it can be used to describe workflows as "composite processes" through its process model. OWL-WS ("OWL for Workflows and Services") [36] is a workflow and service ontology supported by the NextGRID project that is based on OWL-S with various extensions to provide advanced features like the abstract definition of workflows so that a task's bindings (endpoints, etc.) can be specified/located at run time, and higher order workflows so that a task can have a workflow as input and return another workflow as output. Finally the semantic annotation of WSDL [41] is a way to put inline references to user specified ontologies in the WSDL documents. Although the combination of data and metadata in a single entity has clear advantages, we don't consider the current breed of tools to be mature enough to support this "intrusive" approach.

4. Conclusions

Knowledge discovery in clinico-genomic data exhibits interesting challenges for workflow management and execution, which are not fully supported by current Grid workbenches. We have presented an overview of the roadmap in the ACGT project with regards to the KD process in clinico-genomic data using workflows. We propose the combination of several Grid techniques in order to provide an interactive, easy-to-use, yet expressive environment that supports the needs of both domain scientists, e.g. clinicians and biologists, and data miners. These techniques will be further extended and realized in the course of the project taking into account the prior art in the relevant fields of workflow management and enactment and the semantic composition of services while following the user requirements elicitation process.

[4] http://www.myexperiment.org/
[5] http://www.bpmscript.org/
[6] http://www.daml.org/services/owl-s/

Acknowledgements

The authors wish to thank to the ACGT consortium for their contributions with various ideas on which the ACGT project was developed. The ACGT project is funded by the European Commission (Contract No. FP6/2004/IST-026996).

References

[1] K. Yurkewicz, *Accelerating Cancer Research*, Science Grid, June 21, 2006. Available at `http://www.interactions.org/sgtw/2006/0621/cabig_more.html`.
[2] K. H. Buetow, *Cyberinfrastructure: Empowering a "Third Way" in Biomedical Research*, Science Magazine, vol. 308, no. 5723, pp. 821-824, 2005.
[3] M. May, G. Potamias, S. Rüping, *Grid-based Knowledge Discovery in Clinico-Genomic Data*, Proc. Int. Symposium on Biological and Medical Data Analysis (ISBMDA), 2006.
[4] I. Witten, and E. Frank, "Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, 2000.
[5] R Development Core Team: *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2005, ISBN 3-900051-07-0.
[6] PMML Schema Version 2.1: `http://www.dmg.org/v2-1/GeneralStructure.html`
[7] OWL-S; `http://www.daml.org/services/owl-s/`
[8] H. Liu and H. Motoda (eds.), "Feature Extraction, Construction and Selection: A Data Mining Perspective", Kluwer, 1998.
[9] W. van der Aalst et al., *Workflow Mining: a Survey of Issues and Approaches*, in: Data and Knowledge Engineering, 47 (2), pp .237-267, 2003.
[10] K. Morik and M. Scholz, *The MiningMart Approach to Knowledge Discovery in Databases*. in: Ning Zhong and Jiming Liu (eds.), Intelligent Technologies for Information Analysis, Springer, pp. 47 - 65, 2004.
[11] C. Chapman et al., *CRISP-DM 1.0*, `http://www.crisp-dm.org`, 2000.
[12] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, *From Data Mining to Knowledge Discovery: An Overview*. In Advances in Knowledge Discovery and Data Mining, eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Menlo Park, Calif.: AAAI Press, 1-30, 1996.
[13] R. Brachman, T. Anand, *The Process of Knowledge Discovery in Databases: A Human-Centered Approach*. In Advances in Knowledge Discovery and Data Mining, eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Menlo Park, Calif.: AAAI Press, pp. 37-85, 1996.
[14] S. Rüping, *Learning Interpretable Models*, Ph.D. dissertation, University of Dortmund, Germany, 2006.
[15] C. Goble, O. Gocho, P. Alper, D. De Roure, *e-Science and the Semantic Web: A Symbiotic Relationship*,.9th Int. Conf. on Discovery Science (DS2006), Springer, pp. 1-12, 2006.
[16] R. Feldman, J. Sanger, *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*, Cambridge University Press, 2006.
[17] A. Bernstein, F. Provost, *An Intelligent Assistant for the Knowledge Discovery Process* Proc. IJCAI 2001 Workshop on Wrappers for Performance Enhancement in KDD, 2001.
[18] M. Schraefel, G. Hughes, H. Mills, G. Smith, T. Payne, J. Frey, *Breaking the Book: Translating the Chemistry Lab Book into a Pervasive Computing Lab Environment*, in Proceedings of CHI 2004, Vienna, Austria, 2004.
[19] A. Gibson, R. Stevens, R. Cooke, S. Brostoff, M. Schraefel, *myTea: Connecting the Web to Digital Science on the Desktop*, Technical Report.
[20] J. Zhao, C. Goble, R. Stevens, S. Bechhofer, *Semantically Linking and Browsing Provenance Logs for e-Science*, in International Conference on Semantics of a Networked World (IC-SNW), pp. 157-174, Springer, 2004
[21] J.Zhao, C. Goble, M. Greenwood, C. Wroe, R. Stevens, *Annotating, linking and browsing provenance logs for e-Science*, in 1st Workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data, Sanibel Island, Florida, USA, 2003.
[22] Ajax: A New Approach to Web Applications. Adaptive Path (2005-02-18). `http://www.adaptivepath.com/publications/essays/archives/000385.php`

[23] Business Process Execution Language for Web Services version 1.1 http://www-128.ibm.com/developerworks/library/specification/ws-bpel/

[24] ECMAScript Language Specification, Standard ECMA-262, 3rd edition (December 1999) http://www.ecma-international.org/publications/files/ecma-st/ECMA-262.pdf

[25] I. Foster, C. Kesselman, S. Tuecke. *The Anatomy of the Grid: Enabling Scalable Virtual Organizations*. International J. Supercomputer Applications, 15(3), 2001.

[26] I. Altintas, C. Berkley, E. Jaeger, M. Jones, B. Ludäscher, S. Mock *Kepler: An Extensible System for Design and Execution of Scientific Workflows*, 16th Intl. Conf. on Scientific and Statistical Database Management (SSDBM'04), 21-23 June 2004, Greece.

[27] J. Nick, S. Tuecke, *The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration*. Open Grid Service Infrastructure WG, Global Grid Forum, 2002.

[28] Tom Oinn, Mark Greenwood, Matthew Addis, M. Nedim Alpdemir, Justin Ferris, Kevin Glover, Carole Goble, Antoon Goderis, Duncan Hull, Darren Marvin, Peter Li, Phillip Lord, Matthew R. Pocock, Martin Senger, Robert Stevens, Anil Wipat and Chris Wroe. *Taverna: Lessons in creating a workflow environment for the life sciences* accepted for publication in Concurrency and Computation: Practice and Experience Grid Workflow Special Issue

[29] S. Majithia, M. S. Shields, I. J. Taylor, and I. Wang. *Triana: A Graphical Web Service Composition and Execution Toolkit*. Proceedings of the IEEE International Conference on Web Services (ICWS'04), pages 514-524. IEEE Computer Society, 2004.

[30] C Wroe, R Stevens, CA Goble, A Roberts, and M Greenwood. *A suite of DAML+OIL ontologies to describe bioinformatics web services and data*. In International Journal of Cooperative Information Systems, 12(2):197-224, March 2003.

[31] SPARQL Query Language for RDF, W3C Working Draft 4 October 2006, http://www.w3.org/TR/rdf-sparql-query/

[32] SPARQL Protocol for RDF, W3C Candidate Recommendation 6 April 2006, http://www.w3.org/TR/rdf-sparql-protocol/

[33] Resource Description Framework (RDF): Concepts and Abstract Syntax, W3C Recommendation 10 February 2004 http://www.w3.org/TR/rdf-concepts/

[34] Gleaning Resource Descriptions from Dialects of Languages, W3C Working Draft 24 October 2006, http://www.w3.org/TR/grddl/

[35] Jia Yu, Rajkumar Buyya, *A Taxonomy of Workflow Management Systems for Grid Computing*, Technical Report, GRIDS-TR-2005-1, Grid Computing and Distributed Systems Laboratory, University of Melbourne, Australia, March 10, 2005.

[36] Stefano Beco, Barbara Cantalupo, Ludovico Giammarino, Mike Surridge and Nikolaos Matskanis *OWL-WS: A Workflow Ontology for Dynamic Grid Service Composition*, accepted to the 1st IEEE International Conference on e-Science and Grid Computing, Dec. 5 - 8, 2005, Melbourne, Australia

[37] Tim Berners-Lee *Linked Data* Design Issue, http://www.w3.org/DesignIssues/LinkedData.html

[38] Christian Queinnec. *Inverting back the inversion of control or, continuations versus page-centric programming*. Technical Report 7, LIP6, May 2001.

[39] R. Kelsey, W. Clinger, J. Rees (eds.), *Revised5 Report on the Algorithmic Language Scheme*, Higher-Order and Symbolic Computation, Vol. 11, No. 1, August, 1998

[40] D. Manolescu. *Workflow Enactment with Continuation and Future Objects* OOPLSA'02, Seattle, WA, 2002. http://citeseer.ist.psu.edu/manolescu02workflow.html

[41] Semantic Annotations for WSDL W3C Working Draft 28 September 2006 http://www.w3.org/TR/sawsdl/

[42] Workflow Management Coalition Workflow Reference Model, http://wfmc.org/standards/model.htm

This page intentionally left blank

V. State of the Art of the Grid Research and Use at Population Level

This page intentionally left blank

From Genes to Personalized HealthCare: Grid Solutions for the Life Sciences
N. Jacq et al. (Eds.)
IOS Press, 2007

IntegraEPI: A Grid-Based Epidemic Surveillance System

Fabricio A. B. da SILVA[a,1], Henrique F. GAGLIARDI[a,b], Eduardo GALLO[a], Maria A. MADOPE[a], Virgílio C. NETO[c] , Ivan T. PISA[c], Domingos ALVES[d]
[a]*Programa de Mestrado em Informática,UNISANTOS, Santos, Brazil*
[b]*LCCASC,UNISANTOS, Santos, Brazil*
[c] *Departamento de Informática em Saúde (DIS), UNIFESP, Brazil*
[d]*Universidade de São Paulo, Ribeirão Preto, Brazil*

Abstract: The proposal of new analytical techniques has guided innovative methodological developments in public health interventions. The goal of this work is show advances in the development of a large scale system for space-time visualization, monitoring, modeling and analysis of epidemic data using a Grid platform. The resulting virtual laboratory, dubbed IntegraEPI, is expected to provide better epidemic forecasting and to define better strategies to fight the spread of a disease, in which new population-level interventions could be evaluated and iteratively refined using computational simulations, with tangible benefits for real-world epidemic prevention and control efforts.

Keywords: Grid Computing, Epidemiology, Simulation, Data Integration

Introduction

Conventional epidemiology of infectious disease requires extensive collections of population, health and disease patterns data, as well as data related to environmental factors and social conditions. An epidemiologic study may focus on a particular region or a particular outbreak, or it may take as its theme the epidemiology of a condition across a wide area. The range and amount of data required will, therefore, vary depending on the type of study. Moreover, lack of data quality control, lack of definition about the contents to be stored, storage heterogeneity and resource availability are some problems that must be solved to allow more precise and thorough studies in epidemiologic vigilance. Furthermore, analytical studies to identify risk factors related to epidemic development are eventually used by health agencies [1][2]. These studies need several types of data, such as geo-referenced disease cases, space-temporal environmental data relevant to the epidemic prevention and population data based on demographic and geographic information with territory expressiveness.

Considering this scenario, we present in this paper some advances in the development of a large scale system for space-time visualization, monitoring, modeling and analysis of epidemic data studies using a Grid platform [3]. This system, dubbed IntegraEPI, is capable to provide the integration of heterogeneous databases related to epidemic analysis and to make available analytical and computational methods to

[1] Correspondence to: Fabricio Silva, Rua Dr. Carvalho de Mendonça, 144, Santos, SP, Brazil (e-mail: fabricio@unisantos.br). The authors acknowledge CNPq, CAPES and FAPESP for their financial support.

increase the predicting capability of the public heath system, in order to optimize its activities and resources when dealing with epidemic outbreak and prevention.

Particularly, in this work we concentrate ourselves in the data available in two areas in Brazil: the São Paulo State metropolitan coastal region known as "Baixada Santista", composed of nine cities, and a northeastern city of São Paulo State named Ribeirão Preto. Thus, all the conceived systems and infrastructures implemented were initially tested for these Metropolitan regions, carefully incorporating the diversity of the geographic spatial characteristics of the micro-regions. In order to be able to show the system capability in monitoring an epidemic, we work with the local public health system notification data. In this paper we focus on epidemic models for the Dengue fever which is a common disease in Brazil.

This paper is organized as follows: the need for a grid infrastructure is discussed in section 1. The general IntegraEPI architecture is presented in section 2. The services that compose the architecture are detailed in subsections 2.1, 2.2 and 2.3, and section 3 contains our final remarks.

1. Why developing a Grid-Enabled System?

On the beginning of this project many technological decisions were made and the choice of developing grid-based applications was made considering the real nature of the class of applications we should deploy.

In fact, the first feature which IntegraEPI would have to implement is the capability to integrate several health databases containing epidemiologic information with statistical and geographic databases. Such databases contain the necessary data to detect the patterns present on the disease notification process, considering external factors like socio-economic and environmental conditions. The use of grid services to integrate these databases provides transparent and simplified access for geographically distributed databases as if they would be a single and unique virtual database.

Another major reason for developing the IntegraEPI modules as grid-based applications is the forecasting capability provided by the epidemical model simulations, which would require large computational power. The InteraEPI simulator module is a parameter sweep application, and this type of application is especially well-suited for the Grid. The large number of simulations and the amount of computing power needed justify the Grid as the platform of choice to implement the system.

In front of this scenario we have chosen the Globus Toolkit 4 (GT4) middleware [12] to deploy a Grid platform over which the IntegraEPI modules would be implemented. The Globus Toolkit 4 is composed of several services designed for computational grids which provides, for instance, transparent data access [6] and resource virtualization [3][4] in a geographically distributed heterogeneous system. In the next section we should discuss the IntegraEPI system architecture and its main features.

2. The IntegraEPI Architecture

The IntegraEPI system architecture is based on the Open Grid Services Architecture (OGSA) specification [5] and is organized as shown in Figure 1. In fact, there are three main services which were developed: the data integration service Integra-GISE

(IntegraEPI-Grid Data Integration Service) [11] [10], the simulation service Integra-Model and the analysis service Integra-Analysis [9].

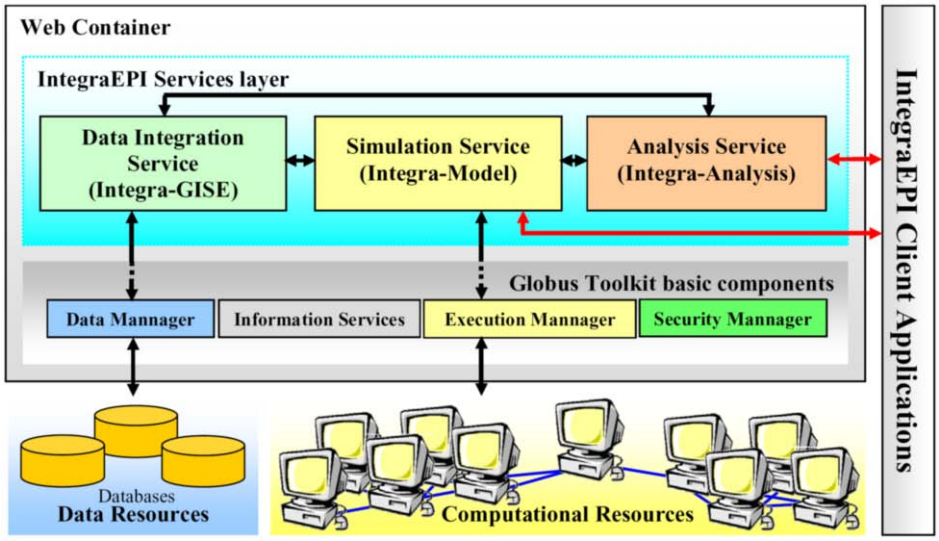


Figure 1. The Integra-EPI System Architecture.

Each one of the IntegraEPI services cover an important aspect of an epidemic surveillance system, such as: the integration of heterogeneous, geographically distributed populational and disease notification databases, a simulation service based on epidemiological models for prediction of the forthcoming trends related the transmission of a particular disease over a city population and, at last, statistical tools to analyze and visualize the status of such diseases, which make possible the inference of risk indicators and to establish epidemiological thresholds to trigger alerts if some risk situation is detected.

Due to the interoperability provided by the use of grid services and the definition of a Common Data Model, the IntegraEPI services are able to interact with each other. The integrated data provided by the IntegraEPI-GISE may be used by both the simulation service Integra-Model and the analysis service Integra-Analysis for estimator inference and detection of risk situations. At the same time, the simulation module can use any of the analysis tools provided by Integra-Analysis to study simulated patterns. Such interoperability allows a large degree of flexibility for service composition and implementation.

In the same figure we can see (on the middle) the Globus Toolkit components which provide services used by IntegraEPI modules such as reliable file transfer, replica location service, security and authorization managers, execution manager components, information services and common runtime components.

After this brief system description, we should describe, in the following subsections, the IntegraEPI main components in more detail.

2.1. Grid Data Integration Service (Integra-GISE)

The data integration service INTEGRA-GISE was developed to obtain data from multiple data sources through a single point of access. Particularly, this service provides an efficient management of the available computational resources. It also makes use of the grid security infra-structure and unique login to access multiple data sources.

Similarly as presented in the general IntegraEPI architecture, the Integra-GISE architecture is divided into layers composed by grid services and resources. The Figure 2 illustrates the GISE architecture, its layers and components.

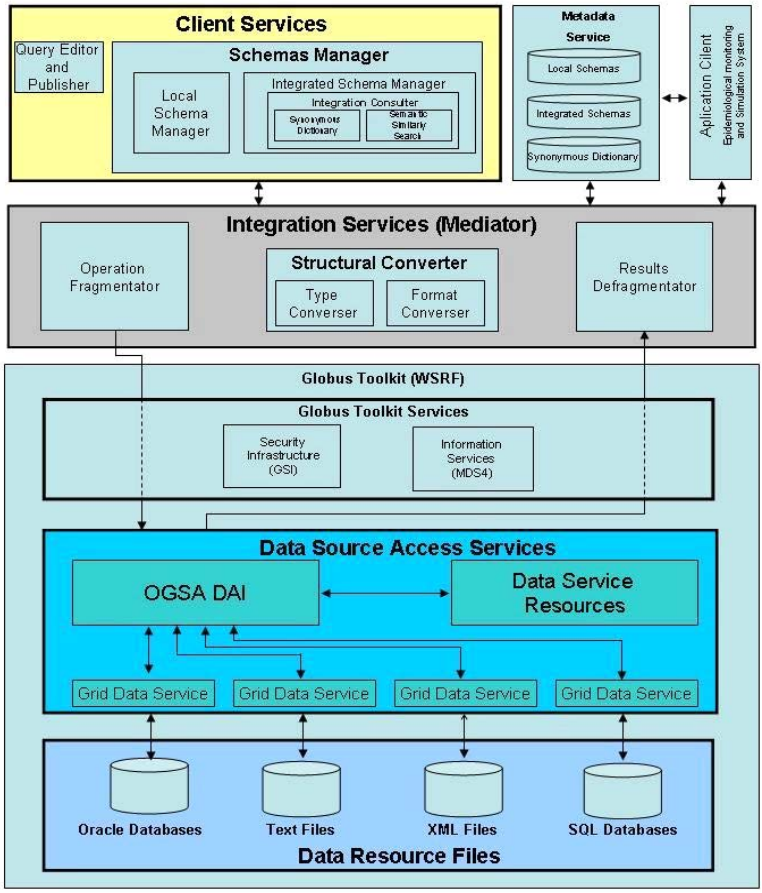


Figure 2. The layers and components of Integra-GISE architecture.

In the simulation and analysis modules, the IntegraEPI system needs to access different data types like maps, epidemic data, geographic information and social-economics data. The required data are geographically distributed in many DBMS such as PostgreSQL, SQL Server, and Oracle. The GISE service provides the necessary data

to other IntegraEPI modules in a transparent manner, by implementing a set of grid data services. In particular, several problems were considered when designing and implementing this module:

a) The number of sources of data is large, complicating still more issues related with the conflict resolution.

b) The available data sources vary dynamically. Therefore the addition and removal of data sources should be made with minimum impact in relation to integrated schemas.

c) The sources of data can have different computational capabilities. The sources of data to be integrated through a grid can vary of simple archives up to parallel DBMS.

For the Integra-GISE we have used the OGSA-DAI (Open Grid Service Architecture - Data Access Integration) [8] services (provided with GT4). The OGSA-DAI services provide wrappers to access data from different sources.

Together with the Integra-GISE service we have also implemented a Metadata Catalogue Service (MCS) (which also access underlying databases using OGSA-DAI services) to provide information about the system data sources which are available through database schemas. Two types of schemas compose the Canonic Data Model of the Integra-GISE service: local data base schemas (GISE-LOCAL-SCHEMA) and integrated schemas (GISE-INTEGRATED-SCHEMA). The integrated schemas may reference several local data base schemas.

The Integra-GISE client layer supplies the necessary components and interfaces used by final users. The layer is composed of the local schema manager, integrated schema generator and the query editor. Furthermore, in this layer it is also implemented the Integration Helper, which is used by the user to solve syntactic conflicts in the process of generating integrated schemas. It is worth noting that, in the Integra-GISE architecture, integrated schemas are generated manually.

The Client Layer builds the query operation through the Query Editor and Publisher, forwarding it to the Integration Services Layer (Mediator). The mediator represents the key element of this integration architecture since it is responsible to split a single query into the many subqueries needed to access the diverse local databases which were referenced by the GISE-INTEGRATED-SCHEMA. The mediator is also responsible for gathering the partial results into a single result set.

When the client application submits a query to the Data Integration Module, some operations are required to process the query. The first step for query resolution is to fetch into the MCS all the information (schemas) related to the local databases referenced at the GISE-INTEGRATED-SCHEMA.

Therefore, the Integra-GISE Service first invokes the MCS service for obtaining the Integrated Schema. The Integrated Schema is a XML-based document containing information about the related databases such as a brief description of them, the databases ID, the databases alias, as well as in which fields the integration occurs.

After the selection of the local databases associated with the query, some specific information is still needed to build all sub-queries with the correct syntax. The Data Integration Service then fetches the description of each individual database (GISE-LOCAL-SCHEMA) in the MCS. Each GISE-LOCAL-SCHEMA is a XML document describing an individual data source.

All the subqueries are executed through the proper data access services and the resulting data sets are joined on a defragmentation process, which can also use the Synonym Dictionary to resolve any database structural or data type conflict, and so the

resultant integrated data set is delivered as a single XML Document to the caller application.

Therefore, environmental, socio-economical, epidemiological, biological, weather and relief data would be retrieved by this service which virtualizes the access to several databases on the grid. Thus, the Integra-GISE importance relies on the data gathering mechanism needed to provide the necessary data for Integra-Analysis and Integra-Model services, which will be presented in the following subsections.

2.2. The Analysis Service (Integra-Analysis)

The analysis module Integra-Analysis is used for spatio-temporal data analysis, through a friendly user interface which provides considerable analysis flexibility. This analytical tool, in the first place, was developed to support a whole set of visualization tools, methods of spatio-temporal cluster detection per area of data, analysis methods of temporal series (self-correlation and spectral analysis) and methods of spatial interpolation for environmental and by area data. Figure 3 shows a particular analysis aimed at the detection of "hot zones" of Dengue spreading in Ribeirão Preto city. The distribution of urban Dengue fever cases over the city map with several clusters of infected individuals can be visualized in (a). To improve the analysis process we can use the *quadrat analysis* method in (b), in which the city map is divided into squares to visualize the 'hot regions' of the city where the epidemic spreading is more critical.

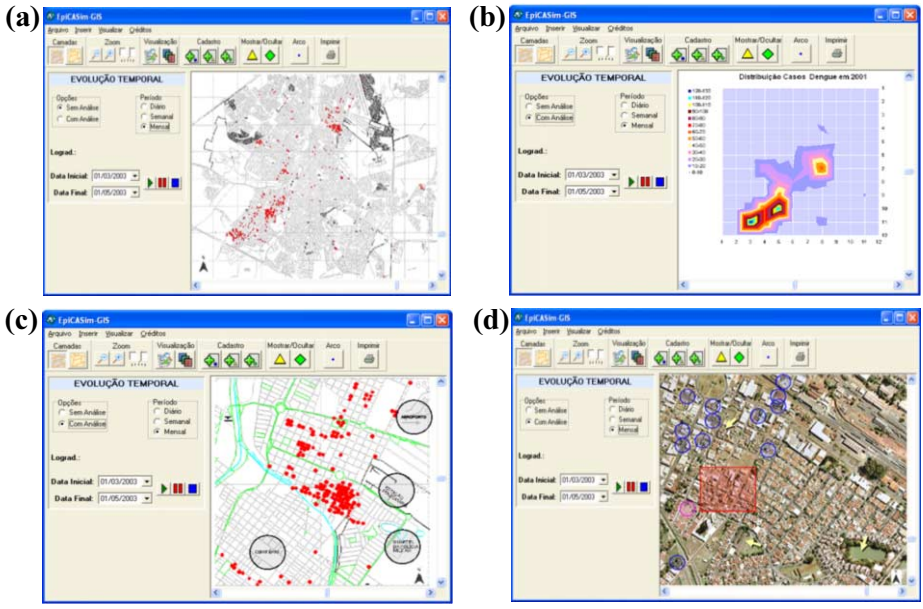


Figure 3. The analysis module features visualized in a client application. This figure illustrates (in a clockwise direction) a monitoring analysis of Dengue cases in Ribeirão Preto, a "hot zone" analysis, an analysis considering strategic points (junkyards, unused terrains) and special landed properties (museums, clubs, public places), and the visualization showing real Dengue fever cases.

After identifying those regions, the analysis module is also capable in (c) to display the most suspicious risk places for mosquito reproduction in the neighborhood (Military Police headquarters, junkyards, airports and cemeteries, for instance). Another visualization tool can be used (d) to identify directly over a real city image the most affected area of that cluster and the possible places of disease spreading due to historical data or behavioral analysis, which helps health agents to act in a more accurate way in the Dengue counter-attack.

In a second moment these analysis tools may used interactively for calibration of both the simulation model and to define thresholds related to alarm and risk indicators, Dengue epidemic dissemination and population social vulnerability. Moreover, this module is able to construct environmental indicators with a scoring methodology to stratify areas in the cities by different levels of risk for Dengue occurrence and transmission.

2.3. The Simulation Service (Integra-Model)

The main goal of the simulation module is to act a forecasting tool for disease spreading through the simulation of epidemiological models based on individual-based networks [7]. This module is used to test the efficacy of several control measures combining richly structuralized GIS networks of the municipal districts, describing the urban structure in multiple scales. It also considers realistic estimates and parameterized populational mobility and interactivity models, as well as disease progress among its hosts.

Generally, this service is capable to collect information through the Integra-GISE integration service and build a virtual city where a disease model will be applied to verify the possible scenarios of disease spreading among a virtual population. In this way, the provided data is processed to define the address location of every infected person in the population network. The additional infrastructural information gathered is used to infer the susceptibility of the population by populational zones, considering poverty, educational level, socio-economical data as well to build the simulation scenario. Therefore, the most important feature presented by this simulation service is the capability to provide a substrate to represent the reality of a given municipal district as a "virtual city".

In this world build from the information provided by Integra-GISE, aspects like weather, relief, demographical density, social caress indexes, historical epidemiologic data, hydrography, urban or rural limits and places with large probability or tendency for the development of some type of vector agent are considered. The data are transformed into model parameters to be submitted as a Bag-of-Tasks[2] (BoT) job on the grid, via the Globus Toolkit service WS-GRAM.

It is worth noting that the Integra-Model service is capable to adapt itself to the reality of a given city, just needing to be parameterized with the data (fetched by Integra-GISE) of the metropolitan region under study, since the inter-individual interactions modeled remain unaltered.

Thus, with a single model we can simulate the behavior of a disease considering various distinct scenarios, and every single disease will have its simulator which needs to be built in a modular and parameterized way. The fact is that every infectious disease

[2] In this work we use the terms "Bag-of-Tasks" and "Parameter-Sweep" interchangeably.

has its particular rules, transitions and modus operandi. Therefore, a specific canonic data model is necessary for every disease model provided on Integra-EPI system.

The simulator itself is a parameter-sweep application, in which an independent task is generated for each different set of parameters, in a *bag-of-task* approach. The model for the dengue fever is in advanced stage of development and it is being used for testing the Integra-Model service. The dengue fever model is described in the following.

The main assumption of our approach is that the dengue transition rules are defined by two sets of states, each one representing the behavior of a distinct population. In the language of state-variable models, the humans are defined by a SEIR (Susceptible, Exposed, Infective and Recovered) model which, when in the exposed state, the individual is infected but still not infective [7]. The mosquito population is represented by a different SEI (Susceptible, Exposed and Infective) model as shown in Figure 4.

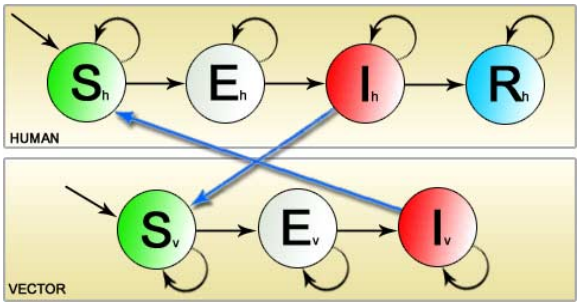


Figure 4. The schematic model of dengue spreading representing the stages of the disease for both populations, where *thick edges* represent the interaction among populations (mosquito bite) and the thin ones the internal state transitions in each population.

The dengue model has a singular characteristic, different from other models based in cellular automata: the use of two overlapping interacting automata cellular grids to represent the human and the vector (Aedes aegypti mosquito) populations. Therefore, the neighborhood of each individual has a different meaning and is not defined around each individual cell in the same population but at an equivalent position at the overlapping population, i.e. a neighborhood of a human cell is defined by the mosquitoes (vector) population and vice-versa. This special characteristic allows for the possibility of a human host both being locally infected by a vector and to infect another mosquito. This interaction (a bite), however, never occurs directly among the neighbors of a same population, because a human may only become infected by the Aedes aegypti bite and a mosquito only becomes infected by biting an infective human. Therefore, we propose a novel framework to model the spread of a dengue outbreak. Besides defining internal state transition rules for each iterative population (human and mosquitoes), we define iterative rules between these two cellular automata (Figure 5), reflecting the interaction between populations.

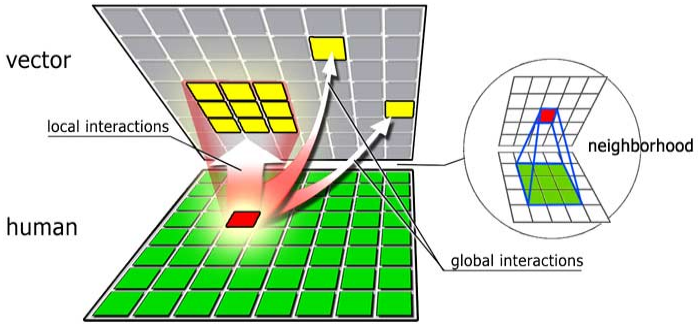


Figure 5. The local and global effects are shown in this figure. The *pointed squares* represent the mosquitoes affected by the local and global human infective influence. The same type of effects occurring in this bottom-up direction for human-vector interactions also occurs for the vector-humans interaction in a top-down direction at each simulation time-step. We can see also in this figure a schematic representation of the *neighborhood* of a single element.

To build the virtual city used in simulations several interrelated layers are used to map city features, like the demographic density, relief, hydrography and weather information. Therefore, considering the urban Dengue fever model [7], each population (human and mosquitoes) should have their own characteristics modeled. However, it is worth noting that some layers are more useful to humans than to mosquitoes and vice-versa. In general, using the Integra-Model service, we expect to be able to identify different levels of risk for a particular disease occurrence and transmission, for a predefined city or metropolitan area, considering the populational groups living in the city. Our model serves as a local strategic complement to other simulation models developed to identify epidemiological interactions within a given county or city.

3. Concluding Remarks

The essential proposal of this research project is to contribute for the modernization of the epidemiologic monitoring system by comparing results of detailed simulations with the observed experimental data related to the spreading of a disease, considering both temporal and geographic aspects. Particularly, the implementation over a computational grid platform open completely new perspectives for gathering data on large populations and - as a consequence - allow stratification of large scale Metropolitan epidemiology studies. Other advantages of this technology are the small deployment cost and high processing and storage capacities. These advantages become even more important when considering the deployment costs of mainframes or supercomputers for countries like Brazil.

Finally, it is important to emphasize that the expected results to be obtained during the development of this project do not apply solely to epidemics. There is a whole class of the public health problems of spatial and temporal nature, over which simulating, detecting, monitoring and visualizing patterns is part of the response to the problem.

4. References

[1] A. Wagstaff. *Socioeconomic inequalities in child mortality: comparisons across nine developing countries*. Bull World Health Organ 2000; 78:19-29.

[2] O. Axelson. *The character and the public health implications of ecological analyses*. In: Disease Mapping and Risk Assessment for Public Health (A. Lawson, A. Biggeri, D. Böhning, E. Lesaffre, J.-F. Viel & R. Bertollini, ed.), pp. 301-309, New York: Wiley and Sons, 1999.

[3] I. Foster. *What is the Grid? A Three Point Checklist*. GRID today, vol. 1, no. 6, 2002. Available at: http://www.gridtoday.com/02/0722/100136.html

[4] I. Foster, C. Kesselman, S. Tuecke. The Anatomy of the Grid: Enabling Scalable Virtual Organizations. International J. Supercomputer Applications, 15(3), 2001.

[5] I. Foster, C. Kesselman, J. Nick, S. Tuecke: *The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration*. June/2002. Available at: http://www.globus.org/research/papers.html

[6] I. Chervenak et al. *The Data Grid: Towards an Architecture for the Distributed Management and Analysis of Large Scientific Datasets*. Journal of Network and Computer Applications, 23:187-200, 2001.

[7] H. F. Gagliardi, F. A. B. da Silva and D. Alves. *Automata Network Simulator Applied to the Epidemiology of Urban Dengue Fever*. Springer-Verlag Berlin Heidelberg: Lecture Notes in Computer Science - ICCS 2006, Part III, LNCS 3993, pp. 297 – 304, 2006.

[8] *Open Grid Services Architecture Data Access and Integration.* Available at: http://www.ogsadai.org.uk/.

[9] A. Rezende, H. F. Gagliardi, R. C. Serafim, F.A.B. Silva, D. Alves. *IntegraEPI-GIS: A Geographic Information System to visualize and analyze the spatio-temporal patterns of the spread and control of epidemics.* Proceedings of the Spatial Epidemiology Conference 2006, London

[10] E. Gallo, H. F. Gagliardi, V. C. Neto, D. Alves, F.A.B. Silva. *GISE: A Data Access and Integration Service for a Grid-Based Epidemic Surveillance System*. Proceedings of the 40th Annual Simulation Symposium, 2007

[11] V. C. Neto, H. F. Gagliardi, A. Rezende, E. S. Gonçalves, E. Gallo, F.A.B Silva, I. T. Pisa, D. Alves. *Data Access Service in a Computational Grid Platform Applied to the Monitoring and Control of Epidemics on Georeferenced Dynamic Maps.* Proceedings of the 2nd IEEE Conference on e-Science and Grid Computing, 2006.

[12] The Globus Alliance. http://www.globus.org. Accessed in February 2007.

From Genes to Personalized HealthCare: Grid Solutions for the Life Sciences
N. Jacq et al. (Eds.)
IOS Press, 2007

Accelerating Medical Research using the Swift Workflow System

Tiberiu STEF-PRAUN [b,1], Benjamin CLIFFORD [b], Ian FOSTER [a,b], Uri HASSON [b], Mihael HATEGAN [b], Steven L. SMALL [b], Michael WILDE [a,b], Yong ZHAO [a,b]

[a] *Argonne National Labs*
[b] *University of Chicago*

Abstract. Both medical research and clinical practice are starting to involve large quantities of data and to require large-scale computation, as a result of the digitization of many areas of medicine. For example, in brain research – the domain that we consider here – a single research study may require the repeated processing, using computationally demanding and complex applications, of thousands of files corresponding to hundreds of functional MRI studies. Execution efficiency demands the use of parallel or distributed computing, but few medical researchers have the time or expertise to write the necessary parallel programs.

The Swift system addresses these concerns. A simple scripting language, SwiftScript, provides for the concise high-level specification of workflows that invoke various application programs on potentially large quantities of data. The Swift engine provides for the efficient execution of these workflows on sequential computers, parallel computers, and/or distributed grids that federate the computing resources of many sites. Last but not least, the Swift provenance catalog keeps track of all actions performed, addressing vital bookkeeping functions that so often cause difficulties in large computations.

To illustrate the use of Swift for medical research, we describe its use for the analysis of functional MRI data as part of a research project examining the neurological mechanisms of recovery from aphasia after stroke. We show how SwiftScript is used to encode an application workflow, and present performance results that demonstrate our ability to achieve significant speedups on both a local parallel computing cluster and multiple parallel clusters at distributed sites.

Keywords. Brain research, Grid Computing, Workflows

1. Introduction

Abundant examples exist of highly computational medical research in domains of vital importance, and of infrastructures focused on supporting such research [10,13]. We describe and advocate here the use of Grid computing technologies toward this end, and present a case study in which these tools are applied to the needs of medical research.

An attractive approach to enhancing the productivity of medical research is to use existing Grid infrastructure. There are several good reasons for this, the most important being its ready availability through global cyberinfrastructure, and the substantial capa-

[1] Corresponding Author: Tiberiu Stef-Praun, 5640 S. Elllis rm. 405, University of Chicago, 60615, IL, USA

bilities that can be employed by Grid users. The Grid can be seen as a large, distributed computing resource used in common by a wide group of scientists. From the end-user's point of view, the Grid is a powerful, multi-user computer, with familiar resource sharing and user access mechanisms. Grid resources are shared according to specified policies or may be reserved upon request. The security of each user's applications and data is enforced by standard Unix permissions and by enhanced access control list security. There are storage services, large scale execution services, easy and efficient data movement utilities, and supporting tools that allow users to take advantage of large amounts of computing power. These capabilities can conveniently address specific requirements of medical research, such as access control to patient data, as mandated by HIPAA rules in the USA; high bandwidth to rapidly transfer large DICOM images from the patient's records [6]; and sophisticated image analysis algorithms to aid in the interpretation of medical conditions.

We describe our success to date in applications in imaging-based neurological research in which we seek to expand the scope and scale of our computing capabilities to study the neurological mechanisms of the process of recovery from aphasia due to stroke. The neuroscience behind the system described below is based on work in permutation tests for clustering analysis [3,12]. The aphasia recovery study (described below) involves the processing of many large functional MRI files through a set of domain-specific applications, and re-running a statistical analysis protocol over a large parameter space.

The benefits of using the Swift workflow system (described in detail below) for this application include large reductions in the data and computing resource management effort that is typpically required in modern scientific research. This is achieved by automating otherwise manual and labor-intensive processes. In addition to providing transparent and on-demand access to Grid resources, our workflows also exhibit reproducibility and provenance tracking of the data results, thus enabling collaboration in the actual research process, not only in sharing the results. The sharing and reuse of the actual research processes, as well as of the data, has already shown benefits in several other domains (e.g., physics, sociology, economics) in which we also have ongoing collaborations.

2. Grid Infrastructure

The purpose of the Grid software stack is to hide the complexity of the vast resources being made available to the users. In a sense this software can be seen as an operating system that provides the user with the desired functionality of transparent distributed data access and application execution.

In large distributed systems, the granularity of the atomic actions changes in scale: instead of CPU-based operations, users describe their work in terms of application invocations. They use existing applications and existing data, customize them, and iterate over them until the sought-after analysis results are obtained. At the next level, the distributed system middleware consists of software that transparently manages the execution of applications on computers that are typically managed by local resource management systems such as Condor [11] or PBS [1]. On top of these computing clusters, the Grid software provides an access layer that gives users uniform interface to distributed resources, and provide tools to manage data transfers, applications execution, and many other features that hide the heterogeneity of resources across distributed clusters. The

current standard for homogenizing the resources is described by the Open Grid Services Architecture (OGSA), and implemented by the Globus Toolkit, among others. At this level of the Grid software stack, users can access remote data and execute remote applications, but they still have to be aware of the networked nature of Grids, and must manage the executions of their applications. A higher-level component that abstracts the distributed nature of Grid resources is the Swift system [7], which maps "virtual local applications" to their corresponding physical installation on remote Grid sites.

2.1. An Application User's View

Scientific researchers do not, in general, want to be aware of computing infrastructure: they want easy to use, high-performance applications that deliver fast and accurate results with little effort, and minimal disruption of their scientific thought process. This dictates that their applications must be reliably and transparently allocated the resources necessary to solve the problems at hand. Our solution to making tools available on the Grid in a fashion which hides the complexities of manually managing different computers is to *virtualize the location* of the tool through a level of indirection. This indirection is implemented in Swift in an internal directory consisting of the application and computing-site descriptors. With Swift, Grid users can consider their applications to be virtually local to them: the selection of the site to execute the application and the transfer of input and output data and parameters to the site is handled transparently by Swift.

In addition to the transparency of accessing the tools, researchers benefit from the ability to create complex functionality by composing simple tools that each solve some subproblem of the researcher's agenda. These gains are further enhanced by the expressivity of SwiftScript, which allows such complex algorithms to be expressed in clear, simple, and high-level logical terms, rather than in low-level physical details.

Treating the existing applications that the scientists use as the *atomic computation units* of the workflow describing the algorithm, we have built a workflow execution engine around the concept of *data flow analysis*: whenever the input data for an atomic computation unit in the workflow becomes available, the engine selects a Grid resource and sends out the computation and the data to that site. At the end of the computation, the engine copys back the results, to make them available to subsequent workflow steps that depend on this result as their input, or copies them to a repository for archival, dissemination, or later analysis.

2.1.1. SwiftScript Language Constructs

SwiftScript extends the earlier Virtual Data Language [7] with support for dataset typing and mapping, dataset iteration, conditional branching, sub-workflow composition, and other advanced features. SwiftScript support for standard programming language control constructs make it easy to script the execution of applications and thus to automate the research process. For example, loops are used to iterate through parameter sweeps, mappers to associate inputs and outputs with actual file names, and arrays to store groups of similar datasets. We describe in the implementation section below how SwiftScript can be used to naturally and effectively express workflows describing neuroscience research tasks.

3. Aphasia and Brain Research Tools

Stroke, in addition to being the third highest cause of death in the United States, is the leading cause of disability among adults. (American Heart Association. (2003). 2003 Heart and Stroke Statistical Update. Dallas, Texas: American Heart Association.) Thus there is intense interest in the clinical research community in understanding the neurological mechanisms involved in recovery from stroke. One such research area, in which we are involved, specifically focuses on analysis of the recovery phase from *aphasia caused by stroke*, and effects on the neurobiological aspects of the patient. In our aphasia-recovery studies, we apply fMRI to analyze neural activity (BOLD response) in the brains of subjects after stroke, in response to various congnitive stimuli.

3.1. The Research Problem

The medical research behind testing the SwiftScript workflow technology on the grid is the study of the stroke recovery process in a set of patients. The study uses fMRI brain images data of the patients subjected to various stimuli to detect neural activations in the brain as a result of the experimental conditions. However, given the practically limited number of patients available for a typical imaging study, the results of the activation detection process are likely to suffer from the uncertainty of random brain activations. Thus, besides the actual activation detection, the research plan also contains of a verification phase to analyze the validity of the results. This step involves assessment of a null hypothesis about the results obtained from the experiment's data using random modifications of the original fMRI readings.

3.2. The Scientific Methodology

In fMRI studies, data are sampled from spatial locations in a resolution measured in voxels. Statistical analysis in a typical experiment with two conditions (e.g., viewing circles vs. viewing faces) is based on the following steps:

1. Spatially align all the brain images from an experimental run.
2. For each subject, for each voxel, establish the activity level (BOLD response) for each condition (2 data points), and save the difference in activity (delta).
3. At the group level, analyze these delta values to establish, for each voxel, whether subjects' deltas differ from zero. This is performed by calculating whether the delta vector for each voxel (of length N = number of subjects) has a mean that is reliably greater than 0, using a t-test (i.e., testing if the two conditions differ reliably).
4. on the group level: Once we establish for each voxel whether there is a reliable difference between the two conditions, find reliable clusters of activity.

Because there are many thousands of voxels in our brain images, some would be "active" just by chance (e.g., if data were randomly sampled). The permutation algorithm identifies which clusters of neural activity are not likely to be found by chance. In brief, the method tests the null hypothesis that the clusters of activation found in the dataset are indeed likely to be found by chance. The null hypothesis asserts that if we were to "switch" the labels of the conditions for one or more participants, and calculate the delta values in each voxel, we would get equally large activations. To test this null hypothesis,

for one or more participants (in all possible combinations), we interchange the labels of the two conditions, re-calculate the reliability of delta in each voxel (step 3), and evaluate the clusters we find. If the clusters in our data are greater than the majority of the clusters found in the permutations, then the null hypothesis is refuted and we conclude that the clusters of activity found in our study are not likely to be found by chance.

