

Grid Portals for Bioinformatics

Lavanya Ramakrishnan

Mark S.C. Reed

Jeffrey L. Tilson

Daniel A. Reed

{lavanya, markreed, jtilson, reed}@renci.org

Renaissance Computing Institute (RENCI)

University of North Carolina, Chapel Hill, NC 27599

Abstract. *The explosive growth of biological and biomedical research requires a new set of software tools and a computational environment that includes high performance computing and large-scale data management. These software environments need to be easy-to-use and scalable to support the diverse requirements of educators and researchers. This paper describes the design philosophy and architecture of a bioinformatics portal that operates atop standard Grid infrastructure and tools such as the Open Grid Computing Environment (OGCE) and Globus toolkit. The Bioportal integrates domain-specific tools and standard community tools to provide an integrated, collaborative environment. We also discuss the experiences with deploying the portal for educational and research users in North Carolina and as part of the NSF TeraGrid.*

1. Introduction

The rise of quantitative biology – the application of quantitative methods to analyze observational data and test hypotheses – and the explosive growth of biomedical research are due in large part to increased involvement of multidisciplinary teams, creation of large-scale computational models, mining of distributed data archives and application of high-throughput instrumentation to capture large volumes of biological and biomedical data. As science and computational science become increasingly synonymous [30], the pace of biomedical discovery and understanding can be further accelerated via judicious use of high-performance computing, Grid infrastructure for resource coupling, federated data management technologies and collaborative tools. However, more robust software tools are needed that simplify access to distributed data resources, provide uniform access to diverse and rapidly growing suites of biomedical applications, and enable the use of collaboration suites for team building and project interactions. Lowering the entry barrier to the use of integrated software environments and (all too often) tool idiosyncrasies can broaden the base of participation and allow researchers, educators and students to focus on research and

education rather than on the computing infrastructure. As in business and consumer domains, web portals provide an integrated gateway to resources and functionality that may be geographically and politically dispersed, but that are logically connected.

This paper describes the design, implementation and preliminary experience with a bioinformatics portal – a Bioportal – that operates atop standard Grid infrastructure [3] and tools but is specialized to the needs of biomedical educators and researchers. As part of a growing movement to build domain-specific science portals using open source toolkits, the Bioportal is one of a suite of “science gateways” that constitute a nascent national cyberinfrastructure [24]. In this distributed model, each member of a disciplinary community hosts some dedicated infrastructure (e.g., a data collection or instrument) and operates a local gateway for access to user-accessible resources, both as individuals and as members of larger, virtual organizations. For example, a group of scientists working on a specific biology problem might access local microarray databases in addition to remote resources (e.g., the NSF TeraGrid [14]) for large-scale data analysis.

The remainder of this paper is organized as follows. In §2-§3, we describe our target communities and the Bioportal design philosophy. In turn, §4 describes the Bioportal architecture and implementation, its software stack and its middleware capabilities. Next, §5 discusses deployment of the prototype as a resource for researchers and educators in North Carolina and as part of the NSF TeraGrid. Finally, §6-§7 summarize related work and concluding observations.

2. Bioportal Target Communities

The Bioportal’s capabilities are based on requirements gleaned from discussions and collaborations with two distinct but related biology and biomedical communities: (a) molecular biology and (b) genetics and genotype-phenotype analysis. The first effort targets researchers and educators in North Carolina [1] and, via the NSF

TeraGrid, biological and biomedical researchers across the United States. The second effort targets data federation and correlation for understanding the interplay of multiple genes in disease [2] – the genetic basis of complex disease phenotypes such as cancer, heart disease and addiction. By exploratory genetic analysis -- testing hypotheses, correlating and statistically analyzing data from multiple sources (e.g., model organisms, research studies and clinical drug trials) – one can identify common traits and characteristics underlying human diseases.

