



CONTRACT NUMBER 508830

**DEISA**  
**DISTRIBUTED EUROPEAN INFRASTRUCTURE FOR  
 SUPERCOMPUTING APPLICATIONS**

**European Community Sixth Framework Programme**  
**RESEARCH INFRASTRUCTURES**  
 Integrated Infrastructure Initiative

**JRA4 – Activity Roadmap: identification of the second set of  
 applications**

Deliverable ID: D-JRA4-4

**Due date: October 30, 2005**

**Actual delivery date: May 11, 2006**

**Lead contractor for this deliverable: IDRIS – CNRS, France**

**Project start date : May 1<sup>st</sup>, 2004**

**Duration: 4 years**

<b>Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)</b>		
<b>Dissemination Level</b>		
<b>PU</b>	Public	
<b>PP</b>	Restricted to other programme participants (including the Commission Services)	<b>X</b>
<b>RE</b>	Restricted to a group specified by the consortium (including the Commission	
<b>CO</b>	Confidential, only for members of the consortium (including the Commission Services)	

## Table of Content

Table of Content.....	2
1. Executive Summary .....	3
2. Introduction.....	4
3. The eDEISA Life Sciences Portal activity .....	5
4. The new DEISA JRA4 work plan .....	6
5. Activity reporting .....	11

## **1. Executive Summary**

This document is an update of the work plan of the JRA4 “Life Sciences” activity.

This document is publicly available.

## 2. Introduction

The JRA4 Life Sciences activity involved initially only one partner organization (IDRIS-CNRS) with a very precise work plan focusing on a well defined set of initial applications for a period of 18 months. The present deliverable, scheduled at PM18, was introduced to provide an update of the initial IDRIS work plan and to identify the set of Life Science applications that would be pursued after PM18.

However, a number of things happened after the project start that have changed this roadmap. In the first place, the DEISA Consortium was enlarged to three new partners at PM12, and one of the new partners (BSC) joined the JRA4 Life Sciences activity with an engagement comparable to that of IDRIS, and reinforced the initial work plan by introducing an important new activity in the area of protein dynamics, starting at PM12.

In the fall of 2005 – at about DEISA PM18 – the DEISA Consortium presented a complementary eInfrastructure proposal to the EU, called eDEISA for “extended DEISA”. This proposal reinforces, among other things, the application strategy of the Consortium. One of the major strategic decisions was to deploy scientific portals as a way of hiding complex supercomputing environments from new user communities, and of enabling interoperability with other Grid environments. The decision of deploying and operating a European wide portal for Life Sciences is part of the eDEISA work program, currently under negotiation with the EU. It was of course part of this strategy to decide to align the DEISA JRA4 activity along this new strategy adopted by the research infrastructure.

The new JRA4 “Life Sciences” roadmap that is presented here is therefore totally different from what was planned at the project start (just an update of applications enabled by the IDRIS partner). It takes into account these two new ingredients that have an important impact in the activity today

- The important contribution from BSC in human resources and know how, together with its expertise and interest in the field of protein dynamics
- The Life Sciences Portal eDEISA project, that establishes a new framework for the JRA4 DEISA activity

We fully understand that eDEISA will be eventually a separate contract and that this deliverable is planned to report on DEISA activity. However, the eDEISA Life Science Portal activity sets the stage for Life Sciences in DEISA. Therefore, we have chosen to describe in the next section the objectives and the scope of the eDEISA Life Sciences Portal, as well as the way in which the DEISA JRA4 activity is integrated in this global strategy. The remaining sections deal with the DEISA JRA4 new roadmap and work plan, the purpose of this deliverable.

The area of Life Sciences is one of the most challenging ones in the context of high performance computing, because in most applications the important raw computing power or the data management facilities provided by the DEISA platforms has to be

interfaced and integrated with external lightweight elements (Web interfaces, lightweight servers, etc) that are the ones that are accessed directly by the end users.

This is the reason why particular attention is being paid in this activity to portals that hide the DEISA environment from end users. The Radiation Therapy Planning application belongs to this class, and is jointly deployed with EGEE.

### **3. The Life Sciences Portal strategy**

The area of Life Sciences is one of the most challenging ones in the context of high performance computing, because in most applications the important raw computing power or the data management facilities provided by the DEISA platforms has to be interfaced and integrated with external lightweight elements (Web interfaces, lightweight servers, etc) that are the ones that are accessed directly by the end users.