4. Grid Implementation

We coded the algorithm described above in SwiftScript and then installed on the Grid the software applications that were previously used on desktop workstatins to solve the original problem.

In the aphasia-recovery study, the main tools used were the *R* [14] Statistical Package, used to generate the data for the null-hypothesis testing, and the *SUMA* [16] tool, part of the AFNI [4] package, for computing the clustering of neural activity levels.

The input files can be separated into two classes. The first group consists of experiment-dependent inputs, such as the files that contain the brain activity measurements from the experiments (the `origBrain` file). The second group consists of files that are required by the tools involved in the processing, such as the full standard brain files `brainFile, specFile` needed by AFNI to map the experimental measurements.

There is a special set of files which result as a by-product of the data-processing focus of the SwiftScript workflow language. These are the intermediary files, that are produced by the application components that make up the final workflow, and which are being fed as inputs to the subsequent blocks in the workflow. In the example below, they have names like `randomBrain, randomCluster, dsetReturn, clusterThresholdsTable`.

Following the *location virtualization principles* described earlier, these file names are *mapped transparently* from real files that exist on the computer running the workflow to logical names that the SwiftScript program uses to describe the workflow data entities.

4.1. SwiftScript Representation of the Aphasia Algorithm

The SwiftScript description of the algorithm first defines the data types of each dataset (file) that participates in the workflow. For clarity, we define a unique type for each file containg syntactically and/or semantically different kinds of data:

```
type file {}
type fileNames{ file f[]; }
type script {}
type brainMeasurements{}
type precomputedPermutations{}
type fullBrainData {}
type fullBrainSpecs {}
type brainDataset {}
type brainClusterTable {}
type brainDatasets{ brainDataset b[]; }
type brainClusters{ brainClusterTable c[]; }
```

Having defined the types of the data entities in the workflow, we define the procedures that process the input files. Some procedures serve as interface wrappers for external programs, and map the input and output parameters used in the SwiftScript workflow to the actual physical arguments of the application program.

```
// Procedure to run R statistical package
(brainDataset t) bricRInvoke (script permutationScript, int iterationNo,
               brainMeasurements dataAll, precomputedPermutations dataPerm){
       app { bricRInvoke @filename(permutationScript) iterationNo
                         @filename(dataAll) @filename(dataPerm); }}

// Procedure to run AFNI Clustering tool
(brainClusterTable v, brainDataset t) bricCluster (script clusterScript,
int iterationNo, brainDataset randBrain,
fullBrainData brainFile, fullBrainSpecs specFile) {
       app { bricPerlCluster @filename(clusterScript) iterationNo
                             @filename(randBrain) @filename(brainFile)
                             @filename(specFile);}}

// Procedure to merge results based on statistical likelhoods
(brainClusterTable t) bricCentralize ( brainClusterTable bc[]) {
       app { bricCentralize @filenames(bc); }}

(brainDataset t) makebrain (brainDataset randBrain,
brainClusterTable threshold, fullBrainData brain,
fullBrainSpecs spec){
        app {   makeBrain @filename(randBrain) @filename(threshold)
                          @filename(brain) @filename(spec); }}
```

Other procedures use more complex language constructs such as iterations and conditional constructs to combine several atomic application invocations.

```
// Procedure to iterate over the data collection
(brainClusters randCluster, brainDatasets dsetReturn) brain_cluster (
fullBrainData brainFile, fullBrainSpecs specFile) {
    int j[]=[1:2000];
    brainMeasurements     dataAll<fixed_mapper; file="obs.imit.all">;
    precomputedPermutations dataPerm<fixed_mapper; file="perm.matrix.11">;
    script                  randScript<fixed_mapper; file="script.obs.imit.tibi">;
    script                  clusterScript<fixed_mapper; file="surfclust.tibi">;
    brainDatasets           randBrains<simple_mapper; prefix="rand.brain.set">;
    foreach int i in j {
        randBrains.b[i] = bricRInvoke(randScript,i,dataAll,dataPerm);
        brainDataset rBrain=randBrains.b[i];
        (randCluster.c[i],dsetReturn.b[i]) =
            bricCluster(clusterScript,i,rBrain, brainFile,specFile);
} }
```

Having declared the data types and the procedures that will process the data, we must define the dynamic mapping of the logical file names used in SwiftScript to actual on-disk file resources. This mapping can range from simple name-to-file mapping to database-select operations or the matching of multiple files by a regular expression, based on the choices available in an extensible library of mapper implementations.

```
fullBrainData       brainFile<fixed_mapper; file="colin_lh_mesh140_std.pial.asc">;
fullBrainSpecs      specFile<fixed_mapper; file="colin_lh_mesh140_std.spec">;
brainDatasets       randBrain<simple_mapper; prefix="rand.brain.set">;
```

```
brainClusters       randCluster<simple_mapper; prefix="Tmean.4mm.perm",
                              suffix="_ClstTable_r4.1_a2.0.1D">;
brainDatasets       dsetReturn<simple_mapper; prefix="Tmean.4mm.perm",
                              suffix="_Clustered_r4.1_a2.0.niml.dset">;
brainClusterTable   clusterThresholdsTable<fixed_mapper; file="thresholds.table">;
brainDataset        brainResult<fixed_mapper; file="brain.final.dset">;
brainDataset        origBrain<fixed_mapper; file="brain.permutation.1">;
```

The actual workflow consists simply of invocations of the high-level procedures defined above:

```
// Main program: launches the entire workflow
(randCluster, dsetReturn) = brain_cluster(brainFile, specFile);
clusterThresholdsTable= bricCentralize (randCluster.c);
brainResult=makebrain(origBrain,clusterThresholdsTable,brainFile,specFile);
```

Note that this simple description, at which level most researchers will work, enhances productives by abstracting and automate many complex tasks. For the scientific research that we described above, the two thousand invocations of the block in the `braincluster` function were determined individualized processing of the `bricRInvoke` function, depending on the parameter `i`. Also, the workflow performs automatic synchronization of the many subtasks involved, waiting for the result of these two thousand executions to finish before continuing with the merging (`makebrain`) procedure.

4.2. The Swift Environment, and Grid Application Deployment

For completeness, we summarize the additional infrastructure that enables the transparent execution of the workflow described above. The application components (containing the problem-solving algorithms) that are invoked in the workflow, must be installed at the sites that are to be involved in the computation. This step is generally done once, as part of application deployment. In our case we installed the applications `bricRInvoke, bricPerlCluster, bricCentralize,` and `makebrain` at several sites, which we recorded in Swift catalogs. Swift uses these catalog entries to choose on which sites, and to what degree of parallelism, to invoke the applications.

Internally, Swift uses Globus [8] software for important functions such as authentication with remote sites, data transfer, and remote task invocation. We run our applications on several sites spanning the Teragrid, Open Science Grid, and independent institution clusters. The architecture of the infrastructure involved in executing one's workflow on the Grid is depicted in Figure 1.

Other Swift facilities allow the user to resume the workflow from the point of any failure, to cluster short-running applications for more efficient remote execution, and to visualize the progress of workflow execution. Figure 2 shows a snapshot of the executing workflow.

5. Results

5.1. Benefits of Grid Computing in Health-related Research

To measure the benefits of using workflow systems to manage research data analysis in Grid environments, we recorded the execution time of the same workflow instance in

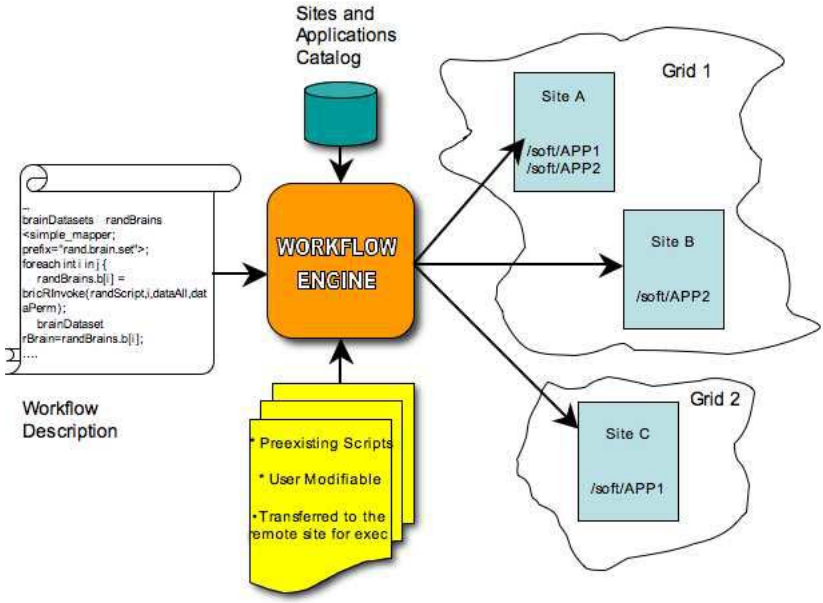


Figure 1. The components of a workflow based application

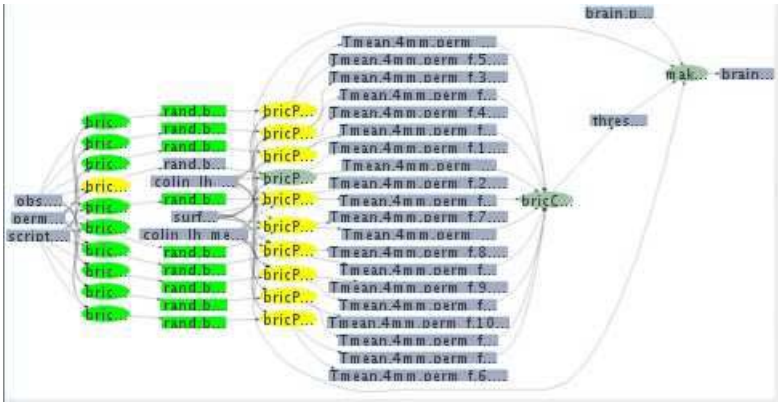


Figure 2. The execution of a small subset of the aphasia study analysis workflow. The colors of the boxes indicate if the task has completed (green) or is executing (yellow). Lines represent data dependencies.

both a local workstation and a distributed Grid environment, and provide initial results in Table 1 below. The performance gains depend primarily on the parallelism that the workflow exhibits (in this case we had two thousand parallel execution threads), on the available number of sites that could execute the applications that made up our workflow (in our case, three sites), and the number of simultaneous jobs executed by those sites (which depends in turn on local cluster sizes, on cluster resource management policies, and on contention on the cluster from the other users that share them).

Other than speedup results, the current implementation allows the researcher to modify the scripts that are used in this workflow, as we chose a model where these scripts,

Table 1. Timing measurements for executing grid versus local execution of the aphasia workflow

Local	Grid
3 min/1 job instance	5 min / single instance run on the Grid
300 min / 100 job instances	50 minutes / 100 job instances

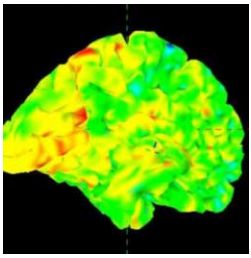


Figure 3. An intermediate stage of activation analysis as processed by the workflow

containing the actual scientific procedures, are deployed on demand, dynamically to the Grid sites.

While we used only three sites in this study, we could increase that number to improve the workflow's speedup significantly. We note a major benefit if this approach here: other researchers using the same tools that were used in this work could readily use the already deployed applications that we used as well (SUMA, R), or simply obtain the current workflow definition and execute it without the need of any special setup. Swift also allows us to visualize both the workflow' s execution (Figure 1) and the "real-time" display of the activations on the brain, displayed in Figure 3.

6. Related Work

Swift has its origins in the GriPhyN Virtual Data System (VDS) [7], originally designed to automate the analysis of the large quantities of data produced by high energy physics experiments. Another VDS component, Pegasus [5], implements specialized strategies for scheduling tasks on computing sites.

Much work on workflow for eScience has focused on the orchestration of web service invocations, as supported, for example, by BPEL and by Taverna [17], which implements a BPEL subset. Kepler [2] is used for similar purposes. We view Swift as addressing a different problem than these systems, namely the orchestration of large numbers of calls to application programs, and their practical and transparent execution in a distributed Grid.

GenePattern [15], like Swift, focuses on the composition of application programs. It differs in its graphical programming approach, and its lack of support for large-scale parallel processing. Google's MapReduce [9], like Swift, focuses on the large-scale analysis of large quantities of data. Swift differs in its support, via XDTM, of diverse file system structures, and its support for task-parallel as well as data-parallel execution.

7. Summary

We have introduced a tool, Swift, that supports the parallel and distributed execution of computationally demanding and data-intensive scientific computations. Using an example from a clinical study of aphasia recovery, we have described how Swift allows (via its scripting language, SwiftScript) for the concise representation of complex algorithms, for the efficient execution of those algorithms on parallel and distributed ("grid") computing systems, and the subsequent exploration and assessment of the workflow's execution history.

References

[1] http://www.openpbs.org/.

[2] C Berkley D Higgins E Jaeger M Jones B Ludäscher, I Altintas. Scientific workflow management and the Kepler system. *Concurrency and Computation: Practice and Experience*, 18(10), 2005.

[3] Suckling J. Stephan Overmeyer Rabe-Hesketh S. Taylor E. Brammer M. J Bullmore, E. T. Global, voxel, and cluster tests, by theory and permutation, for a difference between two groups of structural mr images of the brain. *IEEE Transactions on Medical Imaging*, 18(1):32–42, 1999.

[4] Robert W. Cox. Afni: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res.*, 29(3):162–173, 1996.

[5] Y Gil C Kesselman-G Mehta M Su K Vahi M Livny E Deelman, J Blythe. Pegasus: Mapping scientific workflows onto the grid. In *Grid Computing*, 2004.

[6] Chervenak A Nelson MD-Kesselman C. Erberich SG, Bhandekar M. Dicom grid interface service for clinical and research pacs: A globus toolkit web service for medical data grids. *DICOM grid interface service for clinical and research PACS: A Globus toolkit web service for medical data grids. International Journal of Computer Assisted Radiology and Surgery*, 1, 2006.

[7] M. Wilde Y. Zhao. I. Foster, J. Voeckler. Chimera: A virtual data system for representing, querying, and automating data derivation. *International Conference on Scientific and Statistical Database Management*, 2002.

[8] S. Tuecke I. Foster, C. Kesselman. The anatomy of the grid: Enabling scalable virtual organizations. *International J. Supercomputer Applications*, 2001.

[9] S Ghemawat J Dean. Mapreduce: Simplified data processing on large clusters. In *OSDI*, 2004.

[10] Lynne W. Kakazu K. K., Cheung L. W. The cancer biomedical informatics grid (cabig): pioneering an expansive network of information and tools for collaborative cancer research. *Hawaii Medical Journal*, 63, 2004.

[11] Matthew Mutka Michael Litzkow, Miron Livny. Condor - a hunter of idle workstations. In *Proceedings of the 8th International Conference of Distributed Computing Systems*, June 1988.

[12] Holmes A. P. Nichols, T. E. Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping*, 15(1), 2002.

[13] Fragoso G. Warzel D. Phillips J., Chilukuri R. and Covitz P. A. The cacore software development kit: Streamlining construction of interoperable biomedical information services. *BMC Med Inform Decis Mak*, 6(2), 2006.

[14] R Development Core Team. *R: A Language and Environment for Statistical R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2005.

[15] Gould J Lerner J-Tamayo P Mesirov JP Reich M, Liefeld T. Genepattern 2.0. *Nature Genetics*, 38(5):500–501, 2006.

[16] Argall B. Japee S.-Cox R.W. Saad Z.S., Reynolds R.C. Suma: an interface for surface-based intra- and inter-subject analysis with afni. In *Biomedical Imaging: Macro to Nano, 2004. IEEE International Symposium on*, volume 2, pages 1510– 1513. Dept. of Health and Human Services, Nat. Inst. of Health, Bethesda, MD, USA;, 2004.

[17] M Addis M. N Alpdemir J Ferris K Glover C Goble A Goderis D Hull D Marvin P Li P Lord M R. Pocock M Senger R Stevens A Wipat C Wroe T Oinn, M Greenwood. Taverna: lessons in creating a workflow environment for the life sciences. *Concurrency and Computation: Practice and Experience*, 8(10), 2005.

VI. Security

This page intentionally left blank

From Genes to Personalized HealthCare: Grid Solutions for the Life Sciences
N. Jacq et al. (Eds.)
IOS Press, 2007

Does HealthGrid Present Specific Risks With Regard To Data Protection?

Jean HERVEG
Centre de Recherches Informatique et Droit, Faculté de Droit, FUNDP

Abstract. Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data manages the risks in the processing of personal data in four steps. It provides notably that data processing presenting specific risks must be subject to prior checking beforehand. The paper investigates the theory of risks in Directive 95/46/EC, together with the criteria allowing to recognize those specific risks. Finally, the paper describes the consequences of such specific risks on the data processing.

Keywords. Processing of Personal Data – Specific Risks – European Law – HealthGRID

INTRODUCTION

The introduction of GRID technologies in healthcare arouses numerous legal questions [1]. Among these, one is to know whether the use of HealthGRID technologies could induce specific risks to the rights and freedoms of the data subject concerned by the underlying processing of personal data. Indeed, European Directive 95/46/EC imposes the prior checking of personal data processing presenting such specific risks. This legal issue is relatively important but should be serenely debated. This contribution aims to identify the criteria allowing to recognize these specific risks when using GRID technologies in healthcare.

1. The "Theory of Risks" in Directive 95/46/EC

European Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data [2] pursues a double objective when harmonizing the national legislations of the European Member States. It aims to allow for the free movement of personal data, asserted as necessary to the creation and the operating of the Common Market [3], and to ensure the respect of the rights and freedoms of the natural persons (individuals) concerned by the personal data [4]. The natural persons' rights and freedoms include their right to control in some way their personal data.

In order to remove the obstacles to the free movement of personal data in the Common Market, it is of prime importance to harmonize national legislations, so that all Member States offer an equal but high level of protection towards the rights and freedoms of the

persons regarding the processing of their personal data [5]. After such harmonization, the Member States may not prevent anymore the free movement of personal data for reasons relative to the protection of natural persons' rights and freedoms, these including the right to respect for private life. As the harmonization is limited in its material scope, the Member States may restrict the free movement of personal data for other reasons than those relative to the protection of natural persons' rights and freedoms [6] – without prejudicing the application of articles 95.8 and 95.10 of the Treaty creating the European Community or of any other rules opposing any restriction to the free movement of personal data within Member States or the Common Market.

In order to establish this legal framework shared by all European Member States (although relatively incomplete in a sense) regarding the processing of personal data, in the limits of its legal scope, the Directive results from a quantitative and qualitative assessment of the risks which the personal data processing may cause to the data subjects' rights and freedoms. This assessment has been realized to all levels of the Directive's scope. In this measure, the Directive determines its material scope (cf. Chapter One of the Directive) [7]. It focuses only on situations which require some protection. The latter implies to estimate the risks for the data subjects' rights and freedoms. For example, the Directive only applies to the completely or partially automated processing [8] of personal data [9] and to the non-automated processing of personal data figuring or aiming to figure in a filing system [10]. However, the Directive does not apply to the processing of personal data carried out by a natural person for exclusively personal or domestic reasons [11]. Furthermore, the Directive provides the general conditions for the lawfulness of the personal data processing (cf. Chapter Two of the Directive). It requires the existence of judicial remedies for the protection of personal data and creates a special liability upon the data controller, without omitting the question of sanctions in case of infringement of certain rules (cf. Chapter Three of the Directive). The Directive also rules the transfer of personal data outside the European Union (cf. Chapter Four of the Directive). Finally, the Directive addresses the question of the Codes of Conduct (cf. Chapter Five of the Directive) and establishes special institutions and bodies, such as the national supervisory authorities, the Working Group on the protection of individuals with regard to the processing of personal data (cf. Chapter Six of the Directive) and the Committee composed of the Member State Representatives concerning community implementing measures (Committee 31) (cf. Chapter Seven of the Directive).

Considered in a global approach, Directive 95/46/EC manages the risks presented by the processing of personal data by means of four steps [12]. In a first step, the Directive poses the legal framework applicable to any processing of personal data (including sensitive data [13]). In a second step, the Directive provides special rules to legitimate the processing of sensitive data. It goes without saying that the legal framework developed in the first step applies in addition to the processing of sensitive data. In a third step, the Directive imposes special rules to the processing of personal data presenting specific risks to the data subjects' rights and freedoms. This third approach must also be added to the two previous ones. It is not exclusive of their application for the rest of the data processing. In the fourth and last step, the Directive rules the transfers of personal data outside the European Union.

2. The Management of "Ordinary" Risk in the Processing of Personal Data

The risk management for the data subjects' rights and freedoms relies on a relatively simple principle: The risk does not depend on the informational content of the personal data but on the context in which they will be used [14]. In other words, the risk is linked to the purpose pursued by the processing of personal data. Therefore, the potential or real threat from the processing of personal data has to be assessed with regard to the purpose pursued by the data controller. There lies the reason why personal data are any information relative to an identified or identifiable natural person and not only information susceptible to reveal the intimacy of data subjects. Hence, all information, including the more common ones such as a phone number or a number plate, are personal data as long as they are related to an identified or (reasonably) identifiable natural person because the use of this kind of information may expose data subjects to some risks of infringement of their rights and freedoms, including their right to control in some extent the use of their personal data, with no regard to any specific informational content of the personal data. The aim of the Directive (the management of the risks presented by any use of information relative to identified or identifiable natural persons) explains for the definition of personal data.

3. The Management of "Special Risks" in the Processing of Personal Data

However, the principle relative to the risk management in the processing of personal data is slightly though not completely different with respect to the processing of "sensitive" data, the latter including medical data. Indeed, it is of common knowledge that the informational content of sensitive data is already capable to expose data subjects to some risks of infringement of their rights and freedoms in addition to the risk resulting from the purpose of their processing. In other words, any operation realized upon sensitive data inevitably exposes data subjects to risks of infringement of their rights and freedoms [14]. That is the reason why "sensitive" data require a special protection which has to take into account their informational content as well as the purpose of their processing.

Accordingly, the Directive bans the processing of "sensitive" data [15] because "data which are capable by their nature of infringing fundamental freedoms or privacy should not be processed" [16]. Put otherwise, this ban represents the special protection adopted by the Directive for "sensitive" data, including medical data. Being prohibited, the processing of "sensitive" data is no more susceptible to present any risk for the data subjects' rights and freedoms. Somehow, this policy aims to minimize the risks presented by the processing of "sensitive" data.

Nevertheless, the Directive provides a number of cases in which the prohibition to process "sensitive" data does not apply [19]. In these cases, the legitimacy of the processing of "sensitive" data (their admissibility) is presumed. Indeed, these situations are of nature to justify derogation to the ban to process "sensitive" data without prejudice to the other rules applicable to the processing of personal data. Noteworthy, these exceptions to the prohibition to process "sensitive" data have to be strictly interpreted. Beyond these exceptions, the processing of "sensitive" data is not allowed.

In each of these exceptions, the risk presented by the processing of "sensitive" data is presumed to be adequately under control. It must be immediately stressed that these exceptions do not imply an absence of risk, but balance the interests in presence. This requires assessing the risks for the data subjects' rights and freedoms in order to reasonably appreciate the admissibility of the processing of "sensitive" data [19].

4. The Management of "Specific Risks" in the Processing of Personal Data

The Directive fixes the legal framework applicable in all Member States to the processing of personal data and provides special rules to legitimate the processing of "sensitive" data. Yet, the Directive considers the situation in which, without prejudice to this double approach, some processing of personal data may present some specific risks to the data subjects' rights and freedoms [17].

In 1995, the Directive has indicated that, regarding any processing of personal data in the society, the cases presenting such specific risks should not be very common [18]. More than ten years later and having in mind the vertiginous evolution of the new information and communication technologies, it is not clear that such statement be still valid. By contrast, the number of data processing presenting such specific risks seems nowadays quite significant, especially in healthcare. Indeed, since 1995 the technological evolutions have notably permitted the creation of huge telematic networks linking substantial medical databases and the creation of genetic databases in national or European or worldwide telematic networks. Should we consider that these evolutions have increased the number of data processing presenting specific risks for the data subjects' rights and freedoms?

The Directive provides that the specific risks result from the nature of the data processing, from its range or from its purposes [17]. For example, the Directive cites purposes aiming to exclude persons from the benefit of a right, a service or a contract. These specific risks may also arise from the specific use of a new technology [17]. The latter reminds inevitably the introduction of GRID technologies in healthcare.

Traditionally, the processing of personal data presenting specific risks are those pursued by public authorities and concerning the population (as a whole or in part) or those concerning medical data [20]. Genetic databases and telematic networks in healthcare are further examples of data processing susceptible to present specific risks to the data subjects' rights and freedoms. One should pay attention to the person of the data controller [21], to the sensitivity of the processed data, to the purposes of the data processing, to the range of the data processing, to the categories of data subjects and to the respect of their rights, keeping in mind the transfer of the personal data outside Europe. In short, one should beware anything that could create specific risks to the data subjects' rights and freedoms. But any processing of sensitive data does not necessarily present specific risks and the processing of "ordinary" personal data should not be a priori excluded as it may also present specific risks to the data subjects' rights and freedoms.

Regarding the development of the telematic networks in healthcare, the specific risks result primarily from the fact that patients' data may be processed for multiple purposes. This raises the question whether it is permissible to process medical data for multiple purposes. This also raises the issue of the prior determination of the precise

and real purposes of the data processing. In addition appears the question of further data processing. Indeed, the actual trend aims to not determine anymore on a prior and precise way the purposes of the data processing, but to organize an entire information system combined with a security system in which the data processing purposes will be determined later. Put differently, we witness today the creation of information system with two levels. First, the infrastructure of the information system is created, implying in some extent the collection and the processing of personal data in a virtual complex (notably to identify the actors of the information system – mainly the patients and the health practitioners). Only then, the purposes permitted by the infrastructure are determined, forgetting that these purposes rely on an initial data processing (the creation of the first level of the information system (its infrastructure)). Doing so, the creation of the first level of the information system does not seem to constitute such a risk to the data subjects' rights and freedoms even when this first level is at the origin of the risks. But both the first (the creation of the infrastructure or the network) and further data processing permitted by the infrastructure of the information system have to be assessed. And if the security level helps to assess the risks induced by the data processing, it does not prevent to take into account the other criteria's to legitimate the data processing (for both levels of the information system), especially when the processing concerns sensitive data such as medical data or genetic data.

These new information systems are part of a structural policy aiming at building telematic networks in healthcare. They also indicate the transition from a vertical conception of eHealth to a new conception which is, in a first step, abstract, horizontal and transversal (the infrastructure of the information system) and which, in a second step, becomes vertical and real (the applications - eHealth products and services - using the infrastructure). The mere existence of these new telematic infrastructures in healthcare enables to share scientific databases but implies the identification of the practitioners and patients through special registries, etc. Eventually, these telematic networks will deeply modify the organisation of the public health systems and all actors in healthcare will be concerned and involved: practitioners, patients, institutions and bodies in healthcare and social security, medical laboratories, etc.

But once again, these new information systems differ in their permanency, irrespective of their future applications. Hence, the opportunity to create these infrastructures is no more evaluated regarding their precise and real purposes. Their opportunity is assessed in an abstract way with respect to some categories of purposes whose precise and real content will be determined later. This constitutes a deep change in the required precision and reality to evaluate the purposes pursued by the creation of the telematic infrastructure and its future exploitation.

These new information systems with multiple levels and purposes pose problems regarding the fairness of the data processing since the latter requires to respect the precise and real purposes announced at the beginning of the data processing. It also poses problems with respect to the duty to properly inform the data subject. Indeed, the multiple ramifications of the information system are not transparent, regarding both the technical level as well as the purposes of the data processing ("black box" issue).

However, it must be said that the new information and communication technologies are able to address properly all these issues.

5. Consequences of the presence of « Specific Risks » in the Processing of Personal Data

Member States have a duty to identify the processing of personal data likely to present specific risks to the data subjects' rights and freedoms and to take appropriate measures to ensure the prior checking of the processing of personal data before their starting [22].

The fact that medical data are already subject to special rules due to their sensitive nature does not exclude them from the scope of additional rules relative to data processing presenting specific risks. In other words, the processing of medical data presenting specific risks for the data subjects' rights and freedoms has to be checked prior its beginning. However, any medical data processing does not automatically present specific risks. And processing of "ordinary" personal data may also arouse specific risks.

The prior checking of data processing presenting specific risks may occur in four different ways.

Firstly, the prior checking may be carried out by the national supervisory authority following receipt of the notification from the data controller [23]. The national supervisory authority may, according to the applicable national law, issue an opinion or authorize the data processing [24].

Secondly, the prior checking may be carried out by the data protection official [26]. In case of doubt the latter has to consult the national supervisory authority [23]. With respect to this, the Directive indicates that the data protection official will proceed in cooperation with the national supervisory authority [24].

Thirdly, the Directive provides that Member States may carry out the prior checking in the context of the preparation of a measure of the national parliament, which defines the nature of the data processing and lies down appropriate safeguards [25].

Fourthly, Member States may also carry out to this prior checking in the context of the preparation of a measure based on a legislative measure, which defines the nature of the data processing and lies down appropriate safeguards [25].

6. "Specific Risks" in the Processing of Personal Data and the use of HealthGRID technologies

It is now possible to know whether the use of HealthGRID technologies may induce "specific risks" with regard to data protection. This question is exclusively focused on the use of such technologies and not on the outlines of its implementation project. What could lead to the conclusion that the use of HealthGRID technologies may induce such specific risks?

a. The HealthGRID technologies phenomenal storage capacities are of nature to cause specific risks with regard to data protection. In this case, specific risks may result from the storage of important amount of personal data. More stored data mean more risks. Naturally, the risk is greater in the presence of sensitive data.

b. The extraordinary capacities of HealthGRID technologies to process huge amount of personal data widely disseminated may also open the door to specific risks with regard to data protection. More operations upon personal data mean more risks. Again, the risk is greater in the case of sensitive data.

c. The size of the HealthGRID information system has to be considered, including its inscription in a broad European or international network: the larger, the riskier.

d. A specific risk could result from the data subjects' instrumentalization as they could appear more as informational sources than as patients. It could also lead to discriminations in the provision of healthcare or medicines or treatment or diagnosis.

e. The duration of the HealthGRID information system could also create specific risks to data protection (the "eternity effect").

f. The use of HealthGRID technologies by public authorities or bodies should be considered as inducing specific risks towards data protection.

g. If the exercise of data subjects' rights is more difficult due to the use of HealthGRID technologies, it should be recognized as a specific risk to data protection.

h. Generally, the use of GRID technologies implies the transfer of personal data outside Europe. The complexity of this kind of information system could lead to acknowledge the presence of specific risks towards data protection.

These criteria may naturally be combined, increasing therefore the risks for data subjects' rights and freedoms.

When considering the specific risks that may occur with the introduction of HealthGRID technologies in healthcare, one should not forget to take into account the benefits of its use. Again, it must be stressed that the new information and communication technologies could help to address these issues.

CONCLUSIONS

The risk management in the processing of personal data is deployed in four steps. Firstly, risks are assessed regarding the purposes of the data processing and not regarding the informational content of the processed personal data. Secondly, this principle is slightly though not completely different for sensitive data. For them, risks are assessed regarding their informational content as well as the purposes of their processing. Thirdly, data processing presenting specific risks for data subjects' rights and freedoms must be subject to a prior checking beforehand. The checking may take place in four different ways. Fourthly, transfers of personal data outside Europe are ruled by special rules. Due to some of its characteristics, the use of HealthGRID technologies in healthcare could induce specific risks with regard to data protection. This issue should be carefully monitored by the data controller as well as by the national supervisory authorities. In these situations, it seems more than appropriate to appoint a personal data protection official, to the benefit of everyone in terms of legitimacy, transparency, data subjects' rights and freedoms, confidentiality, security and efficiency. Finally, the presence of these specific risks should not prevent the use of Grid Technologies in Healthcare notably due to their potential but extraordinary benefits for knowledge and healthcare. It should only induce the adoption of

appropriate measures as previously described in order to ensure the respect of data subjects' rights and freedoms to which the entire HealthGrid Community is deeply committed.

Endnotes

[1] For a first overview on these legal issues : J. HERVEG & Y. POULLET, "HealthGRID from a Legal Point of View", in *From GRID to HEALTHGRID*, IOS Publications, Studies in Health Technology and Informatics, 2005, vol. 115, part 5, pp. 312-218.
[2] Journal officiel des Communautés européennes, n° L 281, 23 Nov. 1995, pp. 31 –50. For an in-depth analysis of the Directive : Y. Poullet, M.-H. Boulanger, C. de Terwangne, Th. Leonard, S. Louveaux & D. Moreau, « La protection des données à caractère personnel en droit communautaire », *Journal des Tribunaux de droit européen*, Brussels, Larcier, 1997, p. 121 et s. (in three parts).
[3] Directive 95/46/EC, Recitals 3, 5, 6, 7, and 9.
[4] Directive 95/46/EC, Recitals 2, 3, 10 and 11.
[5] Directive 95/46/EC, Recital 8. See also art. 1 and Recital 9.
[6] Like Public Order.
[7] On Directive 95/46/EC scope : C.J.C.E., 20 May 2003, Rechnungshof & al., C-465/00, C-138/01 and C-139/01 ; C.J.C.E., 6 Nov. 2003, Bodil Lindqvist, case C-101/01, C. de TERWANGNE, « Affaire Lindqvist ou quand la Cour de justice des Communautés européennes prend position en matière de protection des données personnelles », *Revue du droit des technologies de l'information*, Brussels, Ed. Bruylant, 2004, pp. 67-99.
[8] "Processing of personal data" ('processing') means any operation or set of operations which is performed upon personal data, whether or not by automatic means, such as collection, recording, organization, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, blocking, erasure or destruction (Directive 95/46/EC, art. 2.b) (cf. Recital 14).
[9] "Personal data" mean any information relating to an identified or identifiable natural person ('data subject'); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity (Directive 95/46/EC, art. 2.a).
[10] Directive précitée, art. 3.1.
[11] Directive 95/46/EC, art. 3.2 (cf. Recital 12).
[12] J. HERVEG, "La gestion des risques spécifiques aux traitements de données médicales en droit européen », in *Systèmes de santé et circulation de l'information, Encadrement éthique et juridique*, Paris, Dalloz, 2006.
[13] Usually, sensitive data are personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade-union membership, and the processing of data concerning health or sex life.
[14] Convention n° 108, Report, Recital 43.
[15] Directive 95/46/EC, art. 8.1.
[16] Directive 95/46/EC, Recital 33. Convention n°108 is not so explicit in its article 6.
[17] Directive 95/46/EC, Recital 53.
[18] Directive 95/46/EC, Recital 54.
[19] Directive 95/46/EC, art. 8. On the ban and its exception, see: J. HERVEG, "The Ban on Processing Medical Data in European Law: Consent and Alternative Solutions to Legitimate Processing of Medical Data in HealthGrid", in *Challenges and Opportunities of HealthGrids*, IOS Press, Amsterdam, Studies in Health Technology and Informatics, vol. 120, 2006, pp. 107-116.
[20] Y. POULLET, M.-H. BOULANGER, C. de TERWANGNE, Th. LEONARD, S. LOUVEAUX, & D. MOREAU, o.c., *Journal des Tribunaux de Droit Européen*, Brussels, Larcier, 1997, p. 152, n° 62.
[21] For example, a commercial company processing medical or genetic data.
[22] Directive 95/46/EC, art. 20.1.
[23] Directive 95/46/EC, art. 20.2.
[24] Directive 95/46/EC, Recital 54.
[25] Directive 95/46/EC, art. 20.3.
[26] The personal data protection official is a person appointed by the data controller in compliance with the national law which governs him. This official is responsible in particular:
- for ensuring in an independent manner the internal application of the national provisions taken pursuant to this Directive,

- for keeping the register of processing operations carried out by the controller, , containing the items of information referred to in article 21.2,

thereby ensuring that the rights and freedoms of the data subjects are unlikely to be adversely affected by the processing operations (Directive 95/46/EC, art. 18.2).

The presence of a data protection official allows Member States to provide for the simplification of or the exemption from the notification duty (Directive 95/46/EC, art. 18.2).

Selective Bibliography

- BENNETT, B. (ed.), *e-Health Business and Transactional Law*, Washington, BNA Books, 2002, 734 p.
- BOULANGER, M.-H., de TERWANGNE, C., LEONARD, Th., LOUVEAUX, S., MOREAU, D. & POULLET, Y., « La protection des données à caractère personnel en droit européen », *Journal des Tribunaux de Droit Européen,* Bruxelles, Larcier, 1997, p. 121 et s. (en trois parties)
- CALLENS, S. (ed.), *e-Health and the Law*, The Hague, Kluwer Law International, 2003, 183 p.
- CHABERT-PELTAT, C., « La télémédecine », *Revue Alain Bensoussan – Droit des Technologies Avancées*, Paris, 1999, n° 6/3-4, pp. 117-138.
- Commission nationale de l'informatique et des libertés (CNIL-France), Délibération n° 97-049 du 24 juin 1997 portant avis sur la mise en œuvre à titre expérimental d'un réseau de télémédecine sur Internet entre le Centre hospitalier d'Annecy et certains médecins de ville, *Revue Alain Bensoussan – Droit des Technologies Avancées*, Paris, 1999, n° 6/3-4, pp. 169-172.
- DE BOT, D., *Verwerking van persoongegevens*, Kluwer, 2001.
- de TERWANGNE, C., « Affaire Lindqvist ou quand la Cour de justice des Communautés européennes prend position en matière de protection des données personnelles », obs. sous C.J.C.E, arrêt du 6 nov. 2003, Bodil Lindqvist, affaire C-101/01, *Revue du droit des technologies de l'information*, Bruxelles, Ed. Bruylant, 2004, pp. 67-99.
- FLEISHER, L.D. & DECHENE, J.C., *Telemedicine and e-Health Law*, Law Journal Press, 2005.
- HERVEG, J., « HealthGRID from a Legal Point of View », in *From GRID to HEALTHGRID*, IOS Publications, Studies in Health Technology and Informatics, 2005, Volume 112, part 5, pp. 312-318.
- HERVEG, J., "The Ban on Processing Medical Data in European Law: Consent and Alternative Solutions to Legitimate Processing of Medical Data in HealthGrid", in *Challenges and Opportunities of HealthGrids*, IOS Press, Studies in Health Technology and Informatics, 2006, Volume 120, pp. 107-116.
- HERVEG, J. "La gestion des risques spécifiques aux traitements de données médicales en droit européen », in *Systèmes de santé et circulation de l'information, Encadrement éthique et juridique*, Paris, Dalloz, 2006.
- HERVEG, J. & VAN GYSEGHEM, J.-M., "La sous-traitance des données du patient au regard de la directive 95/46", *Lex Electronica*, vol. 9, n° 3, t. 2004, http://www.lex-electronica.org/articles/v9-3/herveg_vangyseghem.htm.
- HERVEG, J., VERHAEGEN, M.-N. & Y. POULLET, « Les droits du patient face au traitement informatisé de ses données dans une finalité thérapeutique : les conditions d'une alliance entre informatique, vie privée et santé », *Revue de Droit de la Santé*, Kluwer, 2002-2003/2, pp. 56-84.
- HERVEG, J., VAN GYSEGHEM, J.-M. & de TERWANGNE, C., *GRID-enabled medical simulation services and European Law*, Final Report on all the Legal Issues related to Running GRID Medical Services, European Research contract IST-2001-37153-GEMSS, 29 February 2005, 341 p.
- IAKOVIDIS I., WILSON, P., Healy J.-Cl*., E-Health: Current Situation and Examples of Implemented and Beneficial E-Health Applications*, IOS Press, Studies in Health Technology and Informatics, 2004, Volume 100, 249 p.
- KAPLAN, G. & Mc FARQUHAR, E., *e-Health Law Manual*, New-York, Aspen Publishers, 2003.
- MIDDLETON, S.E., HERVEG, J., CRAZZOLARA, F., MARVIN, D. & POULLET, Y., « GEMSS : Security and Privacy for a Medical Grid », Stuttgart, Schattaeur, Verlag für Medizin und Naturwissenschaften, *Methods of Information in Medicine*, 2005, 44/2, p. 182-185.
- RIENHOFF, O., LASKE, C., VAN EECKE, P., WENZLAFF, P. & PICCOLO, U., *A Legal Framework for Security in European Health Care Telematics*, Amsterdam, IOS Press, Studies in Health Technology and Informatics, 2000, vol. 74, 202 p.
- RIGAUX, Fr., *La protection de la vie privée et des autres biens de la personnalité*, Bruxelles, Paris, Bruylant, L.G.D.J., 1990.
- RODRIGUES, R.J., WILSON, P. & SCHANZ, S.J., *The Regulation of Privacy and Data Protection in the Use of Electronic Health Information, An International Perspective and Reference Source on*

Regulatory and Legal Issues Related to Person-Identifiable Health Databases, Pan American Health Information, World Health Organization, 2001, 217 p.
- ROGER-FRANCE, Fr., « Informations de santé, télématique et télémédecine, Perspectives d'ensemble à l'horizon 2000 », *Journal de réflexion sur l'informatique*, 1994, n° 30, pp. 7-9.
- ROUSSEAU, A. & HERVEG, J., *Manuel d'informatisation des urgences hospitalières*, Louvain-la-Neuve, Presses Universitaires de Louvain, 2003, 183 p.
- SILBER, D., *The case for eHealth*, Maastricht, Institut Européen d'Administration Publique (ed.), 2003, 32 p.
- STANBERRY, B., *The Legal and Ethical Aspects of Telemedicine*, London, Royal Society of Medicine Press, 1998, 172 p.
- VAN EECKE, P., "Electronic Health Care Services and the e-Commerce Directive", in *A decade of research @ the crossroads of law and ICT*, Gent, Larcier, 2001, pp. 365-379.
- VILCHES ARMESTO, L., « IMS Health : dernier développement de la C.J.C.E. relatif au refus de licence en droit de propriété intellectuelle », note sous C.J.C.E., 29 avril 2004, Brussels, Larcier, *Revue du Droit des Technologies de l'Information*, 2004, n° 20, p. 59 et s.
- WILSON, P., LEITNER, Chr. & MOUSSALI, A., *Mapping the Potential of eHealth, Empowering the citizen through eHealth tools and services*, Maastricht, European Institute of Public Administration (ed.), 2004, 52 p.

From Genes to Personalized HealthCare: Grid Solutions for the Life Sciences
N. Jacq et al. (Eds.)
IOS Press, 2007

Secure and Pervasive Collaborative Platform for Medical Applications

Petr HOLUB [a,c,1] Eva HLADKÁ [b,c] Michal PROCHÁZKA [a,c] Miloš LIŠKA [b,c]
[a] *Institute of Computer Science, Masaryk University, Czech Republic*
[b] *Faculty of Informatics, Masaryk University, Czech Republic*
[c] *CESNET z.s.p.o., Czech Republic*

Abstract. Providing secure, extensible, pervasive and easy to implement collaborative environment for medical applications poses significant challenge for state-of-the-art computer systems and networks. In this paper, we describe such a collaborative environment developed for Ithanet project, based on Grid authentication mechanisms. Significant effort has been put into developing a system, that is capable of deployment across tightly secured networking environments as implemented in vast majority of hospitals. The environment is extensible enough to incorporate Grid-service based collaborative systems like AccessGrid.

Keywords. Secure reliable pervasive collaborative environment, collaborative environment structure, virtual private network, latency performance, Ithanet project

Introduction

Virtual communication environments in the medicine are slowly becoming more popular with advent of reliable high-speed networking. Current trends go in two basic directions: (1) conservative commercially available technologies like voice over IP teleconferencing and H.323/SIP infrastructures, and (2) more advanced and more experimental tools like AccessGrid. While the former group is sufficient for basic communication, we focus in this paper on more advanced extensible environments, that is suitable for advanced collaboration.

The medical network are being protected heavily namely due to requirements on security of data about patients. However, such an "adverse" networking environment has usually sever consequences on collaborative technologies. Firewalls and network address translators (NATs), together with very conservative behavior of network administrators are the major obstacles for deployment of the majority of the conferencing systems, with notable exception of Skype [1], as discussed in Section 1. For collaborative environment, synchronous transmissions audio and video signals is essential. In terms of network transfers, this is usually implemented using RTP/UDP packet distribution. But UDP packet distribution in networks with NATs and firewalls with common settings is not allowed. One way to get around this problem is to change the settings; for this, users

[1] Corresponding Author: Dr. Petr Holub, Institute of Computer Science, Masaryk University, Botanická 68a, 602 00 Brno, Czech Republic. E-mail: hopet@ics.muni.cz.

have not only to ask for an approval of their administrators, but also they need them to make the configuration changes in security-sensitive devices, which is usually very hard to achieve. The second way is to use software for penetrating NATs and firewalls on application layer allowing communication without setting changes. This problem is not limited to collaborative environments, as demonstrated by a recent study by Open Grid Forum [2].

In this paper, we focus on building a secure extensible collaborative platform suitable for medical purposes, that is capable of working in such "adverse" networking environments inside hospitals with minimum requirements on support from network and system administrators. The platform uses Grid-compatible authentication based on PKI infrastructure and is ready to incorporate other Grid extensions based on Grid service oriented architecture.

