



Grid Portals for Bioinformatics

Lavanya Ramakrishnan, Mark S.C. Reed,
Jeffrey L. Tilson, Daniel A. Reed



lavanya@renci.org

Acknowledgements



The UNIVERSITY of NORTH CAROLINA
A 16-campus university

- **Funding**
 - North Carolina : Funded by UNC Office of the President
 - TeraGrid: NSF
- **Renaissance Computing Institute (RENCI)**
 - J. Coyle, K. Gamiel, K. Green, X. Guan, C. Jeffries, G. Kandaswamy, H. Lander, J. McGee, N. Nassar, E. Riehl, J. Reily, S. McLean, M. Rynge, E. Scott, B. Viviano, etc
- **Wake Technical Community College**
 - S. Harrison, T. Chagnon, M. Seda, B. Black
- **UNC-CH Information Technology Services (ITS)**
 - R. Marinshaw, J. Knott, J. Williams, C.D. Poon, M. Shoffner, S. Moffat, Paul Mitchell, etc
- **UNC-CH Center for Bioinformatics**
 - H. Kelkar, T. Randall, D. Fargo



... and many more!



University of North Carolina at Chapel Hill
Center for Bioinformatics

The Challenge and Need

- **Challenge**
 - the rise of quantitative biology
 - burgeoning bioinformatics data
 - complex analysis and modeling problems
 - education and training in new technologies
- **Reality**
 - diverse tools with idiosyncratic interfaces
 - steep learning curves
 - software development by diverse groups
- **Need**
 - integrated, easy-to-use toolset with standard interfaces
 - extensible mechanisms that hide idiosyncrasies
 - tool and bioinformatics training
 - infrastructure support and extension
 - bioinformatics principles and issues
 - economic development and training enabler



Solution: Bioportal

- **The solution**
 - a bioinformatics portal and coupled training
- **Features**
 - access to common bioinformatics tools
 - extensible toolkit and infrastructure
 - OGCE and National Middleware Initiative (NMI)
 - leverages emerging international standards
 - remotely accessible or locally deployable
 - packaged and distributed with documentation
- **Education and training**
 - hands-on workshops
 - clusters, Grids, portals and bioinformatics



North Carolina



www.tgbiportal.org
www.ncbiportal.org

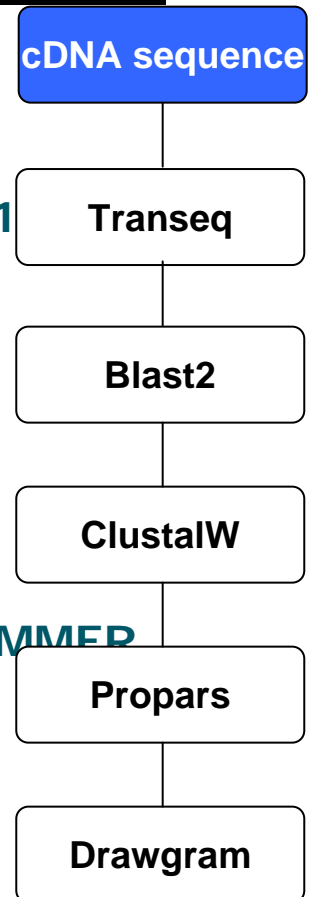
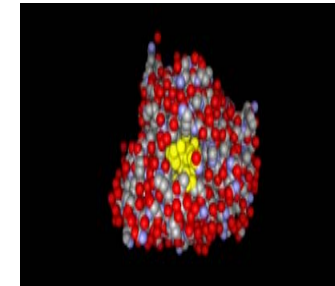
Science Communities

- **North Carolina Bioportal**
 - leverage state-wide investment in bioinformatics and grid
 - undergraduate education, graduate education, faculty research
- **The Carolina Center for Exploratory Genetic Analysis**
 - develop collaborative experiences and plans
 - preliminary data to apply for a P50 grant
 - develop a prototype informatics infrastructure
 - data models, methods, tools and portals
 - facilitate use of best practices for existing projects



