Lightweight Solution for Protein Annotation

Shikta Das, Andrew S McGough, Jeremy Cohen, John Darlington

London e-Science Centre, Imperial College London, South Kensington Campus, London SW7 2AZ, UK
Email: lesc-staff@doc.ic.ac.uk

Abstract. Imperial College e-Science Networked Infrastructure (ICENI) is an end-to-end Grid middleware developed to allow transparent usage of Grid. It consists of both a service-oriented Grid middleware and an application toolkit, using a component-programming model to represent Grid applications. We utilise this infrastructure in the the e-Protein project, a *Biotechnology and Biological Sciences Research Council* pilot project, which provides structure-based annotation of proteins in the major genomes by linking remote resources at three different sites. The annotation utilises homology and fold recognition methods to assign protein structures to the proteomes. We have taken the approach of modularising applications into inter-connected components and running each application through a workflow manager, which executes the components in the specified manner. A web interface provides a user-friendly front end to the Grid system, where users can request their choice for the annotation and check previous annotations from the database.

1 Introduction

Sequenced genomes generate huge amounts of data. With the ever-increasing amount of sequenced genomes available, it is important that this data is not only structured but utilised and integrated with existing knowledge using a well-developed automated annotation system. Annotation of genomes is an example of a complex effort that requires integration of multiple methods for accurate identification of large numbers of genes.

Homologous proteins are two sequences which have similar functions or share a common ancestor. It is possible to predict the structure of a protein to discover the fold and hence information about the probable function of the sequence of a gene about which nothing is known, via homology to a sequence of known structure[1]. About 13% of the human proteome from a homologue of known structure was identified where the sequence alignments provided sufficient sequence for reliable homology modelling.

Structure-based annotation is an important part of this process. The structural data exposes the mechanisms behind biological functions and the evolutionary relationships that are hidden at the sequence level, but may be revealed at the structural level. The use of structural domains may prove a valuable tool for the biologist to design bench experiments (e.g. locating surface exposed residues on the protein for site-directed mutagenesis experiments)[2].

Integration of such large data with the use of many algorithms requires extensive computational resources beyond those available to researchers at a single location. The Grid provides an excellent platform for integrating a large set of resources capable of such large scale pipeline processing and an open architecture for decomposing results for distributed storage and relational retrieval.

The e-Protein project is a distributed pipeline for structure-based proteome annotation using Grid technology. The project aims to generate an automated distributed transparent concept for structure-based annotation of the proteins. The annotation pipeline utilises homology and fold recognition methods to generate 3D models for the protein structure. The objective is to establish local databases with structural and functional annotation and provide it to the biological community through a single web-based distributed system (DAS[3]) based at the European Bioinforamtics Institute in Cambridge. It will use emerging Grid technologies to share computing resources trans-

parently between three sites (Imperial College London - IC, European Bioinformatics Institute - EBI, University College London - UCL)[4].

This project also includes a comparison of alternative approaches for annotation, thereby identifying technological advances and improvements. In this paper we describe our approach towards the development of the project. At London e-Science Centre[5], Imperial College London, one of the institutes involved with the e-Protein project, we utilise the Imperial College e-Science Networked Infrastructure (ICENI), a Grid middleware in order to assist e-Bioinformaticians with the transparent sharing of their resources[6].

We have demonstrated in previous publications, the use of ICENI software in the e-Protein project[7]. The aim of this paper is to describe the extensions we have developed since. We begin by outlining the pipeline and database applied for structure-based proteome annotation. We then analyse the implementation of version 1.4 of ICENI and present the challenges faced during deployment. We have realised the need to incrementally build the annotation pipeline while keeping up with rapid development in the Grid Technology. Finally, we give a detailed description of the upcoming version of ICENI (ICENI II) used in the project. This section deals with the utlisation of ICENI II and how it is integrated with the existing design for the pipeline.