The described system has been designed in Ithanet[2] project as stated in Acknowledgments at the end of this paper. The design reflects our experiences from participation in Ithanet and also EuroCareCF[3] projects, that showed how difficult it is to deploy reliable, secure, and affordable collaborative tools in multi-institutional environment across many hospitals all over the Europe. Both according to surveys we have conducted in these projects and according to our experiences with other medical communities, there is a number of problem to solve: the connectivity among the partners varies to large extent; already existing computer systems are largely based on Microsoft Windows system (often under central administration), making installation of additional software and especially remote support very difficult or virtually impossible; computer support teams at the hospitals often do not have enough experiences to work with more complex collaborative environments. Therefore we have decided to propose and implement a system, that is able to avoid or at least to mitigate these problems.

The rest of the paper is structured as follows: Section 1 summarizes relevant related work, Section 2 gives an overview of the architecture of the whole collaborative environment, Section 4 details proposed client devices and software tools. In Section 5, we give performance measurements of the VPN-based networking solution implemented for the environment and also further experiences with the system. Section 6 concludes the paper with final remarks and directions for future development of the collaborative platform.

1. Related Work

Traditionally, the tele-medicine was implemented using phone-based solutions, later migrating to H.323/SIP solutions. However, both H.323 and SIP are hard to implement in "adverse" networking environments and require serious support for network administrators side.

Skype [1] has thus become very attractive alternative for many users, because of its capabilities to penetrate firewalls and work behind NATs—unless explicitly forbidden by the network administrators, and also rather robust commercial-grade implementation. Providing limited multi-point conferencing support, it is also usable for small teams to communicate. When user signs up for commercial Skype services, it is possible to call also to public telephony network. In recent versions, Skype provides basic video support,

[2] http://www.ithanet.eu/

[3] http://www.eurocarecf.eu/

which is however much less mature than its audio counterpart. However, there are several reasons why this tool doesn't meet the needs of medical communication particularly well and becomes often explicitly forbidden by the hospital management: first, the security model is proprietary and very obscure [3,4] and may change without any prior notification, and second, users participating in the Skype network are automatically providing their computers to be used by other Skype users, thus actually supporting the business of Skype company with their own resources. Further, it is hard to block Skype on the networking layer, because of its firewall/NAT penetration techniques (e.g., even the login messages to Skype servers may be router through the super-nodes in the P2P overlay network as shown in [4]). Being a proprietary solution, Skype is hard to be integrated in more complex Grid-like infrastructures, too.

On the other side of the collaboration tools spectra, there is an experimental open extensible system called AccessGrid [5,6]. Since its version 2, it features service oriented architecture for the collaborative environment based on Grid services [7]. For deployment in the hospital environments, it has two major drawbacks though. First, the data distribution is not suitable for "adverse" networking environments—authors of AccessGrid assumed multicast to be the primary data distribution technology; since the multicast deployment is a problem even in very open academic networks, they added simple unicast bridging technology based on UDP packet reflectors later on. Second, the whole system is rather complicated and hard to deploy without knowledgeable on-site support staff. Despite these issues, AccessGrid features many interesting properties making in worth considering when planning deployment of larger collaborative systems, by at least preparing compatible equipment for its future installation.

Another well known example of collaboration environment, created for high energy physics community that is now becoming more widely used, is the Virtual Room Videoconferencing System (VRVS)[4]. VRVS is based on multicast-like schema and is provided as a service and user data traffic is managed by VRVS administrators. The successor of VRVS called Enabling Virtual Organizations (EVO)[5]—is based on self-organization of system of reflectors, again not empowering the end-user with tools to change the distribution topology. However, for medical problem it has the similar problem as AccessGrid and even worse it is a closed system with the similar consequences like for Skype.

2. Architecture of the Collaborative Stack

When designing the collaborative platform architecture, we have had the following main design principles in mind:

1. The system has to be secure to protect the communicated data. It should utilize Grid-compatible authentication, so that the users can use their existing Grid identity to join the collaborative environment.
2. Support for firewall and NAT penetration to facilitate deployment in "adverse" networking environments in hospitals. The support has to be flexible and auditable, and the network administrators must not be "cheated", (but rather just avoided unless really necessary).

[4] http://www.vrvs.org/
[5] http://evo.caltech.edu/

3. The system has be extensible towards service-oriented architecture compatible with Grids.

The proposed collaborative environment comprises the following layers from the bottom-up perspective: network connectivity layer, data distribution layer, central services layer, client tools and devices layer. Each of the layers is discussed in more detail below.

Network connectivity layer. The network connectivity layer takes care of interconnecting all the elements of the collaborative environment into a continuous network, so that all the element may reach one another or at least some central site or server site when all-to-all connectivity is not desirable for any reason. If the elements can't reach one another over the native network, this may be implemented as an overlay network, be it simple tunnels or more sophisticated VPNs. Assuming Internet protocol stack, the tunnels may run over UDP, TCP, HTTP (including emulation of HTTP and HTML encapsulation), TCP with HTTP proxy, and TCP with SOCKS proxy. For vast majority of the firewall and NAT protected networks, at least one of the mentioned solution works. VPNs are also useful when overlay network privacy is desired based on overlay link encryption.

Data distribution layer. This layer takes care of multi-point data distribution to the connected clients. Assuming the Internet protocol stack, this is usually implemented either using multicast, which is more efficient but much harder to deploy properly esp. in network spanning multiple administrative domains, or using UDP packet reflectors. While less efficient, the UDP packet reflectors are much less error-prone and provide also possibility of data processing even on per-user basis (something theoretically impossible in multicast).

UDP packet reflectors may also be modified to provide the network connectivity layer to the clients and between the reflectors directly, thus merging Network connectivity and Data distribution layers.

Central services layer. This layer comprises services provided to the client on some "server" basis—though the servers may be largely distributed and not limited to one physically central location. The services may include monitoring, virtual rooms or venues (for creating separate virtual spaces for communication of different user groups), wikis, persistent data storage, etc.

Client tools layer. The client layer comprises of tools and hardware devices on client side. The software tools primarily incorporate audio, video, and a chat service, and may include other tools like shared presentation, shared desktop or application window, or shared text editor.

While software tools are traditional when looking on computer based collaborative environments, we have included also the hardware part, as the quality of hardware and a level of its software support is critical for the successful experience with collaborative environments. Often this creates a point of failure for the collaborative environments deployment. The collaborative tools are much harder to deploy compared, e.g., to SETI@Home or similar computation-heavy tasks that are relatively easily distributed because they are only dependent on CPU and very basic OS services.

3. Preliminary Implementation of the Collaborative Stack

3.1. Network Connectivity Layer

Secure communication together with ability of firewall and NAT penetration is achieved using the OpenVPN software. It makes the VPN on the application layer from the ISO/OSI perspective and supports the whole range of methods as discussed in the previous section. It means that there is no need to modify configuration of network elements on the path from the client to the VPN server. Also whole communcation between client can be encrypted.

All client workstations are connected to a VPN server in a point-to-point mode. The set up of the VPN network guarantees that only the traffic belonging to the collaboration services is sent into the VPN tunnel. Originally, we wanted to use one of the private IP address ranges as defined by RFC 1918 for internal VPN addressing. After the partner networking survey we have found it very complicated to avoid conflicts with internal address ranges used at various institutions, especially as new institutions may join. The whole overlay network is therefore addressed using a public IP address range assigned by RIPE, but the addresses are treated as internal address and not distributed outside of the VPN overlay network. As there is no direct traffic between any two partners and thus all the traffic may be filtered on the VPN server.

The OpenVPN server runs in two modes—either over UDP or TCP. The UDP mode is preferred due to better performance, as the VPN is not limited by the TCP congestion control algorithm [8]. The TCP mode can also run over HTTP or SOCKS proxy.

The client side needs OpenVPN software to be able to connect to the OpenVPN server. This software makes the virtual network adapter and sets appropriate routing table records for the client. Clients are authenticated using their personal X.509 certificates—the Grid users may user their existing ones, while others are given new certificates from a dedicated certificate authority. Client software is able to work with certificates stored in the file or on the secure smart card. Second option is strongly preferred because the client certificate can be delivered and kept by the client in a secure way.

3.2. Data distribution layer

The data distribution layer is implemented using modular user-empowered UDP packet reflector [9], which is known to work very well with the taget client media tools—MBone Tools[6]. It is highly configurable with modules loadable in run-time, supporting sophisticated access control policing and even data transcoding for some data formats. Media streams may be encrypted by the client software tools using symmetric encryption in case that the data replication site is not considered trusted enough. Communication based on UDP packet reflector is communication with central replication unit and number of communicated client is limited by capacity of this unit. Solution of this problem is to decentralize reflector by network of reflectors [10].

[6] `http://www-mice.cs.ucl.ac.uk/multimedia/software/`

3.3. Central services layer

Currently, there are UDP packet reflector administration and monitoring run as a central services. When the system is extended to full AccessGrid support, Venue Server could be an example of central service. Another centrally run service is the OpenVPN server supporting users as described in section 3.1. To support the user communication we furthermore provide an IRC daemon (currently IRCD-hybrid[7] run as a central services. A special IRC client daemon which run on the same machine was developed to store and provide the chat history. A Jabber instant messaging server may be provided in the same way as IRC.

4. Client Platform

The collaborative platform client is developed as a mixed HW/SW solution. The client is based on well defined and tested HW with preinstalled operating system and set of collaborative and especially videoconferencing tools. The choice of operating system was done with emphasis on simple modifications of the system, remote management, wide hardware support, security and last but not least user friendly environment. As a result we chose Ubuntu Linux to be the base for the client SW.

The collaborative tools depend on many dedicated hardware devices like sound cards, sound acquisition devices, video capture cards, USB cameras, etc., whose quality and level of support may vary to a large extent. Therefore, we have opted for suggesting a set of hardware: widely available low-cost HP Compaq dc7700 Ultra-Slim Desktop PC with known-to-work and well tested sound and graphics card and video capture cards, together with some other devices like headsets and USB cameras. This provides well defined starting point for efficient distribution of operating system and collaborative tools.

Security and personal configuration of the platform is based on combined USB security and storage token. The token is used for following purposes:

- identity storage which is based on PKI for authentication purposes, especially to connect to VPN server
- customized configuration storage for each users configuration, allowing user to store his contact informations for videoconferencing purposes and start the videoconference in specific mode (e.g., specific reflector address and port or VPN configuration)

Basic videoconferencing capabilities are provided by MBone Tools. Robust Audio Tool (RAT) is used for audio transmission and playback. RAT supports variety of audio codecs and allows to fine tune the audio stream according to quality or bandwidth limitations where necessary. Video communication is provided by Videoconferencing Tool (VIC) providing transmissions of video acquired from video capture card or USB cameras.

Besides VIC and RAT other we provide audio and videoconferencing tools like Ekiga, WengoPhone and even Skype which allow audio and video communication with number of other videoconferencing platforms.

[7] `http://ircd-hybrid.com/`

While the individual tools are rather user-friendly when started, the initialization of the conference itself is not very intuitive step. In order to facilitate this, we are developing an integrating Graphical User Interface (GUI) (see fig. 1 for the platform, that supports easy setup of the conference. The default settings may be stored for future reuse, so in the production state, the user just pushes a single button to start the whole conference. The GUI also monitors all the applications, so that if any part of the system crashes, the user is immediately informed and it also provides a very detailed information for the remote user support.

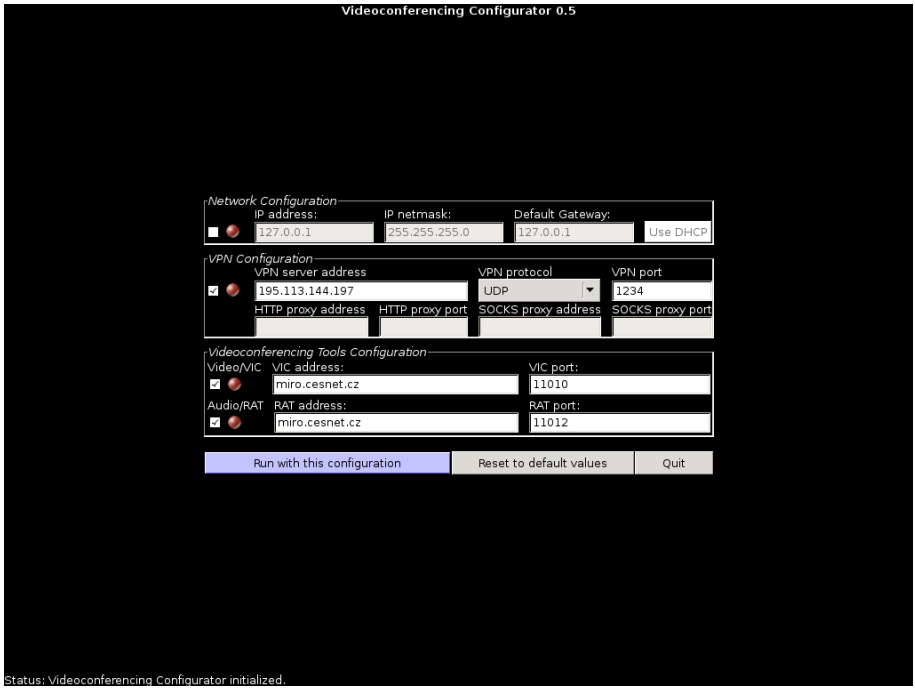


Figure 1. GUI for the collaborative platform client.

To support the collaboration beyond scope of just audio or videoconferencing we provide bidirectional sharing of whole desktop or particular applications between videoconferencing platform clients. Desktop and application sharing is based on VNC protocol [11] and related tools, namely shared-app-vnc[8] and x11vnc[9]. A secondary intent is to provide user with remote control of his videoconferencing machine in case the machine has no display and keyboard or the user wants to control the videoconference from his/her laptop computer. However, using VNC as a software display is considered as an emergency solution only because of high demands on the videoconferencing machine and the latency introduced video displaying.

The client platform is based on SW that is stable yet under development and some new and desired functionality may appear. In a worse case a security hole may be dis-

[8] http://shared-app-vnc.sourceforge.net/
[9] http://www.karlrunge.com/x11vnc/

covered in one or more platform SW components. Thus it is necessary to update the SW base of the platform regularly. The individual videoconferencing tools as well as other platform SW may be updated automatically based on the operating system SW repositories.

Underlying operating system updates are more complicated and failed update may turn whole machine unusable. There may be no system administrator available or skilled enough to perform the operating system update on end users site. That is why we opt for prepare the operating system update as a black box solution. The update is based on bootable CD/DVD with an image of well tested and working updated operating system and platform SW. It is not necessary to distribute the CDs among the end users. More comfortable approach is to create and ISO image of the update CD and make it available for download. The update procedure is performed automatically right after the CD is inserted into the machine. End users don't need to reconfigure their box after each operating system update because all custom configuration is stored on the USB token and thus is not affected by the update.

5. Experiences with Preliminary Implementation

In order to evaluate influence of incorporation of OpenVPN into the collaborative platform, we have measured a number of parameters critical for real-time multimedia communication using different VPN modes. The measurement testbed comprised one client and one VPN server, interconnected with high-speed backbone network link with capacity above 1 Gbps spanning about 250 km. The results of measurements are summarized in Table 1. We can conclude that UDP based VPN is very safe and has minimum impact on the traffic. Slight CPU requirements increase for the UDP based VPN compared to TCP based VPN is due to application-level packet loss recovery and congestion control, which is marginally less CPU efficient compared to kernel-based TCP implementation. TCP-based VPNs also perform very well provided they are on low-latency network with very low packet loss, so that congestion control algorithm doesn't influence the data flow significantly. If the HTTP proxy is of good performance, it has minimum impact on performance, too.

Table 1. Measured comparison between direct communication and communication through various VPN modes as implemented by OpenVPN.

	no VPN	UDP VPN	TCP VPN	TCP VPN + HTTP proxy
pchar latency [ms]	3.51	3.69	3.94	3.93
iperf jitter [μs]	6	6	9	13
pchar capacity est. [Mb/s]	39.8	35.2	20.1	19.8
iperf packet loss @ 30 Mb/s [%]	0.0	0.0	0.0	0.0
iperf CPU idle @ 30 Mb/s [%]	48.9±0.2	41.7±0.4	44.5±0.4	42.6±0.4

When evaluating the performance of this solution subjectively, the media streams are fine and the overall quality is very good. The only problem we were facing is that the users are sometimes very reluctant to buy a new hardware for the client platform, even though it is very cost effective compared to dedicated videoconferencing solutions. However, they are also unhappy about performance of videoconferencing tools self-installed

on their existing desktop computers, as these were not performing well without substantial tweaking because of rather complex interactions with undefined or poor hardware components and existing software. Thus the deployment requires significant work in order to explain principles of the system to the users as described above.

6. Conclusions and Future Work

In this paper, a secure and pervasive collaborative platform for medical applications has been introduced to provide flexible multi-point Grid-compliant collaborative environment. The design and implementation was targeted to create remotely supported system, that is scalable, robust and flexible allowing to collaborate to tens of people. The system has been implemented in year 2006 to the state described in this paper. In 2007, we plan to gather experiences based on wider scale deployment and modify the system accordingly.

The first version of the system, described in this paper, has risen a number of new problems and ideas. The first ideas for future work are concerned with reflector functionality. In the future we plan to utilize per-user processing on the reflector for solving the problem when a single client with a very limited network connectivity limits the quality of the collaboration for the whole group. Currently we are using OpenVPN to traverse NATs and firewalls but in the future, we plan to implement this directly in the reflectors. Such an approach allows for more aggressive failure detection and faster problem recovery. Also as the reflectors may be deployed as a network, it naturally avoids the single point of failure currently imposed by a central VPN server.

Grids and Grid-based systems are now widely developed and used in many areas. We plan to utilize the Grid-services based approach and to incorporate AccessGrid services. On the other side, AccessGrid needs to be modified to work with our advanced reflectors, firewall and NAT penetration techniques and reflector networks for better scalability and robustness. This will be discussed with AccessGrid developers and we will offer our tools to that project. We are at the beginning of practical usage of the proposed platform and its routine operation will definitely bring other new ideas and requirements for the future.

Acknowledgments

This work has been kindly supported byt European Commission project “ITHANET – eInfrastructure for Thalassaemia Research Network”, RI-2004-026539.

References

[1] Zennström, N., Friis, J.: Skype (2003-2007) http://www.skype.com/.
[2] Niederberger, R., Allcock, W., Gommans, L., Grünter, E., Metsch, T., Monga, I., Volpato, G.L., Grimm, C.: Firewall issues overview. Technical Report GFD-I.083, Open Grid Forum (2006)
[3] Biondi, P., Desclaux, F.: Silver needle in the skype. In: BlackHat Europe. (2006) http://www.blackhat.com/presentations/bh-europe-06/bh-eu-06-biondi/bh-eu-06-biondi-up.pdf.

[4] Baset, S.A., Schulzrinne, H.: An analysis of the skype peer-to-peer internet telephony protocol. In: INFOCOM 2006, Barcelona, Spain (2006) http://www1.cs.columbia.edu/~salman/publications/skype1_4.pdf.

[5] Childers, L., Disz, T., Hereld, M., Hudson, R., Judson, I., Olson, R., Papka, M.E., Paris, J., Stevens, R.: ActiveSpaces on the Grid: The construction of advanced visualization and interaction environments. In Engquist, B., ed.: Simulation and visualization on the grid: Parallelldatorcentrum, Kungl. Tekniska Högskolan, seventh annual conference, Stockholm, Sweden, December 1999: proceedings. Volume 13 of Lecture Notes in Computational Science and Engineering., New York, NY, USA, Springer-Verlag Inc. (2000) 64–80

[6] Childers, L., Disz, T., Olson, R., Papka, M.E., Stevens, R., Udeshi, T.: Access grid: Immersive group-to-group collaborative visualization. In: Proceedings of Immersive Projection Technology, Ames, Iowa (2000)

[7] Foster, I., Kesselman, C., Nick, J., Tuecke, S.: The physiology of the grid: An open grid services architecture for distributed systems integration (2002)

[8] Jacobson, V.: Congestion avoidance and control. In: ACM SIGCOMM '88, Stanford, CA (1988) 314–329

[9] Hladká, E., Holub, P., Denemark, J.: An active network architecture: Distributed computer or transport medium. In: 3rd International Conference on Networking (ICN'04), Gosier, Guadeloupe (2004) 338–343

[10] Holub, P., Hladká, E., Matyska, L.: Scalability and robustness of virtual multicast for synchronous multimedia distribution. In: Networking - ICN 2005: 4th International Conference on Networking, Reunion Island, France, April 17-21, 2005, Proceedings, Part II. Volume 3421/2005 of Lecture Notes in Computer Science., La Réunion, France, Springer-Verlag Heidelberg (2005) 876–883

[11] Richardson, T., Stafford-Fraser, Q., Wood, K.R., Hopper, A.: Virtual network computing. IEEE Internet Computing **2** (1998) 33–38

From Genes to Personalized HealthCare: Grid Solutions for the Life Sciences
N. Jacq et al. (Eds.)
IOS Press, 2007

Multi-centric Universal Pseudonymisation for Secondary Use of the EHR

Luigi Lo Iacono
C&C Research Laboratories, NEC Europe Ltd., Germany

Abstract. This paper discusses the importance of protecting the privacy of patient data kept in an Electronic Health Record (EHR) in the case, where it leaves the control- and protection-sphere of the health care realm for secondary uses such as clinical or epidemiological research projects, health care research, assessment of treatment quality or economic assessments. The paper focuses on multi-centric studies, where various data sources are linked together using Grid technologies. It introduces a pseudonymisation system which enables a multi-centric universal pseudonymisation, meaning that a patient's identity will result in the same pseudonym, regardless of which participating study center the patient data is collected.

Keywords. Pseudonymisation, Privacy, Multi-centric Studies, HealthGrid

1. Introduction

The Electronic Health Record (EHR) is a personal medical record in digital format, containing in the first instance information relating to the current and historical health, medical conditions and medical tests of its subject. It is primarily used in the treatment context in which the patient's identity data is needed and protected by medical secrecy. But the EHR also serves as a basis for other purposes denoted as secondary uses, such as clinical or epidemiological research projects, health care research, assessment of treatment quality or health economy.

Characteristic for the secondary use is, that the patient data leaves the control- and protection-sphere of the medical secrecy. Patient data is very sensitive personal data. To make it available outside of the treatment context, various legal and ethical aspects have to be considered [1]. First of all, the patient has to agree to participate with his patient data to a secondary use by giving an informed consent. Then the Personal Identifiable Information (PII) of the patient has to be removed from the EHR before the use in secondary contexts is allowed. Thus, in such a context the use of the EHR is allowed after anonymisation and must be performed whenever possible, but although the identity of the patient does not matter in secondary contexts, it is, however, not always possible to simply anonymise the EHR. In many scenarios of secondary use, the correct association between a single patient and his EHR from distinct sources or distinct points in time is essential. Examples are the provisioning of follow-up data at a later point in time, the withdraw of samples or data on specific patient's request or the quality control of the data such as the checking for double-booking. This usually prevents the application of anonymisation and demands instead for pseudonymisation schemes. In some scenarios even the re-identification of a patient is required, when

clinically relevant information arises during the course of a study which might have a direct impact on the treatment of a patient. In such cases, ethical principles demand recontacting and informing all relevant patients about the findings.

In summary, depending on the kind of research network and its requirements, distinct procedures for anonymisation or pseudonymisation are appropriate. In this paper we focus on research networks where various data sources must be linked together, such as given by a multi-centric study that uses data from EHRs, but also data or samples from biomaterial banks, or follow-up data at a later point in time and where the possible recontact of patients is required. Such requirements are common for studies as carried out for example by the @neurist project[1] funded within the European IST FP6 program. The @neurist project aims to integrate data which spans all length scales, from molecular, through cellular to tissue, organ and patient representations in order to develop advanced decision-support systems to help treat cerebral aneurysms. These data are increasingly heterogeneous in form, including textual, image and other symbolic structures, and are also diverse in context, from global guidelines based on the broadest epidemiological studies, through knowledge gained from disease-specific scientific studies, to patient-specific data from electronic health records.

2. Previous Work

Available pseudonymisation systems can roughly be classified depending on their ability to be reversible or not.

One-way pseudonymisation systems generate pseudonyms in such a way that it is almost impossible to re-identify the patients from the generated pseudonyms. Keyed cryptographic one-way functions [2] are a common technology to implement this property (also referred to as one-way encryption in this context). The pseudonyms could be produced either at clinical centers or through a central Trusted Third Party (TTP) service. For multi-centric studies this step has to be performed by an TTP in order to obtain pseudonyms which are universal to the study, creating links through all information on one patient. An essential prerequisite here is some form of PII, which identifies the patients unambiguously. New information on the same patient would produce the same pseudonym and therefore enable updates to the data. The main advantage to this system is the high level of privacy protection it provides. From the viewpoint of recontacting of patients, the natural disadvantage is that it would be very difficult to trace back to original patients. The conceptually simplest way to enable re-identification in such systems would be to store the association between PII and pseudonyms at the TTP. Such a mapping database is, however, an attractive target for attackers and violates the medical secrecy [3].

Reversible pseudonymisation systems allow the patient to be re-identified through cryptographic mechanisms and the existence of cryptographic keys in particular. The re-identification through decryption of the pseudonym eliminates the need of maintaining a mapping database. In the case of symmetric cryptosystems – which uses one and the same key for encryption and decryption –, one entity holds the key to the pseudonym, creating a level of security very much like the patient/clinician confidentiality relationship. If the clinical centre holds such a key, they are able to re-identify participating patients. A solution to ensure that re-identification of individuals

[1] http://www.aneurist.org

is strictly controlled is to use asymmetric cryptosystems instead of symmetric ones, where the public encryption key is used by the pseudonymisation service, but the private decryption key is only known to and only in possession of the study's medical and ethical advisory board. Such an approach is to the knowledge of the author, however, neither widely used nor discussed. Instead, more commonly an extra pseudonymisation stage is added to the scheme based on symmetric cryptography including a corresponding separate secret key. The second pseudonymisation step of the so-called dual-pass pseudonymisation systems is introduced in such a way that the two pseudonymisation procedures are independent from and do not know anything about each other. This "additional privacy safeguard" allows for much stricter control over linking research information back to a patient, and the situation when the two pseudonyms are brought together can be regulated by strict operating procedures. In single-centric studies the first pseudonymisation step can be performed within the study center [4]. For multi-centric studies, however, both have to be run by distinct TTP in order to obtain the correct linkage between data from different sources. LIPA [5] is an example for such an architecture and the only one the author is aware of. It relies on asymmetric cryptographic techniques and two distinct TTP in order to calculate an universal pseudonym. However, the first TTP receives the patient's National Health Service (NHS) number which reveals PII to the TTP since the NHS number is the common and unique identifier for patients in England and Wales.

For the depicted and targeted research network type and according to the available technologies a reversible dual-pass pseudonymisation system is most desirable. However, to generate universal pseudonym which can link all the diverse data sources on one patient together, the known approaches rely either on a central TTP service which has to obtain at least some form of PII of the patient or are simply not capable to generate such universal pseudonym.

3. Multi-centric Universal Pseudonymisation

To overcome the limitations, that a patient has to be treated always in the same study center or that parts of his PII has to be sent to a central pseudonymisation service in order to obtain the corresponding pseudonym within a particular research network, the following pseudonymisation scheme is proposed. It enables a multi-centric universal pseudonymisation, meaning that a patient's identity will result in the same pseudonym, regardless in which participating study center the patient data is collected. Furthermore, the study centers do not have to reveal the patient's identity data to an TTP service in order to retrieve the corresponding inter-clinic and unambiguous pseudonym. Instead, the first pseudonymisation step is performed locally and a trusted pseudonymisation service then performs a second pseudonymisation step in which the inter-clinic and unique pseudonym is computed based on the output of the locally performed pseudonymisation which does not contain any PII.

The proposed scheme relies on a number theoretic problem usually used for constructing asymmetric cryptosystems. More specifically the Discrete Logarithm Problem (DLP) is used which is the basis for asymmetric cryptographic schemes such as the Diffie-Hellman (DH) [6] key agreement and the ElGamal [7] public key encryption and signature scheme. As for DH and ElGamal, the proposed scheme can be implemented in any group where the DLP is infeasible, including e.g. the group of an elliptic curve defined over a finite field which forms the fundament of Elliptic Curve

Cryptography (ECC) [8, 9]. For the sake of simplicity and clarity, however, the following descriptions are focused on the multiplicative group of a finite field $\boldsymbol{Z}_p$ (where p is prime) only.

The cryptographic techniques used do not allow the reversal of the pseudonymisation to a patient's identity by decrypting the pseudonym. The reverse mapping has to be performed by a corresponding database maintaining the associations between the patient's identifier and the universal pseudonym.

3.1. Symbols and Abbreviations

The following symbols and abbreviations are used to increase the clarity of the paper.

E	Set of all patients.
S	Set of all study centers.
sk^s	Secret key of key pair of study center $s \in S$.
sk^{TTP}	Secret key of key pair of trusted pseudonymisation service.
pk^s	Public key of key pair of study center $s \in S$.
pk^{TTP}	Public key of key pair of trusted pseudonymisation service.
PK	$PK = \{pk^s \mid \forall s \in S\} \cup \{pk^{TTP}\}$. Set of public keys of all study centers and the pseudonymisation service.
PK^s	$PK^s = PK \setminus \{pk^s\}$. Set of public keys of all study centers and the pseudonymisation service, except the public key of study center $s \in S$.
PII^e	Personal Identifiable Information of patient $e \in E$ which are used for pseudonym generation.
ID^e	An identification which identifies the patient $e \in E$ unambiguously and which is produced on its PII^e.
lid^e	Local pseudonym computed for the patient $e \in E$ based on its PII^e or ID^e respectively.
$lid^{e,s}$	lid^e computed by the study center $s \in S$.
gid^e	Global Pseudonym computed for the patient $e \in E$.
$rand[x, y]$	A random number function which chooses randomly a number from the interval defined by $[x, y]$.
p	A large prime number.
g	A primitive element modulo p of the multiplicative group of a finite field $\boldsymbol{Z}_p$ (also called generator).
OWF	An one-way function, such as cryptographic hash-functions like SHA-256 or Whirlpool.

3.2. Initialization Steps

In the initialization phase is based on the system-wide and public parameters p and g, which define a multiplicative group of a finite field $\boldsymbol{Z}_p$ with primitive element g corresponding to DLP-based cryptosystems. The parameters have to be chosen according to current key length proposal by the cryptographic community and the needs of the study [10]. Each participating study center then has to perform the following initialization steps:

1. Select a random number $r = rand[1, p\text{-}1]$.
2. Compute $\rho = g^r \bmod p$.
3. Distribute ρ to the trusted third party (the pseudonymisation service in charge of computing the inter-clinic pseudonym) and keep r secret.

The computed number ρ is the public key pk^s of study center s, the selected random number r is the private key sk^s. The TTP pseudonymisation service has to perform the same steps, but keeps both keys secret (see section 5).

Assuming the following setting for a subsequent example:

- Study Centre Alice: $pk^A = \alpha,\ sk^A = a$
- Study Centre Bob: $pk^B = \beta,\ sk^B = b$
- Study Centre Carol: $pk^C = \chi,\ sk^C = c$
- Study Centre Dave: $pk^D = \delta,\ sk^D = d$
- TTP Service Trent: $pk^{TTP} = \tau,\ sk^{TTP} = t,\ PK = \{\alpha, \beta, \chi, \delta, \tau\}$

After all system participants have performed the described initialization, the process to obtain a research network global and universal pseudonym is divided into two distinct phases. First, a local pseudonym is computed by the study center and second, the global pseudonym is computed by a pseudonymisation service as will be described in the following.

3.3. Computing a Local Pseudonym

The computation of a local pseudonym is composed of two sequential steps (see Figure 1). In the first step, the patient's PII used for pseudonym generation is transformed by the application of an one-way-function (*OWF*) in such a way,

- that the patient is identified unambiguously by the output of the *OWF*,
- that the variable input data is mapped to a fixed- and short-size output, and
- that the reverse operation is not feasible.

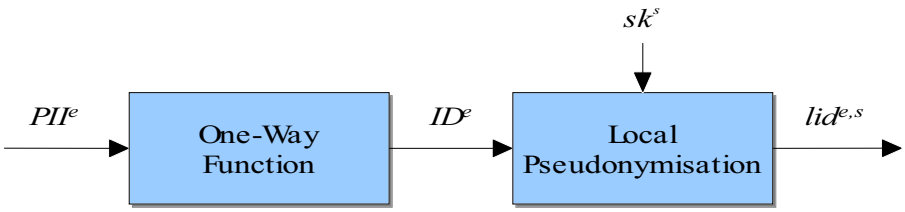


Figure 1. Study center computing a local pseudonym

Cryptographic hash-functions are a good candidate for such OWF. The output of the OWF (which in the case of hash-functions is the hash value) is denoted as $ID^e = OWF(PII^e)$. Note, that this step can be performed by any study center in the research network and will result in an identical ID^e for identical PII^e for a given patient $e \in E$. Thus, an essential prerequisite is the unique patient's PII.

From the generated ID^e the study center s can finally produce the local pseudonym in the second and last step by computing $lid^{e,s} = g^{sk^s + ID^e} \bmod p$, using the research network related private key sk^s generated as described in section 3.2.

3.4. Computing the Universal Pseudonym

The global pseudonym gid^e for the patient $e \in E$ can be computed from the corresponding local pseudonymisation lid^e of the patient e by the central pseudonymisation service as follows: $gid^e = lid^{e,s} \cdot \prod PK^s \bmod p$.

Figure 2 illustrates the computation of the universal pseudonym based on the example setting introduced in section 3.2. If, for example, the study center Bob computes the local pseudonym $lid^{e,B}$ of patient e and sends it to the pseudonymisation service Trent, Trent is able to generate the universal pseudonym gid^e of patient e by performing the following computation: $gid^e = \alpha \cdot lid^{e,B} \cdot \chi \cdot \delta \cdot \tau \bmod p$. Note, that regardless of which system participant (excluding Trent) computes $lid^{e,s}$, the global pseudonym gid^e will always be identical.

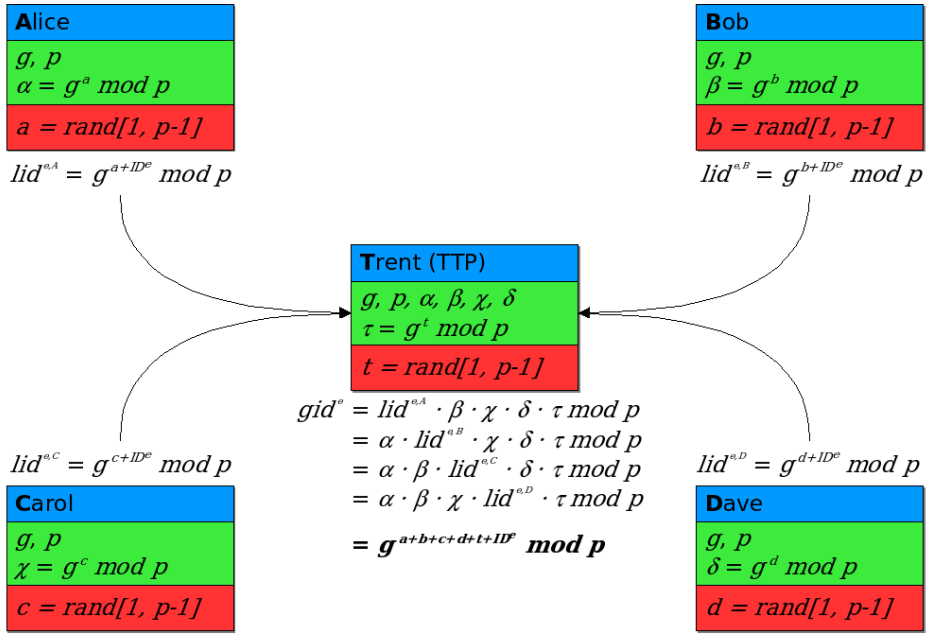


Figure 2. TTP computing a multi-centric universal pseudonym

3.5. Re-Identification

To re-identify the patient from a global pseudonym, mappings between global pseudonyms and local pseudonyms as well as between local pseudonyms and patient identifiers have to be maintained, since the proposed scheme does neither allow the retrieval of the local pseudonym from a particular global pseudonym nor the retrieval of the patient identifier from a local pseudonym.

The maintenance of these mappings can be realized in distinct ways. A common approach would be to keep the mapping between the patient identifiers and the local pseudonyms in the corresponding study centers and to store the mapping between the local pseudonyms and the global pseudonyms in the trusted pseudonymisation service. In such a scenario, to re-identify a patient, the ethical and medical board would first refer to the pseudonymisation service in order to retrieve the local pseudonym(s) for a given global pseudonym and would then ask all the study centers whether one of the local pseudonym(s) belongs to one of their patients.

4. Implementation and Evaluation

An implementation for evaluation purposes has been developed in Java. The pseudonymisation service provides a Web services interface to the study centers. The client side provides a user interface where the patient's PII necessary for the pseudonym generation can be inputted. From this PII the local pseudonym is computed and then transferred to the pseudonymisation service by calling the provided service. The response from the invoked pseudonymisation Web service includes the computed global pseudonym.

When, for example, the global pseudonym of Mr. Peter Patient born on January, 1st 1961 has to be generated, his ID^e is computed using the input data: 'Peter, Patient, 01-01-1961, m'. Since the ID^e is computed by for example hashing the input data in some suitable form, the ID^e is equal in each study center to `84c6fee073ffbe0854049292888b845fdba1d25cb1afb29f593e9823bff0c6aa`, whereas the computed local pseudonyms differ from each other (all values are expressed in hexadecimal form):

- $lid^{e,A}$: local pseudonym computed by study center Alice
  `1e69473a1d59c7969c0c3482e6e855003c850e186e259c35a9768aef13e3fb6ab3ce6cfc8989c12daf6a7ccccaef1e03367c9e033c0527ef4c312150e15650dc98878da34a43787c89819b23c6fd002755c9c01bbb7872599774da81881f9b78702462019185b761dae139a58c5ccd968aaa1df98bd7c148fe7ab2a1f883da42`
- $lid^{e,B}$: local pseudonym computed by study center Bob
  `2c86baf18b58b03743c213f325d0a4fb65cab1fb6b4bf9ca615e4cccfe7c2819ffdc24d2e6ca1266150a1e56bf07ea7e20674656131cfc52f13d1d6422b7bcf25fa2d4af4d33164b89a3d312be1f0c947bbe302df0fde58ce795c66371a3aafa05799ab0b30ff398db1993bb2386d23fa69d62881f61d08e4c21126b39314abb`
- $lid^{e,C}$: local pseudonym computed by study center Carol
  `0086391e496fdb8974ce2b9d64bf760d1c7152d9a323e9db30c6e80909d5e8ce4f56dd3d020a56801addd07b393e25b23221116c`

```
877fd5f6be370c70bc87de85b55be48f8a02f8ba445bcf36d113
443deb7430a735fc9f262eb4c418c8e3329136a4b495549d55de
cefab55f22a0e43b6f9149d823241c0ebd7bd62635043ec7b2
```

- $lid^{e,D}$: local Pseudonym computed by study center Dave

```
2c2baa66c47a79a529ff2bb21071a0f9929a58c1ab7672663c2b
83f49874b47a2fd12d1ffe7c8465f690372c6e2a2ee21cacba2a
05c8263e1b6b4fc528eec7004d9594e5f61d63fa5bfa2276307d
07a9a3a786c031f4377b54676b03c3c13824d29b6065e4197cc6
d3253e5b5dd1801726905d105878119428fb04acc7054da0
```

The global pseudonym computed by the pseudonymisation service out of the supplied local pseudonym and the appropriate public keys from the study centers, however, is again globally unique:

- gid^{e}: global pseudonym computed by pseudonymisation service Trent

```
98e1af507e6f33bf29fe5a1d67d4a05f626b57a5b030d794d85e
d302a0cdb4a1a1296d94d935496219bcc307fcce9a6c670edf03
df15093cc9b243e64e9efaba4571f2660ec2adfff482d463f45c
84abe0f1a5463676462bb3b032be6c00a0d0a14b071f11d90f4a
c1231e1bb514e2ee879f873100c73beaf5e70be0fe030fae
```

To reduce the size of the gid^{e} – which is bound to the cryptographic domain parameters –, it can e.g. be hashed using a cryptographic hash-function.

5. Security Considerations

Through the adoption of DLP-based asymmetric cryptography it is computationally infeasible to derive the random value r from the public domain parameters p and g as well as the public key r. The same holds for the computation of the local pseudonyms lid^{e}. Thus, nobody is able to obtain the ID^{e} from lid^{e} except the corresponding local pseudonymisation service.

In order to prevent anybody from computing the global pseudonyms gid^{e} from the local pseudonyms, the public key of the trusted pseudonymisation service must be kept secret.

The PII used to compute the local pseudonym must be choosen accordingly to render exhaustive search attacks infeasible.

Since the proposed pseudonymisation systems and its included cryptographic techniques are not reversible per se, corresponding mapping databases are required to re-identify a certain patient from its global pseudonym for matters discussed in section 1. Such components have to be safeguarded to prevent unauthorized access and henceforth the unauthorized re-identification of patients.

6. Conclusion and Outlook

The presented pseudonymisation system aims to contribute a possible solution to the problem of providing multi-centric health studies with universal pseudonyms enabling to cross-link the distributed but federated data sources of a study. It remains to be evaluated, however, whether the proposed system can support the real use cases of

multi-centric studies in an appropriate manner. In this context the implications contained in the proposed system concerning for example membership changed during the course of a study as well as related issues such as re-keying and re-pseudonymisation. These aspects will be further investigated and analyzed within the @neurIST project.

Acknowledgment

This work was generated in the framework of the @neurIST Integrated Project, which is co-financed by the European Commission through the contract no. IST-027703.

The author would like to thank Frederik Armknecht, Jun Furukawa, Greg Kohring, Jun Wang, Kazue Sako and Obana Satoshi for many fruitful discussions as well as the anonymous referees for their constructive comments.

References

[1] J. Fingberg, M. Hansen, M. Hansen, H. Krasemann, L. Lo Iacono, T. Probst, and J. Wright: Integrating Data Custodians in eHealth Grids - A Digest of Security and Privacy Aspects, In: Christian Hochberger, and Rüdiger Liskowsky (Eds.), Informatik 2006 - Informatik für Menschen, Lecture Notes in Informatics (LNI) vol. P-93, pp. 695-701, 2006.

[2] S. Bakhtiari, R. Safavi-Naini, and J. Pieprzyk: Keyed hash functions, In: E. Dawson, and Jovan Golic (Eds.), Cryptography: Policy and Algorithms, Lecture Notes in Computer Science, vol. 1029, pp. 201-214, Springer-Verlag, 1996.

[3] K. Pommerening, and M. Reng: Secondary use of the EHR via pseudonymisation, In: L. Bos, S. Laxminarayan, A. Marsh (Eds.), Medical Care Compunetics 1, pp. 441-446, IOS Press, Amsterdam 2004.

[4] D. Kalra, P. Singleton, D. Ingram, J. Milan, J. MacKay, D. Detmer, and A. Rector: Security and confidentiality approach for the Clinical E-Science Framework (CLEF), In: Second UK E-Science All Hands Meeting, Nottingham (UK), 2003.

[5] N. Zhang, A. Rector, I. Buchan, Q. Shi, D. Kalra, J. Rogers, C. Goble, S. Walker, D. Ingram, and P. Singleton: A Linkable Identity Privacy Algorithm for HealthGrid, In: T. Solomonides, R. McClatchey, V. Breton, Y. Legré, and S. Nørager (Eds.), Proceedings of Healthgrid 2005, Studies in Health Technology and Informatics, vol. 112, 2005.

[6] W. Diffie, and M. E. Hellman: New directions in cryptography, In: IEEE Transactions on Information Theory, vol. IT-22:6, pp. 644-654, November 1976.

[7] T. El-Gamal: A public-key cryptosystem and a signature scheme based on discrete logarithms, In: IEEE Transactions on Information Theory, vol. IT-31:4, pp. 469-472, July 1985.

[8] V. Miller: Uses of elliptic curves modulo lage primes, In: Advances in Cryptology – Crypto'85, Springer Verlag, pp. 417-426, 1986.

[9] N. Koblitz: Elliptic curve cryptosystems, In: Math. Comp., vol. 48, pp. 203-209, 1987.

[10] A. K. Lenstra, and E. R. Verheul: Selecting Cryptographic Key Sizes, In: Journal of Cryptology, vol. 14:4, pp. 255-293, 2001.

From Genes to Personalized HealthCare: Grid Solutions for the Life Sciences
N. Jacq et al. (Eds.)
IOS Press, 2007

ICGrid: Enabling Intensive Care Medical Research on the EGEE Grid

Harald Gjermundrød [a], Marios D. Dikaiakos [a], Demetrios Zeinalipour-Yazti [a]
George Panayi [b], Theodoros Kyprianou [b]
[a] *Department of Computer Science, University of Cyprus, 1678 Nicosia, Cyprus*
[b] *Intensive Care Unit, Nicosia General Hospital, Nicosia, Cyprus*

Abstract. Healthcare Information Systems are nowadays among the world's largest, fastest-growing and most information intensive industries. Additionally, Intensive Care Units are widely considered as the most technologically advanced environments within a hospital. In such environments, physicians are confronted with the challenge of storing and analyzing terabytes of data that stream in real-time from inpatients to centralized clinical information systems. In this paper we present the system architecture and early experiences from *ICGrid (Intensive Care Grid)*, a novel system framework that enables the seamless integration, correlation and retrieval of clinically interesting episodes across Intensive Care Units, by utilizing the EGEE infrastructure. ICGrid is based on a hybrid architecture that combines i) a heterogeneous set of monitors that sense the inpatients and ii) Grid technology that enables the storage, processing and information sharing task between Intensive Care Units. ICGrid makes it feasible for doctors to utilize the EGEE Infrastructure through simple, intuitive user interfaces, while the infrastructure itself handles the complex task of information replication, fault tolerance and sharing.

