Data Access and Integration in the ISPIDER Proteomics Grid

Lucas Zamboulis[1,2], Hao Fan[1,2,⋆], Khalid Belhajjame[3], Jennifer Siepen[3], Andrew Jones[3], Nigel Martin[1], Alexandra Poulovassilis[1], Simon Hubbard[3], Suzanne M. Embury[4], and Norman W. Paton[4]

[1] School of Computer Science and Information Systems, Birkbeck, Univ. of London
[2] Department of Biochemistry and Molecular Biology, University College London
[3] Faculty of Life Sciences, University of Manchester
[4] School of Computer Science, University of Manchester

Abstract. Grid computing has great potential for supporting the integration of complex, fast changing biological data repositories to enable distributed data analysis. One scenario where Grid computing has such potential is provided by proteomics resources which are rapidly being developed with the emergence of affordable, reliable methods to study the proteome. The protein identifications arising from these methods derive from multiple repositories which need to be integrated to enable uniform access to them. A number of technologies exist which enable these resources to be accessed in a Grid environment, but the independent development of these resources means that significant data integration challenges, such as heterogeneity and schema evolution, have to be met. This paper presents an architecture which supports the combined use of Grid data access (OGSA-DAI), Grid distributed querying (OGSA-DQP) and data integration (AutoMed) software tools to support distributed data analysis. We discuss the application of this architecture for the integration of several autonomous proteomics data resources.

1 Introduction

Grid computing technologies are becoming established which enable distributed computational and data resources to be accessed in a service-based environment. In the life sciences, these technologies offer the possibility of analysis of complex distributed post-genomic resources. To support transparent access, however, such heterogeneous resources need to be integrated rather than simply accessed in a distributed fashion. This paper presents an architecture for such integration and discusses the application of this architecture for the integration of several autonomous proteomics resources.

Proteomics is the study of the protein complement of the genome. It is a rapidly expanding group of technologies adopted by laboratories around the world as it is an essential component of any comprehensive functional genomics

⋆ Currently at International School of Software, Wuhan University, China.

U. Leser, F. Naumann, and B. Eckman (Eds.): DILS 2006, LNBI 4075, pp. 3–18, 2006.

study targeted at the elucidation of biological function. This popularity stems from the increased availability and affordability of reliable methods to study the proteome, as well as the ever growing numbers of tertiary structures and genome sequences emerging from structural genomics and sequencing projects.

The *In Silico Proteome Integrated Data Environment Resource* (ISPIDER) project[1] aims to develop an integrated platform of proteome-related resources, using existing standards from proteomics, bioinformatics and e-Science. The integration of such resources would be extremely beneficial for a number of reasons. First, having access to more data leads to more reliable analyses; for example, performing protein identifications over an integrated resource would reduce the chances of false negatives. Second, bringing together resources containing different but closely related data increases the breadth of information the biologist has access to. Furthermore, the integration of these resources, as opposed to merely providing a common interface for accessing them, enables data from a range of experiments, tissues, or different cell states to be brought together in a form which may be analysed by a biologist in spite of the widely varying coverage and underlying technology of each resource.

In this paper we present an architecture which supports the combined use of Grid data access (OGSA-DAI), Grid distributed querying (OGSA-DQP) and data integration (AutoMed) software tools, together with initial results from the integration of three distributed, autonomous proteomics resources, namely gpmDB[2], Pedro[3] and PepSeeker[4]. The emergence of databases on experimental proteomics, capturing data from experiments on protein separation and identification, is very recent and we know of no previous work that combines data access, distributed querying and data integration of multiple proteomics databases as described here.

Paper outline: Section 2 gives an overview of the OGSA-DAI, OGSA-DQP and AutoMed technologies and introduces the three proteomics resources we have integrated. Section 3 discusses the development of the global schema integrating the proteomics resources within the ISPIDER project, Section 4 presents our new architecture, Section 5 discusses related work and Section 6 gives our conclusions and directions of further work.

2 Background

2.1 OGSA-DAI and OGSA-DQP

OGSA-DAI (Open Grid Services Architecture - Data Access and Integration) is an open-source, extendable middleware product exposing data resources on Grids via web services [2]. OGSA-DAI[5] supports both relational (MySQL, DB2,

[1] See http://www.ispider.man.ac.uk
[2] See http://gpmdb.thegpm.org
[3] See http://pedrodb.man.ac.uk:8080/pedrodb
[4] See http://nwsr.smith.man.ac.uk/pepseeker
[5] See http://www.ogsadai.org.uk/