2 The Distributed Pipeline

For the e-Protein project at Imperial College London, Fleming *et al* have developed a system for automated large scale structural and functional annotation of proteins from completely sequenced genomes to provide comparative proteome analysis[2]. In the section below, we describe the structure-based annotation and the 3D-GENOMICS database where the annotations are stored. We are prototyping ICENI grid technology around this system to speed up the computationally intensive process of realiably annotating all of the proteomes available. The current system uses the beta version of ICENI II, which distributes jobs transparently across processing clusters and their associated servers. A demo of the distributed 3D-Annotate system may be viewed at http://www.lesc.ic.ac.uk/3D-Annotate.

2.1 The Structural-based Proteome Annotation Pipeline

The annotation pipeline is split into two steps, starting with a single sequence and finishing with a comparative analysis of proteomes. The first step provides information about the basic sequence features. The homologues are identified via BLAST and PSI-BLAST[8], IMPALA[9] and HMMer[10]. The sequence information is derived from different databases such as SwissProt[11], PIR[12], Pfam[13], PDB[14] and SCOP[15].

The Prosite database (i.e. a method for identifying the functions of uncharacterized proteins translated from genomic or cDNA sequences), is used to assign biologically significant patterns and profiles to determine which of the known family of proteins (if any) a new sequence belongs, or which known domain(s) it contains[16]. Further information from the sequence is derived using HMMTOP[17] for transmembrane helices, COILS2[18] for the coiled coils, SEG[19] for contrasting segments of low-complexity and high-complexity of the sequence and PSI-PRED[20] for secondary structure prediction.

Subsequently the aligned sequences are merged according to their regions within the query sequence and the nature of the homologues. This includes known domain structure from SCOP, sequences of known structure from PDB, close homologues from Swissprot, PIR or PDB. The data from the following steps are represented in a tabular summary for the entire genome, which provides information about the number of sequences, residues and regions (merged alignments) that can be assigned to each type of annotation, giving the user an option for comparing each alignment visually instead of tracing each one individually. Statistical summaries for the number of proteins containing SCOP classes, folds and superfamilies, and Pfam domains are also provided. This system specialises in providing various SCOP-based cross-proteome comparisons.

2.2 3D-GENOMICS Database

The proteome annotation is stored in the 3D-GENOMICS database, which provides the structural and functional information for the protein sequences[21]. It has a simple web interface allowing users to query the database in order to find genes within selected organisms that encode proteins with particular three-dimensional folds. At present there are 173 genomes available in the database, out of which homology models are available for the protein sequences of 13 genomes - *Homo sapiens, Saccharomyces cerevisiae, Aeropyrum pernix, Archaeoglobus fulgidus, Bacillus subtilis, Methanococcus jannaschii, Aquifex aeolicus, Chlamydia trachomatis, Escherichia coli K12, Haemophilus influenzae, Listeria innocua, Mycobacterium tuberculosis H37Rv, Mycoplasma genitalium.*

The protein annotation for each organism is loaded into tables within a MySQL relational database. The database is encapsulated by an object oriented software interface that manages the data stored in the database as well as performing sequence and proteome based analysis. All the underlying layers are Perl based. The software interface also allows transparent access to the database without requiring the user to know the structure of the underlying database. The developed system is generic and allows the integration of new analysis methods and source data.

The database is mainly intended for proteomes and is a combination of a portal and a platform for structure-based comparative analyses of proteomes. An important feature of the 3D-Genomics system is the decomposition of the output from the analysis software into several descriptive fields. For example PSI-BLAST output is not stored as a single raw text field, instead the informative parts of the output such as hits(homologues sequences), e-values, scores, sequence identities and alignments are extracted and stored as indexed fields. Relational queries can then be performed on these data-fields, allowing to link and relate results from different analyses. This allows benchmarking and optimization of PSI-BLAST for the recognition of remote homologues with less than 20% sequence identity[22]. Another unique feature of the database is the dynamic generation of constructed tables for statistics of SCOP superfamily composition between proteomes.