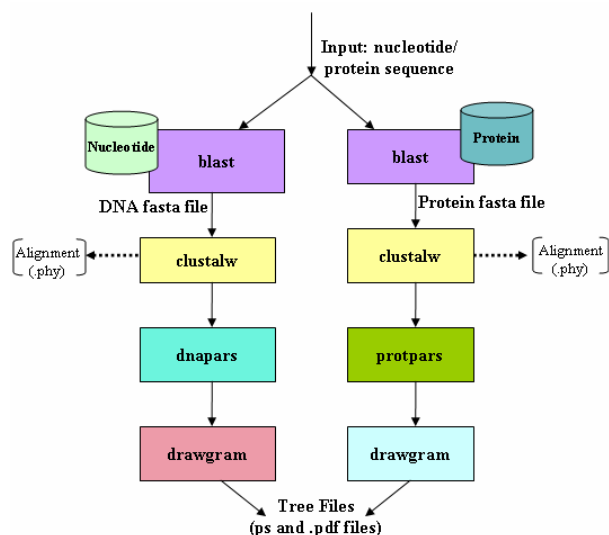


Figure 1 The Gene2Life workflow

As an example of molecular biology analysis, let us consider the Gene2Life workflow that is a multi-step application sequence or *workflow* that is available through the portal. This workflow takes an input sequence, searches databases to find genes matching the sequence. It globally aligns the results and attempts to correlate the results based on organism and function. Figure 1 depicts the steps of the workflow and the corresponding output at each stage. In this workflow the user provides a sequence that can be a nucleotide or an amino acid. The input sequence performs two parallel BLAST[29] searches, against the nucleotide and protein databases respectively. The results of the searches are parsed to determine the number of identified sequences that satisfy the selection criteria. The outputs trigger the launch of ClustalW[31] that is used for the global alignment process to identify relationships. These outputs are then passed through parsimony programs for analysis. The two applications that

maybe available for such analysis are `dnapars` and `protpars`[35]. In the last step of the workflow plots are generated to visualize the relationships, using an application called `drawgram` [35].

Specific communities have disparate needs and only a partial overlap in the underlying tools and databases. Nevertheless, the infrastructure needed to support them is similar – customized access to models, data and resources by individuals and distributed teams. As the above example showed, studying a single sequence requires knowledge and use of different tools, each with a different set of parameters and invocation method. The bioinformatics portal combines access to common databases and these and other computational analysis tools within a standard, “learn once, use many times” interface. In addition, the Bioportal provides educators and researchers the capability to simplify these multi-step processes that are used in the science domain. This requires a new level of integrated data storage, analysis, and exploration, as well as distributed collaboration infrastructure for secure sharing of data and tools. Such integration is the goal of the Bioportal.

3. Bioportal Philosophy and Enabling Technologies

The Bioportal builds on insights and lessons learned from earlier science portals. These include portals constructed as part of the NSF-funded National Computational Science Alliance expeditions [25], which developed domain-specific portals for chemical engineering, atmospheric modeling, and astronomy, as well as related projects [15]. These experiences identified a set of principles that shape the design and implementation of all science gateways, including the Bioportal:

Ease of use. The most important goal is to make the portal and associated infrastructure easy-to-use by educators, students and researchers. The application interfaces must hide the complexities of Grid technologies and focus attention on the details of the task, its inputs and its subsequent analysis.

Scalability. The explosive growth of experimental data and the need for more computing resources to support models and data analysis necessitate use of a scalable, distributed architecture. Hence, the Bioportal is based on Grid technologies and

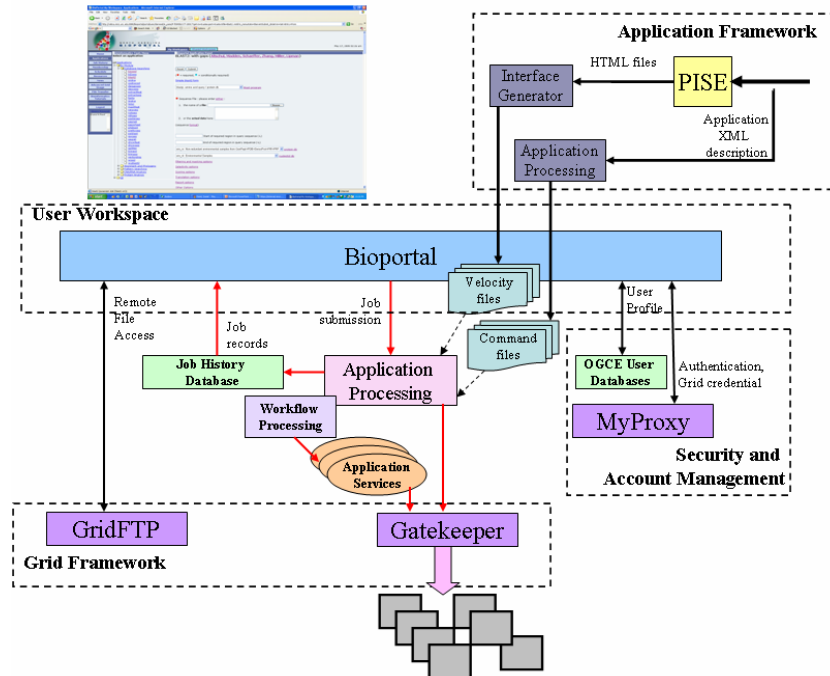


Figure 2 Biportal User and Technology Interaction

leverages Grid and web services standards and tools.