Keywords. Intensive Care Medicine, Grid Computing, EGEE

1. Introduction

An Intensive Care Unit (ICU) is an acute care environment deploying multidisciplinary team skills to resuscitate patients with potentially reversible multisystem dysfunction/failure and in danger of imminent death. The physiological condition of ICU patients is marked by rapidly evolving and frequently life-threatening derangements as well as 'silent' yet important alterations in homeostasis. Therefore, effective and reliable monitoring using multiple, patient-attached sensors is of ultimate importance in order to ensure early diagnosis, timely and informed therapeutic decisions, effective institution of treatment and follow up.

Modern ICU technology supports the continuous monitoring of patients through patient-attached sensors. Data retrieved from these sensors are presented to ICU staff as a continuous real-time flow of screen-displayed numerical values and waveforms. The capability to collect, store, process, and share such data along

with the knowledge and the experience of ICU medical staff can bring tremendous benefits to all aspects of Intensive Care Medicine (practice, research, education), in accordance to the emerging trends and challenges of *e-Science* [13,14]. Nevertheless, current ICU facilities provide limited if any capabilities for longer-term storage, off-line analysis, and easy sharing of data. Such capabilities clearly require access to advanced computational, storage, and software resources. Furthermore, the sharing of data between ICUs requires internetworking that takes into account the security and privacy considerations that go with sensitive medical equipment and data.

In recent years, Grids have emerged as wide-scale distributed infrastructures that support the sharing of geographically distributed, heterogeneous computing and storage resources [10]. Grids represent a promising approach for providing Intensive Care Units with easy access to high-end computational and storage resources. Furthermore, Grid middleware provides services that can be used to support and manage the secure, controlled and coordinated sharing of data deployed on Grid resources [2,3]. These capabilities have been demonstrated successfully in computational- and data-grid applications spanning across a wide spectrum of application areas [4,11,12,18].

In this paper, we present the design and implementation of *ICGrid*, a Grid application that enables the retrieval of data from patient-attached medical sensors found in modern Intensive Care Units, the filtering and annotation of these data by ICU medical staff, and the storage and replication of annotated data-sets on the EGEE infrastructure [1]. Also, ICGrid supports the controlled sharing of annotated ICU data-sets between collaborating hospitals; to this end, it provides functionalities for the distributed searching of stored metadata annotations and the retrieval of data-sets through the Grid. The ICGrid infrastructure (application and data) can be utilized in scenarios such as medical education, early diagnosis and for defining early warning systems that identify when human life is in jeopardy. The remainder of this paper is organized as follows. In Section 2 we present the context of modern Intensive Care Units and the prospects and challenges that arise for medical research from the sharing of ICU data-sets; furthermore, we describe the capabilities offered by the EGEE infrastructure. Section 3 introduces the architecture of the ICGrid system and the key issues that were addressed in the ICGrid design. Section 4 discusses the main implementation and deployment challenges faced. We conclude in Section 5, with a summary of our conclusions and future work.

2. Background: The ICGrid Motivation and Context

In this section we familiarize the reader with the required background and the respective terminology for the two main areas addressed by the ICGrid framework: i) Intensive Care Medicine and ii) the EGEE Grid. We highlight the open challenges with regards to information technology in Intensive Care, summarize the state-of-the-art and then outline the benefits of our proposed system framework.

2.1. Intensive Care Medicine

Clinical Environment: An Intensive Care Unit (ICU) is the only environment in clinical medicine where all patients are monitored closely and in detail for extended periods of time, using different types of *Medical Monitoring Devices (MMD)*. An MMD may be defined as a collection of sensors that acquire the patients' physiological parameters and transform them into comprehensible numbers, figures, waveforms, images or sounds. The acquisition is performed by attaching specially designed invasive or non invasive interfaces, such as cables, transducers, catheters etc., to the inpatients body. This in effect enables the extraction of parameters from the vitals parts of inpatients. The MMD can range from *ventilators* to *drug administration pumps*, *blood gas analyzers* and specialized equipment for *medical imaging*. MMDs are nowadays microprocessor-based and have their own screens and data export capabilities.

The acquired measurements can then be utilized in order to: i) take timely and informed therapeutic decisions, ii) ensure early diagnosis and finally iii) have effective institution of treatment and follow up. Note that the patient monitoring is always provided at the bedside, where the results of the acquisition are displayed on the screens in the form of alphanumeric data and waveforms. In order to assist administration, the operation of this setup can be complemented by centrally administered monitors, which connect through LAN-based technologies to a number of distributed monitors, and which consolidate all the information in a concise representation. This way the patients' acquisition data are saved in trends memories for a specific time period.

Technological Shortages: Taking clinical decisions for the ICU patients based on monitoring can be a very demanding and complex task requiring thorough analysis of the clinical data provided: *even the most skilled physicians are often overwhelmed by huge volumes of data, a case that may lead to errors, or may cause some form of life threatening situation* [7]. Consider for instance a physician that monitors a set of medical monitors in order to identify some abnormal condition in the projected state of an inpatient. Although such monitors enable the physician to *react* to alerts by triggering some medical procedure, they cannot provide means for *proactively* exploiting the real-time signals in order to uncover interesting patterns, predict trends and correlate the state of the inpatient to other similar cases. Providing systems that actively learn from previously stored data and suggest diagnosis and prognosis is a problem that, to our knowledge, has been overlooked in previous Intensive Care Medicine research.

Traditionally, medical research is guided by either the concept of patients' similarities (clinical syndromes, groups of patients) or dissimilarities (genetic predisposition and case studies). Clinical practice also involves the application of commonly (globally) accepted diagnostic/therapeutic rules (*evidence-based medicine* [9]) as well as *case-tailored approaches* which can vary from country to country, from hospital to hospital, or even from doctor to doctor within the same hospital. These different approaches in treating similar incidents produce knowledge which, most of the times, remains a personal/local expertise, not documented in detail and not tested against other similar data. Global sharing of this

cumulative national/international experience would be an important contribution to clinical Medicine in the sense that one would be able to examine and follow up implementation of and adherence to guidelines as well as to get the benefit of sharing outstanding experience from physicians.

Finally, although a number of dedicated and commercially available information systems have been proposed for use in ICUs [8], which support real-time data acquisition, data validation and storage, analysis of data, reporting and charting of the findings, none of these systems is appropriate in our application context because of the following limitations:

- ICUs lack the required high performance and dedicated storage resources for long-term collection of medical data, and the computational power to perform advanced processing and data-mining on these data.
- Most ICUs lack the qualified information technology personnel that can undertake the complex task of running in-house computations.
- The sharing and collaborative processing of medical data collected by different ICUs raises important privacy, anonymity, information integrity challenges that cannot be addressed by existing commercial ICU information systems.

An estimate of the amount of data that would be generated daily is given in the following scenario. Suppose that each tuple is 50 bytes (it is stored as text) and that there are 100 hospitals with 10 beds each, where each bed has 20 sensors. Assuming that each bed is used for 2 hours per day, the data collected amounts to 6.7055 GB per day. After compression, it would be reduced to 675 MB of data per day, but this number only represents the data from the sensors. Additional information includes metadata, figures, etc.

Grids represent a promising venue for addressing the challenges described above, since all of the aforementioned limitations are common advantages of Computational and Data Grids, with success stories in biomedicine, computational chemistry and high energy physics. It is worth noting that Grids have recently been adopted for the storage and sharing of human biological data [20].

2.2. The EGEE Grid

Currently, Grid computing infrastructures assemble an extensive collection of resources and expertise in production-quality operational support. For instance, the *EGEE* Grid (Enabling Grids for E-sciencE) [1], which is the largest Grid infrastructure in operation, at the time of this writing assembles over 200 sites around the world with more than 30,000 CPU's and about 5PB of storage. Access to EGEE resources is managed by the gLite middleware [2]. gLite comprises a variety of data management services, such as the *Storage Element*, which provides an interface to storage resources available at local Grid sites, the *Logical File Catalog*, which holds information about the location of files and replicas held at different Storage Elements, the *File Transfer Service*, which is responsible for replicating files across different Storage Elements, the *Metadata Grid Application (AMGA)*, which manages metadata, the *Encryption Data Service*, which manages the encryption of data stored in Storage Elements, etc.

Users that want to access a Grid infrastructure and make use of its services need to join a *Virtual Organization* (VO) supported by the infrastructure and its resource providers. A Virtual Organization is a dynamic collection of individuals and/or institutions that share resources in a controlled and mutually agreed fashion [10]. More specifically, EGEE users registered within a particular Virtual Organization obtain security credentials for single Grid sign-on that enable them to obtain controlled access to resources belonging to that particular VO, despite the fact that such resources span different EGEE sites across different countries. End-user resource access is obtained via a *User Interface (UI)* machine, which runs the services required to submit jobs and files to EGEE, to monitor job and resource status, to retrieve data from Storage Elements and Logical File Catalogs, etc.

The resources and services available on EGEE and other Grid infrastructures of similar scale are clearly adequate for storing and managing ICU-related data. In our work, we develop the software tools that are required to establish and operate a Data Grid infrastructure dedicated to Intensive Care Medicine, on top of EGEE.

3. ICGrid Architecture

The envisioned Data Grid infrastructure for Intensive Care Medicine will comprise:

- An *ICU Virtual Organization*, bringing together Intensive Care Units, hospitals, medical schools, medical research institutions, and resource providers willing to work collaboratively in order to promote research and education in Intensive Care Medicine through the sharing of data and knowledge produced inside the ICUs.
- *Grid Sites*, providing adequate storage and metadata facilities to securely store and maintain the huge data-sets collected from the medical sensors found in modern ICUs.
- The *ICGrid framework*, which: i) enables the easy retrieval of data-sets from the medical sensors; ii) supports the local buffering, anonymization and annotation of data by authorized personnel inside the ICU, using computing resources isolated from the Internet; iii) supports the uploading and replication of annotated data-sets to a Storage Element and metadata catalog of a nearby Grid site supporting the ICU-VO, and iv) provides members of the ICU-VO with access to searching capabilities and specialized data-mining applications.

In the remainder of this section, we present the architecture of the ICGrid framework and discuss some key design decisions.

3.1. Architecture Overview

Figure 1, illustrates the architecture of the envisioned ICGrid infrastructure, which comprises a number of different sites that are geographically distributed,

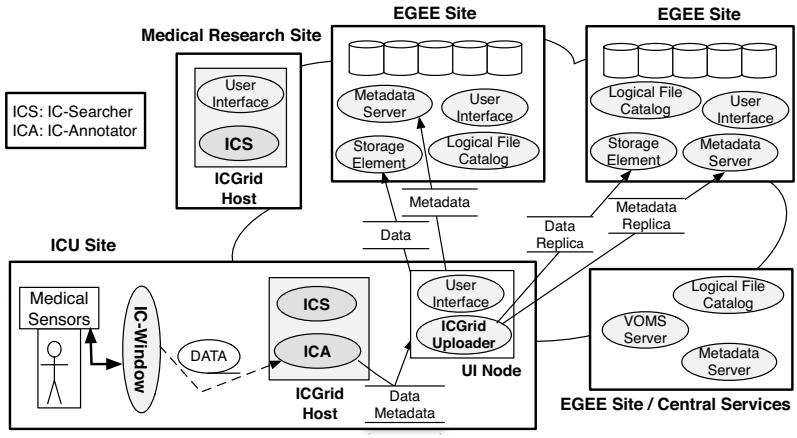


Figure 1. ICGrid System Architecture. White rectangles represent different sites of the infrastructure (each site represents resources of one administrative domain/institution), shaded rectangles represent computer nodes, and shaded ovals depict required Grid services and tools of the ICGrid framework.

belong to different organizations, and are connected through the Internet. In particular: i) *ICU Sites*, which are equipped with the required tools for the retrieval of data from Medical Monitoring Devices, the annotation thereof by expert physicians, and the transfer of collected information to the ICU Grid. A typical ICU Site will be hosted inside the Intensive Care Unit of a hospital that participates to the ICU-VO. ii) *ICGrid Storage Sites*, which support the ICU-VO by contributing their computing facilities (computers, disk farms, tapes) for the storage and processing of data and knowledge published by the ICU sites. iii) *Medical Research Sites*, which participate to the ICU-VO and have access to ICGrid tools for searching and processing data stored throughout the ICGrid Storage Sites. The ICGrid infrastructure is supplemented by one or more *EGEE Sites*, which host the Grid middleware services required for the coordinated operation of the distributed ICGrid resources (VO management, naming services, etc).

The *ICGrid framework* is based on a hybrid architecture that combines a heterogeneous set of monitors that sense the inpatients and three Grid-enabled software tools that support the storage, processing and information sharing tasks: the IC-Window (ICW), the IC-Annotator (ICA) and the IC-Searcher (ICS). The first was developed using Visual Basic while the others were developed in Java and make use of the following services of the gLite middleware of EGEE [2]: i) The *User Interface*, which is the entry point into the Grid infrastructure; ii) The *Storage Element* [19,6], which stores the collected data in data-archive files, replicated to provide fault tolerance, access load sharing, and faster access for the distributed users; iii) The *Logical File Catalog* [16], which is a service that keeps track of the archive files and replicas; iv) The *AMGA* [17] service, which is a metadata storage and retrieval system, complemented with replication, fault-tolerance and access load sharing functionality; and v) the *VOMS* [5] service, which generates the X509 proxy certificates required to authenticate the members of an EGEE Virtual Organization.

The diagram of Figure 1 depicts the acquisition and annotation of parameters of an inpatient at an ICU Site (bottom left) and the transfer of data replicas to two ICGrid Storage Sites. The transfer comprises of the actual sensor data, denoted as *Data*, and the information which is provided by physicians during the annotation phase, denoted as *Metadata*. Note that the *Data* is a selected subset of the acquired signals, those with the highest clinical interest. In particular, the physician that observes the inpatient annotates these signals with metadata in an offline phase. What is considered an interesting incident depends on the subjective opinion of the physician on duty. Consequently, we utilize the notion of a *Clinically Interesting Episode (CIE)* to refer to the captured sensor data along with the metadata that is added by the physician to annotate all the events of interest.

Metadata consists of information about the institution, physician, sensors, patient, intervals of the signals, along with some annotation of the signals. In order to protect patient privacy, information about the inpatient is anonymized. The only information that is collected for the inpatient is height, weight, age, and sex. The metadata is encoded as an XML-document defined by an application-tailored schema. On the other hand, the collected data consists of a set of tab-separated text files, one for each sensor. Each line in these files contains a timestamp, the recorded physical parameter and the state of parameter at the given time-stamp (e.g., if the parameter indicates some alert). All the files that belong to an episode are archived together to form the *Data* archive.

These two files (Data and Metadata) must be transferred to a storage unit that can be accessed by all the authorized and authenticated parties that will be entities of the ICGrid-VO. Thus, the services must satisfy certain security properties, something that is already inherent in the Grid. A Grid user authenticates by presenting an X509 proxy certificate [15] and as a result, the security properties of authentication, authorization, integrity, and non-repudiation are provided. Access to the files that are stored on the Grid is controlled by the readily-available Access Control Lists (ACL).

3.2. Design Decisions

Encryption: Our framework is designed around the concept of open access and sharing of medical knowledge. In order to accommodate the inherent privacy constraints we apply local anonymization procedures at each ICU-site, such that any non-disclosed information is filtered-out, prior to uploading the data to the Grid. This makes encryption redundant to a very large degree. Nevertheless, should such a functionality become a necessity, it could easily be accommodated by the Encrypted Data Storge component that is available by the gLite middleware.
Data Storage: Our current scheme stores data and metadata in plain text and XML format respectively. There is certainly a tradeoff between maximum flexibility and optimality of space utilization. We opted for the former alternative since it facilitates the incorporation of tools that perform searches and do signal processing on the acquired signals. Nevertheless, if storage becomes a concern, we can easily utilize an off-the-shelf relational database or a proprietary binary representation.

Metadata Management: The physical separation of the metadata from the collected data and the utilization of the AMGA Metadata service, provided by the middleware logic, introduces a number of advantages. Firstly, it improves scalability, as the metadata can be distributed across several sites without disrupting the operation of the application logic. Secondly, this distribution certainly improves fault tolerance and optimized access to the indexed information, leading to a simplified software development cycle.

4. ICGrid Implementation

In this section we present the most significant implementation details of the ICGrid framework and procedures for installing and operating the system.

One of the ICGrid framework design criteria was to make the use of the Grid infrastructure completely transparent to the end-user. Therefore we developed a collection of user friendly interfaces which allow physicians to get access to the EGEE resources without any prior knowledge of Grid technologies. In particular, we built three GUI-based tools: *Intensive Care Window (IC-Window)*, *Intensive Care Annotator (IC-Annotator)* and *Intensive Care Searcher (IC-Searcher)*, which enable the task of *acquisition*, *annotation* and *search* respectively.

The IC-Window tool (see Figure 2 (left)) interacts with the a number of patient monitoring devices. We have implemented a full-fledged interface to access the Phillips IntelliVue MP70 medical monitor, one of the most technologically advanced medical monitors on the market. We are also working on providing connectivity to other UDP/IP and RS232-based devices that will supplement our framework with more acquisition capabilities. Our objective is to provide open implementations to the various proprietary and closed protocol of medical monitors, thus contributing at the same time to more general field of medical Informatics.

The IC-Annotator tool (see Figure 2 (right)) assists the end-user with annotating the collected sensor data and uploading this information to the grid. The end-user follows a sequence of steps to annotate interesting incidents that occurred during an episode with the aid of an easy and intuitive GUI. First, the location of the directory where the collected sensor data is stored is specified. Second, the end-user enters the metadata, which is information about the institution, the physician, and minimal patient information. Finally, interesting incidents that occurred during the episode as well as optional HL7 information is entered.

The IC-Searcher tool, is a user interface that supports search over the acquired and stored data. Although IC-Searcher is still at an early stage, it already supports connectivity to the AMGA metadata service for the acquisition of data using free text searches. In the future we will support the automatic identification of similar episodes using high performance timeseries similarity methods executed using grid resources: *Given a real signal from an inpatient, we want to find other signals with a similar temporal movement.* Additionally, we will also support other data mining techniques, such as predicting the future value of a signal and clustering similar inpatient states.

The IC-Annotator and IC-Searcher tools are developed using the Java programming language, which provides platform independence, given that the EGEE

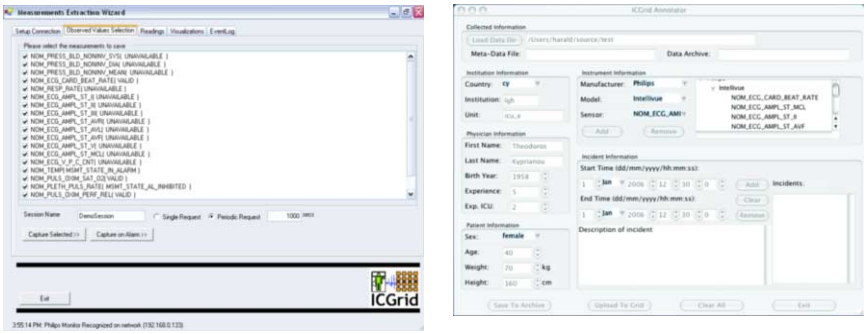


Figure 2.
a) IC-Window: enables physicians to acquire the clinical state of inpatients.
b) IC-Annotator: enables physicians to annotate the respective clinical state.

Grid infrastructure currently only supports a very limited number of Linux operating system distributions. On the contrary, the IC-Searcher is developed with the Visual Basic .NET environment under Windows, as this client is not required to interact with the Grid infrastructure directly. Note that the IC-Window can operate in a stand-alone mode, independently from the rest ICGrid framework, thus creating an invaluable tool for clinical monitoring using real-time graphs and a variety of other data management tools.

Installing the ICGrid framework at an ICU site is a straightforward procedure which is not much different from registering a new EGEE Grid user. In particular, users are issued a private/public key pair that is created and installed on the machine where the IC-Annotator and IC-Windows will operate. This provides a secure channel between the EGEE infrastructure and the ICU-site. This is succeeded by a certificate request from the local Grid certification authority and the issuance of an account on the local Grid entry point. Once the installation procedure is completed, the end-user can interact with the system only with the ICGrid tools. The fact that ICGrid utilizes available services of the EGEE infrastructure once again highlights the importance of Grids as generic infrastructures that can be customized to fit a variety of application scenarios hiding in that way most of the inherent software development and administration complexities.

5. Conclusions and Future Work

In this paper we present ICGrid, a framework that paves the way for Intensive Care Medical Research on the EGEE Grid. ICGrid is based on a hybrid architecture that combines a heterogeneous set of monitors that sense the inpatients and Grid technology that enables the storage, processing and information sharing task between Intensive Care Units. Our preliminary results have so far been extremely encouraging. In the future we plan to extend the framework to use the computational resources provided by the Grid to perform deep-searches in the collected data. Additionally, we want to offer physicians the capability to perform exploratory data analysis in order to maximize insight into the distributed ICGrid data repository, uncover important events, outliers and anomalies.

Acknowledgements

This work was supported in part by the European Union under projects EGEE (#IST-2003-508833) and CoreGRID (#IST-2002-004265).

References

[1] Enabling Grids for E-SciencE project. http://www.eu-egee.org/.
[2] gLite Middleware. http://glite.web.cern.ch/glite/ (accessed June 2006).
[3] Globus project. http://www.globus.org.
[4] WISDOM: Initiative for grid-enabled drug discovery against neglected and emergent diseases. http://wisdom.eu-egee.fr/ (last accessed June 2006).
[5] R. Alfieri, R. Cecchini, V. Ciaschini, L. dellAgnello and?A. Frohner, A. Gianoli, K. Lorentey, and F. Spataro. VOMS, an Authorization System for Virtual Organizations. In *First European Across Grids Conference*, February 2003.
[6] J-P Baud, J. Casey, S. Lemaitre, C. Nicholson, D. Smith, and G. Stewart. LCG Data Management: From EDG to EGEE. In *UK eScience All Hands Meeting*, 2005.
[7] B. Hayes-Roth et al. Guardian: A prototype intelligent agent for intensive care monitoring. *Artificial Intelligence in Medicine*, 4:165–185, 1992.
[8] B.M. Dawant et al. Knowledge-based systems for intelligent patient monitoring and management in critical care environments. In Joseph D. Bronzino, editor, *Biomedical Engineering Handbook*. CRC Press Ltd, 2000.
[9] DL Sackett et al. *Evidence-Based Medicine: How to Practice and Teach EBM*. Churchill Livingstone, 2nd edition, 2000.
[10] I. Foster and C. Kesselman. In I. Foster and C. Kesselman, editors, *The Grid: Blueprint for a New Computing Infrastructure*, chapter 4: Concepts and Architecture, pages 37–64. Elsevier, 2004.
[11] Ian Foster and Carl Kesselman. Scaling System-Level Science: Scientific Exploration and IT Implications. *Computer*, 39(11):31–39, November 2006.
[12] Fabrizio Gagliardi, Bob Jones, Mario Reale, and Stephen Burke. European DataGrid project: Experiences of deploying a large scale testbed for E-science applications. *Lecture Notes in Computer Science*, 2459:480–499, 2002.
[13] Jim Gray. eScience: the Next Decade Will be Interesting! Talk at ETH, Zurich, Switzerland, May 22 2006.
[14] Jim Gray, David T. Liu, Maria A. Nieto-Santisteban, Alexander S. Szalay, Gerd Heber, and David DeWitt. Scientific Data Management in the Coming Decade. *ACM SIGMOD Record*, 34(4):35–41, January 2005.
[15] R. Housley, W. Polk, W. Ford, and D. Solo. Internet X.509 Public Key Infrastructure, RFC 3280. http://www.ietf.org/rfc/rfc3280.txt, April 2002.
[16] P. Kunszt, P. Badino, A. Frohner, G. McCance, K. Nienartowicz, R. Rocha, and D. Rodrigues. Data storage, access and catalogs in gLite. In *Local to Global Data Interoperability - Challenges and Technologies*, pages 166 – 170, June 2005.
[17] N. Santos and B. Koblitz. Distributed Metadata with the AMGA Metadata Catalog. In *Workshop on Next-Generation Distributed Data Management HPDC-15*, June 2006.
[18] Peter M.A. Sloot, Alfredo Tiramo-Ramos, Ilkay Altintas, Marian Bubak, and Charles Boucher. From Molecule to Man: Decision Support in Individualized e-Health. *Computer*, 39(11):40–47, November 2006.
[19] Graeme A. Stewart, David Cameron, Greig A Cowan, and Gavin McCance. Storage and Data Management in EGEE. In *5th Australasian Symposium on Grid Computing and e-Research (AusGrid 2007)*, January 2007.
[20] L. Sun, P. Hu, C. Goh, B. Hamadicharef, E. Ifeachor, et al. Bioprofiling over Grid for eHealthcare. In V. Hernandez, I. Blaquer, and T. Solomonides, editors, *Challenges And Opportunities of Healthgrids: Proceedings of Healthgrid 2006 (Studies in Health Technology and Informatics)*. IOS Press, 2006.

This page intentionally left blank

VII. Imaging

This page intentionally left blank

From Genes to Personalized HealthCare: Grid Solutions for the Life Sciences
N. Jacq et al. (Eds.)
IOS Press, 2007

KnowARC: Enabling Grid networks for the biomedical research community

Henning MÜLLER [a,1], Mikko PITKANEN [a], Xin ZHOU [a], Adrien DEPEURSINGE [a], Jimison IAVINDRASANA [a] and Antoine GEISSBUHLER [a]
[a] *Medical Informatics, University and Hospitals of Geneva, Switzerland*

Abstract. The vast availability of medical patient data in digital format creates the opportunity to use these data in medical informatics research projects. The objective is to improve future care by providing the medical staff with methods for automated data processing, including textual and visual information analysis and retrieval from medical databases. Many medical institutions do not possess a specific research computing infrastructure or the required budget for such an infrastructure to enable processing of these large amounts of data. Still, many institutions have many desktop PCs that could serve for biomedical research during the time they are little used without the need for expensive investments. The *KnowARC* project aims at building a middleware for such a simple–to–install Grid network.

This article reviews requirements for computing Grids in large hospital environments. We use the computing infrastructure in the University Hospitals of Geneva as an example, and then present the solutions that the European Union–funded *KnowARC* project plans to undertake to solve the current problems. Methods currently employed in common Grid middleware distributions are also reviewed and assessed in relation to the goals of *KnowARC*.

The computing infrastructure at the University Hospitals of Geneva is described as well as the needs and requirements for computing and storage services within this domain. A list of requirements for a Grid middleware to employ in such a challenging environment is developed. Finally, the proposed solutions and ideas of the KnowARC project are described in detail to present the project to a larger community. First proof of concept implementations and test results are described to illustrate how Grid networks are expected to become an important supplier of computational resources, which are required in several domains in biomedical research. A continuous process will be necessary to feed in the requirements of the biomedical domain to developers of Grid middleware to make the outcome meet the specific needs of the biomedical community.

Keywords. Grid networks, healthgrid, biomedical informatics research, information retrieval

1. Introduction

The concept of Grid computing networks is currently a hot topic in many domains. However, the fundamental principles are fairly old as already in the late 1980s the Condor [1] project proposed the idea to use idle computer time of simple desktop computers for

[1]Corresponding Author: University and Hospitals of Geneva, 24 Rue Micheli–du–Crest, 1211 Geneva 14, Switzerland, henning.mueller@sim.hcuge.ch

computationally intensive tasks. The concept of Grids over the last years has been driven mainly by the high energy physics community and here in part by the CERN[2] (Centre Européen de la Recherche Nucléaire).

In addition to the early Condor project, several other projects were developed to provide distributed computing middleware packages such as Unicore [2] and Globus [3]. Perhaps the most widely known example about distributed computing is the SETI@home project[3]. This system is based on the BOINC (Berkeley Open Infrastructure for Network Computing) middleware, a fairly simple distribution mechanism for computational tasks.

Similar technology has been used for computing in biomedical domain to study protein folding[4] and to help with research on diseases[5]. However, these projects process data that has no strict requirements for information confidentiality and thus a lightweight middleware can be used. When biomedical (patient–related) information is distributed, more sophisticated security mechanisms have to be employed to ensure information confidentiality. Thus, the needed security guarantees motivate the need for modern Grid computing solution that is more controllable than the BOINC type of Internet computing.

The physicists had requirement for a computing resource that is powerful enough to process the enormous amount of data that the next generation particle accelerator, the LHC (Large Hadron Collider) is expected to provide starting from 2007. For the requirements of the physics community the large European Union project EDG[6] (European Data Grid) was started to advance the technology quickly and provide the community with standard interfaces. EGEE[7] (Enabling Grids for E–science in Europe) provides a follow–up of this project to deploy a worldwide Grid with partners in 57 countries and many application fields, including a work package on biomedical Grids. Many national Grid network initiatives and projects are linked to the EGEE project to base their efforts on well–established standards.

Besides the large global projects, several projects have developed on a smaller scale and with a slightly different focus. The NorduGrid[8] project started in 2000 [4]. The goal was to develop a lightweight middleware that would be non–intrusive for the reuse and coupling of clusters mainly in Northern European countries. One of the objectives was also to make a needs analysis and reuse existing components and standards wherever possible. Today, the result is the ARC (Advanced Resource Connector) middleware that is currently used in over 60 sites in many countries world–wide and is in continuous production use. Non–intrusiveness in this case means that the computers can be used for other software as well and do not require the system to run a single middleware, which is a requirement for many institutions. Dependency on particular versions of an operation system or compiler are also supposed to be as limited as possible.

In recent years, the term computing on demand has also developed fantasies in the computing industry. The recent emergence of services[9,10] that enable users to pay only

[2] http://www.cern.ch/
[3] http://setiathome.berkeley.edu/
[4] http://folding.stanford.edu/
[5] http://fightaidsathome.scripps.edu/
[6] http://www.edg.org/
[7] http://www.eu-egee.org/
[8] http://www.nordugrid.org/
[9] http://www.network.com
[10] http://aws.amazon.com/ec2

for the used computing cycles serve as a prime example of industry interest. Maintaining a computing infrastructure is expensive and paying for only the needed amount is an attractive model that allows specialised providers to develop quality solutions. This model enables high levels of usage and reduces the management cost, which is particularly interesting for medium and small actors both in public and in private sectors. Some companies such as Novartis already use desktop computers of their computing infrastructure over night for computing the testing of new drug docking mechanisms [5] to find candidates for new medications. A common infrastructure for this use is the United Devices[11] middleware that is running on several infrastructures and platforms, including Windows, which is employed in many companies and larger organisations. Most scientific Grids on the other hand are based on Linux. Role model for many of these projects is to have a structure for storage and computing power similar to the information structure of the world wide web.

In the medical domain the healthgrid[12] association has brought together the biomedical informatics research community with health specialists to develop common solutions. So far, it has organised four conferences for this purpose. Several position papers identify a strong potential and the need for biomedical Grid networks in hospitals [6] has been published several times. However, many applications of Grids within hospitals including those of image retrieval are still in an early prototype phase [7,8,9] and not yet used in clinical practice and with routine data. A notable exception is the Mammogrid[13] [10] that has combined routine storage of mammograms in several European countries. A more recent project is health–e–child[14]. Many global visions exist on how to combine large amounts of data from many levels (molecular, cell, organ, individual, population) to really make use of large–scale Grid networks [11] and use the available data up to its full potential. One of the biggest challenges with employing a Grid infrastructure in a non–grid–specialist environment is currently the installation and maintenance of such systems. Many solutions still require access to operating system resources on the used machines and extensive technical knowledge. Most scientific middlewares work only on the Linux operation systems, while failing to install on other widely used systems. In the medical field and in other large administrations there are other factors as well that limit the use of Grid networks, and these are mainly issues related to information security. Many computers contain confidential information, for example patient data or confidential administrative documents. A middleware to be employed on a large scale in an institution needs to take these security issues into account and also find political solutions by integrating the technical deciders. An enormous potential is surely present.

2. Methods

This article reviews the main goals of the KnowARC[15] project from the point of view of the biomedical research community. The situation of the University and Hospitals of Geneva within the project is described in more detail including the challenges concern-

[11] http://www.ud.com/
[12] http://www.healthgrid.org/
[13] http://www.mammogrid.com/
[14] http://www.health-e-child.org/
[15] http://www.knowarc.eu/

ing the infrastructure and also the main research goals. Finally, the first steps towards a hospital–wide computing infrastructure for biomedical research are explained and a short outlook is given.

3. Results

3.1. Computing infrastructure at the University and University Hospitals of Geneva

The University hospitals of Geneva (HUG) have an infrastructure and resource use that is similar to many other administrations. Large databases on servers (mostly on Sun Solaris at the HUG) exist that store the patient records and particularly the large image archive (PACS, Picture Archival and Communication System). The image archive that consists of several Terabytes is stored at the same time in two locations on two SANs (Storage Area Network) to provide a maximum of flexibility and security. With 2'200 beds and 10'000 employees, the Geneva University hospitals are one of the largest hospital groups in Europe. Over 50'000 images are currently produced per day and roughly a million events registered by the fully electronic patient record system. Besides the large servers, most computers in use are simple desktop machines that are centrally managed (for software and hardware distribution) and updated regularly. Roughly 6'000 computers exist in total and these machines are regularly renewed. Currently, a 5–year renewal cycle exists meaning that modern machines are Core2Duo CPUs with 2 GHz and 1 GB of RAM whereas the oldest computer currently in use are Pentium IV machines with 2 GHZ and 512 MB of RAM. A 100 MBit Ethernet network connects the computer on the distributed sites of the hospitals. The network is strongly secured and all outgoing traffic is basically blocked. All WWW traffic has to pass a proxy server for security control.

For research on the other side, no dedicated institutional infrastructure exists at all. Research projects with funding buy their own servers and additional desktop computers that are subsequently used within the project but only very rarely shared on a personal basis. If no funding for servers exists simple desktop computers are used for the research limiting the possibilities in domains such as data mining, text mining [12], and image processing and retrieval [13], where computational power can make a large difference with respect to possible technologies.

3.2. Constraints for using this infrastructure

The computing infrastructures in hospitals have stricter security requirements than usually faced within the University research in computer science faculties. This concerns data confidentiality as well as the use of the infrastructure. On the one hand there are research data sets that can not leave the hospital network to be computed, for example, if the data might allow the identification of the patient. Structures to anonymise the data are available and then, part of the data can possibly be computed outside of the hospital in a computing centre such as the Swiss National Super Computing Center (CSCS[16]). Several constraints were identified early that a to–be–developed middleware has to take into account:

[16] `http://www.cscs.ch/`

- The availability of the infrastructure is not always given and the use of the existing PCs can change quickly (PCs are turned off, they can be used, portable computers can be removed from the network).
- A heterogeneous infrastructure is characteristic for an organisation where computers are only renewed in regular cycles every five years, whereas most Grid networks are much easier to administer and load–balance if the infrastructure contains hardware of the same type.
- Running Clinical applications is the principal goal of the computer network group of the hospitals. The applications need to run 24 hours a day with no interruptions caused by computing research. Research is clearly only an independent little part inside the institution with little decisive power.
- Access to all data on computers is strictly regulated and a process running on a computer with clinical data should at no point have the possibility to access these clinical data.
- No maintenance personnel for the research infrastructure can be allocated at the hospitals as is practice in big computing research centres at Universities. Researchers need to maintain the computers and also adapt their applications for these computers, so simplicity is important.
- No dedicated computers for the computing tasks exist, so the real users of the PCs should not be slowed by these tasks; a user needs to be able to control the amount of computing power that he can make available for research tasks.
- Operating systems in the hospitals are dictated by the organisation. The very large majority of the hospital's computing infrastructure is based on Windows and particularly desktops are, whereas many middleware research projects are only applicable on Linux platforms.

To learn more about the possible challenges that introducing a Grid infrastructure might face in the administrations, we are currently preparing a survey among the personnel of the University hospitals, the University, and the Public Administration of Geneva. This survey is to identify further problems and also to educate the users about the possibilities and challenges that Grid computing networks can offer. Many of the future challenges to surpass are not of technical but rather of political nature.

3.3. Planned solutions to be provided by the KnowARC project

The early evaluation of challenges has also led to possible solutions for many of the issues and several of these solutions are currently being implemented in the KnowARC project:

- Use of a an existing and working middleware as the basis with the need to only develop new components and adapt existing software toward these needs (ARC of NorduGrid, in our case).
- Using virtualisation to run Linux machines within the centrally administered Windows machines. This allows to use Grid software on existing windows desktops. The sandboxing through virtualisation allows also to clearly separate the instance of an operating system running on clinical data and the instance running the research task, which solves one important security issue.

– The installation process needs to become as easy as installing a simple program on windows, so only a few clicks are required. This allows everyone to quickly test the software without the hurdles of complicated installations.
– A simple task submission interface (graphical) is needed to make also the gridification of applications as easy as possible.
– Only a modifiable part of the computing power is used for the Grid, the user can decide on this. This is already enabled by the virtualisation infrastructure that we have deployed on server and desktop machines. Users can easily adjust the virtual machine through a graphical interface to donate only the specified amount of resources to Grid tasks.
– The load–balancer has to deal with the very heterogeneous architecture and has to estimate computation times.

3.4. Image retrieval on a distributed hospital infrastructure and on external Grids

Our current image retrieval system that is to be gridified is based on the GNU Image Finding tool (GIFT[17], [14]). On the currently used desktop computer, the extraction of visual features from a single image takes roughly 1–2 seconds. Databases of image retrieval benchmarks such as ImageCLEFmed[18] [15] contain in the order of 50'000 images, resulting in 14–20 hours calculation time. If various parameters for the features have to be tested the indexing is repeated several times. The retrieval part on the other hand is harder to parallelise but this part is currently still quick. With database sizes of around 50'000 images, response times on a desktop are in the order of 2 seconds. A particular problem is that the current infrastructure limits what we can do with the data to be able to extract features quickly. More complex features can easily take one minute per image instead of 1 second resulting in 94 hours for computing.

3.4.1. Adapting an image retrieval system for KnowARC

The feature extraction component itself is a fairly small program written in C. It does not have any special programming library dependencies. Scripts for database indexing currently take the name of a directory and then search all subdirectories for image files. For these files visual features are extracted sequentially.

A first gridification of the feature extraction takes N images at a time and sends the image files with the feature extraction source code to an ARC server. Then, the source code is compiled for each package and the features are extracted for all images in the package. After the execution the results are downloaded to a local machine for index generation. Overhead times of the job submission are fairly high and it can take 30 seconds before a block is executed. Still, as there can be many parallel packages, the overall computing and communication time is much faster than sequentially computing the visual features of all images of a large collection even without any optimisation.

The initial tests conducted on NorduGrid test sites within the virtual organisation KnowARC in Europe have given the image features 5-10 minutes after the images were sent to the Grid for processing resulting in total time for assembling the info in less than one hour that beforehand took 14 hours on a single machine. These first measurements

[17] `http://www.gnu.org/software/gift/`
[18] `http://ir.ohsu.edu/image/`

clearly illustrate that the overhead of job scheduling and data transfers are in a tolerable range. Moreover, expert opinion has proposed us to use batch sizes that result in roughly one hour of computation when submitted to European wide Grid deployment. This is in line with our initial measurements that small jobs create a fairly large overhead. While the optimal batch size will depend on the use of resources on the time of job execution, an important step has been taken in reducing the time requirements from 14 hours to just a fraction. Thus, more computing power enables us to design systems that provide more rapid response or that can use more complex feature sets, which is the final goal of this project.

3.4.2. Virtualising desktop computers

First steps have also started on virtualising standard hospital desktop computers. The windows computers now run Linux inside a virtual machine and software can be installed securely inside the sandboxed Linux operating system. Currently, we are only allowed to do this on computers of research personnel, which is sufficient for first tests (and the machines have the exact same characteristics). Six machines are currently equipped with a small distribution of Ubuntu that should make the installation of the ARC middleware relatively easy. One problem is currently that the virtual machines do not have their own IP address, which is a requirement for the middleware that we hope can be removed in the near future. Otherwise, we will be obliged to obtain additional IP addresses as well for our tests.

4. Conclusions

To apply a Grid computing infrastructure on networks in hospitals that are equally used for routine clinical practice might still be a few years away, but technology is offering solutions that could change this quickly. The possibility to virtualise even simple desktop machines allows to have a fix separation of the research and the routine parts of the computers removing at least a part of the security concerns. It also allows users of computers to close a virtual machine if this would slow down the work. Of course, much work is still ahead to make efficiency tests for virtualisation solutions to measure the amount of computing power that can really be made available.

One important step for us right now is to imagine computationally intensive solutions that are able to use a possibly unlimited computing power available to improve the current approaches in medical data analysis and medical information retrieval. Researchers in the medical imaging domains do not have the habit of doing so, and this new thinking might open up completely new possibilities.

There will most likely be two surviving business models for hospital Grids. One is the use of an existing infrastructure that allows research projects to profit from computing power partly during the day but mainly during off peak hours by using idle circles of already available CPUs. This model will mainly work for research and non–critical applications. For time–intensive critical application an external computing service provider can be used to perform extensive tasks for example during an operation or while a patient is in intensive care to optimise the outcome. A model for paying these computational requirements has to be developed.

Sharing in the scientific world needs to be extended to more domains than only knowledge with publications. Computing resources and power can equally be exchanged through the use of Grid networks and virtual organisations. Sharing of other resources such as databases for research is equally seen as an extremely important step to make research more efficient and effective [16].

Acknowledgements

This work was partially supported by the Swiss National Science Foundation (Grant 205321-109304) and the EU 6th Framework Program in the context of the KnowARC project (IST 032691).

References

[1] Litzkov, M., Livny, M., Mutka, M.: Condor — a hunter of idle workstations. In: Proceedings of the 8th international conference on distributed computing, San Jose, California, USA (1988) 104–111
[2] Romberg, M.: The UNICORE architecture: Seamless access to stributed ressources. In: Proceedings of the 8th international symposium on high–performance distributed computing, Redondo Beach, California, USA (1999) 287–293
[3] Foster, I., Kesselman, C.: Globus: A metacomputing infrastructure toolkit. International Journal of Supercomputer Applications **11** (1997) 287–293
[4] Ellert, M., Gronager, M., Konstantinov, A., Konya, B., Lindemann, J., Livenson, I., Nielsen, J.L., Niinimäki, M., Smirnova, O., Waananen, A.: Advanced resource connector middleware for lightweight computational grids. Future Generation computer systems (2006 – to appear)
[5] Peitsch, M.C.: Grid computing in drug discovery. In: Sixth IEEE International Symposium on Cluster Computing and the grid, Singapore (2006) 3–13
[6] Müller, H., Garcia, A., Vallée, J.P., Geissbuhler, A.: Grid computing at the university hospitals of geneva. In: Proceedings of the 1st healthgrid conference, Lyon, France (2003) 264–276
[7] Müller, H., Michoux, N., Bandon, D., Geissbuhler, A.: A review of content–based image retrieval systems in medicine – clinical benefits and future directions. International Journal of Medical Informatics **73** (2004) 1–23
[8] Montagnat, J., Breton, V., Magnin, I.E.: Partitioning medical image databases for content–based queries on a grid. International Journal of Supercomputer Applications **44** (2005) 154–160
[9] Costa Oliveira, M., Cirne, W., de Azevedo Marques, P.M.: Towards applying content–based image retrieval in clinical routine. Future Generation Computer Systems **23** (2007) 466–474
[10] Estrella, F., McClatchey, R., Odeh, M., Rogulin, D., Solomides, T.: Experiences of engineering grid–based medical software. International Journal of Medical Informatics (2007 – to appear)
[11] Sloot, P.M.A., Tirado-Ramos, A., Altintas, I., Bibak, M., Boucher, C.: From molecules to man: Decision support in indivisudalized e-health. IEEE Computer **November** (2006) 40–46
[12] Ruch, P., Baud, R., Geissbuhler, A.: Evaluating and reducing the effect of data corruption when applying bad of words approaches to medical records. International Journal of Medical Informatics **67** (2002) 75–83
[13] Müller, H., Lovis, C., Geissbuhler, A.: The medGIFT project on medical image retrieval. In: Medical Imaging and Telemedicine, Wujishan, China (2005)
[14] Squire, D.M., Müller, W., Müller, H., Pun, T.: Content–based query of image databases: inspirations from text retrieval. Pattern Recognition Letters (Selected Papers from The 11th Scandinavian Conference on Image Analysis SCIA '99) **21** (2000) 1193–1198 B.K. Ersboll, P. Johansen, Eds.
[15] Müller, H., Deselaers, T., Lehmann, T.M., Clough, P., Eugene, K., Hersh, W.: Overview of the imageclefmed 2006 medical retrieval and medical annotation tasks. In: CLEF 2006 Proceedings. Lecture Notes in Computer Science (2007 – to appear)
[16] Vannier, M.W., Summers, R.M.: Sharing images. Radiology **228** (2003) 23–25

From Genes to Personalized HealthCare: Grid Solutions for the Life Sciences
N. Jacq et al. (Eds.)
IOS Press, 2007

Globus MEDICUS - Federation of DICOM Medical Imaging Devices into Healthcare Grids

Stephan G. ERBERICH[a,b], Jonathan C. SILVERSTEIN[c], Ann CHERVENAK[d], Robert SCHULER[d], Marvin D. NELSON[a], Carl KESSELMAN[d]
[a]*Department of Radiology, Childrens Hospital Los Angeles, Keck School of Medicine, University of Southern California (USC)*
[b]*Department of Biomedical Engineering, Viterbi School of Engineering, USC*
[c]*Departments of Surgery, Radiology and Computation Institute of the University of Chicago and Argonne National Laboratory*
[d]*Center for Grid Technologies, Division of Advanced Systems, Information Sciences Institute*

Abstract: The Digital Imaging and Communications in Medicine (DICOM) standard defines Radiology medical device interoperability and image data exchange between modalities, image databases - Picture Archiving and Communication Systems (PACS) - and image review end-points. However the scope of DICOM and PACS technology is currently limited to the trusted and static environment of the hospital. In order to meet the demand for ad-hoc tele-radiology and image guided medical procedures within the global healthcare enterprise, a new technology must provide mobility, security, flexible scale of operations, and rapid responsiveness for DICOM medical devices and subsequently medical image data. Grid technology, an informatics approach to securely federate independently operated computing, storage, and data management resources at the global scale over public networks, meets these core requirements. Here we present an approach to federate DICOM and PACS devices for large-scale medical image workflows within a global healthcare enterprise. The Globus MEDICUS (Medical Imaging and Computing for Unified Information Sharing) project uses the standards-based Globus Toolkit Grid infrastructure to vertically integrate a new service for DICOM devices – the DICOM Grid Interface Service (DGIS). This new service translates between DICOM and Grid operations and thus transparently extends DICOM to Globus based Grid infrastructure. This Grid image workflow paradigm has been designed to provide not only solutions for global image communication, but fault-tolerance and disaster recovery using Grid data replication technology. Actual use-case of 40 MEDICUS Grid connected international hospitals of the Childerns Oncology Group and the Neuroblastoma Cancer Foundation and further clinical applications are discussed. The open-source Globus MEDICUS project is available at http://dev.globus.org/wiki/Incubator/MEDICUS.