This is the reason why particular attention is being paid in this activity to portals that hide the DEISA environment from end users. The initial application developed with EGEE on Radiation Therapy Planning application belongs to this class

Portals and Web interfaces are critical to enhance the user adoption of sophisticated supercomputing infrastructures, by hiding from them the complexities of the computational environment. This is a major priority for DEISA, because these user interfaces can play a major role in increasing the outreach of the European supercomputing infrastructures by attracting non-traditional users, and in interfacing the DEISA Grid to other Grid infrastructures.

The fundamental strategic issue to be underlined is that the purpose of Portals and Web interfaces is not simply decorating and enhancing the interactivity of existing applications for existing users. The purpose is to extend the outreach of the DEISA supercomputing infrastructure by reaching new user communities that have already structured their applications strategies around small, discipline oriented grid infrastructures with discipline specific tools. The basic idea is to “connect” the DEISA supercomputing resources as “backend” resources to these discipline oriented interfaces, thereby allowing new user communities to access supercomputing resources without abandoning their working interfaces and environment.

The lines of action in this area – interfacing DEISA to existing cyber-infrastructures - are extremely diversified, given the large number and diversity of discipline oriented grid infrastructures. Adapting to each one of them is obviously not scalable, and a general strategy is needed to define a generic and sustained approach to this problem. The decision taken by the Consortium is to focus initially in one specific discipline, Bioinformatics and Life Sciences, to acquire the required experience to define a more general strategy. The reason of this choice is that this discipline is evolving fast and is reaching the point where the initial computational strategy based on packages running on lightweight platforms is no longer sufficient. There are well identified, specific areas where high performance computing is needed. A substantial number of bioinformatics applications run very efficiently on high end supercomputers, as we have learnt in the DEISA JRA4 (Life Sciences) activity.

The strategy adopted by the DEISA Consortium to set up a global European supercomputing service for Bio-Informatics and Life Sciences, requiring the cooperative activity of several DEISA and eDEISA services:

- The DEISA JRA4 (Life Sciences) and eDEISA eSA4 (Applications enabling, data intensive applications) will port and optimize a number of leading bioinformatics applications to the most adapted platform in the Grid (AIX super-cluster, MareNostrum, Altix SGI or NEC systems).
- The eDEISA eSA3 Service Activity (Middleware) will manage the portal activity: Web presentation layers and the middleware needed for access to the DEISA platforms
- Global file systems will be used to host common databases, shared by all the computing platforms. Indeed, the DEISA Global File System GPFS has been extended to non-IBM clients, which will allow in the near future a global file system configuration in DEISA embracing all the scalar supercomputing platforms, namely, 10 out of the 11 DEISA platforms.

As stated above, the DEISA JRA4 activity will provide support for the first item. The other activities correspond to the (eventual) eDEISA contract.

#### **4. Update of work plan: strategic and organizational aspects**

JRA4 will continue to focus on the subset of Life Sciences scientific areas and applications that have reached a point where traditional lightweight clusters are no longer sufficient for the scientific requirements and supercomputing resources are needed for sustained performance. The protein dynamics activity pursued at BSC is of course in this class.

HPC oriented Life Sciences applications will be ported and fine tuned for all the architectures of the DEISA Grid: shared memory IBM AIX systems, shared memory SGI Altix systems, IBM PPC Linux system (MareNostrum). Obviously, best fits between applications and architectures will be established, so that different applications will be preferably mapped to the best fitted architecture. Previous experience in JRA4 has shown the dramatic importance of the underlying architecture (shared versus distributed memory) for each specific application.

In spite of the fact that the participating partners may have different focus and competences, a cooperative global activity of application enabling and performance tuning will be established, with all the actors accessing in principle the whole set of applications and the whole set of computing platforms. Maintaining a global view of the activity is considered as an absolute necessity for efficiency and impact.

Moreover, the applications enabling JRA4 teams will also contribute to connect the applications to the portal activity, by contributing specifications and recommendations to the middleware experts working on the portal deployment, and by contributing eventually to the presentation layer of the application.

### ***The specific BSC focus on protein dynamics***

One of the priorities of the Computational Biology program at the BSC is the development of computational tools helping users to access in a more efficient way to DEISA computational facilities and especially to MareNostrum.