Bioportal Target Audiences

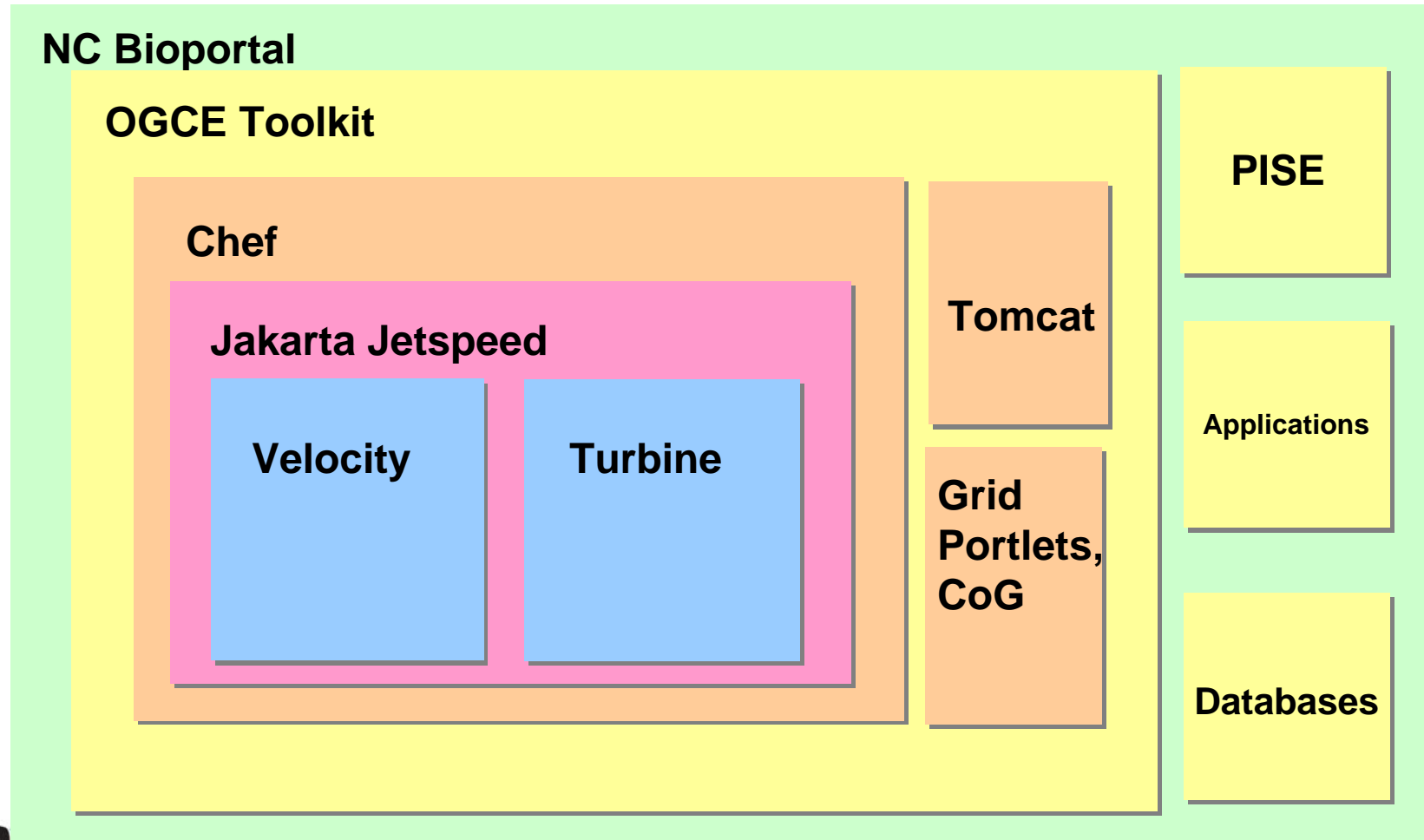
- Three overlapping groups
 - undergraduate education
 - graduate education and research
 - faculty research
- K-12 education opportunities are implicit
 - via faculty-directed enrichment
- Undergraduate education example
 - examine the biophysical properties of the human Cox1
- Graduate education and research example
 - identify protein coding region from DNA sequence
 - examine related proteins using BLAST
 - align related proteins
 - examine phylogenetic relationships
 - generate tree
- Faculty research example
 - tools for identification of prokaryotic genes using GLIMMER
 - use the *Bacillus cereus* genome
 - as a template for gene discovery in *Bacillus anthracis*



Outline

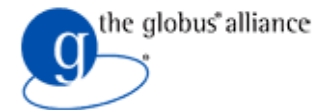
- ✓ Overview
- Bioportal Architecture
 - Technologies
 - Workflow Support
- Experiences
- Conclusions and Future Work