Keywords: DICOM Grid, Medical Image Grid, Protected Health Information, Grid PACS

Introduction

The Digital Imaging and Communications in Medicine (DICOM) standard (www.nema.org) defines the image and meta-data format and the network protocol between medical imaging devices. DICOM has become the de-facto standard for medical imaging adopted by all medical equipment vendors.

The latest DICOM standard, release v3, defines the image format and a peer-to-peer network transport protocol. DICOM objects contain meta information about patient, study, series, and image attributes. Together, image format, meta data, and network model, define a medical image specific application domain, referred here as DICOM domain.

Healthcare image management comprises primarily (i) storage, (ii) availability (fault-tolerance and disaster recovery), and more recently (iii) interoperability and information exchange. However DICOM does not provide image management, but leaves it to the vendors to implement. As a consequence, we see a variety of silo implementations today making it almost impossible to achieve interoperability. The Integrating the Healthcare Enterprise (IHE, www.ihe.net) initiative has be created to address the issue of interoperability with the approach to interface existing silos with significant success in recent years. The latest interface, the cross-enterprise document exchange (XDS) and its medical imaging cousin XDS-I are taking shape.

However XDS and XDS-I only address interoperability. Still missing is an open standards based architecture that provides an overarching and scalable infrastructure to implement all aspects of healthcare image management (i-iii). Such an infrastructure must keep the integrity of DICOM intact to preserve compatibility to existing and future investments into imaging devices, Picture Archiving and Communication Systems (PACS), and display workstations – a $1.6 Billion annual business.

We believe Globus Alliance Grid technology (1,2) is one such open architecture. It provides reliable industry standards for the most challenging problems associated with network collaborative environments: (i) high-speed reliable data transport utilizing high-bandwidth networks, (ii) enterprise level security (data, authentication, authorization), (iii) large scale data management and replication, and (iv) publication, discovery, sharing, and use of distributed, independently owned and operated computational, storage, and data resources federated in the Grid as web-services (WS, www.oasis-open.org). The Grid paradigm spawns a virtual organization (VO) over public and/or private networks between resource providers (e.g. imaging devices) and consumers (e.g. radiologists and scientists).

The significance of the MEDICUS project (Medical Imaging and Computing for Unified Information Sharing) described here, extending previous work (3-5), is the seamless integration of DICOM devices into Grids. As such it builds on the overarching Grid infrastructure to implement fault-tolerant storage with disaster rccovery and interoperable exchange. In the Grid, medical images become transparently available anywhere in the VO (e.g. a Regional Health Information Organizations (RHIO) of hospitals or practices) and among VOs.

Because the images are available within the Grid infrastructure, standards based WS (e.g. storage, image processing, data mining) become easier to develop and to deploy. We believe that such an overarching open standards infrastructure, like

MEDICUS, will enable image availability at the point-of-care and thus helps to eliminate redundant imaging to the benefit of the patient and subsequently reduces cost.

1. Methods

The design of the Grid image management is based on the gateway concept, bridging between the DICOM domain and the VO (Figure 1). This concept provides the benefit of a lightweight extension, without changes to the existing DICOM equipment.

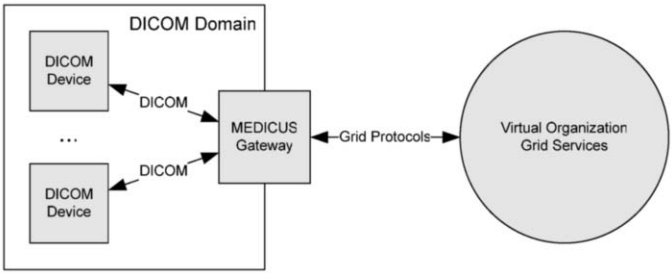


Figure 1. Gateway concept: Transparent connection between DICOM devices, e.g. PACS, and Globus Toolkit Grid services. The Gateway resides in the LAN or DMZ of the DICOM domain (e.g. hospital).

The MEDICUS gateway has four major functions: (i) DICOM to Grid Protocol translator, (ii) discovery (metadata and logical file), (iii) image data transport, (iv) image data caching. The components are vertically integrated from existing Globus technologies: (i) MyProxy for delegated credential creation and Grid authentication management (grid.ncsa.uiuc.edu/myproxy), (ii) Gridshib (6) and Internet2 Shibboleth (shibboleth.internet2.edu) to state user assertions for role-based service authorization, (iii) OGSA-DAI (7) as meta catalog to reference image attributes and storage keys, (iv) Replica Location Service (RLS, dev.globus.org/wiki/Replica_Location) to map image keys and physical storage locations, and (v) GridFTP for data transport and as storage provider (dev.globus.org/wiki/GridFTP).

1.1. DICOM Grid Interface Service

The DICOM Grid Interface Service (DGIS) (8) implements these functions and acts as the conduit between DICOM and Grid domains. Seamless communication between both domains requires carefully orchestrated image workflows and Grid methodologies (Figure 2), integrating security (data and access), data storage, image publication and discovery, and fault-tolerance. DGIS enables image workflows for (i) send (publication) and (ii) query/retrieve (discovery) of images which are described as follows.

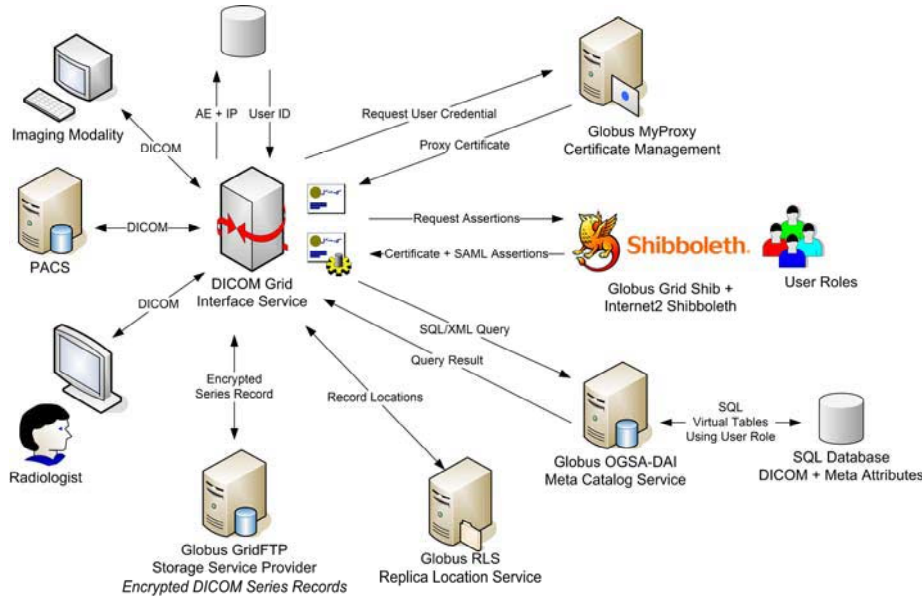


Figure 2. Healthcare image management with standards based Grid open architecture comprising user authentication (MyProxy), user-role authorization (SAML assertions), DICOM meta data query (OSGA-DAI), Grid image series discovery (RLS), and storage/retrieve (GridFTP).

1.1.1. Image Send Workflow

A typical image send workflow of a DICOM modality, e.g. a PACS, pushes images to DGIS using DICOM TCP/IP communication. DGIS accepts the connection and validates the application entity title and internet protocol address. Next the calling modality initiates a DICOM C-Store operation and DGIS acts as DICOM Storage Class Provider (SCP) to receive the images.

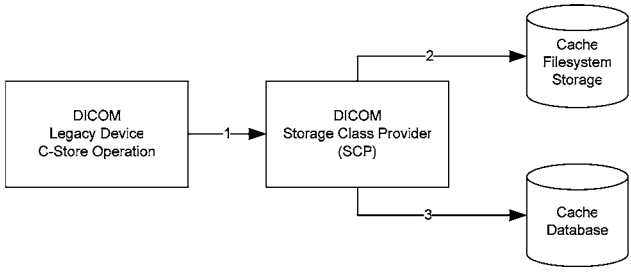


Figure 3. DICOM image C-Store operation from DICOM legacy devices stores images in study, series, and image hierarchy on the DGIS host file system.

DGIS is multithreaded and capable to receive multiple requests at a time (Figure 3.1). Images are stored as part-10 DICOM object on the host file system (Cache) (Figure 3.2) with meta-data reference in the Cache DB (Figure 3.3).

Next the Grid Scheduler finds and schedules new series for Grid publication. Each scheduler entry lists the series, Grid storage and the Meta Catalog URLs. Multiple storage and/or catalog servers at a time can be configured.

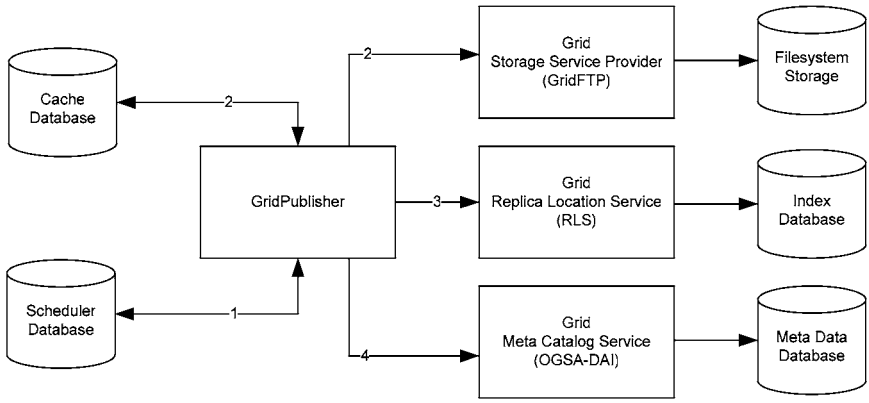


Figure 4. DGIS Grid Publisher thread executes scheduled DICOM image series updates to the Grid. Data (compressed series records) and meta data (DICOM attributes) are stored separately.

Next the Grid Publisher thread (Figure 4) executes the scheduled Grid requests (Figure 4.1). First a compressed series record using the Java zip container is created using the loss-less Lempel-Ziv-Welch (LZW) algorithm. Compressing the images of one series into a single series record provides about 60% in data reduction in average and a new atomic data element used in the Grid. This is of importance because Grid data transport is optimized for large files and single small image file transfer used by DICOM proved to be inefficient. However DICOM query and retrieve at the image level from a modality is supported by DGIS in conformance to the DICOM standard.

The unique identifiers (UID) of study and series are used as keys in MCS and in RLS to reference a series. In future versions of Globus MEDICUS, these UIDs will be replaced with new DICOM compatible ones to ensure anonymity of the series information (even when unassociated with specific identifiers, these identifiers also exist in the individual hospitals' PACS, so trail re-identification is possible).

Now the Grid Publisher publishes series records via GridFTP to the remote Grid SSP (Figure 4.2). Next the RLS is updated with the new Grid location (Figure 4.3). The VO's MCS (Figure 4.4) is invoked to execute an XML/SQL update statement of relevant DICOM attributes (e.g. patient, study, series, images attribute levels). After successful Grid publication the series request is removed from the DGIS Scheduler Database. Now the series is discoverable in the Grid.

1.1.2. Image Receive Workflow

A DICOM legacy device (e.g. a PACS Display Workstation) queries DGIS using standard DICOM C-Find operation to obtain images from the Grid. DGIS translates C-Find into a Grid request, an XML embedded SQL query schema, and sends it to the MCS. The MCS executes the statement on the back-end database (e.g. MySQL, Oracle, or DB2). The query result is returned to DGIS and translated into a DICOM C-Find respond object and sent to the calling device. The device parses the result object and lists the query results to the application.

Once a C-Find is obtained, a selection of images at the patient, study, series or image levels is prepared and returned to DGIS. Both synchronous DICOM C-Get and asynchronous DICOM C-Move operations (Figure 5.1) are handled by DGIS and translated into XML embedded SQL queries. Study and series UIDs are obtained from the MCS (Figure 5.2) and compared to existing series in the Cache DB (Figure 5.3). If

not in cache, the Grid storage location is requested from RLS (Figure 5.3) and a GridFTP transfer is invoked (Figure 5.4). Then a C-Get/Move response (C-Store SCU role) is created and images are send to the destination device of the DICOM retrieve (Figure 5.5). DGIS provides streamed operations - series images are sent immediately to the DICOM device while remaining requested series are still in Grid transfer. This allows immediate image review before the entire study is retrieved. DGIS preserves study and series order for C-Find and C-Get/Move operations.

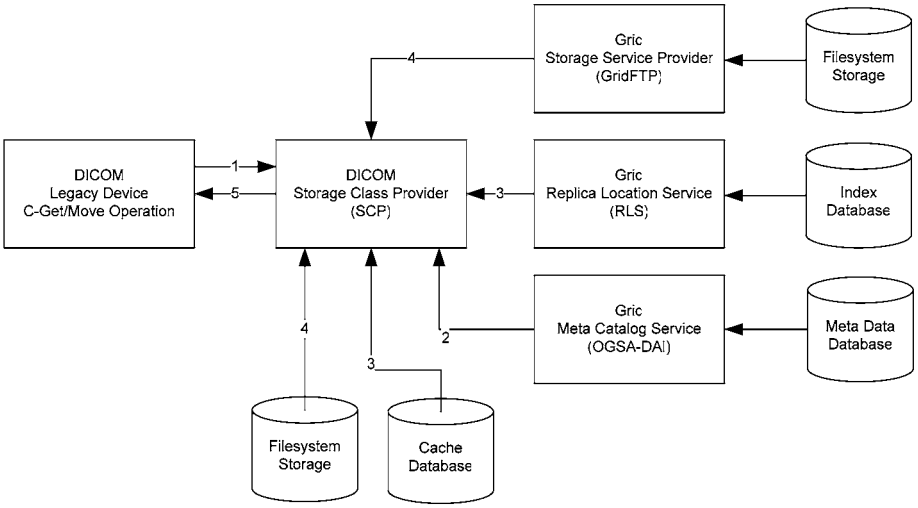


Figure 5. DGIS C-Get/Move DICOM operation. Image locations are discovered from the Grid MCS and delivered either from the local file system or from a Grid SSP.

Using DB back-ended handlers for publication and discovery workflows ensures fault tolerance of the DGIS. Thus a system failure and restart of the system does not affect the operational state of DGIS which will continue where interrupted. DGIS logs all activities in its DB providing an audit trail.

1.2. Security

Sharing medical images with embedded meta-data across medical facilities raises concerns about protected healthcare information (PHI) defined by HIPAA (Health Insurance Portability and Accountability Act of 1996, www.hhs.gov/ocr/hipaa).

Therefore Grid image workflows must be strictly controlled by user authentication and access authorization with audit logging. Globus Grid technology provides two industry standard methodologies (i) X.509 certificates for authentication and (ii) SAML certificate extensions (www.oasis-open.org) for access authorization.

MEDICUS integrates these methodologies to ensure privacy-preserving through a three-layered process (Figure 2). The first layer is that DGIS obtains authentication from an identity provider (IdP). Here we use the Internet2 Shibboleth as trust anchor with GridShib translation and obtain SAML (www.oasis-open.org) assertions (Figure 2). The second layer is the MyProxy X.509 certificate authority (CA) of the VO which creates a limited lifetime credential with or without SAML embedded assertions. Here

the CA is operated by the VO administrator (e.g. the hospital network), the IdP at the federal level.

The third layer is that all images stored in the Grid (series records) are de-identified in compliance to DICOM using a message digest one-way encryption. Now the original patient identifiers (e.g. patient name, patient ID) are stored in the MCS database along with the encrypted identifiers. To access the original identifiers one needs to bear matching SAML assertions in the X.509 certificate. GridShib receives the proxy credential of the first-layer and adds SAML assertions about the user, e.g. "Dr. No is member of the Keck School of Medicine USC, Los Angeles".

The layered security mechanism is comparable to a passport (X.509) and a visa (SAML) which allows entrance and privileges. A user needs to apply for both, passport and visa, separately increasing the overall security.

Table 1. Example of SAML assertions stored in the MCS database.

ID	User	Department	Institution
0815	Dr. No	Radiology	Keck School of Medicine, USC, Los Angeles, USA
0816	Dr. Yes	Neurology	Sick Childerns Hospital, University of Toronto, Toronto, Canada

Now DGIS implements this workflow by either obtaining a user credential (e.g. PACS administrator credential) or by matching the calling DICOM device. The latter uses the calling application entity title (AE) and internet address (IP) as user identification (e.g. Dr. No uses Display Workstation "Dr. No's WS" with a static IP "192.168.0.42"). Now SAML assertions and proxy credentials are obtained. The SAML embedded proxy credential is now presented to all MCS activities, publication and discovery. MCS exposes WS operations – update/insert and query. In both cases SAML assertions are incorporated into the SQL statement at the MCS. Therefore the DGIS SQL statement is never directly executed, but only used to define the update/search space.

DGIS removes PHI when performing publication and restores them when needed for query/retrieve. When images are published to the MCS, user, department, and institutional assertions of the publisher are associated with the patient. If data of the same patient from multiple institutions is added, additional entries exist for all image providers allows them to access the PHIs (Table 1). Now accessing RLS or SSP one only requires first-layer security - a valid proxy credential of the VO, because the information stored here does not contain PHI.

The layered security model has two significant benefits: (i) it is PHI preserving under the control of a VO independent trust anchor allowing multiple security domains on-top of VOs, (ii) it allows access to the image data (with encrypted PHI) without PHI clearance. The latter is particularly useful if images must be shared, but PHI should not (e.g. commercial SSP, commercial image processing services, teleradiology, and VO external image review). Because SAML assertions are added to the certificate, one can collect assertions from other security domains and aggregate access to several VOs.

1.3. Data-Storage, Fault-Tolerance, and Disaster Recovery

Because image storage is external and supplemental to existing PACS in the MEDICUS model, FT and DR is inherited by the core replication capabilities of Globus Toolkit (9). As such smaller healthcare providers without PACS (> 400 beds)

can use Globus MEDICUS as open-source Grid PACS with image caching at the gateway. Larger providers with PACS can deploy their SSP(s) on campus or remote with high-speed network access. In both scenarios FT and DR is achieved by the flexible and scalable nature of the Grid model. A GridFTP SSP is a light-weight storage unit requiring only the GridFTP server on a standard PC hardware (currently ~$800/TB). This allows proportional scaling with application requirements (e.g. Linux based PC up to multi-processor data center). Obviously one can achieve significant increase in data safety and FT, when using data replication at multiple deployed Grid connected SSPs in combination with RLS indexing the SSPs content, shifting PACS operations toward a pure network based approach - already the trend in the industry.

1.4. Web-Service Image Management

Besides image publication and discovery through DICOM legacy devices, one can take advantage of the fact that images are now discoverable directly with in the Grid - direct access to the data (meta and image data) through standards based WS interfaces. Therefore no separate SDK or API is needed for application development.

In order to provide a global image management paradigm, one wants to be able to control DGIS from the Grid. The next release of DGIS will provide a WS interface accessible from Grid allowing reverse – Grid initiated – workflows. This is of particular interest in combination with GridPortal technology (e.g. GridSphere), because it enables a pure WS image management.

2. Results

We made two observations using MEDICUS in the research image workflows within the Children's Oncology Group (COG) and Neuroblastoma Cancer Foundation (NANT) Grids, connecting 40 international medical centers.

(i) MEDICUS was well accepted by PACS technologists and physicians because it hides Grid specifics and does not require changes to the existing DICOM environment and workflow. As a result MEDICUS has been incorporated seamlessly into the research image workflow at the deployed sites.

(ii) One anticipates better network performance using the more efficient compressed series records and GridFTP vs. DICOM for WAN Enterprise operations. We setup a benchmark test to measure these network performance differences using six representative nodes using Linux 2Ghz systems within the COG Grid (CHLA/USC Los Angeles; Stanford Univ. Medical Center, Palo Alto; Univ. of Washington, Seattle; Mayo Clinic, Rochester; Univ. of Toronto, Toronto; St. Justine Hospital, Montreal).

A test suit comprising real image data representing typical clinical and cancer research data types was used: CT: 1 chest series, 42MB; MRI, 13 brain series, 255MB; PET/CT, 2 abdomen series, 45MB. The test suit was transferred 10 times (about 3.4GB) between nodes (one at a time) during office hours at 9am, 12pm, and 3pm pacific standard time. The average performance gain measured for MEDICUS Grid transfer using compressed series record compared to direct DICOM transfer to the six nodes is about factor 2.3 (avg.: 20.6 Mbps Grid; 8.6 Mbps DICOM). No data encryption was used for both protocols.

We also investigated the performance difference per data type (MRI, CT, PET/CT) to compare DICOM vs. Grid protocol overhead. Data types with small data elements

result in many small 2D files. These have a proportional large meta-data overhead (e.g. low resolution PET data) (10). Initial results measured between Los Angeles-Seattle showed Grid performance remains more constant (std.dev: 2.75 Mbps Grid; 3.19 Mbps DICOM using public networks).

3. Discussion

Sharing of clinical radiological image data across institutions has proven challenging for clinicians, clinical scientists and even patients. Although the DICOM standard has proven very useful for vendors to build compatible systems and components within radiology departments at hospitals and other large clinical care and research settings, the standard is bulky or limited in regard to scalability, fault-tolerance, security, and communication performance. Difficulties faced by users due to these limitations include the inability to routinely and predictably manage clinical image data obtained from one institution at another institution. Images are often transported by hand on films or on compact disk, released only via individual patient consent for that particular image series and thereby are viewable only in the limited settings in which the patient physically participates in obtaining them and providing them to the user or workstation. We believe this is below current expectations of patients. Similarly, in the research setting, separate archives (usually de-identified, which limits future usefulness) are typically created under IRB protocol for each particular use. This creates redundancy and waste. Ideally, clinical image users would participate in virtual organizations in which integration and sharing of images across institutions would be easier, more universal, and still meet the strictest security standards (11-13).

Grid technologies in general and Globus in particular were designed to enable virtual organizations to integrate and share data across institutions by decomposing the systems into specialized compatible components. Globus MEDICUS addresses the underlying technical issues that enable integration of today's PACS across institutions via a technical decomposition of them into resource consumers and providers. Specifically, Globus MEDICUS can be deployed as a master archive into which multiple sites can provide data with specific access privileges gaining FT and security features and improved performance only; or by deploying DGIS and SSPs at each institution allowing the user to virtually aggregate images across multiple SSPs (without pre-negotiated inter-institutional sharing agreements, rather with sharing agreements with individuals or VOs). In these architectures, no specific individual identifier is required for each patient because fuzzy logic can be applied to identify patients at multiple institutions or data can be requested only from the specific places in which it is expected to exist.

Naturally, building an enabling infrastructure alone is insufficient to meet user demands. Social issues such as legal requirements and work-flow preferences within each institution and across institutions must be addressed to build workable global healthgrids. However, what MEDICUS does for DICOM is enable virtually any management policy imaginable to be executed while providing FT and improved performance.

4. Conclusions

The Globus MEDICUS project addresses several major barriers to establishing global clinical image sharing: (i) Integration of multiple sites into distributed scalable, FT, global PACS via trivial cost and open technology; (ii) Role-based, user-based and/or anonymous data security according to Grid and other open HIPAA compliant standards; (iii) Integration of DICOM devices and PACS via DICOM protocols; (iv) Loss-less image compression to substantially increase communication performance. Initial deployment across 40 medical centers in North America is concluded and will serve as ample test-bed to prove the value of the Grid model. Others are encouraged to use Globus MEDICUS and contribute toward its open development at: dev.globus.org/wiki/Incubator/MEDICUS

5. Acknowledgment

The authors would like to acknowledge Gaurang Mehta for his generous help with GT4 coding, installing, and advice. Thanks to Dr. David Clunie for his support with DICOM and his open-source PixelMed Java Dicom toolkit (www.pixelmed.com). This work is supported by a grant of the National Institute of Health, UO1-CA97452, and by an award to the New Approaches to Neuroblastoma Therapy (NANT) consortium from the Children's Neuroblastoma Cancer Foundation (cncf-childcancer.org).

References

[1] The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration. Open Grid Service Infrastructure WG, Global Grid Forum, June 22, 2002

[2] Globus Toolkit Version 4: Software for Service-Oriented Systems. IFIP International Conference on Network and Parallel Computing, Springer-Verlag LNCS 3779, pp 2-13, 2005

[3] An OGSA Middleware for managing medical images using ontologies. J Clin Monit Comput. 2005 Oct;19(4-5):295-305

[4] MammoGrid: large-scale distributed mammogram analysis. Stud Health Technol Inform. 2003;95:194-9

[5] Grid Based Medical Image Workflow and Archiving for Research and Enterprise PACS Applications. Proceedings SPIE Medical Imaging 2006, Volume 6145, pp. 203-208 (2006)

[6] Identity Federation and Attribute-based Authorization through the Globus Toolkit, Shibboleth, Gridshib, and MyProxy. In *5th Annual PKI R&D Workshop*, April 2006.

[7] The design and implementation of Grid database services in OGSA-DAI. Concurrency and Computation: Practice & Experience, Vol17 , Issue 2-4 (February 2005), p357-76

[8] DICOM grid interface service for clinical and research PACS: a globus toolkit web service for medical data grids. Int Journal of Computer Assisted Radiology and Surgery, 2006, 1:87-105; p100-102, Springer, Heidelberg

[9] Wide Area Data Replication for Scientific Collaborations. Proceedings of 6th IEEE/ACM International Workshop on Grid Computing (Grid2005), November 2005

[10] PACS-based functional Magnetic Resonance Imaging. SG Erberich. Comput Med Imaging Graph. 2003 Mar, 27(2-3):229-40

[11] New trends in the virtualization of hospitals--tools for global e-Health. Stud Health Technol Inform. 2006;121:168-75

[12] An extensible framework for sharing clinical guidelines and services. Conf Proc IEEE Eng Med Biol Soc. 2004;5:3155-8.Stud Health Technol Inform. 2006;120:381-91

[13] TRENCADIS--a WSRF grid MiddleWare for managing DICOM structured reporting objects. Stud Health Technol Inform. 2006;120:358-67

From Genes to Personalized HealthCare: Grid Solutions for the Life Sciences
N. Jacq et al. (Eds.)
IOS Press, 2007

Alchemist Multimodal Workflows for Diabetic Retinopathy Research, Disease Prevention and Investigational Drug Discovery

Adina RIPOSAN [a, b,1] Ian TAYLOR [c, d] David R. OWENS [e] Omer RANA [c] and Edward C. CONLEY [c, e]

[a] *Department of Applied Informatics, Military Technical Academy*
[b] *Contact Net Ltd, Bucharest, Romania*
[c] *School of Computer Science, Cardiff, UK*
[d] *Center for Computation and Technology, LSU, USA*
[e] *Diabetes Research Unit, School of Medicine, Cardiff University*

Abstract. In this paper we present mechanisms for imaging and spectral data discovery, as applied to the early detection of pathologic mechanisms underlying diabetic retinopathy in research and clinical trial scenarios. We discuss the *Alchemist* framework, built using a generic peer-to-peer architecture, supporting distributed database queries and complex search algorithms based on workflow. The *Alchemist* is a domain-independent search mechanism that can be applied to search and data discovery scenarios in many areas. We illustrate *Alchemist*'s ability to perform complex searches composed as a collection of peer-to-peer overlays, Grid-based services and workflows, e.g. applied to image and spectral data discovery, as applied to the early detection and prevention of retinal disease and investigational drug discovery. The *Alchemist* framework is built on top of decentralised technologies and uses industry standards such as Web services and SOAP for messaging.

Keywords. Biomedical Informatics, Biomedical integration, Biomedical imaging, Spectroscopy, Diabetic Retinopathy, Search engine, Resource discovery, Grid, HealthGrid, P2P, peer-to-peer, WS-RF, WSPeer, Ubiquitous computing

1. Introduction

The majority of Internet search engines are based primarily on crawling the Web to create large index databases which are sorted according to a sophisticated ranking system. An end-user searches these massive databases via a Web site and retrieves potentially hundreds of thousands of results, sorted according to the given ranking system. The *Alchemist* infrastructure described here uses an alternate approach that allows users to *proactively* push information into a decentralised "search database" using standardised

[1] Email: adina.riposan@contactnet.ro, Web: www.contact-net.ro

Web Services interfaces, developed within the business and Grid computing communities, and hosted on a peer-to-peer (P2P) infrastructure.

The *Alchemist* is a project at Cardiff University that is creating a framework to provide a P2P layer for supporting pluggable network discovery and caching overlays coupled with the ability to execute distributed workflows [1]. The framework is being built on an existing middleware system, called WSPeer [2], which provides a SOAP messaging layer (using Web or WS-RF Services) within a P2P network that supports a super-peer topology of rendezvous or advert caching peers to support the scalability of the discovery and access to information on the network as a whole. This system already interfaces with existing Grid middleware (e.g. Globus) through Web Services interfaces but is also able to provide access to such capabilities within a decentralised environment. For example, rather than relying on centralised discovery mechanisms e.g. UDDI or similar, we can search super peers for WSDL files to provide access to the distributed services. In this way, the network can cope with far more transient participants, it can support exchanging roles and is capable of scaling proportionately with the number of peers.

The *Alchemist* infrastructure can be customised to be used in particular fields and particular user-environments and scientific communities, such as the biomedical research environment, astrophysics research, audio research, etc. This paper here presents an analysis of how such technologies may be used in the field of diabetic retinopathy by aggregating different sources of information at different levels in order to achieve more reliable and accurate results for the patients. The next section describes the application of *Alchemist* to data discovery and diabetic retinopathy workflows; we propose two scenarios - one for early detection and prevention of retinal disease, and one for retinal drug discovery. In section 3, we discuss the *Alchemist* framework and how the distributed interactions take place.

2. Workflows for Data Discovery in Diabetic Retinopathy

Here we propose the study of diabetic retinopathy as a focus for customised utilisation of the *Alchemist* multimodal search/workflow system - one that spans both fundamental biomedical research and routine clinical (screening) practices.

The development of *Alchemist* applications relevant to diabetic retinopathy is being guided by expertise within the Diabetic Retinopathy Screening Service for Wales (DRSSW), and the Diabetes Research Unit (DRU, Llandough Hospital) with the School of Optometry and Vision Sciences (SOVS), Cardiff University. Within the DRSSW and DRU a large body of high quality research data has been collected relating factors associated with specific graded outcomes of diabetic retinopathy (DR). Additional data is currently being obtained from primary care settings. Anonymised retinopathy data collected over time with specific ethical permissions for re-use represents a unique research resource. Data can include high-resolution retinal images plus associated quantitative physiological, demographic and other variables of relevance to an individual's risk of disease progression. Expert manual analysis of selected data within the DRU suggests the existence of progression *patterns* (progression risk signatures) that underlie common disease mechanisms (relevant to several disease endpoints, not limited to DR). Enough has been learned from manual inspection of *disease progression patterns* to suggest the feasibility of computer-based progression risk identification. Disease progres-

sion *data discovery* and distributed *pattern matching* across whole populations has the potential to radically improve efficiency in patient selection, intervention effect evaluation, optimal frequency/cost of screening. However, to realise such advanced retinopathy search paradigms in *Alchemist*, significant work applying focused ontologies needs to be done, introducing formal representations, structured domain terminologies (e.g. from SNOMED-CT constrained by CDA) using a defined messaging specification (such as HL7 V3). Such structures are absolutely necessary to permit interoperability of data amongst families of composite applications in the diabetes domain such as *Healthcare@Home* [3].

2.1. Scenario 1: Early Detection and Prevention of Retinal Disease

A key application of data discovery using *Alchemist* would be in early detection and prevention of disease (in later stages of retinal disease, interventions have diminishing beneficial effects). By performing pattern searches at multiple levels and transformations of signal data *Alchemist* has the potential to help discover - in the outwardly disease-free population - disease associative variables. These could take many forms - signal frequency shifts, consistent patterns, inconsistent patterns, image pixel patterns, physiological variable patterns - that are present in those people that on the timeline *succumbed to retinopathy versus those that did not*. The purpose of *Alchemist*, therefore, could be to discover the variable patterns intrinsic to the disease at the observable pathology level (macro), the molecular and cellular level (micro) and at the nano level. Molecular-level disease-intrinsic biomarkers (e.g. biophysical or chemical probes that generate reliable quantitative of qualitative indicators of disease in combination with an instrument sensor technology) would have the most potential impact worldwide if they can be made inexpensive and reliable. Currently, practitioners are limited with observable indirect pathology that is the result and not the cause of retinopathy.

Physiological domain data patterns from the *Healthcare@Home* project [3] could also reveal disease-associative risk from individual primary care case data (e.g. hypertensive or not, treatments people are on, lipids problems, smoking status, obesity status, HbA1c status, type of diabetes, and duration of treatments). *Healthcare@Home* is the Cardiff University-based research demonstrator for diabetes data interoperability, monitoring, evaluation and pathway-based decision support (see Figure 1). *Healthcare@Home* composite applications provide a research model for data aggregation services and patient-centred care that operate on a large scale using standardised clinical datasets. The *Healthcare@Home* project [3] enables both “push” and “pull” based mechanisms to enable data from mobile devices and/or dedicated home-based network servers to one or more analysis engines. We intend to make these and other data available to *Alchemist* through the *Healthcare@Home* applications so that the search and associative processes can work on them at an individual level. The target system will match continuous data collection from service inputs with intervention episode outcome analysis on the life timeline. For this disease prevention measure, DRSSW has systematically collected data related to retinal grading outcomes on individual timelines (albeit the limited physiology data from primary care sources is currently not integrated). We believe this *evaluated outcome-based* approach to disease prevention that takes into account quality data at multiple levels of observation is an ideal test scenario for the metadata multimedia search potential of *Alchemist*.

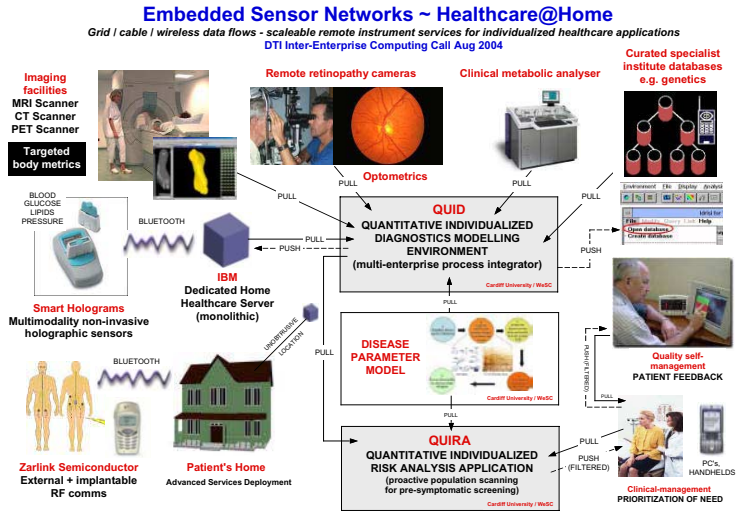


Figure 1. Rich picture illustrating information flow within Healthcare@Home - a research demonstrator for patient-centred healthcare services [3]

2.2. Scenario 2: Retinal Drug Discovery

There is a clear lack of a unified search framework for tracing mechanistic effects of drugs (beneficial and deleterious) on the nervous and vascular cells of the human eye. In its applications to drug discovery and systems biology *Alchemist* has potential to start bridging these wide gaps — e.g. relating drug effect signals observed in relatively simple in vitro cellular test systems with effects in highly complex cellular systems (in man). This comparison framework has applications in reduction of animal testing, especially for drugs whose effects can be tested on the vascular and neural signals measurable in retinal studies. Comparisons could be made between drug efficacy/toxicity effect via (sensor-based) signals on human cellular/molecular dissociated systems versus complex human cell co-culture systems versus in vivo effects recorded within investigational drug settings. A consistent multi-level predictive search framework that could link mechanistic observations at genomic, molecular, cellular, intercellular (tissue), organ and organism levels may better predict failures in drug discovery pipelines. Obtaining good evidence for a drug failure decision early in the pipeline is always preferable to failing a drug later (e.g. when it creates toxicity problems in human subjects).

2.3. Vertical Data Integration in Diabetic Retinopathy

Diabetic retinopathy (DR) is a disease that accounts for c.80-90% of cases of blindness due to diabetes in the UK (c.13% in the overall working population of 19-65 years - Evans Report, HMSO 1991). All persons with diabetes are at risk of developing DR, and its progression to a sight-threatening stage is often not detected. Observable vascular disease in DR - microaneurysms, haemorrhages and cotton-wool spots - have associated *functional* defects that often precede (and therefore can predict) the positional incidence of gross pathology. As reviewed by Antonetti et. al (2006) [4], vascular and neural cell

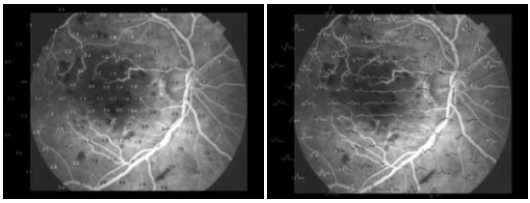


Figure 2. Multifocal electroretinograms (MfERG's) recorded with the visual evoked response imaging system (VERIS) can superimpose localised electrical response amplitudes/latencies and waveforms over digital fundus images and angiograms. For credit see footnote

deficits are inter-dependent in DR, and both are key to understanding the mechanism of vision impairment (see Figure 2[2]).

Moreover, a coordinate marking system should provide for rendering the data comparable between examinations, and correlated within a single time-scale. This could be achieved by an initial point of reference - relating this to the times when the different stages of retinal disease are categorically identified. *At the clinical level*, *retinal fundus grading* is an established method for morphologic assessment of diabetic retinopathy. Grading (i.e. a standardised classification for severity of disease) is generally based on the number, location, and type of *discrete microvascular lesions* in the fundus of the eye. Analysis is generally static for higher lesion densities. At low lesion densities, a chronological sequence of appearance, number, maturation, and disappearance of fundus lesions can be made. Semi-quantitative classifications (based on comparison with standard photographs at various stages of retinopathy) are available but are time consuming and subject to error (i.e. borderline elements can confuse human graders). Detailed lesion mapping / counting and dynamic time series analysis of fundus photographs can be more easily achieved by digital photographic techniques [5].

Although observed microvascular changes are certainly related to retinopathy, *proliferative* retinopathy is better predicted by loss of oscillatory potentials on *electroretinograms* than by methods such as (i) vascular lesions observed on fundus photographs or (ii) capillary non-perfusion visualised by fluorescein angiograms. It is most important to take into consideration the fact that the retina is substantially a vascularised *neural tissue*, not merely a network of blood vessels. *Neural dysfunction* accompanies and precedes visually observable cellular pathology. Numerous reports using electroretinography, dark adaptation, contrast sensitivity, and colour vision tests demonstrate that neuroretinal function is compromised *before* the occurrence of vascular lesions in the retina. From these findings it will be clear that inclusion of data and metadata trails from examinations *at the functional level* is of high significance for developing systems capable of *early detection* and (through early intervention) *prevention* of disease [4].

2.3.1. Automated Retinal Lesion Detection in Diabetic Retinopathy Workflows

Metadata generation in the context of a retinopathy workflow at the *pathology/clinical level* (e.g. colour retina fundus photography grading) consists of contextual observations made by a professional grader. These observations form part of an examination's work-

[2]Permission to use photograph kindly given by Dr. Erich E. Sutter of Electro-Diagnostic Imaging Inc., Redwood City, California, USA

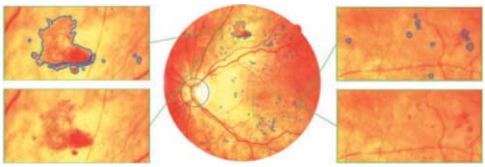


Figure 3. Recognition of retinal lesions by an automated image-analysis algorithm (from Larsen et. al., [5])

flow, e.g. to mark vascular lesions. The grading is then registered on a structured annotation system (see Figure 3) [5].

Figure 3 shows a digitised colour fundus photograph representing non-proliferative diabetic retinopathy with annotation markers outlining red microangiopathic lesion types (haemorrhages and microaneurysms) identified by an automated image-analysis algorithm. Studies have been performed with the objective of developing methods for automated detection of haemorrhages and microaneurysms and to compare this automated lesion detection corresponding to visual identification of diabetic retinopathy. In a key study [5] the feasibility of using automated detection of red lesions in diabetic retinopathy was demonstrated, but the accuracy could be further raised by adjustment of a single user-supplied parameter determining a balance of screening priorities (essentially a piece of metadata). The standard arbiter for automatic lesion detection in the Larsen et. al. [5] study was defined by classifying each patients disease status as having (or not having) diabetic retinopathy (based on overall visual grading of the digitised transparencies). Further assessment in the same patient would indicate whether there has been a change in retinopathy status (based on lesion appearance).

In these and other degenerative pathologies, search functions of *Alchemist* have the potential to reveal if characteristics of graded disease are present. In cases where they are present, *Alchemist* can indicate how they vary over a defined reference timeline.

2.3.2. Systems Biology-based Interpretations of Functional Tests for Retinopathy Status

At the *cellular level*, diabetes alters the function and structural integrity of all retinal cell types. There is evidence for neural function changes in the retina as one of the earliest detectable changes in diabetes. The highly sensitive and complex *trophic* interactions between neuron-neuron and neuron-effector cell signalling in diabetic retinopathy requires a more sophisticated *systems biology* approach to understand. Systems biology approaches (in general) take into account complex interactions of gene, protein, and cell elements to explain system-emergent properties (in the most sophisticated form by biologically realistic modelling of measured outputs of underlying components). For the present time, functional analyses of intact - *in vivo* - systems in the clinic still rely on macroscopic-level techniques. Highly informative molecular-level and cellular-level (microscale) models generally require system-dissociative - *in vitro* - techniques prior to gathering data via (for example) microelectrode and patch-clamp electrophysiology. While these techniques deliver the most detailed mechanistic information, the experimental effects observed do not necessarily scale to the next level up. There remain significant gaps in system understanding essentially due to the dissociative process. The *Alchemist* has some potential to bridge these gaps, but realistically, for the present, data originating from various positional and non-positional clinical optometry techniques have much potential for new search paradigms e.g. multifocal electroretinography

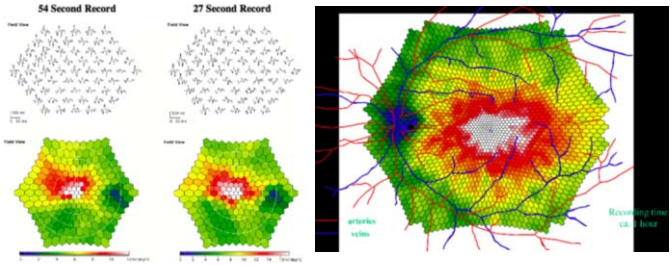


Figure 4. Left: Examples of retinal functional response mapping by electrophysiology in mfERG. Right: Overlay of responses superimposed with positions of arteries and veins. For credit see footnote

(mfERG), multifocal oscillatory potentials (central/peripheral visual field), microperimetry, visual acuity, contrast sensitivity measurements and macula-restricted colour vision.

Investigative ophthalmic techniques as potential data sources for *Alchemist*: *Short Wavelength Automated Perimetry (SWAP)* can detect progressive visual field loss prior to conventional (white-on-white, W-W) perimetry. SWAP is applied in a number of conditions: glaucoma, diabetic macular oedema, various neuro-ophthalmic disorders [6].

For study of intervention treatment efficacies on a longitudinal basis (i.e. analysing individual patients data over time) *2- and 3-dimensional maps of retinal function* represent a powerful direct method. Stimulus-record techniques such as multifocal electroretinography (mfERG) that use *light stimulus patterns* directed across the retina in a predefined 2-d matrix (fig. 4 [3]) can have their corresponding electrophysiological responses recorded within that positional matrix.