Extensibility. No design can anticipate all user needs. A portal infrastructure should allow users to add applications to the framework using a simple XML mechanism. A default set of tools and data sets are included in the Biportal, with an extension interface for user addition of new bioinformatics tools.

Sharing. Ready access to standard tools and public data enables scientists to solve new research problems. The Biportal enables sharing of the resources and data, allowing collaboration across many communities.

With this background, the Biportal builds on the Open Grid Computing Environment (OGCE) [8], the PISE [6] application description infrastructure, standard Grid middleware from the Globus toolkit [4], common bioinformatics software and databases, and workflow extensions based on Taverna [33]. An outgrowth of the NSF Middleware Initiative (NMI) [26], OGCE is a toolkit of reusable portal components that can be combined to form a common portal container. Using PISE, the Biportal integrates a biological application framework with OGCE, making available a large and diverse suite of biological and biomedical applications. The framework is modular and extensible, allowing both site-specific

policies and addition of new applications. Via Taverna [33], the Biportal supports distributed web services for biomedicine, allowing composition of services for complex workflow analysis. Collectively, these middleware tools provide

- standard access to commonly used bioinformatics applications and data for education and research,
- access to distributed cyberinfrastructure such as Grids and web services,
- easy integration of new applications and data into the framework, and
- a scalable open source software stack that can be distributed to the bioscience community.

4. Biportal Architecture

Figure 2 shows the four primary components of the Biportal and their interaction: the user workspace, application framework, grid framework, and security and account management. As the name suggests, the user workspace is the interaction point for user actions with backend services. The application framework allows one to describe the interface and processing logic needed to interpret the user's inputs. In turn, the Grid, security and account management components coordinate authentication, job submission and file transfer.

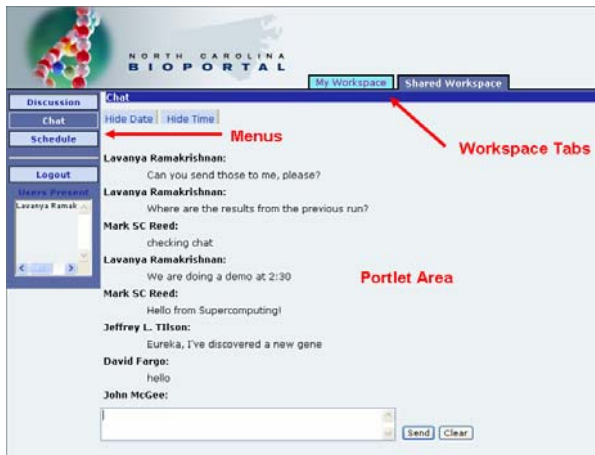


Figure 3 Bioportal Portlet-Driven Interface

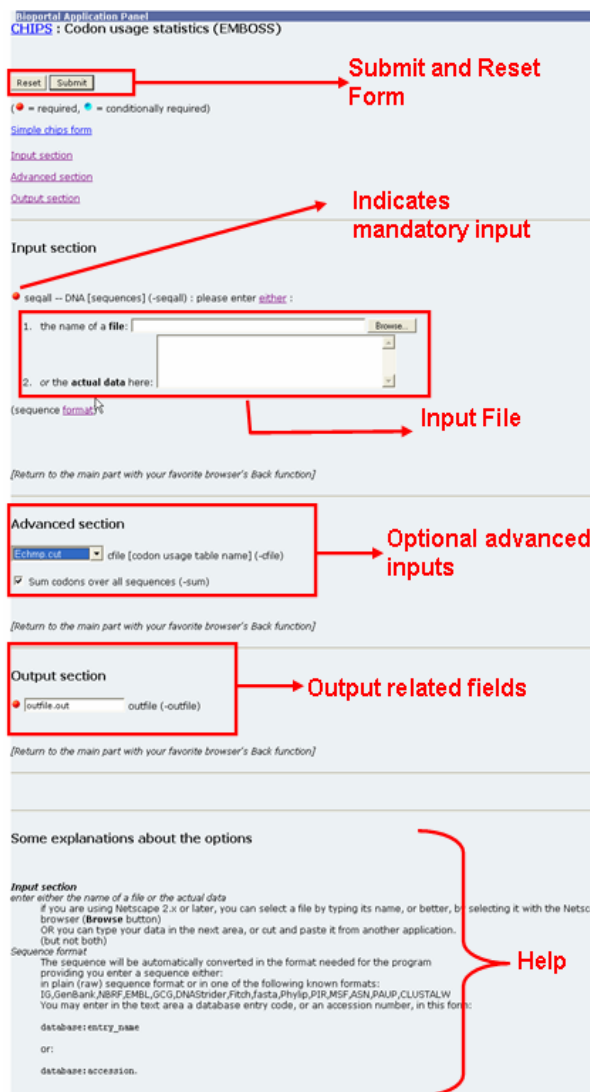


Figure 4 Application Form for Application "Chips"

Below, we describe the technology components and their interaction in detail.