2.3 The Challenges with Version 1.4 of ICENI

ICENI is a service-oriented middleware. It utilises open and extensible XML schemas to encapsulate meta-data relating to resource capability, service interface and application behaviour. ICENI is used extensively within the e-Protein project to capture the workflow, and map it to a multitude of resources. Each Workflow element within ICENI is referred to as a component, which may have numerous implementations. Each implementation is annotated with meta-data describing its design, performance and requirements. This meta-data allows ICENI to select the most appropriate implementation of the component to ensure that the workflow is executed within the requirements specified by the user.

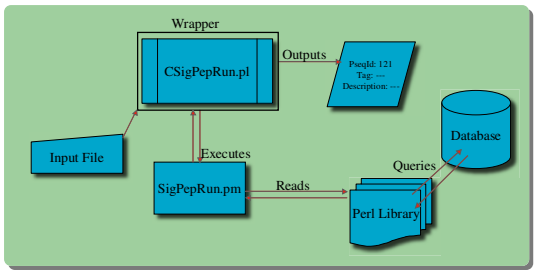

Fig. 1. The ICENI Workflow Pipeline

ICENI provides two main services in the e-Protein project, lauching and scheduling components onto resources. The inital prototype utilised version 1.4 of ICENI. The protein annotation pipeline consists of a series of programs that together form a Workflow. Each element of this workflow represents a program defined by a single operation in the protein annotation pipeline. ICENI treats each module as a binary component. A binary component wraps an existing executable and provides information about the execution and handling of input and output streams. A module can be decomposed into a perl executable, a pre-compiled binary executable and an input file (see Figure 1). For example, figure 1 represents

a module SigPepRun which runs the SEG application in the pipeline. The SigPepRun module is executed by a CSigPepRun wrapper which accepts an input file and presents the output to standard output. The perl executable (SigPepRun.pm) may potentially access the database through the perl database access library. A Job Description Markup Language (JDML) document is used to describe the location for each wrapper, input file and output file and presents the executable as a binary component in ICENI.

During the course of the project we have identified several problems in using ICENI (version 1.4) in the e-Protein architecture. We discovered that the implementation of ICENI 1.4 had become overburdened due to software decay, although it was able to perform the required job. It was also noted that ICENI 1.4 is complex to install and each feature in ICENI 1.4 required a full ICENI installation. The major problem was the execution of jobs over remote resources. ICENI (version 1.4) uses Java Jini as the underlying distributed architecture and communication channels are troublesome if firewall protection is in place.

With the Grid community moving towards adopting Web Services as the distributed architecture, there was a need to make certain changes in ICENI before it seems somewhat outdated by its original selection of the JINI architecture. Therefore, Lee *et al* attempted to restructure the main architecture for ICENI to incorporate web services and it is now superseded by ICENI II, which is a web service oriented middleware[23].

2.4 The Implementation of ICENI II

ICENI II is modular in design and is being developed on top of Web Services which gives us the option of deploying it from any remote location as long as it fullfills the underlying requirements. This has lead to the developement of a lightweight version of ICENI with Web Service job Submission and monitoring services based on an original prototype WS-JDML[24]. For this project, we have used one of the services featured in ICENI II, called GridSAM, which uses the upcoming Job Submission Description Language(JSDL) evolved through collaboration with the Global Grid Forum(GGF)[24]. Figure 2, explains the architecture used for e-Protein with GridSAM.

GridSAM takes a standard-based approach to job submission, funded by the Open Middleware Infrastructure Institute managed programme, as one of the first systems to adopt the Job Submission Description Language(JSDL) and Web Services[25]. GridSAM is an efficient tool for submitting jobs transparently onto the Grid and existing Distributed Resource Management(DRM) systems such as *Condor* [26] and *Sun Grid Engine* [27]. It can be easily deployed and configured on any Java Servlet compliant container, such as Apache Tomcat.

JSDL is an evolving specification standardising the vocabularies and schema for describing a job[25]. JSDL has emerged from Job Description Markup Language (JDML), a common job description language originally developed for use within the European Union Datagrid[29] based around Condor(ClassAds), which was used in the project[24]. A JSDL document is arranged into sections to describe the application and file staging. The application section describes the job to be executed, the environment for the execution, the path for the execution files and the path for the result files. This document can be extended to include information relevant to particular DRMs. The data staging section of a JSDL document provides the path for the file which contains information about the execution of the job and where to store the output files. The generated files can be transfered by GridFTP, FTP and HTTP transfer.