There are understandable and sometimes valid issues of lack of trust in data generated by automated routines of any kind. This gap inhibits the development of high-productivity machine learning-based approaches. A significant part of the problem is the lack of *validated datasets* that are of sufficient quality to be used in algorithm performance testing. Consistent outcome datasets that are *trusted* could actually be used for testing and performance ranking (validation) of many algorithms submitted by different computational research groups over the internet (e.g. on a competitive basis for the best blind outcome prediction performance). The purposeful generation of large volumes of interoperable, quality-controlled data is an essential evolutionary step for *consensus pattern searches*, *pattern matching* and *artefact recognition*. Ultimately, a trusted data status for computer-based recognition means signals can be taken as representative of the biological system (and/or the effect of a specified system perturbance).

3. The *Alchemist* Framework and Application to Retinopathy

The *Alchemist* is a P2P framework and an associated workflow toolkit (called the *Alchemist* Integrated Toolkit (AIT)) that provides a flexible way of deploying P2P overlays and allows workflows to be distributed across the system at various levels of granularity. The current system is built on existing well-tested technologies, such as WSpeer, P2PS [4] and the Triana workflow environment [7].

[3] Permission to use photograph kindly given by Dr. Erich E. Sutter of Electro-Diagnostic Imaging Inc., Redwood City, California, USA

[4] http://www.trianacode.org/p2ps/

Triana[5] is a well known workflow environment that can integrate within a number of different distributed environments for allowing true heterogeneous computing across different Grids and distributed paradigms. Such an approach has led to a number of different bindings to underlying middleware and therefore a number of possible modes of operation. For example, Triana on the one hand has a full binding to Java GAT interface, capable of invoking tools and services such as Condor, GridFTP, GRAM, etc., whilst on the other has integrated with service-based middleware, such as Web Services, WS-RF, Jxta and P2PS. It also has the capability to dynamically wrap applications remotely behind Web Services interfaces so that existing software can be easily integrated.

WSPeer has been Triana's Web Services toolkit for the past three years and many projects have used this combination to specify their distributed course-grained service workflows [8]. For example, Triana has been in this capacity[6] for: radio astronomy, astrophysical simulations, data mining, biodiversity problems, grid-enabled medical simulations, environmental science. P2PS has been used as the underlying P2P environment during this time and it has been successfully used in many domains, such as gravitational wave analysis , audio processing and distributed music information retrieval (MIR), distributed P2P simulations and e-health.

WSPeer has a binding to P2PS enabling Web Services to be hosted within decentralised environments and the *Alchemist* builds on this to provide dynamic overlays of peers (like super-peers) that are capable of caching application-specific data (i.e. not just the conventional discovery information) such as metadata or even science data. In *Alchemist* such overlays can be created on-the-fly for the application at hand and deployed onto the peers on the network through the use of P2P groups, used to specify security for participating overlay (e.g. secure authentication, altruistic or reward-based participation, etc). This is necessary for some applications that have sensitive data but wish to employ the use a data-caching overlay of untrusted peers for scalability; the usefulness of such an approach has already been demonstrated through simulations of cycle-scavenging applications, such as BOINC [9]. Here, we are using these mechanisms dynamically propagate retinopathy package workflows (stored as jar files) onto the network using an overlay of package repositories. This scenario is illustrated in our existing DART project, shown in figure 5, illustrating how workflow packages can be propagated through a decentralised layer of package repository cachers (created by *Alchemist's dynamic overlay mechanism*). In DART, users can then query the super peers for music recommendations, being more sophisticated than super peers since they not only cache information but are capable of comparing it too. *Alchemist* makes it very easy for these roles to be defined.

Such a scenario can be adapted easily for retinopathy research and Healthcare@Home, where each Healthcare@Home node propagates information to the network of retinopathy super peers, then can aggregate this information with other available information in order to create higher quality results and healthcare predictions. *Alchemist* can contribute to providing innovative search mechanisms; for example, a multi-modality metadata-assisted search may enable a pathologist to link observations at the level of gross retinal pathology (e.g. vessel bleeds), to physiological dysfunction (e.g. lack of retinal responsiveness), to cellular dysfunction (e.g. hyper-secretory dysfunction), to molecular dysfunction (e.g. constitutive receptor activation), to a genetic

[5] http://www.trianacode.org/

[6] See http://www.trianacode.org/collaborations/index.html for a complete list

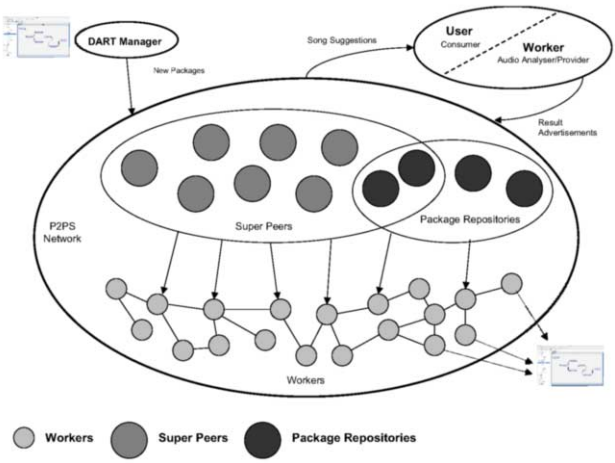


Figure 5. A high level overview of DART and its use of the *Alchemist*, showing the connectivity of its peers.

cause (a specific mutation in the germ-line). In this example, a biomarker that detects the molecular or genetic level could give an unambiguous disease-intrinsic marker that can be applied in future screenings. This benefits the vertical data integration discussed in 2.3, which refers to data representation at multiple levels that might in combination explain origins of pathology (e.g. clinical, functional, morphological, molecular, cellular, genomic). These explanatory combinations may be represented in workflows by analytical algorithms within data discovery workflows or in decision support tools that allow users to apply discovered knowledge. Patterns and signals from combinations of observational modalities (spectral, electrophysiological, retinopathy gradings and annotations/metadata from expert assessments) can help predict what is commonly observed - at the macroscopic level - as a pathological outcome. The high dimensional searches enabled by *Alchemist* also have clear potential here for novel biomarker discovery discussed in 2.3.2 - i.e. identification of signals, technical effects, risk factors or combinations that directly underlie any observation (either in non-exceptional - i.e. normal - functions or in exceptional - i.e. pathological - functions).

By integrating applications, data providers, digital content, and algorithms, the *Alchemist* toolkit enables the simple composition of mixed-media queries for combinational searches to interpret heterogeneous data sets in a logically defined order. Fusing metadata from distinct, yet related, sources into a contextually aware environment allows *Alchemist* to multiplex search results and produce rich metadata that goes beyond what can be traditionally harvested from a single object. At the core, *Alchemist* provides application developers with a framework for specifying complex search algorithms, using a series of logical search steps, within a graphical workflow builder. By building on *Alchemist*'s extensible architecture, developers do not need to write custom software algorithms from scratch, and are able to create complex queries and data fusion techniques in a modular and pluggable fashion.

4. Conclusion

We discuss the *Alchemist* framework, which provides a novel search paradigm based on workflows that can be applied to many application domains for the indexing and search-

ing of digital content. The *Alchemist*, unlike traditional search mechanisms, is based on distributed peers rather than a centralised database. Peers in the network act together to form a decentralised database, where metadata generation and indexing is performed by those peers, avoiding traditional processing bottlenecks. Using the use case scenario for biomedical images and spectral data discovery in diabetic retinopathy, we have described how the *Alchemist* can add value to this community. Annotations added by clinicians or researchers can be used later in complex searches combined with more typical indexing data to produce more detailed and finer grained search results. Annotations by domain experts add value to the stored data, allowing incongruous results to be discarded and non-obvious results to be included. Built on established industry standards such as Web services and SOAP messaging, the *Alchemist* framework and tools are both extensible and interoperable. Reliance on a distributed database and remote processing removes some of the traditional bottlenecks associated with large centralised services or clusters of services.

5. Acknowledgements

We would like to thank our sponsors, PPARC, EU and EPSRC for supporting this *Alchemist* effort through the GridOneD, GridLab and EP/C006291/1 project, respectively. We would also like to thank the researchers and programmers (Eddie Al-Shakarchi, Andrew Harrison, Ian Kelley, Matthew Shields and Ian Wang) at Cardiff for realising these software components. We also thank Healthcare@Home participants at Cardiff University, Smart Holograms, Zarlink Semiconductors, IBM and Llandough Hospital.

References

[1] M. Shields and I. Taylor, "Alchemist: User Driven Searching in Ubiquitous Networks," in *Proceedings of (ADPUC'06) Advanced Data Processing in Ubiquitous Computing*, ACM Press, New York, 2006.

[2] A. Harrison and I. Taylor, "Wspeer - an interface to web service hosting and invocation," in *HIPS Joint Workshop on High-Performance Grid Computing and High-Level Parallel Programming Models*, 2005.

[3] M. Subramanian, A. Shaikhali, O. Rana, A. Hardisty, and E. Conley, "Healthcare@Home: Research Models for Patient-Centred Healthcare Services," in *IEEE 2006 JVA International Symposium on Modern Computing, Sofia, Bulgaria*, IEEE, 2006.

[4] D. A. Antonetti, A. J. Barber, S. K. Bronson, W. M. Freeman, T. W.Gardner, L. S. Jefferson, M. Kester, S. R. Kimball, J. K. Krady, K. F.LaNoue, C. C. Norbury, P. G. Quinn, L. Sandirasegarane, and I. A.Simpson, "Perspectives in Diabetes, Diabetic Retinopathy, Seeing Beyond Glucose-Induced Microvascular Disease," *Diabetes*, vol. 55, pp. 2401–2411, 2006.

[5] M. Larsen, J. Godt, N. Larsen, H. Lund-Andersen, A. K. S. lie, E. Agardh, H. Kalm, M. Grunkin, and D. R. Owens, "Automated Detection of Fundus Photographic Red Lesions in Diabetic Retinopathy," *IOVS, Investigative Ophthalmology and Visual Science*, vol. 44, pp. 761–766, February 2003.

[6] J. M. Wild, "Short wavelength automated perimetry," *Acta Ophthalmologica Scandinavica*, vol. 79, p. 546, 2001.

[7] I. Taylor, "Triana Generations," in *Scientific Workflows and Business workflow standards in e-Science in conjunction with Second IEEE International Conference on e-Science*, (Amsterdam, Netherlands), December 2-4 2006.

[8] I. Taylor, M. Shields, I. Wang, and O. Rana, "Triana Applications within Grid Computing and Peer to Peer Environments," *Journal of Grid Computing*, vol. 1, no. 2, pp. 199–217, 2003.

[9] D. T. Pasquale Cozza, Carlo Mastroianni and I. Taylor, "A Super-Peer Protocol for Multiple Job Submission on a Grid," in *Proceedings of the CoreGRID Workshop on Grid Middleware Workshop in conjunction with Euro-Par, Dresden. To be Published.*, Springer, August 28-29 2006.

VIII. Posters

This page intentionally left blank

From Genes to Personalized HealthCare: Grid Solutions for the Life Sciences
N. Jacq et al. (Eds.)
IOS Press, 2007

Tissue MicroArray: a Distributed Grid Approach for Image Analysis

Federica VITI[a], Ivan MERELLI[a,1], Antonella GALIZIA[b], Daniele D'AGOSTINO[b], Andrea CLEMATIS[b] and Luciano MILANESI[a]
[a] *Institute for Biomedical Technologies, National Research Council, Segrate (Milan), Italy*
[b] *Institute for Applied Mathematics and Information Technology, National Research Council, Genoa, Italy*

Abstract. The Tissue MicroArray (TMA) technique is assuming even more importance. Digital images acquisition becomes fundamental to provide an automatic system for subsequent analysis. The accuracy of the results depends on the image resolution, which has to be very high in order to provide as many details as possible. Lossless formats are more suitable to bring information, but data file size become a critical factor researchers have to deal with. This affects not only storage methods but also computing times and performances. Pathologists and researchers who work with biological tissues, in particular with the TMA technique, need to consider a large number of case studies to formulate and validate their hypotheses. It is clear the importance of image sharing between different institutes worldwide to increase the amount of interesting data to work with. In this context, preserving the security of sensitive data is a fundamental issue. In most of the cases copying patient data in places different from the original database is forbidden by the owner institutes. Storage, computing and security are key problems of TMA methodology. In our system we tackle all these aspects using the EGEE (Enabling Grids for E-sciencE) Grid infrastructure. The Grid platform provides good storage, performance in image processing and safety of sensitive patient information: this architecture offers hundreds of Storage and Computing Elements and enables users to handle images without copying them to physical disks other than where they have been archived by the owner, giving back to end-users only the processed anonymous images. The efficiency of the TMA analysis process is obtained implementing algorithms based on functions provided by the Parallel IMAge processing Genoa Library (PIMA(GE)^2 Lib). The acquisition of remotely distributed TMA images is made using specialized I/O functions based on the Grid File Access Library (GFAL) API. In our opinion this approach may represent important contribution to tele-pathology development.

Keywords. Tissue Microarray, Grid platform, image analysis

---

[1] Corresponding Author: Ivan Merelli; E-mail: ivan.merelli@itb.cnr.it

Introduction

In recent years, microarray technology has increased its potential and now it produces a large amount of genomic, transcriptomic and proteomic data. The data become interesting and informative only if they are screened, filtered, statistically analyzed, correlated to previously existing information and, above all, scientifically validated, to produce accurate and biological consistent predictions.

Gene expression microarrays are used, for example, to determine which genes are differentially expressed in a pathological tissue and in a healthy one. They rely on DNA/oligonucleotides probes, synthesized or spotted on a glass coated chip, which are hybridized with retro transcribed RNA from different cells. This is a useful high throughput technique to deliver a first large-scale screening of the transcriptomic products of cells. The output is a large amount of gene data, which must be analyzed to identify the most representatives transcripts for the cell conditions: only these genes will be valuated and studied.

The Tissue MicroArray technique represents a good validation of these selected data [1]. In fact, while microarrays use tens of thousands of probes as input, and give back only some tens of gene sequences (those expressed) as output [2], TMA can take hundreds of tissues as input to evaluate a single biological entity in a parallel way, concerning many tissues of the same paraffin block [3]. Researchers can check gene expression microarray output using probes to highlight results directly on tissues, validating the different quantity of genes or proteins in case-control samples. This technique allows to analyse at once many tissues, with a decrease in costs and time and with a statistical enrichment of biological profiles. Accurate analysis of the biological structure present in each sample of TMA can be obtained using different molecular biology techniques: most common analyses are IHC (immunohistochemistry), ISH-RNA and ISH-DNA (in situ hybridization for RNA and DNA), and FISH (fluorescent in situ hybridization).

The combinatorial explosion in analyzing data from TMA experiments places a strain on computational and storage resources. From each TMA block we can process slides with different techniques and handle a large number of high resolution images. Moreover, considering that a TMA experiment can include hundreds of tissue samples it is clear the importance of an infrastructure not only to assure patient data privacy, but also as a resource for storage and computing to process images. Grid platforms could solve these problems by providing storage facilities and high computational power under a solid security architecture.

At present, pathology institutes store images and associated metadata - clinical situation, sample treatment information, paraffin block creation and slide reaction type - relating them to patients through anonymous identifiers. This is a good method for a local storage, but is not safe for sharing data with the scientific community. The Grid middleware enables the development of a secure method of storage and analysis in order to promote TMA images and metadata sharing.

In this paper we present a new approach for processing remote TMA images using the Enabling Grids for E-sciencE (EGEE) Grid infrastructure. The image analysis is performed using the Parallel IMAGE processing GEnoa Library (PIMA(GE)^2 Lib). The

privacy of the sensible data is ensured through the use of specialized I/O functions based on the Grid File Access Library (GFAL) API, that is provided by the EGEE framework.

1. The Grid platform

In this work we considered the use of the Enabling Grids for E-sciencE (EGEE) Grid infrastructure, a wide area platform for scientific applications that relies on the Globus Toolkit, as middleware, which operates as an ideal communication layer between the different Grid components. Let us provide a short overview of EGEE and its way of managing data.

1.1. The EGEE distributed environment

This Grid infrastructure is a network of several Computing Elements (CE), that are gateways for the Worker Nodes (WN) on which jobs are performed, and a number of Storage Elements (SE) on which the data are stored. Through the User Interface (UI) it is possible to submit jobs, monitor their status and retrieve the outputs if jobs are terminated successfully or resubmit them in case of failure.

Due to the use of remote computational resources, the Grid communication software must offer an efficient security system. Security is guaranteed by Grid Security Infrastructure (GSI) which uses public key cryptography (asymmetric cryptography) to recognize users. GSI certificates are encoded in the X.509 format, a standard established by the Internet Engineering Task Force (IETF) [4], and accompany each job to authenticate the user. Moreover, in order to submit jobs, users must be members of a Virtual Organization (VO), a group of Grid users with similar interests and requirements who are able to work collaboratively and share resources (data, software, etc.), regardless of geographical location. This is another aspect of Grid supervision because users request accounting must be authorized by the VO manager.

The distribution of the computational load is performed by the Resource Broker (RB) delegated for routing each job on an the available CEs. Jobs are submitted from a UI using the Job Description Language (JDL) scripts, which specify the necessary directives to perform the task. Important features specified by JDL are the files to be accessed on the SE, the data that have to be exchanged between the UI and the WN and the software requirements for the job execution.

1.2. The data and metadata handlers

The EGEE middleware provides a set of tools in order to manage data in a remotely distributed way. They provide accessibility to the physical location of the files and to the LCG File Catalog (LFC) system used to associate a physical identifier to the Logical File Name (LFN). The LFC contains a GUID (Globally Unique Identifier) as an identifier for a physical file and combines logical and physical mappings for the file in the same

database. To upload a file on an SE users must specify the hostname of the Grid node where they want to store images: once the file has been uploaded the server gives back an id that univocally corresponds to the file. At the same time the file is registered in the LFC in order to be accessible through its LFN. Both, the GUID and the correspondent LFN can be used to access data. Uploaded files are visible from everyone belonging to the same VO, even if the owner can manage permissions, avoiding the access by non-authorized users.

A very important feature is that in the EGEE Grid it is possible to handle the access to files archived on the SEs by performing computations without copying data on the WNs. This is possible using GFAL API, a library available with interface for different programming languages (C, Java, etc.) that allows to access files which are located in remote SEs, anywhere in the Grid. In this way we avoid to have multiple physical copies of raw data without compromising the security level of sensitive data.

Presently we disregard all the metadata information. We plan to manage the metadata using the AMGA system [5]. It is based on a client/server architecture, and it can be easily integrated into the Grid environment. The metadata are recorded in an AMGA table in relation to the file's GUID. In the same way as the other Grid components, the AMGA catalogue preserves data security because it supports VO authentication.

2. The use of PIMA(GE)^2 Lib

We implemented the image processing algorithms using the functions of PIMA(GE)^2 Lib [6] library functions. This library has been developed in order to satisfy the scientist's requirements in the image processing field, ensuring robust and high performance execution. It implements the most common image processing operations, according to the classification provided in Image Algebra [7].

The operations provided by the PIMA(GE)^2 Lib have been grouped in 'conceptual objects', that are not intended to prescribe how an operation is performed but to underline the operation similarity and to help in the definition of an effective and efficient interface. Furthermore, even from the user's point of view, the conceptual objects allow an easier management of the library operations, since the user is no longer involved with a large number of functions, but he/she has to consider and handle a small set of well-defined objects. Operations are grouped together according to different rules, such as the nature similarity, or the data structures processed: they are collected together following 'algorithmic pattern'. The single image processing operation can be called by instantiating algorithmic pattern with the proper parameters, including the function to be applied to the data elements. Following this approach the core of PIMA(GE)^2 Lib is represented by a set of eight objects:

- I/O Operations: operations that perform the I/O and memory management;
- Point Operations: operations that elaborate one image applying a unary function, i.e. square root, absolute value, sine;
- Image Arithmetic: operations that elaborate two image, i.e. addition, product, etc;

- Reduce Operations: operations that elaborate one image calculating a collective value, i.e. the maximum, minimum pixel value of an image;
- Geometric Operations: operations that elaborate one image applying a Region of Interest (ROI) or a domain function, i.e. rotation, translation and scaling;
- Neighbourhood Operations: operations that elaborate one image applying a convolution function that involves a neighbourhood of each pixel: percentile, median;
- Morphology Operations: operations that elaborate one image applying a convolution function that combines together two arithmetic functions: Gaussian convolution, erosion, dilation;
- Differential Operations: operations that elaborate one image and perform differential operators, i.e. Hessian, gradient, Laplacian.

The PIMA(GE)^2 Lib operations can be performed both in a sequential and in data parallel fashion. Till now we considered only a sequential use of the library operations, however we plan to exploit the EGEE resources for parallel executions, like MPI, when considering more time consuming implementations.

In this work, we added specialized functions to the object of I/O operations in order to ensure the privacy preservation. It has been shown in the literature [8] that GFAL API for the Grid platform provides a secure system for image manipulation. We moved the library software to the Grid environment acquiring properties of distribution, extensibility and dynamicity. Once data have been accessed they can be processed as usual by PIMA(GE)^2 Lib operations.

3. Testing our approach

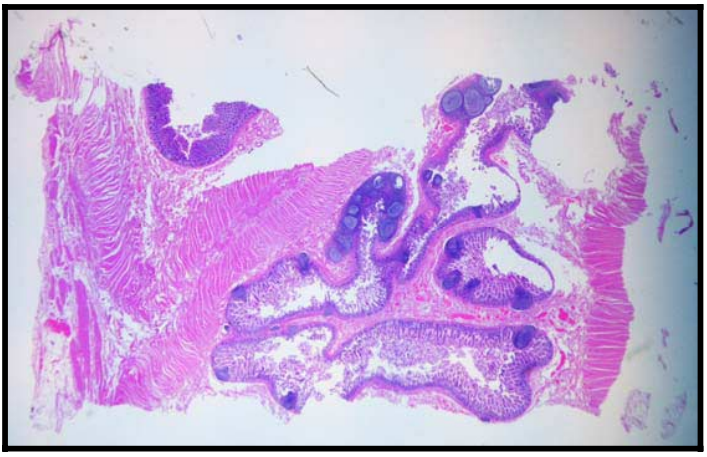
We used a modified version of PIMA(GE)^2 Lib to perform algorithms of edges detection and segmentation of images stored on Grid SE. The implemented code uses GFAL API to open images in tiff and bmp format: in fact a standard format for saving TMA images does not exist at the moment, but these two packaging methods ensure that the image are stored in a lossless form. Both image management and analysis submission are possible from the UI. The JDL scripts describe how the job has to be computed on the remote WN, copying the image to its memory buffer, processing data using the selected algorithm and retrieving the output on the UI.

As a test case we stored 10 images in a set of SEs. The algorithm we implemented accesses the data distributed on the Grid, acquires one image at a time using the new I/O functions, processes it on a remote WN, then produces the resulting image, a masked image from which it is not possible to reconstruct original data, to ensure patient privacy, and which is more informative than the raw images concerning pathology.

An example of our analysis is shown in Figure 1. Figure 1.A is the input image and it represents an haematoxylin-eosin reaction: this coloration differentiates basic and acid entities in the tissue slice. Basic elements (like cytoplasm) are highlighted in a red shade

while acid elements (like nuclei) are in blue. Figure 1.B represents the result of a first attempt of the edge detection algorithm.

A.



B.

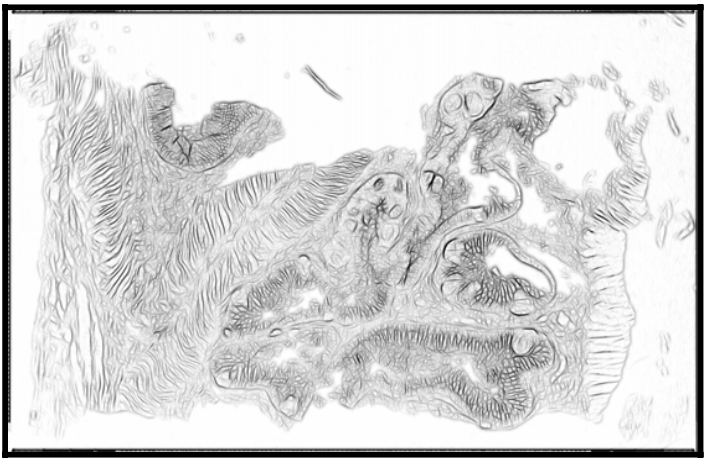


Figure 1. Input (A) and output (B) of remote image analysis using the edge detection algorithm

We are also developing functions which work on thresholds and segmentations. PIMA(GE)^2 Lib provide interesting operations to help detection of cellular structures and tissue architectures, i.e. Morphology and Differential objects.

4. Conclusions

In this paper we presented an approach for processing remote TMA images in the EGEE environment. We enabled the use of PIMA(GE)^2 Lib in the EGEE Grid by implementing a set of specialized functions in order to access distributed medical images without moving them and preserving the privacy of sensible data. In particular we implemented an edge extraction algorithm.

Up to now images have not been associated with any supplementary information and analyses have been performed on manually chosen images. Without the associated metadata images become meaningless. Users should have the possibility of choosing images to work with according to their associated information. The community of pathologists (API, Association of Pathology Informatics) [9] developed an XML language to facilitate TMA data exchange [10] and we are creating a metadata classification compliant with this acquired specification.

Positive test results about the feasibility of our work brings us to consider future developments including features to automatic data access using metadata and the implementation of more informative image elaboration algorithms.

Acknowledgments

Thanks are due to the Italian FIRB project LITBIO (Laboratory for Interdisciplinary Technologies in BIOinformatics), (http://www.litbio.org).
Thanks also to EU BioinfoGRID no.:026808 project.
A special thanks to BioLab Laboratory of the Department of Computer Science, Control Systems and Telecommunications of University of Genoa.

References

[1] S. Mousses, L. Bubendorf, U. Wagner, G. Hostetter, J. Kononen, R. Cornelison, N. Goldberger, A.G. Elkahloun, N. Willi, P. Koivisto, W. Ferhle, M. Raffeld, G. Sauter, O.P. Kallioniemi, Clinical validation of candidate genes associated with prostate cancer progression in the CWR22 model system using tissue microarrays, *Cancer Research* **62** (2002), 1256-1260.

[2] M. Schena, D. Shalon, R.W. Davis, P.O. Brown, Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science* **270(5235)** (1995), 467-470.

[3] J. Kononen, L. Bubendorf, A. Kallioniemi, M. Barlund, P. Schraml, S. Leighton, J. Torhorst, M.J. Mihatsch, G. Sauter, O.P. Kallioniemi, Tissue microarrays for high-hroughput molecular pro_ling of tumor Specimens, Nature Medicine **4** (1998), 844-847.

[4] http://www.globus.org/security/overview.html

[5] N. Santos, B. Koblitz, Metadata services on the grid, *Proc. of Advanced Computing and Analysis Techniques*, ACAT'05, Zeuthen, Berlin, May 2005.
[6] A. Clematis, D. D'Agostino, A. Galizia, The Parallel IMAGE GEnoa processing Library: PIMA(GE)^2 Lib, *Technical Report,* IMATI-CNR-Ge, N.22/2006.
[7] G. Ritter, J.Wilson, Handbook of Computer Vision Algorithms in Image Algebra, *CRC Press*, 2nd edition, Inc, 2001.
[8] S. Bagnasco, F. Beltrame, B. Canesi, I. Castiglioni, P. Cerello, S.C. Cheran, M.C. Gilardi, E. Lopez Torres, E. Molinari, A. Schenone, L. Torterolo, Early Diagnosis of Alzheimer's Disease Using a Grid Implementation of Statistical Parametric Mapping Analysis, *HealthGRID 2006*, Valencia, June 2006.
[9] http://www.pathologyinformatics.org/
[10] J.J. Berman, M.E. Edgerton, B.A. Friedman, The tissue microarray data exchange specification: A community-based, open source tool for sharing tissue microarray data, *BMC Medical Informatics and Decision Making* **3(5)** (2003).

From Genes to Personalized HealthCare: Grid Solutions for the Life Sciences
N. Jacq et al. (Eds.)
IOS Press, 2007

A Grid Framework for Non-Linear Brain fMRI Analysis

Rodolfo ANDRADE[a,1], Ilídio OLIVEIRA[a,b], José Maria FERNANDES[a,b], João Paulo Silva CUNHA[a,b]
[a] *IEETA – Universidade de Aveiro, Portugal*
[b] *Dep. of Electronics, Telecommunications and Informatics, Universidade de Aveiro, Portugal*

Abstract. Functional magnetic resonance imaging (fMRI) is an imaging technique that can be used to characterize brain physiological activity, usually presented as 3D volumes in function of time. In the context of our previous work in nonlinear association studies in electroencephalogram (EEG) time series, we were able to identify clinical relevant features useful in clinical diagnosis. The use of a similar approach in fMRI, now adapted for 3D time series, is both appealing and new. Such time series analysis imposes challenging requirements regarding computational power and medical image management. In this paper we propose a grid architecture framework to support the typical analysis protocol of association analysis applied to fMRI. The system, implemented using the gLite middleware, provides the necessary support to manage brain images and run non-linear fMRI analysis methods.

Keywords. Non-linear fMRI analysis methods; brain function; medical imaging; Grid computing.

Introduction

fMRI is a relatively new MRI based technique that allows monitoring of the brain activation patterns by measuring the magnetic variation induced by changes in the blood flow associated to brain neural activity. The neural activity produces an hemodynamic response —an increase in blood flow, with a delay of about two seconds— richer in oxyhemoglobin to compensate the increase in oxygen consumption. This change in oxyhemoglobin is called the Blood Oxygen Level Dependent response or BOLD effect [1]. When BOLD activations and deactivations are time related with specific events, they can be correlated to the event's metabolic response in the brain [2]. This is a valuable tool for studies ranging from brain diseases (e.g. Alzheimer [3], Parkinson [4]) to study normal brain function in aging [5].

One major problem of this technique is that BOLD changes are small, in the order of 3-5% from the background MRI imaging signal [6]. A methodological problem with fMRI analysis is that there is no unique model for the hemodynamic response function (HRF) [7]. This may compromise the fMRI interpretation since it depends on the HRF activation model used.

[1] Corresponding author: Rodolfo Andrade, IEETA – Universidade de Aveiro; E-mail: randrade@ieee.org.

The traditional approach to analyze fMRI series is based on statistical parametric mapping (SPM) [8]. In SPM it is assumed that through structured paradigms with clearly distinguishable states (usually activation/no activation paradigms related with the studied brain function) it is possible to infer related brain areas to BOLD changes in fMRI through general linear models (GLM) [9]. This is done by finding fMRI voxels which are statistically and linearly correlated with the paradigm activation model in time. One example among several, is given by Powell et al. [10] in which SPM is used to infer memory related brain areas.

The major drawback of SPM is that it assumes and uses unproven response models and linear assumptions in the analysis, in which no model clearly relates studied tasks and BOLD response in the brain [7]. This is worsen by the fact that BOLD response has nonlinear contributions that should be also taken into account [11]. For these reason, SPM fMRI analysis is dependent on several factors to obtain statically significant results [12] and demands expertise to extract relevant information from the analysis results [13]. In this context, non-linear models, parametric and non-parametric, are attractive approaches that deserve attention.

Following our previous work in nonlinear association studies in electroencephalogram (EEG) time series [14-16], where we were able to identify clinical relevant features useful in clinical diagnosis, the use of a similar approach in fMRI, now adapted for 3D time series, is both appealing and a novelty.

Such computer-based analysis imposes high-requirements with respect to medical image management (raw data and results) and processing times, as detailed in the next sections.

In this paper we propose a grid architecture framework to support the typical analysis protocol of association applied to fMRI. The system, implemented using the gLite middleware, provides the necessary support to manage brain images and run non-linear fMRI analysis methods.

1. Background on fMRI Time Series Association Analysis

The basic procedure for association analysis between time series consists in determining the association level of two time series using an association measure. In case of this association measure being non-linear we are performing a non-linear association analysis. The extension to 3D fMRI time series analysis is straight forward by performing a voxel time series pair-wise association analysis. The drawback is that, when considering 3D volumes in time, the number of pairs to analyze increases, increasing both the computational complexity of the overall process and the management of the results.

To illustrate this fact, consider a sequence of 5 minutes of fMRI with a volume acquisition of 16 64x64 slices acquired from top to bottom of the head at each 3 seconds where the final result consists of 200 volumes of 64x64x16 voxels. Performing association analysis as described would imply:

- (64x64x16)x(64x64x16) volumes, in the worst case, where the association is not symmetric (e.g. using the h^2 association measure [14]) and for each voxel a volume is generated containing the association coefficient with all other voxels.
- The overall computational complexity would be at least supra linear (several association measures have supra linear complexity).

- In addition, considering time delay analysis would increase any of the previously referred issues.

2. The BImageG Grid Framework

2.1. Overall System Architecture

Grid infrastructures are being successfully used in medical image processing to handle the demanding requirements of large images storage and communication, and to enable complex analysis workflows [17-19]. They provide the ability to seamless aggregate distributed computational power, extensive storage resources and high-bandwidth networking. In addition, Grids also ensure a proper level of security, both at identity (digital certificates) and access (Virtual Organizations management) levels. Building upon the state of the art Grid middleware, we propose a Grid enabled framework to provide the computational power and data access needed to run non-linear analysis applied to fMRI (Figure 1).

The association analysis of fMRI can be applied to different parts of the 3D volume independently. This enables a natural parallelization of the association analysis protocol. In a grid environment, the analysis of each volume partition can map to the creation of one job.

Users harness from the computing resources pool using a web portal. In this portal, the user (typically a biomedical engineering researcher) is able to select the intended fMRI analysis protocol, though currently only the non-linear time series association is implemented.

The portal delegates the user requests in the BImageG application layer, which provides the necessary services to generate grid job workflows (according to the fMRI research protocol chosen by the user), suitable to the grid environment, allowing a

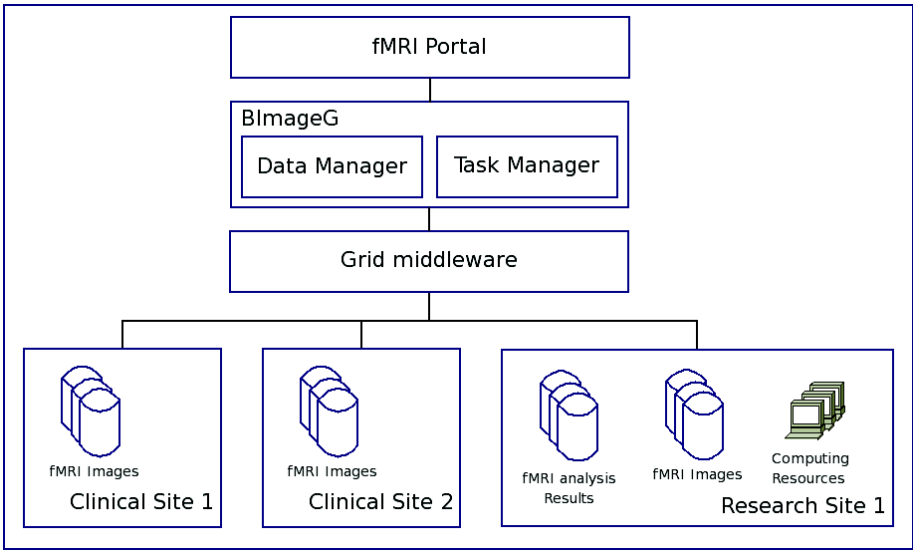


Figure 1. fMRI analysis framework, building on top of Grid services.

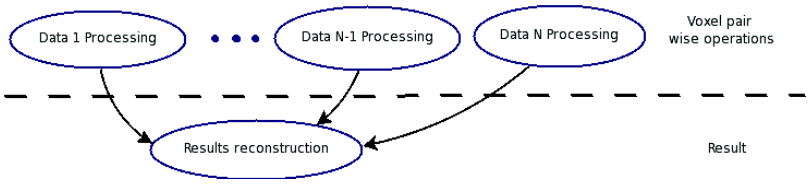


Figure 2. Grid job workflow of an fMRI association analysis method.

efficient use of the resources available (Figure 2). In our system, BImageG stands for brain imaging grid services.

The BImageG layer also provides grid-aware data management services for fMRI, such as volume or time-series extraction and normalization. Data management is being deployed at a single research site but, in the future, BImageG will be used to securely access fMRI repositories from several sites. Besides integrating the "source" fMRI image repositories, the system will rely on additional repositories, distributed among different sites, to store fMRI analysis results and temporary data that can be generated in the course of analysis workflows.

2.2. Supporting Research Workflows on BImageG

The BImageG holds the two main system components: the Task manager and the Data manager.

The Task Manager generates the jobs that will run in the grid and then implements the specific job workflow, submitting each job at its due time. This involves the jobs state continuous monitoring and a correct management of the data generated during the job workflow. In the association analysis procedure we divide the data and then apply the process (e.g.: a voxel pair wise association) to each segment of data. Each one of these tasks is a job that will be run in the grid. The partition of data and generation of jobs takes into account the availability of computing resources in the grid (Figure 2).

The research workflow is often pre-defined. Nevertheless, we provide a flexible

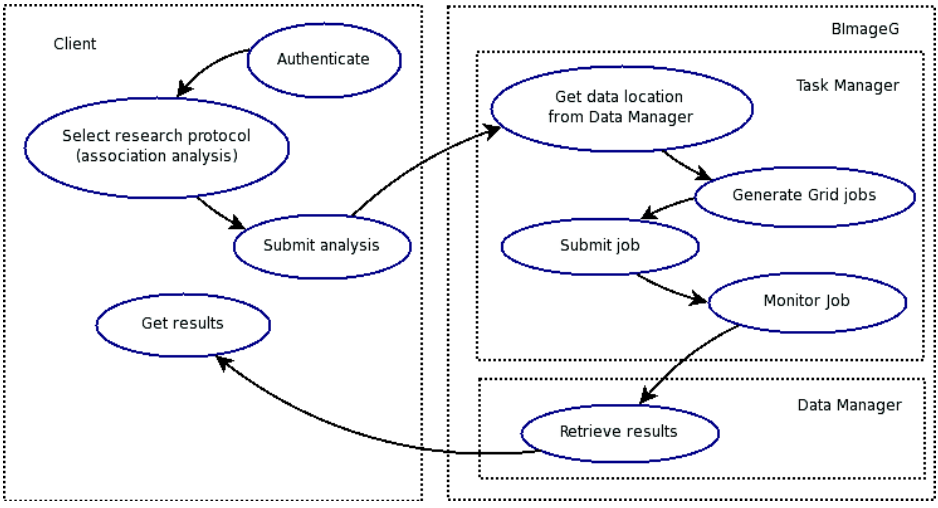


Figure 3. Typical workflow when running an fMRI association analysis.

approach for researchers to create and submit their own custom research workflows. To accomplish this, the fMRI portal allows the user to select a research problem specification, by importing a XML file which describes the intended job workflow. Later, the task manager can interpret the XML specification and generate the Grid job workflow that will be run on the target execution environment. When the job workflow is complete, the results are retrieved and made available to the user (Figure 3).

The Data Manager will be responsible for the fMRI specific data operations, such as normalization, fMRI 3D volumes data extraction, data conversion and retrieval.

BImageG is currently built on top of gLite[2] middleware, using the Java APIs available for data and job management, running on the pan-European EGEE pre-production Grid infrastructure.

gLite is continuously evolving and integrating new services; in the biomedical area, a Medical Data Manager that provides anonymization and security necessary to work with medical data and the integration with DICOM servers is under development [20]. This could be a major development to our objectives since image repositories in clinical sites are usually available through DICOM servers.

3. Preliminary Results

We have implemented the services necessary to test the protocol of association analysis over gLite middleware. In this pilot experiment, we use a fMRI sequence of an epileptic patient to calculate voxel pair wise operations in a fMRI. The sequence is composed of 3D volumes of 16 64x64 slices acquired from top to bottom of head at each 3 sec, for 5 minutes.

The use of the framework is illustrated for a specific brain area (Figure 4), represented by a particular voxel (a). The result of the association analysis in relation to a specific voxel (b) is a volume. Here we present two axial views where is possible to visualize clearly spatial areas that present positive correlation (warm colors) and negative correlation (blue tones). These correlations can be related used to find time dependences between a specific brain area and other brain regions. The next step is assessing the value of these findings within clinical models, supported in fMRI acquisition protocol design.

Input **Results**

a) a specific brain area as a voxel b) Correlation generated by BImageG.

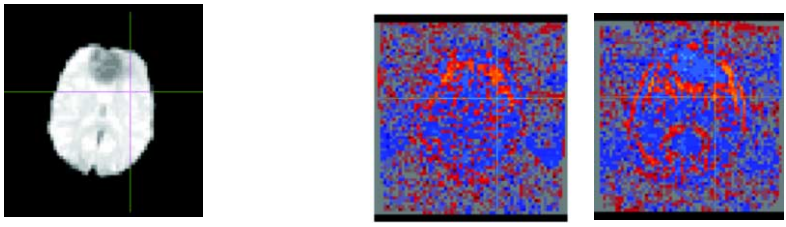


Figure 4. fMRI analysis results: the correlation between brain areas, represented as association maps (b), and specific brain areas, represented in the example by a localized voxel (a), can be used to infer time dependencies between different brain areas. These dependencies can be later interpreted within clinical models. In (b) warm colors represent positive correlations while blue tones negative.

[2] http://glite.web.cern.ch/glite/

4. Discussion and Conclusions

Other projects have been successfully using the Grid to analyze fMRI data. The Virtual Laboratory for e-Science [21] project is working in a infrastructure to facilitate the management and analysis of fMRI data and is using the LCG-2 grid middleware. Other preeminent example is the north-American Biomedical Informatics Research Network (BIRN) [22] which deploys a specific testbed on fMRI, the FBIRN, working on the development of methods to analyze multi-site fMRI data.

We implemented a framework to support the association analysis and illustrated its use over a fMRI sequence. While this target has been achieved, other issues are under development, such as the summarizing and analysis of the results. Both issues have efficiency implications at grid level: the summarizing of results may imply further computational processing to provide friendly human readable results.

For that reason the current framework was centered in the support to the association calculations and not on post processing, although it is able to handle – not in the most efficient way – with the overwhelming size of result volumes.

The infrastructure created was able to run the fMRI association analysis using gLite middleware. Further studies must be accomplished to evaluate the most efficient way of generating the grid job workflows in order to minimize the grid middleware overhead. One possible way of increasing the efficiency of this infrastructure would be to transfer services provided by the data manager in the BImageG to the repositories' sites (move computation to data). This would decrease the transfer time of data over the grid.

References

[1] S. Ogawa, T. M. Lee, A. R. Kay, and D. W. Tank, "Brain magnetic resonance imaging with contrast dependent on blood oxygenation," *Proc. Natl. Acad. Sci. USA,* vol. 87, pp. 9868-9872, 1990.

[2] K. Uludag, D. J. Dubowitz, E. J. Yoder, K. Restom, T. T. Liu, and R. B. Buxton, "Coupling of cerebral blood flow and oxygen consumption during physiological activation and deactivation measured with fMRI," *Neuroimage,* vol. 23, pp. 148-55., 2004.

[3] S. A. Rombouts, R. Goekoop, C. J. Stam, F. Barkhof, and P. Scheltens, "Delayed rather than decreased BOLD response as a marker for early Alzheimer's disease," *Neuroimage,* vol. 26, pp. 1078-85, Jul 15 2005.

[4] A. O. Ceballos-Baumann, "Functional imaging in Parkinson's disease: activation studies with PET, fMRI and SPECT," *J Neurol,* vol. 250 Suppl 1, pp. I15-23, Feb 2003.

[5] M. D'Esposito, L. Y. Deouell, and A. Gazzaley, "Alterations in the BOLD fMRI signal with ageing and disease: a challenge for neuroimaging," *Nat Rev Neurosci,* vol. 4, pp. 863-72, Nov 2003.

[6] R. Turner and T. Jones, "Techniques for imaging neuroscience," *Br Med Bull.,* vol. 65, pp. 3-20., 2003.

[7] D. A. Handwerker, J. M. Ollinger, and M. D'Esposito, "Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses," *Neuroimage,* vol. 21, pp. 1639-51., 2004.

[8] K. J. Friston, P. Fletcher, O. Josephs, A. Holmes, M. D. Rugg, and R. Turner, "Event-related fMRI: characterizing differential responses," *Neuroimage,* vol. 7, pp. 30-40., 1998.

[9] P. McCullagh and J. A. Nelder *Generalized Linear Models*. London: Chapman and Hall, 1989.

[10] H. W. Powell, M. J. Koepp, M. P. Richardson, M. R. Symms, P. J. Thompson, and J. S. Duncan, "The application of functional MRI of memory in temporal lobe epilepsy: a clinical review," *Epilepsia,* vol. 45, pp. 855-63., 2004.

[11] K. L. Miller, W. M. Luh, T. T. Liu, A. Martinez, T. Obata, E. C. Wong, L. R. Frank, and R. B. Buxton, "Nonlinear temporal dynamics of the cerebral blood flow response," *Hum Brain Mapp,* vol. 13, pp. 1-12, May 2001.

[12] G. K. Aguirre, J. A. Detre, E. Zarahn, and D. C. Alsop, "Experimental design and the relative sensitivity of BOLD and perfusion fMRI," *Neuroimage,* vol. 15, pp. 488-500., 2002.