4.1. User Workspace

The Bioportal is based on the notion of a "portlet," a portal server component that controls a user-configurable pane in the user's web browser. A portal server supports a set of web browser frames, each containing one or more portlets that provides a user service. This portlet component model allows one to construct portals by instantiating a portal server with a domain-specific set of portlets, complemented by domain-independent portlets for collaboration and discussion.

Concretely, the Bioportal extends the OGCE framework, which itself builds on the *CompreHensive collaborativE Framework* (CHEF) [22], Jetspeed [27] and Velocity [28]. We use collaborative portlets from CHEF to support community building and information sharing, the Jetspeed portal services for user interface customization, caching, persistence and user authentication, and Velocity's template engine for application specification. The choice of the portal software was guided by the choices available at the inception of the project and the stability needed for a production system. OGCE provided us the ability to leverage portlets deployed in collaborative production environments while providing a clear path for transition to standards compliance through OGCE2.

Figure 4 shows the Bioportal user interface and the different regions in the user workspace. Typically the user has access to one or more workspaces as a virtue of group membership. The workspace tabs are at the top of the screen. In each workspace, a set of portlets available to the user are shown via a menu on the left. Conceptually each user has a "personal workspace" and a "shared workspace." The personal workspace, akin to a personal desktop, gives access to desktop utilities (e.g., the latest bioinformatics news through RSS feeds and a personal calendar access to the applications). In turn, the shared workspace is accessible by all community users and is intended to foster collaborations through chat, discussion and group schedule portlets.

Finally, the user can access application-specific bioinformatics tool interfaces (Figure 4). These interfaces are rendered in Velocity, which allows users to select from among the available

applications, then configure and submit jobs through the application panel. The job history portlet allows users to view previously submitted jobs and thus access corresponding input and output files. Together, the combination of a personal and shared environment allows users to run applications on distributed resources, then share and discuss the results with other colleagues and collaborators.

```
<pise> <head>
<title>CHIPS</title><description>
Codon usage statistics(EMBOSS)
</description>
<category>nucleic:codon
usage</category>
<doclink>http://www.uk.embnet.org/Software/EMBOSS/Apps/chips.html</doclink>
</head>
<command>chips</command>
<parameters> &emboss_init;
```

Table 1 PISE XML Header (CHIPS Example)

4.2. Application Framework

The Velocity-based application framework allows users to select, configure and launch applications through the portal. Figure 4 illustrates an application panel that displays the application input

form. These applications may execute locally or on any remote resource where the user has valid Grid credentials.

This extends the desktop metaphor by coupling powerful, though distributed resources to support large-scale computations, while hiding the details of job scheduling and resource allocation.

The bioinformatics interfaces are based on PISE [6], a tool that generates web interfaces for biomedical applications using XML descriptions of the application inputs. The Pasteur Institute's PISE distribution includes XML descriptions for roughly 200 applications for biological database search, comparison programs, gene finding and modeling, RNA analysis, phylogeny and pattern discovery. This extensible framework also simplifies addition of new applications by writing a single XML description file. In the PISE framework, an XML specification describes the interface and the logic needed to process application inputs. In the Biportal, the HTML forms synthesized by PISE are transformed and imported as Velocity templates. These XML descriptions are processed once when the portal is deployed and a configuration file describing each

```
<parameter type="Paragraph"> <paragraph><name>input</name>
<prompt>Input section</prompt>
<parameters><parameter type="Sequence" ismandatory="1" issimple="1"
ishidden="0">
  <name>seqall</name>
  <attributes>
    <prompt>seqall -- DNA [sequences](-seqall)</prompt>
    <format>
      <language>perl</language><code>" -seqall=$value
-sformat=fasta"</code>
    </format>
    <group>1</group><seqtype><value>dna</value></seqtype>
    <seqfmt> <value>8</value></seqfmt>
    <pipe><pipetype>seqsfile</pipetype>
    <language>perl</language><code>1</code>
    </pipe>
  </attributes>
</parameter></parameters></paragraph></parameter>
```

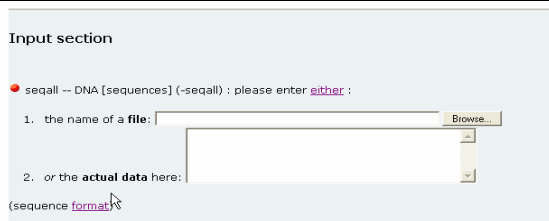


Table 2 PISE XML Configuration (CHIPS Example)

application's command line interface is generated. When the user submits an application form, the user input is processed in conjunction with the preprocessed configuration file to generate the executable command line for the job.