Analysis of the needs of scientists working in the area of life sciences has shown to main problems, for which supercomputers have to provide solutions:

- Sequence related problems. Here the computational problem is simple and the need to use supercomputer is due to the need to perform the same calculation many times. Often calculations are done as workflows and user likes to use many computer but do not spend time modifying internally the code.
- Structure related problems. These are more expensive calculations, where a correct optimisation of the code and the selection of the correct compilation directives can reduce dramatically the cost of the calculation. The effective use of computational resources by external community implies here:
  - Optimization of executable codes to improve their performance on a given platform of the DEISA supercomputing infrastructure.
  - Benchmarking of available programs to determine the best suited code for a given biological problem
  - Help the user in the generation of input files in the management of output files and in the inter-change of information between programs.

The JRA4 work will be guided by the analysis above. The objectives are:

- the developed of a general parallelization utility,
- the development of interface for the automatic set-up of molecular dynamics simulation,
- the installation, optimisation and benchmarking of MD codes,
- the installation, optimisation and benchmarking of docking programs, and
- the development of a front-end for the analysis of trajectories obtained from different platforms and the input $\leftrightarrow$ output and output $\leftrightarrow$  output translation.

## 5. Update of work plan: description of planned activities on applications

### *5.1 - Protein dynamics: Implementation, optimisation and refinements of codes for protein-protein docking and scoring.*

**Objectives:** Determining the interaction network of whole organisms has become a major theme of functional genomics and proteomics efforts and it can be safely expected that vast amounts of supercomputer resources will be focused in helping in the determination of the “interactome”. Protein docking is an emerging technique in this field, in which the objective is to predict the complex formed when proteins physically interact using the atomic coordinates of two individual proteins. Due to the very large number of degrees of freedom involved in the calculations, protein docking is one of the most computationally demanding fields in bio-informatics.

**Activity: generation and tests of a multi-program docking platform:** The purpose is to analyze the most representative and the (a priori) better programs in order to run a benchmark on them and determine which ones provide the best results from a purely biological point of view. Particularly, the objective is, not only to determine the average quality of each method, but also to determine the performance of the methods to deal with problems of different difficulty (this mostly depends on the conformational changes they suffer, the size of the interface area or the kind of relation that proteins establish such as enzyme/inhibitor or antigen/antibody). The final objective will be to advice the community in which of the methods is better for each of the possible conditions. We can even think in expert programs able to manage and re-score the different docking solutions provided by the different programs based on their expected quality to solve the specific docking problem.

### *5.2 – General purpose tool deployment: MpiGrid, an easy way to write parallel programs.*

**Objective:** The efficient usage of the DEISA supercomputing platforms (and especially of MareNostrum) in Bioinformatics projects requires often doing the same task many times or a group of weakly coupled tasks in parallel. MpiGrid is an easy way to implement parallelism in certain special cases, for example embarrassingly parallel codes with no dependencies in their tasks. The usage of MpiGrid will reduce the total time in development, maintenance and upgrade these programs. MpiGrid also allows the parallelization of codes with dependencies but this feature is less powerful than the traditional MPI API.

MpiGrid is a library implemented in C with MPI support and it use light-threads to implement the communication management. The program is designed to offer the user with a friendly and simple interface, which allows him to parallelize programs in fast and simple way, even for scientist without especial technical skills. Previous knowledge of MPI is not required.

Typical bioinformatics cases where MpiGrid can help include:

- Performing several executions of the same program with different inputs.
- Executing a lot of different programs at same time
- Multiple executions of workflows (consecutive executions of programs)

These are of course all examples of capacity computing, which is not the primary focus of the DEISA infrastructure. However, in some well selected cases MpiGrid can indeed operate as a useful tool for DEISA users.

**Activity: deployment of MpiGrid on the DEISA platform.** This activity includes deployment of the library on the DEISA supercomputers, as well as the deployment of a user guide inside the DEISA Primer with a small number of examples or test cases illustrating its usage. This is a lightweight activity, that should be completed in a short lapse of time. The later usage by users of the library is not part of the JRA4 activity, this should be handled by the traditional user support services.

### ***5.3 – Deployment and performance tuning of bio-informatics applications of general usage for the DEISA Life sciences portal.***

**Objective:** To maintain the effort put into making general genomics codes optimized and available to the community directing special attention to a new scientific area: phylogeny. Two of the main genomic software packages have already been made available on the AIX super-cluster. The phylogenics field in some European countries (like France) does not dispose of sufficient resources and support. This is why the domain will be browsed in order to determine which codes gather most of the users demand.