[13] S. J. Kiebel and K. J. Friston, "Statistical parametric mapping for event-related potentials (II): a hierarchical temporal model," *Neuroimage.,* vol. 22, pp. 503-20., 2004.
[14] J. P. Cunha and P. G. de Oliveira, "A new and fast nonlinear method for association analysis of biosignals," *IEEE Trans Biomed Eng,* vol. 47, pp. 757-63, Jun 2000.
[15] J. P. S. Cunha, P. Guedes de Oliveira, J. Ramalheira , J. Lopes, and A. Martins da Silva, "Is Background EEG Analysis Relevant on Predicting Epilepsy Evolution?," in *8th European Congress of Clinical Neurophysiology*, Munich, Germany, 1996.
[16] A. Martins da Silva and J. P. S. Cunha, "Scalp EEG Recordings: Inter Ictal / Ictal Location and Spreading of epileptiform events," *Acta Neurol Scand,* vol. suppl 152, pp. 17-19, 1994.
[17] V. Breton, K. Dean, T. Solomonides, I. Blanquer, V. Hernandez, E. Medico, N. Maglaveras, S. Benkner, G. Lonsdale, S. Lloyd, K. Hassan, R. McClatchey, S. Miguet, J. Montagnat, X. Pennec, W. De Neve, C. De Wagter, G. Heeren, L. Maigne, K. Nozaki, M. Taillet, H. Bilofsky, R. Ziegler, M. Hoffman, C. Jones, M. Cannataro, P. Veltri, G. Aloisio, S. Fiore, M. Mirto, I. Chouvarda, V. Koutkias, A. Malousi, V. Lopez, I. Oliveira, J. P. Sanchez, F. Martin-Sanchez, G. De Moor, B. Claerhout, and J. A. Herveg, "The Healthgrid White Paper," *Studies in health technology and informatics,* vol. 112, pp. 249-321, 2005.
[18] J. Montagnat, F. Bellet, H. Benoit-Cattin, V. Breton, L. Brunie, H. Duque, Y. Legré, I. E. Magnin, L. Maigne, S. Miguet, J. M. Pierson, L. Seitz, and T. Tweed, "Medical Images Simulation, Storage, and Processing on the European DataGrid Testbed," *Journal of Grid Computing,* vol. 2, p. 387, 2004.
[19] S. R. Amendolia, J. Galvez, D. Manset, W. Hassan, F. Estrella, C. Del Frate, M. Odeh, T. Hauer, and R. McClatchey, "Deployment of a grid-based medical imaging application," *Stud Health Technol Inform,* vol. 112, pp. 59-69, 2005.
[20] J. Montagnat, D. Jouvenot, C. Pera, A. Frohner, P. Kunszt, B. Koblitz, N. Santos, and C. Loomis, "Bridging clinical information systems and grid middleware: a Medical Data Manager," *Studies in health technology and informatics,* vol. 120, pp. 14-24, 2006.
[21] S. D. Olabarriaga, R. G. Belleman, J. G. Snel, and A. J. Nederveen, "Towards a virtual laboratory for FMRI data management and analysis," *Stud Health Technol Inform,* vol. 120, pp. 43-54, 2006.
[22] J. S. Grethe, C. Baru, A. Gupta, M. James, B. Ludaescher, M. E. Martone, P. M. Papadopoulos, S. T. Peltier, A. Rajasekar, S. Santini, I. N. Zaslavsky, and M. H. Ellisman, "Biomedical informatics research network: building a national collaboratory to hasten the derivation of new understanding and treatment of disease," *Studies in health technology and informatics,* vol. 112, pp. 100-109, 2005.

From Genes to Personalized HealthCare: Grid Solutions for the Life Sciences
N. Jacq et al. (Eds.)
IOS Press, 2007

From HealthGrid to SHARE: A Selective Review of Projects

Mark OLIVE, Hanene RAHMOUNI and Tony SOLOMONIDES[1]
CCCS / CEMS Faculty / UWE, Bristol / Coldharbour Lane / Bristol BS16 1QY / UK

Abstract. The SHARE project (Supporting and structuring Healthgrid Activities and Research in Europe) is an EC funded specific support action to define a roadmap for future healthgrid research, highlighting opportunities, obstacles and potential bottlenecks. The aim is identify technical, legal and ethical issues that would affect the wide deployment of healthgrids in the medical research community. The initial technical roadmap has proposed a series of technology, standards and deployment milestones, starting with the testing and development of a reference implementation of grid services, and ending with the deployment of a knowledge grid for medical research.

In this paper we review a number of projects from pre-healthgrid to the second generation of healthgrid projects and retrospectively consider their achievements and issues in the light of the first SHARE technical road map.

Keywords. healthgrid, e-health, grid applications

Introduction

In order to address the key technical issues that would affect the wide deployment of healthgrids for medical research, two technology milestones have been defined by SHARE for the development of grid middleware and services, two more milestones as examples of required grid standards for bioinformatics and medical informatics, and three deployment milestones increasing in complexity and scope.

In order to identify issues that would affect the development and deployment of healthgrids, we have examined a number of key national (UK) and European projects. This included medical applications designed to run on early generic grid environments (or 'pre-healthgrid'), the first true generation of healthgrid projects tackling generic medical problems that presented novel challenges, and the current second generation.

This paper will include technology and development issues identified in open publications from the following projects: GEMSS (Grid-Enabled Medical Simulation Services), CLEF (the Clinical e-Science Framework), ECIT (the ESHRE Classification of Infertility Taskforce), eDiaMoND (Digital Mammography National Database), IBHIS (Integration Broker for Heterogeneous Information Sources), GIMI (Generic Infrastructure for Medical Informatics), WISDOM (Wide In Silico Docking On Malaria), Health-e-Child and Integrative Biology.

This work has been carried out as part of the SHARE project. Work on the technical road map, and in particular the phasing into technology and deployment

[1] Corresponding Author: Tony Solomonides, CCCS, CEMS Faculty, University of the West of England, Coldharbour Lane, Bristol BS16 1QY, United Kingdom; E-mail: Tony.Solomonides@uwe.ac.uk.

milestones, has been led by our colleagues at CNRS-IN2P3. We are especially grateful to Vincent Breton and Yannick Legré at CNRS-IN2P3, Vicente Hernandez and Ignacio Blanquer at UPV, Petra Wilson formerly at EHMA, and Isabelle Andoulsi and Jean Herveg at FUNDP.

1. Milestones

A total of seven technical milestones have been defined for the first roadmap. Three deployment milestones were defined for the creation of demonstration production environments for a computational grid (MD1), a data grid (MD2) and a knowledge grid (MD3) for medical research. Two milestones dealing with specific standardisation efforts for medical imaging (MS1) and electronic health records (MS2) for grids were also defined. There is some overlap between the standardisation milestones and the second deployment milestone of a data grid production environment to represent concurrency and a cyclical relationship.

Additionally, two ongoing technology research efforts were identified and have been added as technical milestones prior to the creation of the first integrated roadmap, which will also incorporate ethical, legal and socio-economic considerations. These were the testing of grid middleware (based on web services, and with sufficient job and data management) with medical applications for robustness, scalability and compliance with EC country laws regarding reliability, recovery and backup (MT1), and the production of a robust, user-friendly and open source distribution of healthgrid services with appropriate user support (MT2).

In what follows, we consider the projects we have studied in depth with reference to these seven milestones.

2. Review of Projects Against Milestones

2.1. Milestone MD1

We begin by considering the first SHARE road map, whose first deployment milestone is the development of a demonstration computing grid production environment for the medical research community. The importance of a robust, well maintained infrastructure was stressed by eDiaMoND, with appropriate redundancy to address issues of availability. This is a necessity for any sustainable grid infrastructure, and so would naturally be a requirement of the first deployment milestone.

The necessity (and cost) of providing a suitable network infrastructure will certainly be an issue for the deployment of data grids in medical research centres. Due to the volume and size of digital mammograms, which must satisfy both legal and practical requirements, eDiaMoND required gigabit Ethernet, expensive monitors and high speed, high capacity storage. There is also the issue of medical data not typically residing in a convenient central location, but distributed among firewall protected hospital databases. Another concern was how hospital executive managers could be convinced to allow external systems to penetrate their firewalls.

The necessity of providing a user friendly interface for non technical users was recognised by a number of the projects examined. In particular, Integrative Biology, GEMSS and WISDOM saw hiding grid complexity and grid mechanisms as important.

However, WISDOM noted that considerable knowledge of grid mechanisms was required in order to quickly resolve unforeseen errors and failures, which would be a concern for projects where jobs are being submitted by users with limited computing skills, as may be the case in medical research centres or hospitals. Additionally, human supervision was considered mandatory, which could suggest that scalability would be an issue.

2.2. Milestone MT1

MT1 is the testing of grid middleware(s) with medical applications for scalability and robustness. A good example of a high throughput computing grid application is WISDOM, which demonstrated the capabilities of the EGEE (Enabling Grids for E-sciencE) infrastructure. However, a number of errors and efficiency issues were identified by both the project and EGEE reports. Data management has been identified as a bottleneck for biomedical data challenges on grid infrastructures that only support one Replica Location Service (RLS), such as the biomedical VO (Virtual Organisation) framework on the EGEE grid at the time of the WISDOM data challenge.

There were significant problems with automatic job resubmission both within WISDOM and the EGEE middleware at the time. WISDOM described this as a 'sink-hole' effect, which led to a significant number of aborted jobs and excessive job execution times. Checking, cancellation and resubmission of jobs had to be performed manually as a result, and automatic job resubmission was still absent from the second (avian flu) WISDOM data challenge.

WISDOM also highlighted a number of issues with grid Information Systems (IS) and Resource Brokers (RB) that will need to be addressed before the reference implementation can be finalised. The IS holds pertinent data to optimise performance and the RB makes decisions on the basis of this information. Ideally, an RB would use the IS data to select the best computing element for any job that arises. But if the IS fails or does not have the required information, the submission of jobs will be inefficient or the user will have to expend significant effort manually correcting the failure.

A number of improvements to the EGEE middleware were proposed, including better configuration/policy discovery by the grid IS, suggestions for improving the reliability of RBs, and better handling of errors/failures. These should be addressed by future versions of the gLite middleware.

2.3. Milestone MT2

MT2 is the production of a reference distribution of healthgrid services, using standard web services technology. An important issue for this milestone will be the security mechanisms employed to protect medical and commercial data in a grid environment.

The earliest project examined, GEMSS, identified internal staff as the main security threat for grid-based healthcare. Through the use of use case scenarios, the eDiaMoND project went on to identify five different classes of user that represented a potential threat:

- Insiders making innocent mistakes, causing the accidental disclosure of confidential information
- Insiders who abuse access privileges

- Insiders who knowingly access information through spite or for profit
- An unauthorised physical intruder gains access to information
- Vengeful employees and outsiders

There is a need for each individual entity within a virtual organisation to be able to describe the access control policies associated with that site's data. The utilisation of the OASIS eXtensible Access Control Markup Language (XACML) is a partial solution to this problem, although the verbosity and complexity of XACML descriptions makes tool support essential. Another requirement is that it should be possible to give temporary access to data on a healthgrid. It is important that people given temporary access are not able to make unauthorised copies of restricted data. To address this, eDiaMoND proposed that data should have a lifetime, and would be deleted after use. If this were the case, it would then be following the fifth principle of the Data Protection Act. If temporary access to data is to be allowed, then either the policies must contain time-based information, or a secondary process would have to change the policies at the appropriate time. GIMI later proposed the use of Digital Rights Management (DRM) technologies to address this issue.

An aim of many grid projects, including Health-e-Child, is enabling 'single sign-on' for users, using a single identity credential that is valid for many infrastructures, communities and organisations. This is simply an issue of practicality; it would take extra time and effort for the user, and if they had to remember multiple passwords for different grid resources they could be tempted to write them on paper, opening the door to a security breach. However, eDiaMoND identified a number of issues with credentials that will need to be addressed. It is important that credentials are portable, so that researchers are able to move around their own centre or to access data when visiting remote sites without carrying a computer with them. Another problem associated with credentials is propagation – external services need to be trusted to pass users' credentials to other services. The typical grid mechanism for this task is the use of a proxy certificate, however the user has to rely on the intermediate services to use the proxy certificate as intended. Finally, it must be possible to revoke a set of credentials. This is especially important if doctors are carrying their credentials around with them. Current systems often rely on the service actively requesting lists of revoked certificates, leaving a window of opportunity for misuse. Use of the Online Certificate Status Protocol (OCSP) may reduce the window to a minimum but this has the drawback of requiring frequent communication.

Role Based Access Control (RBAC) is a common approach to authorisation in healthgrid projects. RBAC assigns permissions based on the 'role' of the user, e.g. 'medical researcher', 'district nurse', 'local doctor', etc. However, many healthgrid projects including IBHIS have found that RBAC, which might involve simply mapping a grid user to a local user account, is too inflexible for the health domain and does not provide fine-grained rules for access control, including how to support users with multiple authorisation profiles. The second IBHIS prototype used the Tees Confidentiality Model for finer grained rules, which was also used for mapping between security domains. The project also mentioned the need for managing changes in security policies and requirements, with user roles and authorisation levels changing dynamically to reflect organisational restructuring. eDiaMoND also mentions RBAC and proposed that the co-existence of local and global policies should be investigated, to support a flexible 'situated role-based access', where skills are delegated on a needs basis.

This leads to the issue of how to provide flexible, fine-grained access control, a recurring issue in grid based research. As mentioned by GIMI, this becomes problematic when securing services that are provided by a third party vendor that have their own access control mechanisms. These may not provide sufficiently flexible or fine-grained access control, and coordinating access control policies for many resources could be extremely difficult. Therefore, GIMI proposed that all access control for resources at a node should be determined by a single set of coordinated policies, and all access to existing vendor-provided web services will have to comply with these policies.

Another major task for this milestone will be the development of a stable and mature web services framework. IBHIS publications made several comments about a key component of web services, the Web Service Description Language (WSDL), used to describe the interface of a service. WSDL only describes a service in functional terms – its data types, methods, parameters, message format, etc. The WSDL standard lacks flexible, semantic, non-functional descriptions required for a dynamic service-oriented environment, such as descriptions of a service's security requirements and quality of service. Without semantic and non-functional descriptions, there can be confusion in the meaning of service and parameter names, and certain security considerations – a key concern when dealing with medical data – could be neglected. Ontology-based description languages, such as the DARPA Agent Mark-up Language for Services (DAML-S, now OWL-S) will provide a much more complete service description, but these are not fully mature and have limited tool and/or registry support.

IBHIS also noted a number of issues with the current version of UDDI (Universal Description, Discovery and Integration), which provides a registry for service discovery. The search functions in UDDI have only limited support for making automatic service selection decisions, cannot facilitate matching at the service capability level, and a key limitation of UDDI is that it does not provide semantic searching; it is essentially limited to keyword-based searching. It also does not capture relationships between entities, and is not able to infer relationships using semantic information.

Finally, IBHIS noted that the existing web services stack framework requires clarification in order to determine which technologies are usable at which level, and which are compatible with each other. The updated framework proposed by IBHIS is composed of protocols that use or extend WSDL, have roots in the semantic web's resource description framework and DAML-S, and include ebXML specifications.

Further development is currently underway in these areas, with WSDL-S supporting semantic descriptions, and WSDL 2.0 promising to include non-functional requirements. To facilitate semantic searching, IBHIS notes that UDDI's capabilities can be extended using OWL-S, and ways to address the other limitations mentioned are being examined for upcoming versions of UDDI.

2.4. Milestone MD2

The next deployment milestone is the development of a demonstration data grid production environment for medical research, including distributed storage and the querying of medical data at a distance. The distributed nature of the grid will be of particular relevance for this milestone, including how to deal with remote data sources, issues concerning the grid-based storage of medical data, and dealing with dynamic data providers and sources.

In addition to communication between clients and services, Integrative Biology discussed integration within and interaction with infrastructures that have their own security models. The authorisation mechanism for healthgrids should complement rather than conflict with these. The healthgrid projects examined here have typically chosen to use the authorisation mechanism that best suits their purpose rather than standardising on any particular one, which if not addressed could become an obstacle for inter-grid authorisation, or at least for single sign-on.

ECIT identified two security concerns with distributed storage. Local privileges may allow a user inappropriate access to sensitive data either locally or remotely, and data replication management can result in sensitive data being replicated and stored without permission, possibly from one country to another. According to ECIT, this may mean data on the grid will need to be censored rather than just encrypted to mitigate against data 'leakage'.

eDiaMoND notes that most system designers concerned with security and trust requirements fail to take organisational trust relationships into account. As a result, protection mechanisms may obstruct the effective use of the system and the activities it is designed to support. Where physical artefacts such as records and images are involved, the activity taking place can be overseen by others in the same domain – the work provides a "natural, locally visible account of itself"[7]. People also have a biographical familiarity with those they work with. Replicating this situation for digital artefacts is an important consideration, as it affects procedures for accountability, visibility of actions taken, informal practices and discussions, effective team working, and people's trust in the reliability and credibility of data and decisions. In particular the project recommended facilitating informal communication and collaboration between remote users.

2.5. Milestones MS1 and MS2

The standards milestones that begin during MD2 and continue until the start of the final deployment milestone are required for the sharing of electronic health records (EHRs) and medical images (in this case, using DICOM) on the grid. These milestones are only two examples of the bioinformatics and medical informatics standards that may need to be modified or extended in order for them to be used on a healthgrid.

eDiaMoND identified a large number of standards, sometimes conflicting and often overlapping, that must be disambiguated and adhered to for a grid deployed in the UK. There are NHS-wide standards such as the NHS data dictionary and compatibility with NHSnet, standards endorsed by the National Programme for IT (NPfIT) including STEP standards, e-GIF standards, BS7799 for information security management, and HL7. For example, the DICOM (Digital Imaging and Communications in Medicine) standard for medical images is endorsed by e-GIF, but overlaps with HL7 v2.

For medical imaging, DICOM as used by eDiaMoND and GEMSS has already been accepted by UK and worldwide bodies as the accepted standard for the acquisition, connection and storage of images. SMF (Standard Mammogram Form) has also become something of a de-facto standard for the standardisation of mammograms with different procedures, film types and processing systems.

Although this paper focuses on the technical roadmap, it is important to note that a number of important legal and ethical issues, such as pseudonymisation and the recording of patient consent, both examined by CLEF, will also require standards.

2.6. Milestone MD3

The final deployment milestone is the development of a demonstration knowledge grid production environment for medical research. The synthesis of knowledge from data will require sophisticated data integration, data mining, and image processing applications, and may also involve the use of techniques from artificial intelligence to derive relationships between data from different sources and in different contexts.

According to Health-e-Child, the integration of biomedical information will be complicated due to the fact that there is no universally accepted biomedical data model. An integrated health record such as the one proposed by Health-e-Child will need to make use of data that is not just semantically and syntactically heterogeneous, but also conceptually and temporally heterogeneous, posing a major challenge for data integration. Managing the distribution, acquisition, normalisation, aggregation, and curation of distributed heterogeneous data are all issues requiring further attention. How to cope with missing, incomplete, conflicting and uncertain data is also an issue that will need to be addressed for this milestone.

Image processing is another important technology for knowledge grids, and eDiaMoND has demonstrated that there has been considerable success in this area already. The normalisation of images (using SMF from Mirada, for example) has been successfully used to correct differences in images due to contrast, tube voltage, equipment and other variations. This also allows images to be compared more easily, which can aid radiographers in their diagnosis.

Although primarily an activity for the medical research community, the implementation of medical ontologies will be hugely important for a medical knowledge grid. These will allow relationships between concepts and nuances in meaning to be captured, greatly enhancing the opportunities for communication, knowledge sharing and machine reasoning. Ontology mapping discovery, mapping between ontologies, and semantic integration of biomedical data have been identified as bottlenecks for generating integrated case data, and are currently being explored by several projects, including Health-e-Child.

3. Conclusion

Mapping the issues identified in a selection of existing healthgrid projects against the SHARE project's initial technical roadmap has shown that there are considerable challenges for each of the milestones described. However, important progress has already been made, with the current generation of projects promising to deliver early examples of data grids and even knowledge grids for particular medical research areas.

Ethical, legal and socio-economic issues were not examined in this paper, but will continue to be a concern. Further research and standardisation in this area will certainly be required to ensure these do not hinder the efforts of future projects.

Given that development is occurring in all areas within individual projects, closer collaboration and technology transfer between EC projects will be essential for rapid progression through the milestones described here. Also, raising the profile of grids and publicising examples of successful projects in the wider medical community will be vital for the future adoption of grids for medical research.

References

[1] T Solomonides et al (Eds). From Grid to HealthGrid, IOS Press, 2005
[2] S Cox & D Walker (Eds). Proceedings of the UK e-Science All Hands Meeting 2005, EPSRC, 2005
[3] D Kalra. CLEF – De-identifying the EHR: building a resource for research, AHM 2003
[4] A Rector, A Taweel, J Rogers, D Ingram et al. Joining up Health and BioInformatics: e-Science meets e-Health, AHM 2004
[5] P Jeffreys. UK, Oxford. Life Sciences and Healthcare, Forum Engleberg 2004
[6] J Declerck. MammoGrid & eDiaMoND: grid-enabled federated databases of annotated mammograms, Mirada Solutions Ltd (Siemens)
[7] M Jirotka, R Procter, M Hartswood et al. Collaboration and Trust in Healthcare Innovation: The eDiaMoND Case Study, Computer Supported Cooperative Work (2005) 14, Springer
[8] M Brady, D Gavaghan, S Harris, M Jirotka et al. eDiaMoND: The Blueprint Document, 2005
[9] J Fenner, R Mehrem, V Ganesan, P Melas, & L Walton. Practical Experience of Grid-enabled Health Computing with the GEMSS Grid, AHM 2004
[10] S Benkner, G Berti, G Engelbrecht, J Fingberg, G Kohring, S E Middleton & R Schmidt. GEMSS: Grid-infrastructure for Medical Service Provision, Healthgrid 2004
[11] GEMSS Deliverable D1.3b – Final Evaluation & Validation, GEMSS Consortium, 2005
[12] J Fingberg et al. GEMSS Deliverable D6.3b – Edited Final Report, GEMSS Consortium, 2005
[13] D M Jones, J Fenner, G Berti, F Kruggel, R A Mehrem, W Backfrieder, R Moore & A Geltmeier. The GEMSS Grid: An Evolving HPC Environment for Medical Applications, Healthgrid 2004
[14] S Benkner, G Engelbrecht, W Backfrieder, G Berti, J Fingberg, G Kohring, J G Schmidt et al. Numerical Simulation for eHealth: Grid-enabled Medical Simulation Services, ParCo 2003
[15] A Simpson, D Power, M Slaymaker & E Politou. GIMI: Generic Infrastructure for Medical Informatics, Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems (CBMS'05)
[16] A Simpson, D Power, M Slaymaker & E Politou. Towards fine-grained access control for health grids, Proceedings of the Ottawa Workshop on New Challenges in Access Control, 2005
[17] Health-e-Child Proposal – Part B, March 2005
[18] J Freund, D Comaniciu, Y Ioannis, P Liu, R McClatchey, E Morley-Fletcher, X Pennec, G Pongiglione et al. Health-e-Child: An Integrated Biomedical Platform for Grid-Based Pediatrics, Healthgrid 2006
[19] Enabling Grids for E-sciencE II (EGEE-II) Proposal – Part B, August 2005. Available at: http://delphi.pd.infn.it/~mmazzuca/proposal-25-8/EGEE-II-PartB-25-08-2005.pdf
[20] IBHIS 2nd Annual Review (2003-2004). February 2004
[21] I Kotsiopoulos, J Keane, M Turner et al. IBHIS: Integration Broker for Heterogeneous Information Sources, 27th Annual International Computer Software and Applications Conference, Dallas, 2003
[22] M Turner, F Zhu, I Kotsiopoulos, M Russell, D Budgen et al. Using Web Services Technologies to create an Information Broker, 26th International Conference on Software Engineering, Scotland, 2004
[23] K Bennett, N Gold, P Layzell et al. A Broker Architecture for Integrating Data using a Web Services Environment, First International Conference on Service-Oriented Computing, Italy, 2003
[24] M Turner, D Budgen & P Brereton. Turning software into a service, IEEE Computer, Vol. 36, 2003
[25] F Zhu, M Turner, I Kotsiopoulos, K Bennett, M Russell, D Budgen et al. Dynamic Data Integration Using Web Services, 2nd International Conference on Web Services, USA, 2004
[26] D Gavaghan, S Lloyd, D Boyd, P Jeffreys, A Simpson et al. Integrative Biology – exploiting e-Science to combat fatal diseases, AHM 2004
[27] D Mac Randal, D Gavaghan, D Boyd, S Lloyd, A Simpson & L Sastry. Integrative Biology – Exploiting e-Science to Combat Fatal Diseases, ERCIM News No. 60, January 2005
[28] IBVRE Infrastructure – Initial Design Report, October 2005
[29] F Jacq, F Harris, V Breton, J Montagnat, R Barbera et al. First revision of EGEE application migration progress report (DNA4.3.2), EGEE Collaboration, September 2005
[30] V Breton, N Jacq & M Hofmann. Grid added value to address malaria, Proceedings of the 6th IEEE International Symposium on Cluster Computing and the Grid (CCGrid), 2006
[31] H C Lee, J Salzemann, N Jacq, H Y Chen et al. Grid-enabled High-throughput in silico Screening against influenza A Neuraminidase, IEEE Transactions on Nanobioscience, Vol. 5, N°4, 2006

From Genes to Personalized HealthCare: Grid Solutions for the Life Sciences
N. Jacq et al. (Eds.)
IOS Press, 2007

An Agent Approach for Protein Function Analysis in a Grid Infrastructure

Alessandro ORRO [a,b,1] and Luciano MILANESI [a]
[a] *Institute for Biomedical Technologies - Italian National Research Council*
Via Fratelli Cervi 93, I-20090 Segrate (MI) – Italy
[b] *Consorzio Interuniversitario Lombardo per L'Elaborazione Automatica*
Via R. Sanzio, 4 - Segrate (MI) – Italy

Abstract. Many tasks in bioinformatics can be faced only using a combination of computational tools. In particular, functional annotation of gene products can be a very expensive task that may require the application of many analysis together with a manual intervention of biologists. In this area, the phylogenomics inference is one of the most accurate analysis methodologies for functional annotation that is not yet widely used due to the computational cost of some steps in its protocol. This paper discusses the implementation and deployment of such analysis protocol in a distributed grid environment using an agent architecture in order to simplify the interaction between users and the grid.

Keywords. bioinformatics, protein function, grid computing, agent systems.

Introduction

Nowadays, the huge amount of biological sequences produced in sequencing projects increases the need for automatic tools for structural and functional annotation. Manual annotation supported by experimental analysis is often very accurate, but it is not able to keep up with the flow of data in high throughput sequencing experiments.

The simplest method for functional characterization of proteins is the sequence similarity analysis. Unfortunately, when an evident similarity is not observed, it is difficult to adopt a transitive assignment. On the other hand there are many known proteins that share the same function but are dissimilar in their amino acid sequences. Many methods have been developed in order to improve the similarity-based function assignment. Let us recall OntoBlast [1], GOBlet [2] and GOtcha [3] that weights the functions represented in similar sequences using the E-value derived from a blast search. The GOFigure [4] software uses a similar approach but the amount of possible functions is reduced by minimizing the graph of represented nodes. GOAnno [5] annotates a sequence by propagating the GO terms through subfamilies in a hierarchical multiple alignment.

[1] Corresponding Author: Alessandro Orro, Institute for Biomedical Technologies – National Research Council, via Fratelli Cervi 93, 20090 Segrate (MI), Italy; E-mail: alessandro.orro@itb.cnr.it

However, as for many other bioinformatics tasks, protein function assignment usually requires the use of several programs organized in computational pipelines. In this context the manual intervention of a biologist is often needed in order to choose, in each step of the pipeline, the best strategies to improve the quality of classification.

In this paper we describe the implementation, in the EGEE Grid environment, of a protein function classification method based on the phylogenomics pipeline. The implementation makes use of software agents in order to hide some of the complexity of the grid and exhibit to the users only a simplified user oriented interface.

1. Materials and Methods

Methods based on assignment from highly similar sequences perform quickly and therefore are suitable especially for a preliminary annotation of a whole genomes but the underlying model is known to have significant drawbacks mainly due to evolutionary events (for example gene duplication and shuffling). Phylogenomics protocol [6] [7] has been conceived to remove these limitations by using evolutionary information contained in a phylogenetic tree built from the set of sequences related to the target. Functional annotations are transferred only if they are consistent with the tree structure and the genetic events found in the tree.

The general schema of the phylogenomics analysis is shown in Figure 1. The target sequence is queried against a database by similarity in order to find out potential related sequences that are globally aligned to highlight similar regions. The alignment is used to produce a phylogenetic tree that is annotated with functional and evolutionary information. Finally an inference rule is applied in order to transfer functions from the nodes in the tree to the target sequence.

Although phylogenomic analysis is deemed to be one of the most accurate in the inference of sequence functions, it is not easy to adopt on large scale. In fact some computation in the analysis, like the construction of the multiple alignment and the phylogenetic tree can be very expensive. Moreover, to take full advantage of such methodology, all steps of the pipeline need to be accurate enough to exclude false positives in the set of homologous sequences and not leave out important ones.

1.1. Grid Infrastructure

The EGEE Grid Infrastructure [8] adopted in this work is a network of computing and storage resources connected together with the gLite middleware [9] that provides a framework for the management of grid nodes and a complete command line API for access the distributed resources (job and data).

Figure 2 shows the main components of the EGEE Grid and their interactions. A job, described in the JDL format, is submitted by the user from a *User Interface* (UI) to the *Resource Broker* (RB) that processes the job in order to find a *Computing Element* (CE) matching user requirements. When the Computing Element receive a job from the Resource Broker, executes it in a cluster of *Worker Nodes* (WN).

Security is handled through *Personal Certificates* that allow users to be recognized in the infrastructure and to execute their jobs for a limited authentication time. Usually users belong to a *Virtual Organization* (VO) that represents a logical group of users that share common research interests and projects.

1.2. Execution Framework

The Grid Infrastructure is not an integrated system in which the correct overall behaviour is guaranteed but it is a network that connects independent machines each one providing a particular service. Therefore, due to synchronization or overload problems, there are a number of aborting jobs or staying scheduled for a long time. Moreover, the execution time of a complex pipeline often takes a time longer than the personal certificate time life.

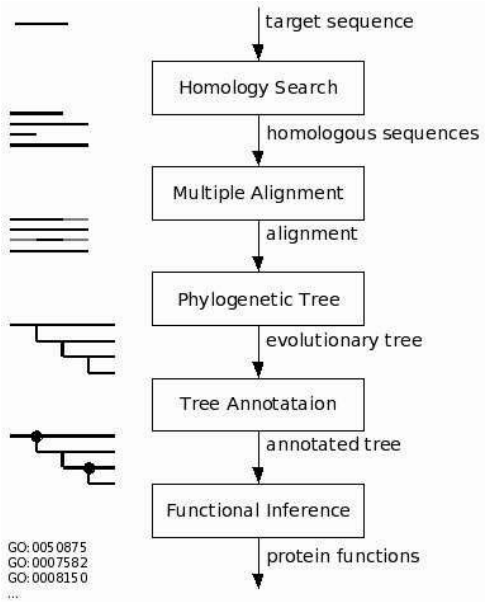


Figure 1. Phylogenomics pipeline from the target sequence to the functional annotation of GO terms.

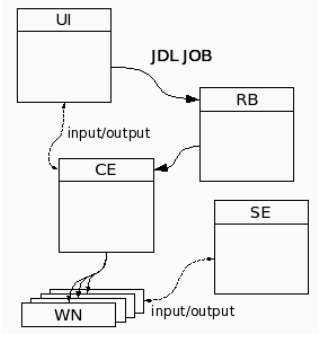


Figure 2: Main components of the EGEE Grid Infrastructure.

The proposed agent framework makes the execution of job more robust hiding the complexity of the underline grid infrastructure. In the agent paradigm [10] a program (the agent) is not directly invoked by the user or other program, but it is able to decide how and when to perform its action. In fact the actions performed by agents are mainly goal oriented i.e. based on an assigned goal rather than enabled by a function call.

The use of agent systems in the bioinformatics field has been focused on several aspects: genome annotation [11] [12], text mining [13], resource integration [14] [15] and the management of webservices-based grid environment (MyGrid) [16].

The proposed agent system is mainly devoted to enable users to easily and effectively interact with the grid environment and to manage job execution. All tasks related to the authentication, data management and low level job scheduling are left to the underline grid environment while the agents wrap the grid services in order to export to the user a more application oriented view of the computational resources. Four types of agents have been implemented in the system:

- *UserAgent* is devoted to the user interaction. Its main duty is to take care of all the communications needed to ask the user about some choice during the execution of the pipeline. In particular interactions occur when the user is alerted when a job finishes or when a partial output has to be checked and approved.
- *GridAgent* is devoted to all the operations concerning the Grid Environment (execution of jobs, proxy management, read/write files in storage element). In particular it interacts with the Grid through primitive commands of the gLite framework.
- *NodeAgent* is the node that wraps a single application in the pipeline and takes care of the job distribution.
- *ExecuteAgent* executes the whole pipeline interacting with the *NodeAgents* and updating the execution environment.

Figure 3 depicts the agent infrastructure. Each node of the pipeline is wrapped by a NodeAgent able to execute local or remote programs. The most part of NodeAgents performing computational intensive jobs execute programs through the grid distributed environment. In this case the NodeAgent executes jobs by requesting one or more services to the GridAgent. It also takes care of distributing the jobs between the nodes in grid by splitting application jobs in JDL jobs and reconstructing the output of each job from the outputs of the related JDLs.

ExecuteAgent manage the execution of a pipeline of application jobs each one represented by a NodeAgent. Its main duty is to coordinate the execution of the NodeAgents by implementing some strategies in order to better exploit the features of the grid and automate some routine work. In particular

1. enforcing resubmission of failed jobs
2. changing CE destination in case of repeated failures
3. periodic refreshing of status of the jdl jobs
4. retrieving outputs and cleaning temporary folders and files

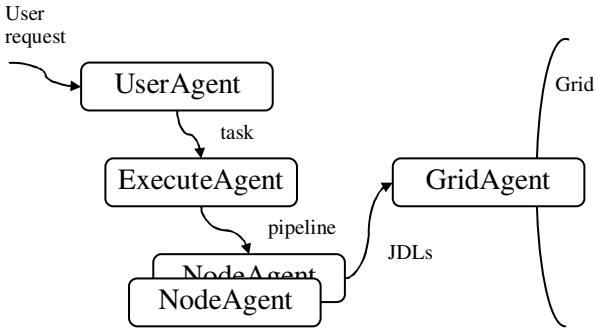


Figure 3. Agent infrastructure.

Results

The infrastructure has been tested in a particular implementation in which the pipeline analysis is composed by the following steps

1. the target sequence is searched with the blast program in order to find out all sequences similar to the target one. Sequences gathered at this step are stored in a relational database allowing user to manually remove entries from the set and providing to the next analysis step possibly only "true" homologs
2. homologs sequences are aligned with the MAFFT multiple alignemt program [17] that allow a good compromise between efficency and quality. Like the first step the resulting alignment can be edit by the user and orphan sequences removed or realigned.
3. phylip program [18] has been chosen for building the phylogenetic tree from the multiple alignment.
4. sequences in the resulting tree are annotated locally by quering external database (depending on the data source used in the blast search), running the Forester program [19] for labelling duplication/speciation nodes.

Outputs of all the programs are collected in files as partial outputs and stored in grid storage in order to reduce network load between the user interface and the grid environment. When a user intervention is required on the data, they are also parsed and stored in a relational database.

Steps 1, 2 and 3 generally associated with different jdl jobs. In particular distribution strategy enforce grouping of single jobs in sets of jdl jobs that are executed sequentially in a single worker node. This allows to make the overall time of each job longer enough and to better exploit the overhead due to the grid middleware.

Steps 4 concerns above all the visualization and integration of the results produced in previous steps and the functional annotation of the sequences in the tree (especially the target one). This is an operation that often requires user intervention and it difficult to automate. Some inference rules have been implemented (like the N-best-hits, and other function weighing methods [3]) in order to make automatic inference and other rules can be added by users. Threshold values are used for alert user that the prediction

could be not accurate and another intervention is required. All these operation are performed locally.

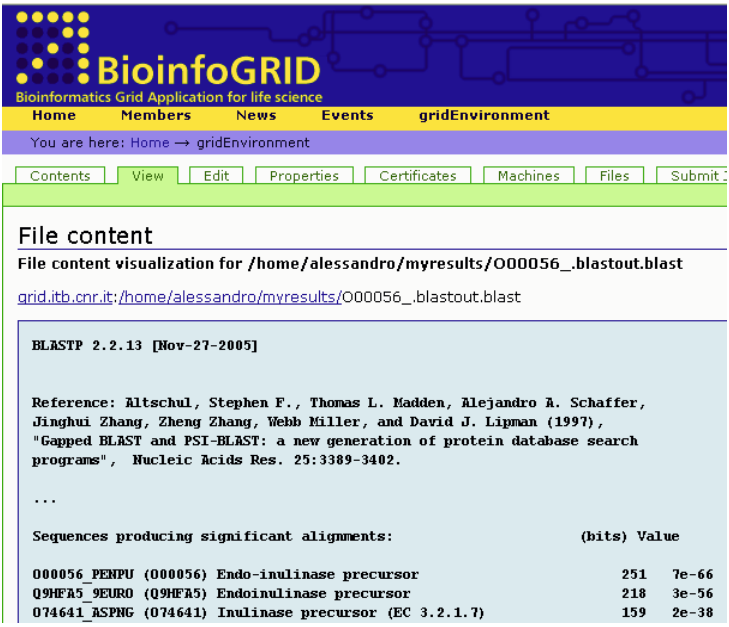


Figure 4. A screenshots of the Web portal for the management of grid-oriented application.

All the agents are implemented as active objects in a web portal that follow user actions or modification in the grid (for example change of the job status).

The UserAgent takes input from the users through a web portal built with the Zope Framework [20]. Users can navigate in storage elements and user interface and visualize both local and remote files that represent partial of final results of the submitted jobs. In particular outputs can be shown in different view modes (ascii text, xml, graphical) depending of the meta type of the file. For example, figure 4 shows the visualization of an ascii file in the user interface containing the result of a blast search.

The ExecutionAgent acts like a cron job that periodically refresh the status of all JDL jobs and automatically download the output of finished jobs. All these tasks are done by holding some information about each submitted job JDL in a relational database. These information include the grid identifier, the start time, the computing element executing the job, the job status and the name of input and output files.

Each time a user runs an application, the corresponding NodeAgent generates a set of job JDLs and puts them in the database of jobs managed by the ExecutionAgent. Figure 5 show an example of creation of a blast job that will be splitted in 5 independently jdls jobs each one running in a (potentially) different computing elements and processing a different portion of the input file.

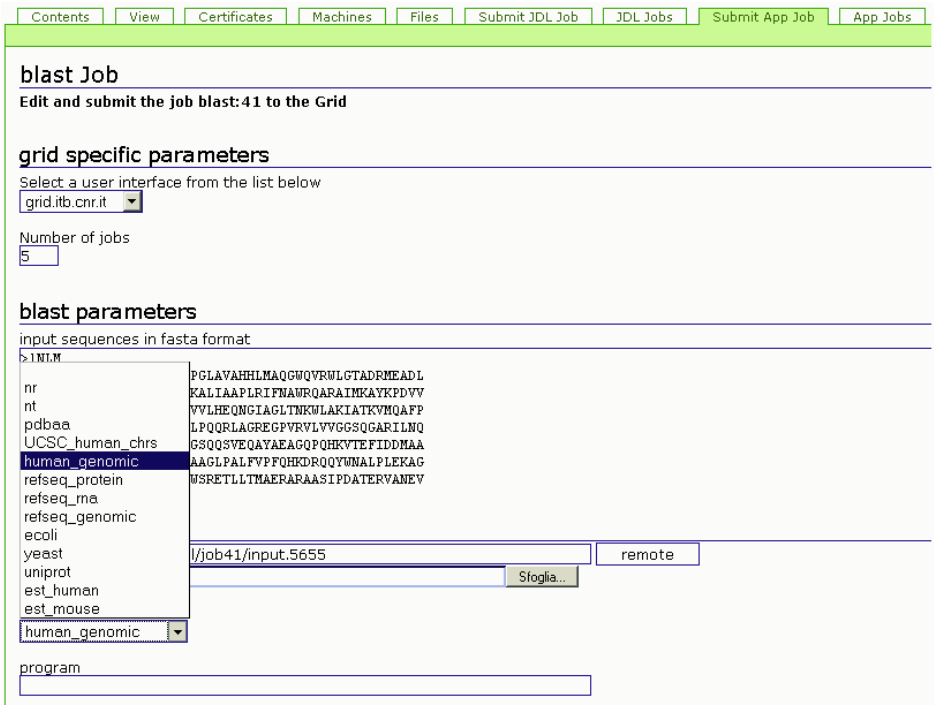


Figure 5. Parameters for the blast application job splitted in 5 jdls jobs

Conclusion

In this paper an agent system for the management of a distributed grid pipeline has been presented. It allow the user to access the computational resources exposed by the EGEE Grid in a application oriented way. This is done by hiding to the users some details that usually are performed in command line mode in the grid user interface.

Acknowledgements

This work has been supported by the Italian FIRB-MIUR project “LITBIO - Italian Laboratory for Bioinformatics Technologies” and by the European Specific Support Action BioinfoGRID and EGEE projects.

References

[1] G. Zehetner, OntoBlast Function: From Sequence Similarities Directly to Potential Functional Annotations by Ontology Terms, *Nucleic Acids Research* **31** (2003), 3799–3803.

[2] S. Hennig, D. Groth and H. Lehrach, Automated Gene Ontology Annotation for Anonymous Sequence Data, *Nucleic Acids Research* **31** (2003), 3712–3715.
[3] D.M. Martin, M. Berriman and G.J. Barton, GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes, *BMC Bioinformatics* **5** (2004) 178.
[4] S. Khan, G. Situ, K. Decker and C.J. Schmidt, GoFigure: automated Gene Ontology annotation, *Bioinformatics* **19** (2003), 2484–2485.
[5] F. Chalmel, A. Lardenois, J.D. Thompson, J. Muller, J.A. Sahel, T. Leveillard and O. Poch, GOAnno: GO annotation based on multiple alignment, *Bioinformatics* **21** (2005), 2095–2096.
[6] J.A. Eisen, C.M. Fraser: Phylogenomics: intersection of evolution and genomics. *Science* **300** (2003),1706–1707.
[7] K. Sjolander: Phylogenomic inference of protein molecular function: advances and challenges, *Bioinformatics* **20** (2004), 170–179.
[8] F. Gagliardi, B. Jones, F. Grey, M.E. Begin, M. Heikkurinen, Building an infrastructure for scientific Grid computing: status and goals of the EGEE project, *Philosophical Transactions: Mathematical, Physical and Engineering Sciences* **363** (2005), 1729–1742.
[9] gLite: Lightweight Middleware for Grid Computing, http://glite.web.cern.ch/glite
[10] S.J. Russel and P Norvig, Artificial Intelligence: A Modern Approach, *Prentice Hall,* 2002.
[11] K. Bryson, M. Luck, M. Joy, D Jones: Applying agents to bioinformatics in Geneweaver, In Cooperative Information Agents IV, Lecture Notes in Artificial Intelligence, (2000), 60–71.
[12] K. Decker, S. Khan, C. Schmidt, G. Situ, R. Makkena, D. Michaud: BioMAS: A Multi-Agent System for Genomic Annotation, *International Journal Cooperative Information Systems* **11** (2002), 3:265–292.
[13] A. Doms and M. Schroeder: GoPubMed: exploring PubMed with the Gene Ontology, *Nucleic Acids Research* **33** (2005) 783–786.
[14] K.A. Karasavvas, R. Baldock, A. Burger: A criticality-based framework for task composition in multi-agent bioinformatics integration systems, *Bioinformatics* **21** (2005), 3155–3163.
[15] K. A. Karasavvas, R. Baldock, A. Burger: Bioinformatics integration and agent technology. *Journal of Biomedical Informatics* **37** (2004) 3:205–219.
[16] L. Moreau et al.: On the Use of Agents in BioInformatics Grid. CCGRID, Proceedings of the 3st International Symposium on Cluster Computing and the Grid, 2003.
[17] K. Katoh, K. Misawa, K. Kuma, T. Miyata: MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform, *Nucleic Acids Research*, **30** (2002) 14:3059–3066.
[18] J. Felsenstein: PHYLIP (Phylogeny Inference Package) version 3.6 (Department of Genetics, University of Washington, Seattle, WA). 2002.
[19] C.M. Zmasek, S.R. Eddy. A Simple Algorithm to Infer Gene Duplication and Speciation Events on a Gene Tree. *Bioinformatics* **17** (2001), 821–826.
[20] www.zope.org

From Genes to Personalized HealthCare: Grid Solutions for the Life Sciences
N. Jacq et al. (Eds.)
IOS Press, 2007

Shortcomings of Current Grid Middlewares Regarding Privacy in HealthGrids

Yassene MOHAMMED[1, *], Ulrich SAX [*], Fred VIEZENS, Otto RIENHOFF
Department of Medical Informatics, University of Goettingen, Germany

Abstract: Although grid computing middlewares are in research use since many years, they lack of particular security features for biomedical applications. The analysis of the common Globus middleware reveals several security-related shortcomings. As a result, extended security measures for HealthGrids have been identified. They include tools for auditing, tracking, fine grained access control for structured documents, trust and trust delegation. The German MediGRID project is facing this with an "Enhanced Security" package intending to bridge the gap between current legal, data protection as well as data security requirements and the available grid technology.