As an example, Figure 4 shows the interface generated for EMBOSS 'chips', a program that gives codon usage statistics [34]. The figure shows the input panel the user sees, generated from the XML configuration file. In more detail, Table 1 shows the header of the XML file that describes the application. The document starts with the <prise> tag, followed by HTML specifying the application's name, description, broad menu classification and URL for more information. The <command> tag indicates that the executable code for this application is named "chips."

Table 2 shows the XML for the application input and the corresponding, synthesized interface. The <prompt> tags define the text to be displayed in the HTML form in the appropriate location (the heading "Input section" and the text "seqall -- DNA [sequences] (-seqall)"). The following XML segment identifies that the input is a sequence file and the value entered by the user (either as an entry in the textbox or by uploading a file) will be substituted for \$value in the <code> segment of "-seqall=\$value".

Similar configuration specifications define advanced inputs, output file disposition and other configuration details. User inputs from the synthesized GUI are processed in conjunction with the details from the XML specification to generate the application's command line, for example:

```
chips -seqall=input.dat -sformat=fasta -sum -outfile=outfile.dat -auto -stdout
```

4.3. Workflow Support

One of the goals of the Bioportal is to provide a framework that will allow users to compose the available application as a workflow. However, most times users like to repeatedly use pre-configured workflows with different inputs. Hence we provide some offline-composed workflows to the users. The workflow backend is implemented using the Taverna [33] open-source workflow tool. Taverna is a workflow tool developed as part of the ^{my}Grid[36] project that allows users to compose and run complex analysis using components available both locally and at remote locations

through standard web service interfaces. The Bioportal uses Taverna and a web service toolkit called the Generic Service Toolkit[32] to deploy the applications dynamically with a web service interface. In turn, these web services are composed into workflows. The workflow input interface is described using PISE thus treating the workflow as a meta-application. One such workflow that has been integrated into the portal is Gene2Life that is described in Section 2. The scientific community is being engaged to deploy additional workflows in the Bioportal from different research domains. While the workflows may be from slightly different scientific domain, the generality of the infrastructure allows rapid development and deployment of additional workflows.

4.4. Security and Account Management

The Bioportal provides security at multiple levels, beginning with the Secure Socket Layer (SSL) to encrypt user communications with the portal server. For access to remote services, we leverage the Grid Security Infrastructure (GSI) and its public key cryptography. Globus uses GSI to authenticate and map user identities to user accounts. In turn, Globus X.509 certificates are stored in MyProxy [9], a credential repository that allows users to access credentials from any resource. Finally, we have built an auditing and accounting system for tracking user accounts and resource use.

To access the portal and grid services, every user needs a portal account, a user certificate and accounts on the underlying computing resources. Because we wished to hide the details of remote resource authentication and access, we built a simple web account management system that allows users to select a username and password during account creation.

This system generates a public-private key pair and stores it in the account database. An email is then sent to the user asking him/her to confirm the request. After the request is confirmed, the administrator is notified. If the user has a valid email from an educational institution or a non-profit institution, he/she is granted an account. Once the administrator approves the account, a UNIX account is created, a user certificate is loaded in MyProxy and the OGCE databases are populated with both user account data and access rights defining the set of accessible portlets.

4.5. Grid Framework

The Bioportal uses existing grid technologies and protocols to manage jobs and files, namely the Grid Security Infrastructure (GSI) [17], Grid Resource Allocation and Management (GRAM) [4] and GridFTP [16]. As described earlier, user credentials are stored in a globally accessible credential repository. The OGCE portal user interface uses the MyProxy server to authenticate users and obtain proxy certificates (short-lived certificates) for use with Globus. Thus, the MyProxy server provides single sign-on capability from the portal.

The portal performs Grid operations on behalf of the user using the proxy. When the user submits an application, we select an available resource provider based on a load balancing scheme. The Globus gatekeeper at the selected resource provider then manages job submission, execution and monitoring using GRAM. The gatekeeper can optionally connect to site-specific batch schedulers (e.g., PBS or LSF). The GRAM protocol also allows users to query the status of the jobs or associate listeners for invocation when the job status changes.

