Mapping of Scientific Workflow within the e-Protein project to Distributed Resources

Angela O'Brien, Steven Newhouse and John Darlington

London e-Science Center, Department of Computing,
Imperial College London, London SW7 2BZ,
lesc-staff@doc.ic.ac.uk

Abstract

The e-Protein project, a BBSRC pilot project, aims to examine the issues in building a structure-based annotation of the proteins in the major genomes by linking resources (computing, software and databases) at three sites using Grid technologies. This paper describes the implementation of the Imperial College annotation pipeline (3D-GENOMICS) within ICENI. The scientific problem of large-scale protein analysis is addressed along with a solution for mapping workflow within the project to components on distributed resources.

1 Introduction

The recent developments of projects such as the sequencing of the genome from several organisms and high-throughput X-ray structure analysis, have brought to the scientific community considerable amounts of data about the sequences and structures of several thousand proteins. This wealth of information requires exploitation in order to understand the biological functions of the gene products. With recent progress in computer technology genome technology has been dramatically advanced. Bioinformatics research has benefited from the availability of inexpensive desktop computers, which have been used to build large processor farms to support the analysis of this data deluge. This does bring its own problems as the physical demands for space, power and cooling place a common limit on the maximum size of these clusters that may be supported by a particular research group or institution. The results of analysis conducted by scientists are accumulated and managed separately in each research institute all over the world. Therefore, there is an ever increasing demand for high computational capacity to deal with storing such analyzed data into databases, searching gene sequences and homology from experimental results, and simulation of chemical compound. Sharing of computing resources across different sites provides the obvious solution especially when given the computational load for certain steps within the protein annotation pipeline, such as homology searching using PSI-BLAST[1]. Up until now, the lack of appropriate management and security tools has not allowed competent develop-

ment of this resource sharing. Grid technology tools (such as Globus and ICENI) are now able to provide the first step towards transparent resource sharing.
The e-protein project (http://www.e-protein.org) is funded by the Biotechnology and Biological Sciences Research Council (BBSRC -28/BEP17014 March 2002) through their e-Science programme. The e-science objectives of the e-Protein project are to provide a structure-based annotation of the proteins in the major genomes linking resources at three sites (European Bioinformatics Institute - EBI, Imperial College London and University College London - UCL) by Grid technology. The objectives are to:

- Create local databases with structural and functional annotations.
- Advertise this proteome annotation to the general biological community though a single web-based distributed system, DAS [2].
- Share computing resources transparently between sites using Grid middleware.
- Use the finished system for comparison of alternative approaches for annotation and thereby identify technological advances and impovements.
- to provide a working system after two year.
- Integrate relevant bioinformatics resources into the pipeline through Grid resources.

This paper describes our progress, thus far, in delivering the objectives listed above to the biological community. We begin by outlining the particular issues involved in protein annotation and the line of investigation these efforts will allow the biological scientists to explore. We describe the data and local databases that are being used. We then detail the development of component based applications, the mapping of workflow within e-protein to distributed resources and the integegration of such a capability within the Imperial College e-science Networked Infrastructure (ICENI), an integrated Grid middleware [3][4] for component based applications. We then move on to the scientific results gained and conclude with an outline of future developments and improvements.

2 The Scientific Problems

2.1 Protein annotation

Many model proteomes or "complete" sets of proteins of given organisms are now publicly available with much effort being invested in the functional and structural annotation of those proteomes. 40-60% of proteins in a genome can be predicted solely based on similarity with other proteins, but more detailed analysis results in higher prediction rates [5]. In order to achieve this indepth analysis a range of methods may be used including: (i) identification of sequence motifs/profiles that characterize structure or function, (ii) determination of sequence features such as coiled coils and transmembrane regions, and (iii) the identification of homologous proteins for which a known structure and function has already been identified. Normally when deciding whether a protein should be given a certain annotation, a score threshold is set with annotation being given to proteins scoring higher than the threshold. Obviously some annotation mistakes may occur and it is necessary for the outcome of various methods to be examined in order to produce a more accurate result. With this in mind, it is important to note that in order to achieve this us-

ing Grid technology several key technical problems will need to be overcome on the biological as well as the grid infrastructure side:

- Since no single UK group has the tools to handle the needed annotation for the proteomes of dozens of major species its is necessary to mix areas of annotation expertise to achieve this.
- Proteome annotation is not a well established methodology and the comparison of results from using different available annotation procedures is the only route to the identification and resolution of problems.
- Enhanced structural and functional annotation employ predictive algorithms
  and comparison of results from different approaches is necessary to obtain estimates of the reliability of the predictions
- New methodologies are constantly being developed and need to be evaluated and incorporated if necessary Several groups working together will identify these new strategies more quickly by virtue of their different contacts

2.2 E-science issues involved

The project presents a complex Grid infrastructure of distributed computational resources with different databases and specialised analysis software that may not be deployed on every resource. It is therefore essential that the Grid infrastructure is able to accurately represent the state of the software and hardware resources at each site. This heterogeneous capability needs to be effectively exploited by the complex workflow presented within the protein annotation pipeline. The capture of this workflow and the mapping of its components to the distributed resources are the key e-science issues within this project. The database will require regular, and frequent, updating with the sharing of computing resources essential given the computational load for certain steps.