Keywords: Data Protection, Privacy, Security, GSI, GT4, Enhanced Security

Introduction

There are several challenges the biomedical community has to face until biomedical grids will be largely in use. Beyond the problem of retrieving the relevant data sets using the metadata description, data access control is of paramount importance, as the owners of the data are foremost patients. Due to the heterogeneity of the data an additional ontology process is needed to homogenize the data. Figure 1 shows the grid data-flow for biomedical applications differing from usual grids by the need of retrieval, authorization and homogenizing steps.

In contrast to conventional grid applications, medical applications typically use high dimensional data. Biomedical data are not only heterogeneous; rather they contain different information types and different levels of privacy. They vary from aggregated data describing population and diseases (epidemiology, clinical practice, clinical trials), to more granular patient data and pathological descriptions (health record, clinical history, physical exams) and to cellular and molecular data (histology, genetic test results and genomic data) [1-3]. Given semantic data interoperability, the researcher can correlate and analyze the data using suitable biomedical informatics methods and tools. On the other hand having this data online with the suitable tools to correlate, merge and analyze creates new challenges for data protection and data security [4].

[1] Corresponding Author: Yassene Mohammed, Department of Medical Informatics, Georg-August-University of Goettingen, 37075 Goettingen, Germany. E-Mail: ymohammed@med.uni-goettingen.de.
[*] Authors contributed equally.

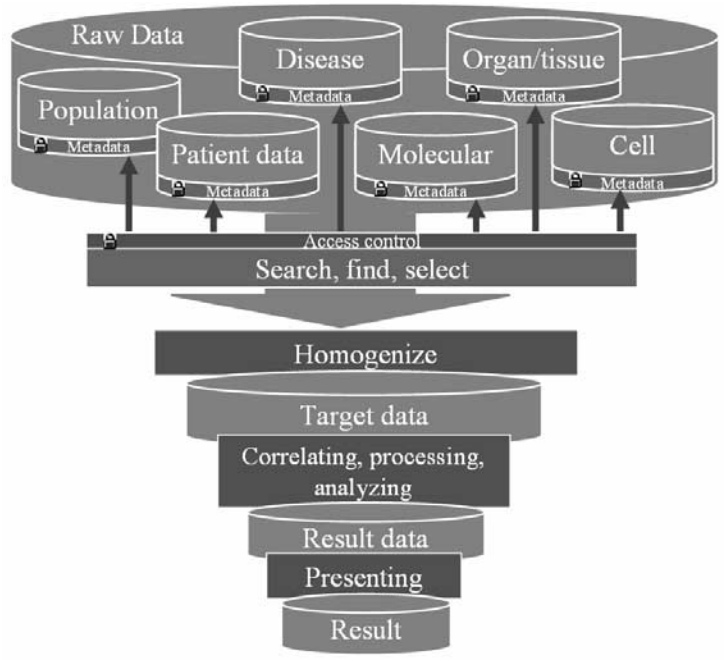


Figure 1. Data flow in MediGRID as an example for a HealthGrid: numerous data formats as well as high dimensional data in medical applications are the rational for an additional homogenizing step, before the usual eScience data processing can be started.

1. Methods

In order to analyze the privacy needs we examined the current security in grid middleware focusing on the grid security infrastructure in the Globus Toolkit. Although there are quite some grid middlewares like gLite [5] and UNICORE [6], the Globus Toolkit 4 (GT4) is widely used and could be considered as the "standard" Grid middleware for biomedical research. The security tools in GT4 deal with [7]:

- authentication: establishing the identity of users or services,
- communication security
- authorization: determining who is allowed to perform what actions, and
- other supporting functions such as managing user credentials and maintaining group membership information.

GT4 provides distinct web services (WS) and pre-WS authentication and authorization capabilities [7-10]. Both use standard X.509 certificates and proxy certificates [11], which are used to identify persistent entities such as users and servers and to support the temporary delegation of privileges to other entities.

Following the Globus design model, which intends to use current internet technologies with as less modifications as possible and the hour glass model for new standards [12], the Globus Security Team implements security as a "five layers grid security infrastructure (GSI)" [8] (see Table 1) based on standard X.509 certificates.

Table 1. The five layers in Grid Security Infrastructure (GSI) as presented in GT4 [8]

Authorization	Grid-Mapfile/ SAML(Security Assertion Markup Language)
Delegation	X.509 Proxy Certificates
Authentication	X.509 ID Certificates
Message	WS-Security/ WS-SecureConversation
Message Format	SOAP (Simple Object Access Protocol)

The Grid Security Infrastructure (GSI) builds the core for security in the GT4 middleware. The use of this security infrastructure in combination with job submission, data management, and execution management contrives the secure Grid infrastructure.
In this context, GT4 provides Data Management tools [13] for

- Data Movement including GridFTP and Reliable File Transfer (RDT),
- Data Replication including Replica Location Services (RLS), and
- Higher Level Data Services -Data Replication Services (DRS).

These tools are designed to work in combination with the GSI, which leads indeed to suitable confidentiality of communication and to data integrity required for networks for biomedical research - HealthGrids. In Figure 2, this fulfills the data security requirements for the first step – the Upload service.
In contrast to "uploading" data in the Grid, the second step – retrieval – requires more comprehensive and advanced data management. To some degree this could be introduced by available "plugins" designed to work with GT4. With tools like Storage Resource Broker (SRB) – a data grid management system – [14, 15] and Data Access and Integration Services (OGSA-DAI) [16-18] one can achieve the necessity of data availability.

While web services provide the ability to access and manipulate data, there is a need to define conventions for managing data. This led to the development of the WS-Resource Framework (WSRF) [19, 20]. WSRF, Grid Resource Allocation Management (GRAM) [21] and Monitoring and Discovery System (MDS) [22], representing the execution and information management in GT4, provide confidentiality within applications.
Some HealthGrid projects, namely the French MEDIGRID [23], implemented their own light weighted "µgrid" middleware [24, 25], suitable security for this middleware - "Sygn" [26] and an encrypted storage mechanisms of medical data on grids [27]. The aspects of fine grained authorization with respect to user-organization relationship were discussed and implemented in Sygn. Sygn was designed to be more efficient than the Community Authorization Service (CAS) developed by the Globus team [28] and than the Virtual Organization Membership Service (VOMS) [29]. MammoGrid project [30] handled security as a service 'on the Grid' and build it on the top of the GT4-GSI tools [31, 32]. GEMSS project [33] considers security for the case of medical simulation and image processing on the grid and reflects in the implementation the legal security requirements [34, 35].
While most grid projects follow the common grid middleware in focusing on security, less work has been done regarding data protection in grids. A legal framework for the protection, security and transport of personal data as well as patient data is introduced

in different EU directives. E.g. directive 95/46/EC concerns processing of personal data and free movement of such data, directive 97/66/EC regards the protection of privacy in the telecommunications sector, directive 99/93/EC describes a framework for electronic signatures, and directive 2002/58/EC deals with privacy and electronic communications. Country specific implementations vary among the EU countries [36].
The legal framework implies special requirements regarding data security and data protection [36-38] already beeing included in most grid middlewares: (1) Confidentiality of communication and application, (2) Integrity and authenticity, (3) Data availability, and (4) Personal responsibility of data processing.
Additional data protection requirements arise dealing with personal data especially in the health care sector:

- Data necessarity principle: disclose all medical and medical-relevant data of a patient, but not more than needed data for the treatment and provision of that patient.
- Context of treatment: medical and medical-relevant data of a patient should be disclosed only to the personals participated in his treatment and only the information related to this treatment is allowed to be disclosed.
- Patient consent: the patient should formally agree on the storage of his medical and medical-relevant data
- The guarantee of patient rights: the possibility of rectification, blocking, deletion of his personal data should be presented.

Offering services to fulfill these requirements on HealthGrids helps the developer to implement her application according to the legal data protection and security level.

2. Results

As a Result of this Analysis, extended security measures for HealthGrids have been identified. Beyond anonymization and pseudonymization, which are procedures to be accomplished before uploading sensitive/patient data (see Figure 2 first step), the above mentioned technologies fulfill to a good degree the requirement for data security in the grid. Anyhow we still need to know who did what when and why, namely to follow the responsibilities on the grid in order to completely fulfill the legal data security and protection requirements. On the other hand, and especially because grid middlewares are yet developed not for the special use by the biomedicine community, the requirements of data protection should be considered as well.
Several security extensions have been discussed in MediGRID [39], the biomedicine community grid project in the German national grid infrastructure D-Grid [40] funded by the Federal Ministry of Education and Research (BMBF). Our analysis shows the essential “Enhanced Security” elements for a HealthGrid:

- **Auditing** (a posteriori): an audit trail consists of log files and activity protocols. Auditing is crucial for any privacy regulation assessment. Beyond the relevant user and machine data, especially valid time stamps and time periods are needed for an efficient audit. Further dimensions of auditing are data provenance and data annotation.

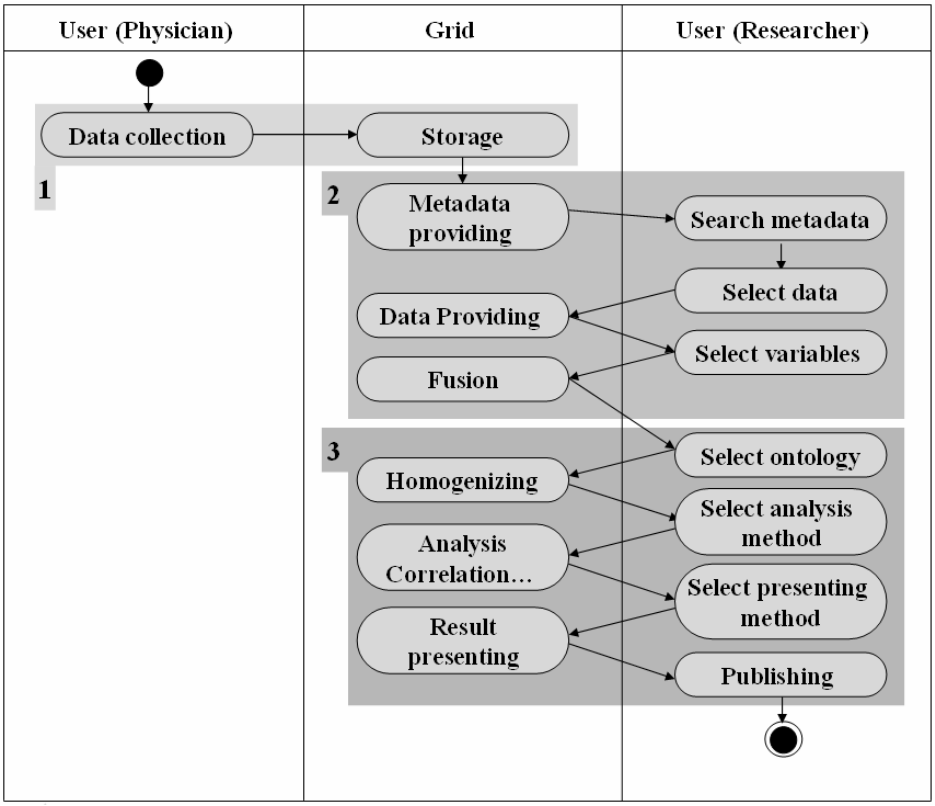


Figure 2. Activity diagram of the service flow in MediGRID as an example for a HealthGrid: 1- Upload on the grid 2- Retrieval: the user (researcher) can retrieve and select the data he needs for his work or research - the researcher prepares the data for processing, anyhow the data it self is not changed yet, 3 – Processing: here the researcher will use algorithms and processing power available on the grid to process and analyze the data intending to receive the needed results.

- **Trackability** (a priori): Additionally to auditing, trackability requires knowledge about where transfers, transactions, calculations and storage of person related data take place. This has to be part of the informed consent process between doctor and patient, as the patient data leave the doctor-patient confidentiality.
  → Auditing and tracking-possibilities cover the requirement to retrace responsibilities and retain the separation of identification data and medical data in order to preserve anonymity or pseudonomity.
- **Access rights and control**: in addition to the authentication and authorization, biomedical grid computing needs fine grained access control with respect to access rights within medical documentations, which means, that the current access control on file level (e.g. Grid-Mapfiles) not suffice, as structured medical documents [41, 42] provide different sections with a different degree of confidentiality.
- **Confidentiality**: In addition to fine grained access control in structured documents, fine grained confidentiality services have to be modified for grid computing.

→ Fine grained access rights and control as well as fine grained confidentiality fulfill the requirements of releasing only necessary data and retain the doctor-patient confidentiality.

- **Trust and trust-delegation:** trust relations and delegation as well as trust hierarchies from every day life have to be set up electronically. Using the data of a minor or a person with dementia requires that an authorized person signs electronically on behalf of those persons (eConsent). These workflows are described in some projects [43, 44], but have to be adapted for grid usage.
- **Safety:** security of data in possibly dynamic grid environments requires policies for data storage and policies for data management.
  → Safty reflects the need to develop and adopt suitable policies for the use and storage of data; a complementary safeguard principle when intending to use sensitive data considering the availability concept in time (long term archiving) and place (replicas).

The elements of the Enhanced Security consider after all the current requirements of data protection and data security intending to make grid technologies better suitable for the biomedicine community. In the future, Enhanced Security should be also flexible to fulfill future legal requirements and new developments in the medical area, e.g. genome wide association studies.

3. Conclusion and Outlook

The development of standards for data protection and data security in grids is crucial for the success of grid computing in many grid communities. Current grid middlewares lack standards and have technological shortcomings in regard to fulfill basic data protection and data security requirements. The need for a secure grid is not only an issue of computing in biomedicine. Within the German D-Grid communities there is a notable interest in the different security aspects especially in the automotive sector concerning intellectual property protection. Meanwhile the "classic" grid communities - for example climate researchers - aim for similar security standards as well. This means a long development process until biomedical and intellectual property related grid computing can make full use of the grid [4, 45, 46].
The Enhanced Security package in MediGRID is rather a one step towards enabling grid technology to be used by the biomedicine community than a complete solution. In biomedicine applications sustainability should be guaranteed. That means we need to deal with two further dimensions for a more suitable solution: future development of the grid technologies and legal framework, and international collaborative work on the country specific (legal) requirements.

The 26th international conference on privacy and data protection in Wrocław 2004 resulted in a resolution about a „Privacy Framework Standard". The resolution urges the International Standards Organzation (ISO) to work on privacy and Data Protection standards: „Development from Privacy Law into Privacy Standards". The "Privacy Enhancing Technologies" (PET) [47, 48] are of interest for the future ISO Privacy-Standard [49]. This has to be closely monitored in the interest of the biomedical grid community in order to set up a sustainable grid infrastructure.
Each change in the legal framework or in the technology in regard to grid-computing use by the biomedical community should take these standards into account. A

"converging" between the legal framework and the technical solutions of data protection and data security to the common ISO privacy standards should be considered [4]. As it is not expected to have them before 2008 [49], we need to keep track of the development of the ISO privacy standards in order to keep the converging time later as short as possible.

Acknowledgment

This work was supported by the e-Science research program of the Federal Ministry of Education and Research (BMBF) under the promotion number: 01AK803H

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] Sax, U. and Y. Mohammed. *Data-related Challenges of Genotype - Phenotype studies*. in *GMDS 2006 - Genome Wide Association Studies in Complex Traits*. 2006. Leipzig.
[3] Sax, U. and S. Schmidt, *Integration of genomic data in Electronic Health Records--opportunities and dilemmas.* Methods Inf Med, 2005. **44**(4): p. 546-50.
[4] Mohammed, Y., Viezens, F., Sax, U., Rienhoff, O. , *Rechtliche Aspekte bei Grid-Computing in der Medizin*, in *Rechtliche Aspekte der Telemedizin*, W. Niederlag, C. Dierks, and H.U. Lemke, Editors. 2006. p. 235-245.
[5] *gLite*. [2006 Dec. 27]; Available from: http://glite.web.cern.ch/glite/.
[6] *UNICORE*. [2006 Dec. 27]; Available from: www.unicore.org.
[7] The_Globus_Security_Team. *GT 4.0 Security*. 2006 [2006 Dec. 27]; Available from: www.globus.org/toolkit/docs/4.0/security/.
[8] Foster, I. and C. Kesselman, *The grid : blueprint for a new computing infrastructure*. 2nd ed. The Elsevier series in grid computing. 2004, Amsterdam ; Boston: Morgan Kaufmann. xxvii, 748 p.
[9] The_Globus_Security_Team. *Globus Toolkit Version 4 Grid Security Infrastructure: A Standards Perspective*. 2005 [2006 May 31]; Available from: http://www.globus.org/toolkit/docs/4.0/security/GT4-GSI-Overview.pdf.
[10] Welch, V., et al. *Security for Grid Services*. in *Proceedings of the 12th IEEE International Symposium on High Performance Distributed Computing (HPDC'03)*. 2003 IEEE Computer Society
[11] IETF. *Internet X.509 Public Key Infrastructure (PKI), Proxy Certificate Profile*. 2004; Available from: www.ietf.org/rfc/rfc3820.txt.
[12] The_Globus_Security_Team. *GT 4.0 Security: Key Concepts*. 2006 [2006 Dec. 27]; Available from: www.globus.org/toolkit/docs/4.0/security/key-index.html
[13] The_Globus_Alliance. *GT 4.0 Data Management* [2006 Dec. 27]; Available from: www.globus.org/toolkit/docs/4.0/data/.
[14] Rajasekar, A., et al., *Storage Resource Broker - Managing Distributed Data in a Grid.* Computer Society of India Journal, 2003. **33**(4): p. 42-54
[15] San_Diego_Supercomputer_Center_(SDSC). *Storage Resource Broker* 2006 [2006 Dec. 27]; Available from: www.npaci.edu/DICE/SRB/.
[16] Karasavvas, K., et al., *Introduction to OGSA-DAI Services*, in *Scientific Applications of Grid Computing, First International Workshop*. 2005, Springer Berlin / Heidelberg: Beijing. p. 1-12.
[17] Herrero, P., M.S. Pérez, and V. Robles, *Scientific applications of grid computing : first international workshop, SAG 2004, Beijing, China, September, 20-24, 2004 : revised selected and invited papers*. Lecture notes in computer science, 3458. 2005, Berlin ; New York: Springer. x, 208 p.
[18] Open_Middleware_Infrastructure_Institute_UK_(OMII-UK). *Open Grid Services Architecture - Data Access and Integration Services*. [2006 Dec. 27]; Available from: www.ogsadai.org.uk/.
[19] The_Globus_Alliance, IBM, and HP. *WS-Resource Framework (WSRF)*. 2004 [2006 Dec. 27]; Available from: www.globus.org/wsrf/.
[20] OASIS_Web_Services_Resource_Framework_(WSRF)_Technical_Committee. *Web Services Resource Framework (WSRF)*. 2006 [2006 Dec. 27]; Available from: www.oasis-open.org/committees/tc_home.php?wg_abbrev=wsrf.

[21] The_Globus_Alliance. *GT Execution Management: Grid Resource Allocation Management (GRAM)*. [2006 Dec. 27]; Available from: www.globus.org/toolkit/gram/.
[22] The_Globus_Alliance. *GT Information Services: Monitoring & Discovery System (MDS)*. [2006 Dec. 27]; Available from: www.globus.org/toolkit/mds/.
[23] *MEDIGRID*. [2006 Dec. 27]; Available from: www.creatis.insa-lyon.fr/MEDIGRID/.
[24] Montagnat, J., V. Breton, and E.M. I, *Partitioning medical image databases for content-based queries on a Grid.* Methods Inf Med, 2005. **44**(2): p. 154-60.
[25] MEDIGRID. *µgrid*. 2005 [2006 dec. 28]; Available from: www.creatis.insa-lyon.fr/MEDIGRID/.
[26] Seitz, L., et al., *Authentication and Authorisation Prototype on the µgrid for Medical Data Management*, in *From Grid to Healthgrid - Proceedings of Healthgrid 2005*, T. Solomonides, et al., Editors. 2005, IOS Press: Amsterdam ; Washington, DC. p. 222 - 233.
[27] Seitz, L., J.M. Pierson, and L. Brunie, *Encrypted storage of medical data on a grid.* Methods Inf Med, 2005. **44**(2): p. 198-201.
[28] Pearlman, L.W., V. Foster, I. Kesselman, C. Tuecke, S. , *A Community Authorization Service for Group Collaboration*, in *Third International Workshop on Policies for Distributed Systems and Networks, June 5-7, 2002, Monterey, California, USA : proceedings*, IEEE Computer Society. TC on Distributed Processing., Naval Postgraduate School (U.S.), and United States. Office of Naval Research., Editors. 2002, IEEE Computer Society: Los Alamitos, Calif. p. 50-59.
[29] R. Alfieri, R.C., V. Ciaschini, L. dellAgnello, Á. Frohner, A. Gianoli, K. Lõrentey, F. Spataro3, *VOMS, an Authorization System for Virtual Organizations*, in *Grid Computing - First European Across Grids Conference*, G. Goos, J. Hartmanis, and J.v. Leeuwen, Editors. 2004, Springer Berlin / Heidelberg. p. 33-40
[30] MammoGrid, *www.mammogrid.com.*
[31] Estrella, F., et al., *Experiences of engineering Grid-based medical software.* Int J Med Inform, 2006.
[32] Estrella, F., R. McClatchey, and D. Rogulin, *The MammoGrid Virtual Organisation - Federating Distributed Mammograms.* Stud Health Technol Inform, 2005. **116**: p. 935-40.
[33] GEMSS, *www.gemss.de.*
[34] Herveg, J.A.M., et al. *GEMSS: Privacy and security for a Medical Grid*. in *HealthGRID*. 2004. Clermont-Ferrand, France.
[35] Herveg, J.A.M. and Y. Poullet. *Directive 95/46 and the use of GRID technologies in the heathcare sector: selected legal issues*. in *Healthgrid 2003*. 2003. Lyon.
[36] Sax, U., Mohammed, Y., Viezens, F., Rienhoff, O., *Grid-Computing in der Biomedizinischen Forschung – Datenschutz und Datensicherheit*. Medizinische Informatik, Biometrie und Epidemiologie. Vol. 90. 2006, München: Medizin und Wissen Publ.Comp.
[37] Sax, U., *Modellierung ausgewählter Sicherheitsaspekte der "ambulant-stationären Verzahnung" im deutschen Gesundheitswesen (Dissertation)*, in *Mathematisch-Naturwissenschaftliche Fakultäten*. 2002, Georg-August-Universität Göttingen: Göttingen. p. 126.
[38] van_der_Haak, M., *Architekturkonzepte für einrichtungsübergreifende elektronische Patientenakten am Beispiel des Tumorzentrums Heidelberg/Mannheim*, in *Medizinische Biometrie und Informatik*. 2006, Universitaetsklinkum Heidelberg: Heidelberg. p. 231.
[39] MediGRID, *www.medigrid.de.* 2005.
[40] D-Grid, *www.d-grid.de.* 2005.
[41] Dolin, R.H., et al., *HL7 Clinical Document Architecture, Release 2.* J Am Med Inform Assoc, 2006. **13**(1): p. 30-9.
[42] HL7, *HL7 Receives ANSI Approval of Three Version 3 Specifications Including CDA, Release 2.* 2005.
[43] TMF. *Telematikplattform für Medzinische Forschungsnetze*. 2006 [2006 Dec. 28.]; Available from: www.tmf-ev.de.
[44] TMF. *AG Datenschutz*. 2006 [2006 March 16.]; Available from: www.tmf-ev.de/site/DE/int/AG/DS/container_ag_ds.php.
[45] Rienhoff, O., *The SIREN legal workshops: list of urgent legal actions for telemedicine.* Stud Health Technol Inform, 1999. **64**: p. 61-4.
[46] Rienhoff, O., *A legal framework for security in European health care telematics*. Studies in health technology and informatics, v. 74. 2000, Amsterdam ; Washington, DC: IOS Press. x, 192 p.
[47] Hes, R. and J.J. Borking, *Privacy-enhancing technologies : the path to anonymity*. Rev. ed. Achtergrondstudies en verkenningen ; 11. 1998, The Hague: Registratiekamer. 54 p.
[48] Federrath, H. *Privacy Enhanced Technologies: Methods – Markets – Misuse* in *Trust, Privacy and Security in Digital Business* 2005. Copenhagen, Denmark: Springer Berlin / Heidelberg.
[49] Borking, J.J. *Privacy Standards for Trust*. in *27th International Conference on Privacy and Personal Data Protection*. 2005. Montreux, Switzerland.

From Genes to Personalized HealthCare: Grid Solutions for the Life Sciences
N. Jacq et al. (Eds.)
IOS Press, 2007

Benefits of Provenance in Home Care

Ákos HAJNAL, Tamás KIFOR, Gianfranco PEDONE, László Zsolt VARGA
System Development Department, Computer and Automation Research Institute, Hungary

Abstract. Electronic information processing is more and more common in healthcare applications and involves the handling of patient healthcare records, medical procedures and decisions. The distributed and heterogeneous nature of healthcare institutions results in disconnected information islands which hinder the treatment of patients. Better and user centred healthcare services require that the information islands are connected and medical professionals are able to view the treatment process as a whole. Healthcare record exchange standards help pulling together the data. The concepts of provenance and process documentation help to analyse the way results are produced. Given the technologies of patient record exchange and process documentation, we are investigating how these technologies can help to improve healthcare services provided in the home care environment of the K4Care EU project. We present the service oriented informatics environment of the K4Care platform and show that making this system provenance aware enables medical professionals to easily find the reasons why certain procedures were followed in a treatment and find out how the procedures can be improved.

Keywords. Home care, service oriented architecture, provenance, reunification of treatment process, analysis and improvement of treatment process.

Introduction

It is more and more common in healthcare applications that practitioners cooperate and share patient data electronically. Electronic information processing in this context involves three major fields: the patient healthcare record, the medical procedures carried out on the patient, and the crucial events in the medical procedures. These are handled by several heterogeneous and distributed information systems, possibly connected into a healthcare grid. These information systems form disconnected *islands of information* and are under the authority of different healthcare actors such as general practitioners, hospitals, hospital departments, etc. Although the different information systems provide services to each other and their clients in order to allow the sharing of information and distributed decision making, the distributed nature of healthcare institutions sometimes hinders the treatment of patients, because the documentation of the healthcare history and therapy of a patient cannot easily be collected from the distributed independent healthcare institutions.

The requirement to provide better, user-centred healthcare services demands that the different pieces of patient treatment are viewed as a whole. This means not only pulling together the different pieces of the healthcare record of a patient, but also discovering the logical links between the different pieces of the treatment processes executed at different places and time. From the informatics point of view the treatment processes are supported by software services connected through the network which can

be viewed as a service oriented architecture forming a grid. In this informatics infrastructure the individual services record information on the patient locally and then later when this information is needed by another service, the electronic healthcare record (EHCR) of the patient is transmitted using some standard format like ENV13606 [1] or HL7[2]. However these data exchange standards do not solve the reunification of the treatment process completely, because they support the data exchange, but they do not give support for locating the relevant information and finding the causal relationship between the record items. Therefore, it is not uncommon for doctors to depend on the patients themselves in order to find and include relevant data from previous treatments and tests.

The Grid Provenance EU project investigated the concept of *provenance* and developed an architecture to support provenance in information systems, especially in grid systems [3]. Making electronic systems *provenance aware* enables users to trace how a particular result has been arrived at by identifying the individual and aggregated services that produced a particular output. Answering questions related to the provenance of a particular output can be done with the help of the *process documentation* which is created by the provenance aware services participating in a distributed process execution. In healthcare systems the process documentation is the basis of an integrated view of the execution of treatment processes in order to a) analyze the performance of distributed healthcare services, b) to be able to carry out audits to assess that, for a given patient, the right decisions were made and the right procedures were followed and c) to help medical professionals to determine the relevance of the patient's medical record to the current treatment.

Given the above mentioned technologies of EHCR exchange and process documentation, we are going to investigate how these technologies can help to improve the healthcare services provided in a home care environment. Our investigations are within the framework of the K4Care EU project [5]. The goal of the K4Care project is to develop an informatics platform to manage the information needed to guarantee improved home care service for the new EU society and for the increasing number of senior population requiring personalized home care assistance. Within this environment we show the possible benefits of provenance awareness and how medical professionals can use process documentation to improve their activities in home care services.

In section 1 we introduce the K4Care home care environment and the service oriented informatics concepts that we developed. In section 2 we summarize provenance in service oriented architectures and grid systems. In section 3 we describe how provenance can be used in the K4Care home care environment. Finally in section 4 we conclude that making the K4Care environment provenance aware enables medical professionals to easily find the reasons why certain procedures were followed in a treatment and find out how the procedures can be improved.

1. The K4Care Home Care Environment

It is increasingly important to develop informatics eHealth applications to support people involved in providing basic medical care like physicians, nurses, patients, relatives, and citizens in general. The care of elderly people, who usually have chronic illnesses and sometimes are disabled, involves life long treatment under continuous expert supervision. In case of elderly people with chronic illnesses it is widely accepted that being cared for in hospitals or residential facilities may be unnecessary and even

counterproductive, while their treatment at home increases their life quality and reduces costs, therefore it is important to develop informatics support for home care.

While the healthcare information system of a single medical centre may be rather centralised, medical assistance in home care naturally needs distributed service oriented architecture. In order to support home care, the K4Care project develops a platform to manage the information needed to guarantee an ICT Home Care service. The K4Care Home Care platform will integrate information of different types and from different sources; be integrated with information and communication technologies whilst ensuring private and customized data access; use ontologies to define the profile of accessing subjects (e.g. physicians, patients) and objects (e.g. disease, case study); have a mechanism to combine and refine the ontologies to personalize the system taking into account the way a physician works and the individual patient characteristics; incorporate 'know-how' from geriatric clinical guidelines as Intervention Plans (IPs); generate IPs from the healthcare centre databases if clinical guidelines do not exist or are inappropriate for a particular situation; configure a knowledge-based decision support tool that can supply eServices to all subjects involved in the Home Care model; extract evidence from real patients and integrate it with published evidence derived from randomised control trials.

While the K4Care platform will contain a generic model for home care as well, in this paper we focus on the informatics aspect of the platform which is outlined in the following.

1.1. The K4Care Informatics Service Model

The K4Care platform will provide services to its users. Typical user categories are patients, family doctors, physicians in charge, nurses, head nurse, social workers, etc. Each user will achieve its goals with the help of a set of services specific to his or her user type. A set of services specific to a user type will be incorporated in a software component which we call *agent*. The agents may be distributed in the computer network. The services will invoke other services and thus the K4Care platform will have distributed service oriented architecture. Some of the services will correspond to medical processes and their execution processes will be based on medical guidelines, while other services will correspond to administrative or technical processes related to the operation of the platform or the home care centre.

The K4Care service model is derived from the processes of home care and is based on the following concepts:

- Service: it is an abstract notion of a complex activity which is typically accomplished in collaboration with several actors (see also task below). A service is identified by a unique name (or service id) and it may have several instantiations which are called *procedures*. Different procedures instantiating the same service may be for example different localisations of the same service, e.g. in different countries or medical centres. In a given K4Care platform installation each service has one and only one procedure instantiation.
- Procedure: it is a formal description of a set of *tasks* organized in some workflow (sequential, parallel, preconditions, etc.). The procedure may be the instantiation of a medical or any general process in the medical centre. If the procedure is the instantiation of a medical process, then we call it *Intervention Plan* (IP). The workflow control structures of the procedure are described in

some formal medical guideline language like Proforma [6][7], Asbru [8] or a specific language developed for the K4Care platform. Tasks can invoke the services of another agent in the system, therefore a procedure may be some composition of services. Procedures and IPs are created by humans, e.g. medical centre managers or physicians.

- Task: it is an execution step in a procedure and it is usually a request to execute another service. The task is described by an **n-tuple**: **task = (subject, object, service or action, doc, ….)**. The "*subject"* is the type of agent which is expected to execute the *"service"* or the *"action"*. Subject can be e.g. the agent of a specific nurse or the physician in charge. The "*object"* is the actor on which the service is expected to be executed (e.g. a specific patient). The "*doc"* is a document relating to the "*service*". All actors are expected to document their activities in this document. There may be other optional parameters. If the subject corresponds to the same agent that executes the current procedure, then the service is executed internally.
- Action: it is any activity that can be executed by the agent on its own. The set of actions that an agent can perform can be considered as the agent's skills. Each action of the agent is provided as a service for other agents. When an action is executed, no procedure description retrieval is needed. The set of services that a given agent type is capable of carrying out is described by the Actor Profile Ontology (APO) which describes the actors and their agent representation within the medical centre. Actions can be imaged as a piece of Java code that implements the action.

1.2. The K4Care Platform and EHCR Systems

In the process of supporting the home care activities by providing home care services as described above, the K4Care platform must have to retrieve data from EHCR systems as well, because the medical data of the patient from previous treatment may be necessary for the current home care activity. If the data was produced in a most recent treatment, for example just after the patient was released from the hospital to his or her home, and the home care centre is in interaction with the institute where the data was produced, then the needed data can be easily located and retrieved through EHCR data exchange standards, because the location and the reference of the data is known. However if the data was produced in a former treatment, even years ago in a hospital, when home care was not envisaged for the patient, then locating the data is difficult, because there is no direct interaction between the hospital and the home care centre. In this case the physician does not know and even the patient might not remember that there is some relevant data in that hospital. This is when provenance awareness can help to realise the importance of some EHCR data and locate that data as described in [4].

Figure 1 shows that agent Bd executes a procedure and in action A1 retrieves some data about patient Oy in the Home Care (HC) centre. This is done through the "HC Centre EHCR Store" which in turn connects to the "Hospital 3 EHCR Store" and retrieves the relevant information. Depending on the retrieved data either service S2 is executed by agent Be or service S3 is executed by agent Bf.

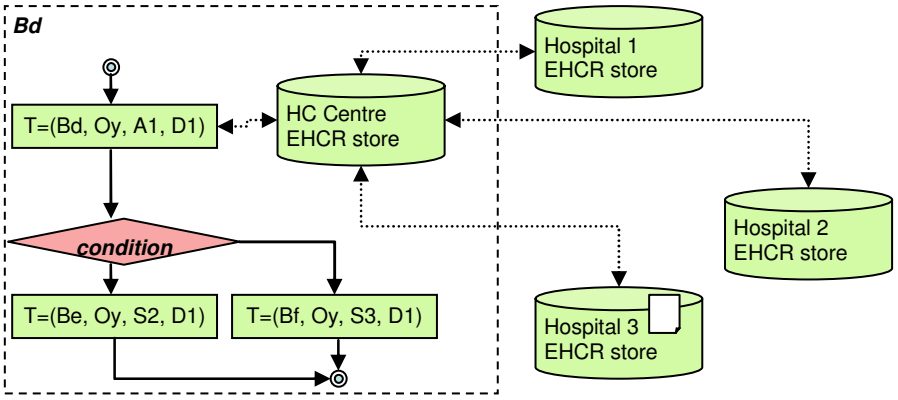


Figure 1. K4Care services and EHCR stores

1.3. K4Care Home Care Environment Summary

In the above we have described the service based informatics environment of the K4Care platform. We have also seen that this platform might use the services of EHCR stores. All together this environment can be seen as a service oriented architecture spanning over a grid of computing nodes inside and outside a home care centre. This operating environment provides the facilities needed for the usual operation of the home care centre together with the EHCR stores, but it does not support per se some analysis and verification functions.

For example if the procedure of Figure 1 is a procedure to be executed by a nurse and later the physician in charge would like to check this procedure and ask the following question "**Why was service S3 invoked?**", then it would be difficult for him or her to find the relevant information, because this is not directly present in the system. One can find out from the procedure description that action A1 was executed and than a decision was made and the execution of service S2 or S3 depends on the decision, but the relationship between these actions, services and data are not explicitly represented anywhere. In the following we will focus on provenance awareness and how provenance awareness can support answering such questions.

2. Provenance and Process Documentation

As described in [3], the concept of *provenance* is well known in fine art where it refers to the trusted, documented history of some work of art. This concept of provenance may also be applied to data and information generated within a healthcare information system. In accordance with the provenance architecture [9] developed by the Grid Provenance EU project [10], the provenance of a piece of data is represented in a computer system by a *process documentation* which is a suitable documentation of the process that produced the data. Provenance is investigated in open, large-scale systems typically designed using a service-oriented approach [11]. In this abstract view, interactions with services (seen as *actors* in the provenance architecture and realized as *agents* in the K4Care platform) take place using messages that are constructed in

accordance with service interface specifications. Such services are brought together by composition into a process to solve a given problem.

The process documentation is represented in a computer system by a set of *p-assertions*, which are assertions made by the actors involved in those processes, documenting some step of the process. The explicit description of the flow of data in a process is recorded by two kinds of p-assertions: *interaction p-assertions* and *relationship p-assertions*. An interaction p-assertion is an assertion of the contents of a message by an actor that has sent or received that message. A relationship p-assertion is an assertion about an interaction, made by an actor that describes how the actor obtained output data or the whole message sent in that interaction by applying some function to input data or messages from other interactions. An interaction p-assertion links together the actions of two actors in a process, while a relationship p-assertion links together multiple actions by a single actor. In addition, *actor state p-assertions* are assertions made by an actor about their internal state in the context of a specific interaction. The *provenance store* is the long-term facility for storing, managing and providing controlled access to process documentation.

The *provenance lifecycle* is composed of four different phases. First, actors *create p-assertions* that are aimed at representing their involvement in a computation. After their creation, *p-assertions are stored* in a provenance store, with the intent they can be used to reconstitute the provenance of some data. After a data item has been computed, users or *applications can query* the provenance store. At the most basic level, the result of the query is the set of p-assertions pertaining to the process that produced the data. More advanced query facilities may return a representation derived from p-assertions that are of interest to the user. Finally the provenance store and its contents *can be managed* (subscription management, content relocation, etc).

In the case of a healthcare information system, by recording all the medical processes related to a given patient one can get an explicit representation of the distributed processes that take place and re-construct the treatment history of the patient. Therefore, making a healthcare information system *provenance-aware* provides a way to have a unified view of a patient's medical record with its provenance.

3. Provenance in the K4Care Platform

In the following we describe how the procedure of Figure 1 is executed in a provenance enabled K4Care platform. The same procedure of Figure 1 and the graphical representation of p-assertions are shown on Figure 2. The provenance aware execution of the procedure consists of the following steps:

a) When agent Bd executes action A1 and invokes the "HC Centre EHCR store", then stores interaction p-assertion IPA1 in the provenance store. On Figure 2 the interaction p-assertion is represented by an arrow between the interacting parties: agent Bd of the first task of the procedure and the EHCR store.
b) The "HC Centre EHCR Store" uses EHCR system provenance information to find out that there is relevant data in "Hospital 3 EHCR Store". The way how provenance information is created and used by the EHCR system is described in [4], and this is not shown on Figure 2. The EHCR provenance information was created when the patient was treated in Hospital 3 possibly long time before the current retrieval.

c) The "HC Centre EHCR Store" retrieves the relevant data from "Hospital 3 EHCR Store" and stores interaction p-assertion IPA2 in the provenance store. This interaction p-assertion is about the interaction between the two EHCR stores.
d) The "HC Centre EHCR Store" stores a relationship p-assertion RPA1 in the provenance store. This relationship p-assertion is between IPA1 and IPA2, and represents that the data retrieval from "Hospital 3 EHCR store" is the consequence of the request from agent Bd.
e) When the requested EHCR data is returned to agent Bd, then the agent makes a decision based on the returned data.
f) As a consequence of the decision, agent Bd asks agent Bf to execute service S3 and stores interaction p-assertion IPA3 in the provenance store about the interaction between agent Bd and Bf.
g) Agent Bd stores a relationship p-assertion RPA2 in the provenance store. This relationship p-assertion is between IPA1 and IPA3, and represents that the invocation of service S3 is the consequence of the EHCR retrieval from "HC Centre EHCR store".

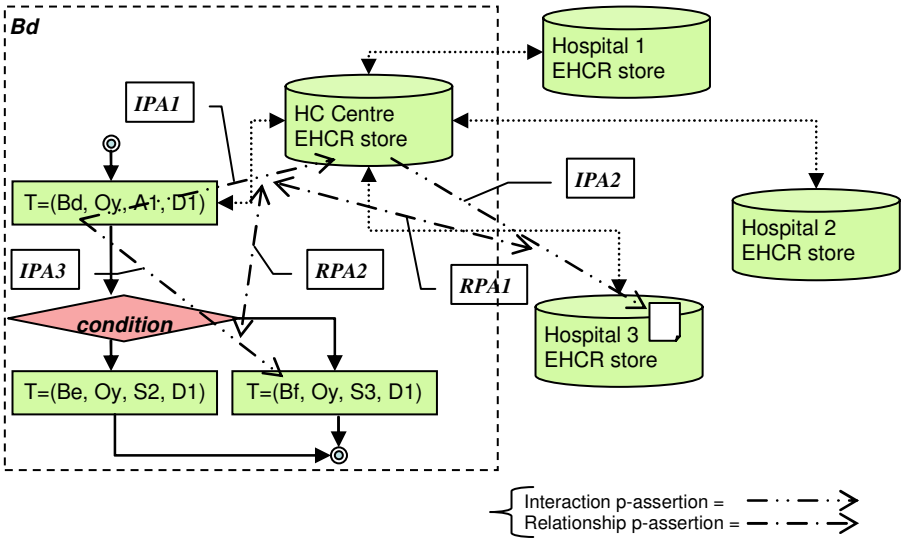


Figure 2. Provenance enabled execution of a K4Care procedure

The p-assertions recorded during the execution of the above procedure are the process documentation which can later be used to analyse the execution of the above procedure. So the question of section 1.3 can be answered by the provenance system. The physician in charge may sit in front of the analysis tool of the provenance system and pose the question. The provenance system can trace the graph of p-assertions and present to the physician in charge that service S3 was invoked because of the EHCR data returned by the "HC Centre EHCR store". The provenance system can also trace that the data returned by "HC Centre EHCR store" in fact originates from a treatment executed long time ago and stored in "Hospital 3 EHCR store".

Based on the information provided by the process documentation, the physician in charge can find out the reason why service S3 was executed and determine whether it was really needed.

The process documentation can also be used for the improvement of the processes themselves by analysing decisions and determining whether the procedure is correct or if it has to be redesigned.

4. Conclusions and Acknowledgements

We have developed and presented here a service oriented conceptual architecture for the K4Care home care environment. The concepts of this informatics environment are based on the home care processes investigated in the K4Care EU project. We have pointed out that the analysis and overview of the medical processes are not supported by the service oriented architecture in itself, however provenance awareness can help to answer analysis questions. We have shown how a provenance enabled K4Care platform works and enables medical professionals to easily find the reasons why certain procedures were followed in a treatment and find out how the procedures can be improved.

The provenance concepts and architecture was developed in the Provenance EU project [10]. The investigations presented here are part of the K4Care EU project [5].

References

[1] *ENV 13606-1:2000 Health Informatics—Electronic Healthcare Record Communication*, European Committee for Standardization, Tech. Committee 251 (Health Informatics), Work Group I (Information Models), 2000.

[2] *HL7 Version 3 Normative Edition, 2006*, Health Level Seven, Inc., 3300 Washtenaw Ave., Ann Arbor, MI 48104, https://www.hl7.org/library/bookstore/

[3] Luc Moreau et al., "The Provenance of Electronic Data," to be published in *Communications of the ACM*, 2007; www.ecs.soton.ac.uk/~lavm/papers/cacm06.pdf.

[4] Tamás Kifor, László Z. Varga, Javier Vázquez-Salceda, Sergio Álvarez, Steven Willmott, Simon Miles, Luc Moreau, "Provenance in Agent-Mediated Healthcare Systems," *IEEE Intelligent Systems*, vol. 21, no. 6, pp. 38-46, Nov/Dec, 2006.

[5] K4Care: Knowledge-Based HomeCare eServices for an Ageing Europe, IST-2004-026968, Specific Targeted Research or Innovation Project funded within Sixth Framework Programme of the European Commission (FP6), http://www.k4care.net/

[6] Jonathan Bury, John Fox, David Sutton, "The PROforma guideline specification language: progress and prospects," *Proceedings of the First European Workshop, Computer-based Support for Clinical Guidelines and Protocols* (EWGLP 2000), Leipzig 13-14 Nov. 2000.

[7] John Fox, Nicky Johns, Ali Rahmanzadeh, "Disseminating Medical Knowledge: the PROforma Approach," *Artificial Intelligence in Medicine*, vol. 14, no. 1-2, pp. 157-182, 1998.

[8] Robert Kosara, Silvia Miksch, "Metaphors of Movement: A Visualization and User Interface for Time-Oriented, Skeletal Plans," *Artificial Intelligence in Medicine*, vol. 22, no. 2, pp. 111-131, 2001.

[9] Paul Groth, Sheng Jiang, Simon Miles, Steve Munroe, Victor Tan, Sofia Tsasakou, Luc Moreau, "An Architecture for Provenance Systems," *Technical report*, ECS, University of Southampton, February 2006; http://eprints.ecs.soton.ac.uk/12023/

[10] Provenance: Enabling and Supporting Provenance in Grids for Complex Problems, IST-2002-511085, Specific Targeted Research or Innovation Project funded within Sixth Framework Programme of the European Commission (FP6), http://www.gridprovenance.org/

[11] M.P. Singh and M.N. Huhns, *Service-Oriented Computing: Semantics, Processes, Agents*, John Wiley & Sons, 2005.

This page intentionally left blank

From Genes to Personalized HealthCare: Grid Solutions for the Life Sciences
N. Jacq et al. (Eds.)
IOS Press, 2007

Author Index

This page intentionally left blank