**Activity: Deployment and operation of selected software packages.** The table below shows the software packages that have been identified as potentially adapted to operation in a supercomputing environment. Some of them will be selected to provide the first set of applications that will be deployed in the DEISA Life Sciences portal.

**Table 1: Software packages to be deployed in the DEISA environment**

<b>Domain</b>	<b>Software name and source</b>	<b>Description</b>
Phylogeny	<b>Mrbayes</b> mrbayes.csit.fsu.edu	Bayesian inference of phylogeny: based upon a quantity called the posterior probability distribution of trees, which is the probability of a tree conditioned on the observations
Phylogeny	<b>Treepuzzle</b> www.tree-puzzle.de	TREE-PUZZLE is a code used to reconstruct phylogenetic trees from molecular sequences data by a maximal likelihood method
Phylogeny	<b>Phyml</b> atgc.lirmm.fr/phyml	PHYML is a software implementing a new method for building phylogenies from DNA and protein sequences using maximal likelihood
Phylogeny	<b>IQPNNI</b> www.bi-uni.duesseldorf	An efficient tree reconstruction method is introduced to reconstruct a phylogenetic tree based on DNA or amino acid sequence data
Alignment and sequence	<b>BLAST</b> www.ncbi.nlm.nih.gov	Provides a method for rapid searching of nucleotide and protein databases.

comparison		
Alignment and sequence comparison	<b>MPI-BLAST</b> www.ncbi.nlm.nih.gov	Provides a method for rapid searching of nucleotide and protein databases.
Alignment and sequence comparison	<b>ClustalW-MPI</b> web.bii.a-star.edu.sg/~kluobin/clustal	Tool for aligning multiple protein or nucleotide sequences.
Alignment and sequence comparison	<b>FASTA</b> fasta.bioch.virginia.edu	Compares a protein sequence to another protein sequence or to a protein database, or a DNA sequence to another DNA sequence or a DNA library
Motif search	<b>HMMER</b> hmmmer.wustl.edu	Profile hidden Markov models (profile HMMs) is used to do sensitive database searching using statistical descriptions of a sequence family's consensus
Linkage analysis	<b>GeneHunter Twolocus</b> <a href="http://www.staff.uni-marburgh">www.staff.uni-marburgh</a>	Performs parametric and non-parametric multi-marker linkage analysis of dichotomous traits with two autosomal diallelic disease loci.
Molecular simulation	<b>NAMD</b> www.ks.uiuc.edu/Rese	Parallel molecular dynamics code designed for high performance simulation of large biomolecular systems

**Table 2: Status of the software packages to be deployed in the DEISA environmentt**

Software name	Parallelization state	DEISA status
<b>Mrbayes</b>	MPI version available	Being installed at IDRIS (AIX super-cluster)
<b>Treepuzzle</b>	MPI version available	Not yet tested
<b>Phyml</b>	In progress at IDRIS	Not yet tested
<b>IQPNNI</b> www.bi-uni.duesseldorf	MPI software available	Not yet tested
<b>BLAST</b> www.ncbi.nlm.nih.gov	OpenMP version produced at IDRIS	In operation on the AIX super-cluster
<b>MPI-BLAST</b> www.ncbi.nlm.nih.gov	MPI software available	Not yet tested
<b>ClustalW-MPI</b> web.bii.a-star.edu.sg/~kluobin/clustal	MPI version available	Being installed at IDRIS (AIX super-cluster)
<b>FASTA</b> fasta.bioch.virginia.	MPI version available	Not yet tested

edu		
<b>HMMER</b> hmmmer.wustl.edu	OpenMP version produced at IDRIS	In operation on the AIX super-cluster
<b>GeneHunter Twolocus</b> <a href="http://www.staff.uni-marburgh">www.staff.uni-marburgh</a>	MPI software available	Being installed at IDRIS (AIX super-cluster)
<b>NAMD</b> www.ks.uiuc.edu/ Rese	MPI software available	Not yet tested

## 6. Activity reporting

The activity work plan presented here spans the last two years of operation of the project (from PM 24 to PM 48). The proposed reporting is:

PM 30: Midterm status report on the three activities described here

PM 36: Annual status report, and update of the work plan

PM42: Midterm status report

PM 48: Final Activity report