The portal framework also manages the data associated with submitted applications. A unique directory is created for every application execution, and the portal manages the uploaded input files and any application output. Both the job history and output files are accessible through the job history portlet. Additionally, GridFTP can be used for high-performance, secure, reliable data transfer.

5. Current Deployments and Experiences

Our Bioportal infrastructure has been operational for over a year, supporting both local infrastructure and more recently the NSF TeraGrid. Below, we describe (a) our hosting infrastructure, the application suite and databases and (b) experiences and enhancements to support the TeraGrid.

5.1. Default Portal Infrastructure

The North Carolina instantiation of the Bioportal is deployed on a small (34 node) Linux cluster (a head node, 32 compute nodes and a storage node) each with two 3.06 GHz Xeon processors, 4 GB of memory/node and a Gigabit Ethernet interconnect. A complementary 1.73 TB storage array hosts the bioinformatics databases. The cluster configuration

is similar to one supported by most Grid sites today.

The current Bioportal deployment supports roughly 140 bioinformatics applications, including suites such as EMBOSS (European Molecular Biology Open Software Suite)[34], GLIMMER (Gene Locator and Interpolated Markov Modeler) [10], HMMER (Hidden Markov Model program for profile-based sequence analysis) [11], NCBI (National Center for Biotechnology Information) [9], PHYLIP (PHYLogeny Inference Package for inferring phylogenies) [35] and other applications such as ClustalW [31] and FASTA [37]. The bioinformatics applications depend on a set of standard databases of about 300G that are regularly updated on the host systems.

5.2. TeraGrid Science Gateway

Recently, we extended the Bioportal to act as an NSF TeraGrid Science Gateway. This extension provides access to high-end computing systems at the TeraGrid resource providers. Specific enhancements included improved data staging, community account management and stronger audit tracking.

To enable remote job submission to distributed TeraGrid sites, we enhanced our job submission module to handle input and output data staging. Initially the input files are staged to the remote resource and the job execution is started. When the job completes, a listener retrieves the files and stores them on the local portal file server. One of our design principles for the TeraGrid science gateways is simplified access to high-end computing resources to broaden the base of scientific computing users. To support such scaling, science gateways provide “community accounts” for TeraGrid use without registration at each resource provider. Much as one can browse the products on many e-commerce portals without registration, a TeraGrid community account allows users to run small jobs without cumbersome registration. To support this, the Bioportal framework manages user verification, authentication and authorization. Each Bioportal user has a login and password that allows us to track each action in the portal to a particular user. When a job is submitted to the TeraGrid, the backend services retrieve the community account proxy. This associates the community account with a user’s job while preserving the user identity for the remainder of the session. As described earlier,

the data belonging to a user is only stored on the local portal file server, where there is stronger identity association, preserving user data privacy.

In addition, the Bioportal framework must maintain audit tracking data should a security incident on a resource need to be tracked and a user identified. An administrator at a TeraGrid resource provider can log into the portal and query the history to match a community account job to the actual user who submitted that job. More standardized mechanisms for accounting are being investigated in the context of the TeraGrid.

5.3. Experiences

The Bioportal cluster was first deployed for community use during the summer of 2005. Since then, it has been repeatedly extended and enhanced with new applications, integrated workflows and access to TeraGrid resources. The Bioportal has users drawn from diverse biomedical and bioinformatics communities as well as use education and outreach. The Bioportal deployment, not unlike other science gateways, raised interesting policy questions regarding user job and data management. Key questions include (a) the balance between encouraging open access and community building and strong authentication for access to unique national resources, (b) load balancing for complex workflow execution, (c) data replication, mirroring and staging for execution efficiency and (d) the balance between simplicity of use and expressive power. We believe the basic Bioportal architecture is sufficiently flexible to accommodate a diverse set of policy solutions to these questions.

Among the science gateway development community, philosophical debates continue regarding how much user knowledge should be required about underlying resources. Our design principle has been to hide that complexity as we work to extend use of high-performance computing in the biological and biomedical communities. Our experience to date has validated this design principle. Biomedical researchers prize simplicity of use and the ability to focus on complex data analysis. While hiding the resource complexity is important, it is equally important that the users have access to application-specific data from the resource. Examples include knowing when databases were last updated or what version of each application is available on a certain resource.

With a single entry point gateway, the aggregated information is readily available to the user.

6. Related Work

Domain-specific portals are increasingly recognized as critical enablers for scientific discovery. Below, we describe our work relative to other bioinformatics tools and grid technologies.