The architecture of GridSAM is composed of a Job Management Library and a Web Service interface, where both are independent of each other. The job management library is a JSDL consumer managing a job launching pipeline. This provides an adaptable framework that can be configured to interact with a a variety of DRM systems and cater for different availability requirements such as databases or file-based presistence of job states.

The Web Service interface is divided into two related port types for job submission and monitoring. The job submission port type consumes the submission action specified in a JSDL document. The submission port responds with a Uniform Resource Identifier (URI) that uniquely identifies the job that can be passed to the monitoring port to

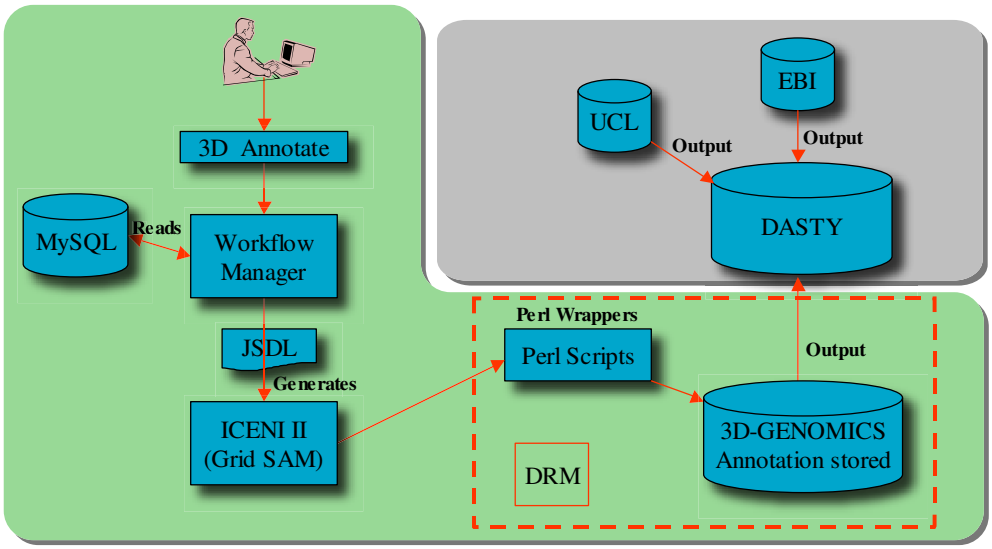

Fig. 2. The Architecture for e-Protein

retrieve job status. The job monitoring port type allows users to obtain information about the job status.

This infrastructure provides an efficient bridge between an e-Bioinformatician and the Grid for submitting a job transparently onto the Grid. Each job is submitted to GridSAM via a set of command-line interfaces. Alternatively, third-party clients can interact with the GridSAM services through the Web Service protocols. A control program has been developed to submit multiple jobs to GridSAM services according to the stages in the annotation pipeline. It reads in the output from each execution and pipes it to the next component specified by the user according to the pipeline. A web interface providing users with the ability to browse through different input options for the annotations and a link to DASTY viewer[30] have also been developed.

3 Discussion

Utilisation of ICENI II in e-Protein has provided an excellent platform to work on. It is an efficient bridge between users and the Grid for submitting jobs transparently onto heterogeneous resources. Since GridSAM can be deployed in a number of Java Servlet compliant platforms, this opens up the choice of vendor support depending on the performance and cost requirements. ICENI II is being developed into a number of separated composable toolkits, each of which can be used separately to perform Grid-related tasks, this feature of ICENI II provides us the opportunity to develop a tailor-made system for the e-Protein project.