3 Protein Workflow

At each of the 3 sites involved in the project sites there is a local database providing the local contribution to the proteome annotation. Different strategies are used at each site to assign protein structures to the sequences.Each database will have an associated web page.The distributed annotation system (DAS) at the EBI will provide the common links to the three web pages for the users in the community.

At Imperial there is a MySQL-based relational database called 3D-GENOMICS [6] that includes protein structure and function annotation across different genomes. The analysis pipeline currently has a focus on protein sequences for which several steps of analysis using various applications are performed such as: Identification of transmembrane regions, coiled-coils,low complexity regions, Prosite-patterns, PFAM and SCOP [7]domains, repeats, homologous sequences and secondary structure prediction. Structural information (fold classification) is assigned to sequences of the genomes via homology (using BLAST [8], PSI-BLAST and local software 3D-PSSM [9] to recognise remote homologies missed by PSI-BLAST).See figure 1.

The capture of this workflow and the mapping of its components to the distributed resources is the priority within this project. ICENI allows the scientist

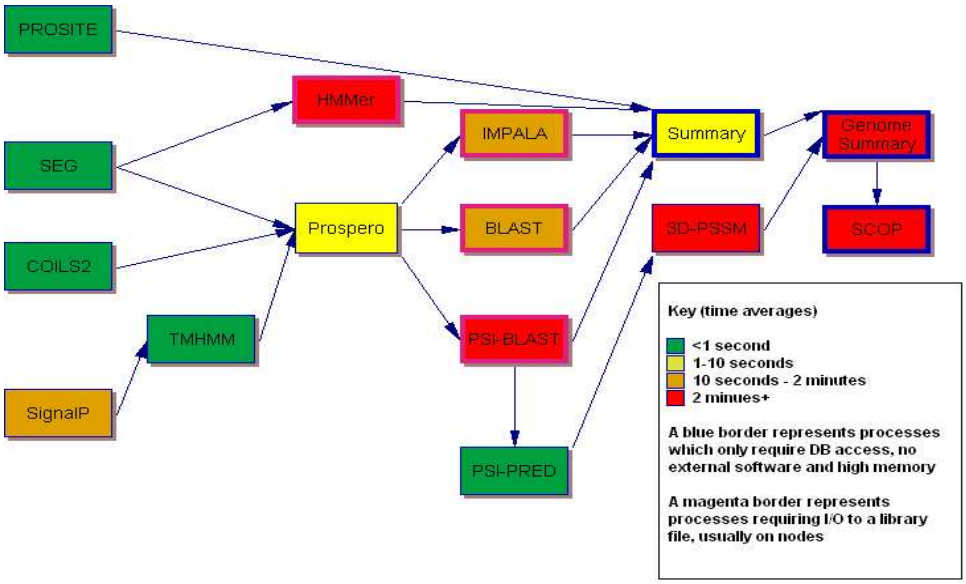

Figure 1: Imperial college e-protein workflow.Arrows describe the direction of data flow between components.

to define the required workflow within a graphical environment by dragging-and-dropping components to build their required workflow. This application specification is then mapped to the best currently available resources through a scheduling infrastructure.

4 The ICENI Middleware

The ICENI middleware used within the project has a rich meta-data structure that will allow the current state of the resources to be captured [10]. This meta-data allows the different library versions and application programs to be accurately represented. This information is defined in an XML schema. Two of the main services provided by the middleware are the Launching Framework and the Scheduling framework. By making use of the componentised nature of the ICENI applications, along with the rich metadata held within the system and the workflow of the given application, the scheduler examines the currently available services (both software and data sources) and evaluates the capability of the free resources to meet the requirements specified by the user [11]. ICENI uses a Launching Framework service in order to deploy work onto Grid resources. This service has two main tasks, firstly of advertising the resource(s) available through the launcher; and secondly of taking a JDML doccument, transforming it into a locally understood format before executing it upon the appropriate resource. Many Launching Framework services may be available within an ICENI based Grid, with each service representing one or more resources on that Grid. It is the responsibility of the Launching Framework to stage any files that may be required for job execution to the resource, and to then stage any necessary files back afterwards.