Many tools exploit Common Gateway Interfaces (CGI) to access bioinformatics applications and databases, including National Center for Biotechnology Information (NCBI) and Biology Workbench [18]. The Biology Workbench allows biologists to search popular protein and nucleic acid sequence databases and access analysis and modeling tools. These portals provide application-specific interfaces but do not provide mechanisms to store results and view previous jobs. The next generation Biology Workbench is being designed and implemented to support tools in a way similar to the Bioportal's application framework.

Portal toolkits, including GridSphere [20] and OGCE [8], are used to build portals for specific scientific communities. The OGCE toolkit and its components have been used to build portals for some of the largest distributed science and engineering infrastructure projects, including the NSF TeraGrid, the NEESgrid earthquake engineering Grid [21], the LEAD atmospheric sciences Grid [12] and the DOE Science Grid. The Bioportal is similar to other Grid-based portals in providing access to Grid resources using Globus services. However, the Bioportal is customized for a community with applications integrated using XML descriptions. This provides extensibility, allowing users and sites to add specific applications of interest.

Finally, there is some overlap in our account management system and the Portal-based User Registration Service (PURSE). PURSE [7] defines a user registration and certification process but does not allow users to manage account details (e.g., changing passwords after creation). The Bioportal provides an integrated account management system with site-specific policies.

7. Conclusions and Future Work

The Bioportal provides tools for access to distributed data and resources via a standard and extensible infrastructure for application integration and execution. The PISE XML architecture enables addition of applications for specific user

groups, maintaining a standard look and feel while supporting tool diversity. This reduces the learning curve for new tools and leverages distributed environments. As the Biportal user community continues to grow, we will add new workflows and tools, in response to user feedback. In addition, we are working with teachers and other educational outreach groups to expose new communities to biology and biomedical research. We also plan to transition to newer technologies including support for Globus 4.0 and porting to JSR168 compliant portlals.

Acknowledgements

The Biportal was developed in part with seed funding from the University of North Carolina's Office of the President, which targeted development of advanced research and education applications in high-performance computing, information systems and computational and computer science. Biological and biomedical enhancements were funded in part by the National Science Foundation and the National Institutes of Health under grants 05-03697 and 5-P20-RR020751-01-02, respectively. The Biportal development was a team effort. The authors extend their thanks to the entire Biportal team, which includes members of the Renaissance Computing Institute (RENCI), UNC-Chapel Hill's Information Technology Services (ITS) and Center for Bioinformatics, and Wake Technical Community College.

References

1. North Carolina Biportal, 2006. <http://www.ncbiportal.org>
2. The Carolina Center for Exploratory Genetic Analysis (CCEGA),2006. <http://www.renci.org/projects/ccega.php>
3. I. Foster, C. Kesselman and S. Tuecke, "The Anatomy of the Grid: Enabling Scalable Virtual Organizations," *International Journal of Supercomputer Applications*, 15(3), 2001
4. I. Foster and C. Kesselman, "Globus: A Metacomputing Infrastructure Toolkit," *International Journal of Supercomputer Applications*, 11(2):115-128, 1997
5. J. Novotny, S. Tuecke and V. Welch, "An Online Credential Repository for the Grid: MyProxy," *Tenth International Symposium on High Performance Distributed Computing* (HPDC-10), August 2001
6. C. Letondal, "A Web Interface Generator for Molecular Biology Programs in Unix," *Bioinformatics*, 17(1), pp 73-82, 2001
7. PURSE, Portal Based User Registration Service, 2006. <http://www-unix.grids-center.org/r6/ecosystem/security/purse.php>
8. Open Grid Computing Environment, 2006, <http://www.collab-ogce.org/nmi/index.jsp>
9. National Center for Biotechnology Information (NCBI), 2006 <http://www.ncbi.nlm.nih.gov/>.
10. Gene Locator and Interpolated Markov Modeler (GLIMMER), 2006. <http://cbcb.umd.edu/software/glimmer/>
11. HMMER(Hidden Markov Models), 2006.<http://hmmer.wustl.edu>
12. K. K. Droegemeier et. al., , "Service-Oriented Environments In Research And Education For Dynamically Interacting With Mesoscale Weather," *IEEE Computing in Science and Engineering*, Volume 7, Issue 6, pp 12-29 November 2005
13. J. Novotny, "The Grid Portal Development Kit, Grid Computing: Making the Global Infrastructure a Reality", in The Grid Portal Development Kit, Grid Computing, John Wiley, March 2003.
14. C. Catlett, "The Philosophy of TeraGrid: Building an Open, Extensible, Distributed Terascale Facility," *Second IEEE/ACM International Symposium on Cluster Computing and the Grid* (CGRID2002), p. 5, 2002
15. M. Russel, G. Allen, I. Foster, E. Seidel, J. Novotny, J. Shalf, G. von Laszewski and G. Dues, "The Astrophysics Simulation Collaboratory: A Science Portal Enabling Community Software Development. Proceedings of High-Performance Distributed Computing," *Tenth International Symposium on High Performance Distributed Computing* (HPDC-10), pp. 207-215, 2001
16. W. Allcock, J. Bester, J. Bresnahan, A. L. Chervenak, I. Foster, C. Kesselman, S. Meder, V. Nefedova, D. Quesnal and S. Tuecke, "Data Management and Transfer in High Performance Computational Grid Environments," *Parallel Computing*, 28 (5), pp. 749-771, May 2002
17. I. Foster, C. Kesselman, G. Tsudik and S. Tuecke, "A Security Architecture for Computational Grids," *Fifth ACM Conference on*