Another advantage of using ICENI II is the DRMConnector abstraction that provides a pluggable submission pipeline to connect to a variety of Distributed Resource Management systems or novel Grid launching mechanisms (e.g. super-scheduler). The current GridSAM distribution supports launching mechanisms, such as *Condor*, *Secure Shell*, *Forking* and *Globus*. It also provides clustering and fault-tolerance by building on a range of industrial and open-source tools.

Thus with the developement of the lightweight ICENI II, the cumbersome nature of ICENI is removed making it easily transportable to multiple resources. By making the elements in ICENI decoupled, it reduces the resource footprint and the installation burden.

The web service interface of ICENI II uses HTTPS transport security and the WS-Security framework to protect message exchange as well as authenticating and authorising users. In a separate study, GridSAM has been shown to provide a

scalable and highly available job management solution[25]. Fault-tolerance is ensured by the long-term persistence of job states in the database.

4 Conclusion and Futher Work

In this paper we have discussed many issues that arise in managing the complex annotation process using existing Grid middleware in the e-Protein project. We have then described the implementation of the ICENI II toolkit, which provides a practical approach to Grid deployment, making it easier for the end users. We have tried to develop the system so that the user can easily use the Grid and is able to run their application on heterogenous resources.

The deployment of ICENI II in the e-Protein project has enabled e-Bioinformaticians to share and run their applications transparently across administrative domains. In future, we intend to provide regular updates of the database with new proteomes and versions of the software . Additionally, we would also like to investigate issues concerning network security and usability. Since the requirements of the project have changed over the course of the project, it is difficult to apply just one methodology to it. Our approach so far has provided us with a lot of flexibility in our work, although we hope to incorporate more advance features in the future.

5 Acknowledgements

The authors would like to thank all other members of the London e-Science Centre, especially William Lee for his help with the description of the architecture for GridSAM and Oliver Jevons for his help with the diagrams. We would also like to thank Prof Michael Sternberg and Dr Keiran Fleming at Structural Bioinformatics Group, Imperial College London for their help and expertise. e-Protein is a Biotechnology and Biological Sciences Research Council(BBSRC -28/BEP17014 March 2002) funded project under its e-Science programme.

References

1. *J.Gough, K. Karplus, R. Hughey and C. Chothia. Assignment of Homology to Genome Sequences using a Library of Hidden Markov Models that Represent all Proteins of Known Structure. J. Mol. Biol.,2001, 313, 903-919*
2. *K. Fleming, A. Muller,R.M. MacCallum, M.J.E. Sternberg. 3D-GENOMICS: A database to compare structural and functional annotations of proteins between sequenced genomes. Nuc. Acid Res., 2004, 32, D245-D250*
3. *http://www.biodas.org*
4. *http://www.e-protein.org*
5. *London e-Science Centre. http://www.lesc.ic.ac.uk.*
6. *Imperial College e-Science Network Infrastructure Project. http://www.lesc.ic.ac.uk/iceni/.*
7. *A. O?Brien, S. Newhouse, J. Darlington. Mapping of Scientific Workflow within the e-Protein project to Distributed Resources. UK e-Science All Hands Meet, 2004, ISBN 1-904425-21-6.*
8. *S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nuc. Acid Res., 1997, 25, 3389-3402.*
9. *A.A. Schaffer, Y.I. Wolf, C.P. Ponting, E.V. Konnin, L. Aravind, S.F. Altschul. IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. Bioinformatics, 1999, 15, 1000-1011.*
10. *S.R. Eddy. Profile hidden Markov models. Bioinformatics,1998, 14, 755?763*
11. *B. Boeckmann, A. Bairoch, R. Apweiler, M.C. Blatter, A. Estreicher, E. Gasteiger, M.J. Martin, K. Michoud, C. O'Donovan, I. Phan, et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nuc. Acid Res., 2003, 365-370.*
12. *C.H. Wu, L.S. Yeh, H. Huang, L. Arminski, J. Castro-Alvear, Y. Chen, Z. Hu, P. Kourtesis, R.S. Ledley, B.E. Suzek, et al. The Protein Information Resource. Nuc. Acid Res., 2003, 31, 345-347.*
13. *A. Batmen, E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S.R. Eddy, S. Giffiths-Jones, K.L. Howe, M. Marshall, E.L. Sonnhammer. The Pfam protein families database. Nuc. Acid Res., 2002, 30, 276-280.*
14. *H.M Berman, T. Battistuz, T.N. Bhat, W.F. Bluhm, P.E. Bourne, K. Burkhardt, Z. Feng, G.L. Gilliland, L. Iype, S. Jain, et al. The Protein Data Bank. Acta Crystallogr., 2002, D, 58, 899-907.*