4.1 Binary Components

Most of the components used in the e-protein workflow were wrapped as binary components. The use of the Binary Component requires that the application is run from within an ICENI component with that component providing the necessary metadata required to schedule and launch it using the aforementioned frameworks.

Every Binary Component is connected to the binary executable that the component represents, plus a JDML file which describes how the application is to be executed and the arguments that it may take [12]. Input and output data may be passed from other components to the Binary Component in ICENI allowing a set of arguments to be passed to the binary executable, or the output of the application to be passed back to ICENI (through the stdout and stderr ports). This is an essential part of the project as it allows the output of one application to be used as the input for the next application thus allowing a continuous steam of analysis to be done on a protein sequence. See figure 2.

The benefits of this type of application model for the e-protein application workflow are immediately obvious. It promotes code reuse and reduces the task of Grid application development, an important issue for the biologist, more interested in results than programming wishing to create their own Binary Components. The components may be replaced and linked up with ones of the user's choosing allowing a different aspect of the protein analysis to be achieved. To facilitate this a netbeans based client has been developed to allow end-users to browse and monitor available services upon ICENI whereby components can be dragged-and-dropped onto a composition pane. These are connected together to visually represent the workflow of the application.

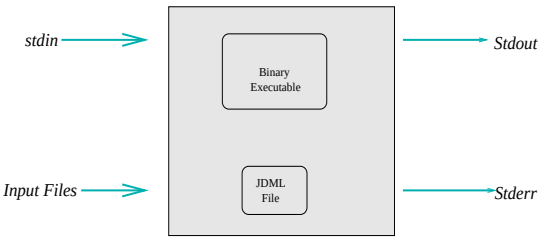

Figure 2: Model of a Binary Component used within ICENI. The arrows depict the direction of data flow.

5 Further work

An area that requires more development and is one of the major issues involved within the project is data format. Since the format of the input and output data is predetermined by the application it is necessary to come up with a method to ensure each application recieves the correct format. For the imperial side of the project a number of perl scripts have also been componentised to deal with this formatting problem. However this is more a short time solution and some method must be determined to standarize this data format process to extend the functionality of the present setup. Alse currently, it is very difficult for researchers not specialized in IT to use Grid technology. The motivation it is to provide an interface for researchers who are not specialized in IT to be able to easily make use of grid technology - the final goal of operationality being to use Grid technology without being aware of it.

6 Conclusions

As biological/bioinformatics applications challenge organizations to rethink their computing resources, grid computing offers a scalable, efficient infrastructure that

can help enable collaboration and accelerate research. We have demonstrated the possibility and advancement to this level in this paper. We have used the ICENI software to mask the complexity of managing data at different locations that is stored on multiple hardware systems and protected by a range of security models. By doing this, the e-protein can deliver a single environment with a unified data catalog that allows scientists to find and access the data available to them quickly and easily - so they can focus on their research, rather than spending time tracking down information.

References

[1] Altschul SF, Koonin EV. Iterated profile searches with PSI-BLAST-a tool for discovery in protein databases. /it Trends Biochem Sci. 1998 Nov;23(11):444-7.

[2] http://www.biodas.org

[3] http://www.lesc.ic.ac.uk/iceni

[4] N.Furmento,W.Lee, S.J. Newhouse, J.Darlington. Test and Deployment of ICENI, An Integrated Grid Middleware on the UK e-Science grid. *Nottingham, UK: All Hands Meeting, September 2003.*

[5] M.Y.Galperin, E.V. Koonin. Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement, and operon disruption. *In Silico Biol. 1, 0007, 1998.*

[6] K.Flemiing,A.Müller, R.M.MacCallum,M. J. E.Sternberg. 3D-GENOMICS: a database to compare structural and functional annotations of proteins between sequenced genomes *Nucleic Acids Res. Jan 1;32 Database issue:D245-50, 2004.*

[7] L.L.Conte, et al. SCOP: a structural classification of proteins database. *Nucleic Acid Res. 28, 257-259, 2000.*

[8] S.F.Altschul, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res. 25 (17), 3389-3402, 1997*

[9] L.A.Kelley, et al. Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol. 299, 501-522, 2000*

[10] M.Y.Gulamali, A.S.McGough, S.J. Newhouse, J.Darlington. Using ICENI to run parameter sweep applications across multiple Grid resources. *Berlin, GlobalGrid Forum 10 , March 2004.*

[11] L.Young, A.S.McGough, S.J. Newhouse, J.Darlington. Scheduling within ICENI. *Sheffield,UK e-Science All Hands Meeting, September 2003.*

[12] A common Job Description Markup Language written in XML: Http://www.lesc.ic.ac.uk/projects/jdml.pdf