- Computer and Communications Security*, pp. 83-92, 1998
18. S. Subramaniam, "The Biology Workbench: A Seamless Database and Analysis Environment for the Biologist," *Bioinformatics, Proteins*, 32, 1-2, 1998
 19. K. Michalickova¹, G. D. Bader, M. Dumontier, H. Lieu, D. Betel, R. Isserlin, and C. Hogue, "SeqHound: biological sequence and structure database as a platform for bioinformatics research" *BMC Bioinformatics*, 3:32, 2002.
 20. J. Novotny, M. Russell and O. Wehrens, "GridSphere: A Portal Framework for Building Collaborations", 1st International Workshop on Middleware for Grid Computing," *ACM/IFIP/USENIX Middleware*, June 2003.
 21. L. Pearlman, C. Kesselman, S. Gullapalli, B.F. Spencer Jr, J. Futrelle, K. Ricker, I. Foster, P. Hubbard and C. Severance, "Distributed Hybrid Earthquake Engineering Experiments: Experiences with a Ground Shaking Grid Application," *13th IEEE International Symposium on High Performance Distributed Computing (HPDC-13 '04)* pg. 14-23.
 22. P.A. Knoop, J. Hardin, T. Killen and D. Middleton, "The Comprehensive collaborative Framework (CHEF)", *AGU Fall Meeting Abstracts 2002*
 23. W. Allcock, S. Tuecke, I. Foster, A. Chervenak and C. Kesselman. "Protocols and Services for Distributed Data-Intensive Science," *ACAT2000 Proceedings*, pp. 161-163, 2000
 24. D. Atkins, "Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure," *Technical Report*, <http://www.nsf.gov/od/oci/reports/atkins.pdf>, 2003
 25. Science Portal Alliance Expedition, 2006. <http://www.extreme.indiana.edu/alliance/>.
 26. NSF Middleware Initiative, 2006. <http://www.nsf-middleware.org>
 27. The Jakarta Jetspeed Project, 2006. <http://portals.apache.org/jetspeed-1/>.
 28. The Jakarta Velocity Project, 2006. <http://jakarta.apache.org/velocity/>.
 29. S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", *Nucleic Acids Research*. 25:3389-3402, 1997.
 30. Computational Science: Ensuring America's Competitiveness, Report of the Computational Science Subcommittee, *President's Information Technology Advisory Committee (PITAC)*, June 2005
 31. J.D. Thompson, D.G. Higgins, and T.J. Gibson, CLUSTALW: Improving the Sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22:4673-4680, 1994.
 32. G. Kandaswamy, L. Fang, Y. Huang, S. Shirasuna, S. Marru, and D. Gannon. Building Web Services for Scientific Grid Applications. *IBM Journal of Research and Development*, 50(2/3):249-260, 2006.
 33. T. Oinn, M. Addis, J. Ferris, D. Marvin, M. Greenwood, T. Carver, M.R. Pocock, A. Wipat, P.Li. Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, 20, 3045-3054.
 34. P. Rice, I. Longden, A. Bleasby. EMBOSS: the European Molecular Biology Open Software Suite, *Trends in Genetics*, 16, (6) pp276-277, 2000
 35. J. Felsenstein, PHYLIP Phylogeny Inference Package (Version 3.2). *Cladistics* 5: 164-166., 1989.
 36. R. Stevens, A. Robinson, and C.A. Goble. myGrid: Personalised Bioinformatics on the Information Grid. *Proceedings of 11th International Conference on Intelligent Systems in Molecular Biology*, 2003.
 37. W.R. Pearson. Flexible sequence similarity searching with the FASTA3 program package, *Methods Mol. Biology*. 132:185-219, 2000.