15. *L. LoConte, S.E. Burenner, T.J. Hubbbard, C. Chothia, A.G. Murzin. SCOP database in 2002: refinements accomadation strucutral genomics. Nuc. Acid Res., 2002, 30, 264-267.*
16. *L. Falquet, M. Pagni, P. Bucher, N. Hulo, C.J. Sigrist, K. Hofmann, A. Bairoch. The PROSITE database, its status in 2002. Nuc. Acid Res., 2002, 30, 235-238.*
17. *G.E. Tusnady, I. Simon. The HMMTOP transmembrane topology prediction server. Bioinformatics, 2001, 17, 849-850.*
18. *A. Lupas, M. Van Dyke, J. Stock. Predicting coiled coils from protein sequences. Science, 1991, 252.1162-1164.*
19. *J.C. Wootton, S. Federhen. Analysis of compositionally biased regions in sequence databases. Methods Enzymol.,1996, 266, 554?571.*
20. *L.J. MuGuffin, K. Bryson, D.T. Jones. The PSIPRED protein structure prediction server. Bioinformatics, 2000, 16, 404-405.*
21. *http://www.sbg.bio.ic.ac.uk/3dgenomics/*
22. *A. Muller, R.M MacCallum, M.J. Sternberg. Benchmarking PSI-BLAST in genome annotation. J. Mol. Biol.,1999, 293, 1257?1271.*
23. *W. Lee, A.S. McGough, J. Darlington. ICENI II. Proceedings of UK e-Science All Hands Meet, 2005.*
24. *W. Lee, S. McGough, S. Newhouse, J. Darlington. A Standard Based Approach to Job Submission Through Web Services. UK All-hands e-Science Conference, 2004, September 2004.*
25. *W. Lee, A.S. McGough, J. Darlington. Performance Evaluation of the GridSAM Job Submission and Monitering System. Proceedings of UK e-Science All Hands Meet, 2005.*
26. *The ClassAd Language Reference Manual Version 2.1. http://www.cs.wisc.edu/condor/classad/refman//.*
27. *http://www.sun.com/software/gridware/.*
28. *Job Submission Description Language Working Group. https://forge.gridforum.org/projects/jsdl-wg.*
29. *The DataGrid Project. http://www.eu-datagrid.org/.*
30. *http://www.e-protein.org/e-proteindastypr.html.*
31. *J.M. Chandonia, N.S. Wlaker, L. LoConte, P. Koehl, M. Levitt, S.E. Bernner. ASTRAL compendium enhancements. Nuc. Acid Res., 2002, 30, 260-263.*
32. *A. Muller, R.M. MacCallum, M.J. Sternberg. Benchmarking PSI-BLAST in genome annotation. J. Mol. Biol., 1999,293, 1257-1271.*
33. *N.J. Mulder, R. Apweiler, T.K. Attwood, A. Bairoch, D. Barrell, A. Bateman, D. Binns, M. Biswas, P. Bradley, P. Bork, et al. The InterPro Database, 2003 brings increased coverage and new features. Nuc. Acid Res., 2003, 31, 315-318.*
34. *M. Clamp, D. Andrews, D. Barker, P. Bevan, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, et al. Ensembl 2002: accomodating comparative genomics. Nuc. Acid Res., 2003, 31, 38-42.*
35. *J. Cohen, A.S. McGough, J. Darlington, N. Furmento, G. Kong, A. Mayer. RealityGrid: An Integrated Approach to Middleware through ICENI. Royal Society, 2005, in press.*