Putting the Technologies Together



Biportal Technologies

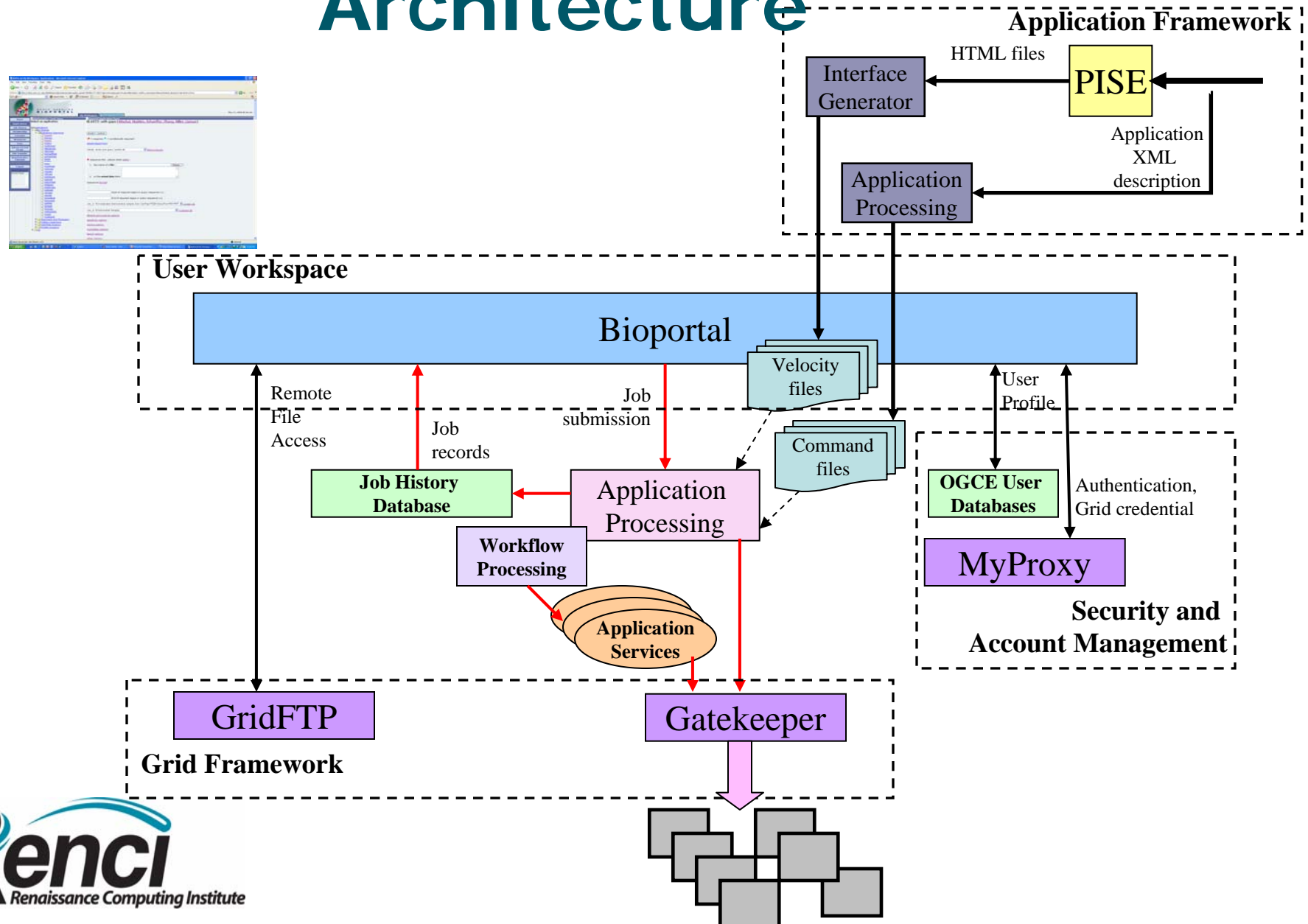
- **Globus**
 - Job Submission through Gatekeeper
 - GridFTP for file transfer
 - MyProxy credential repository
- **Open Grid Computing Environment (OGCE)**
 - access to other Globus functionality through Java
 - many application uses
 - LEAD, NEES, TeraGrid, NC Bioportal
- **Michigan Chef/Sakai**
 - collaborative course tool
 - later used for distributed communities (e.g., NEE)
- **Tomcat, Jetspeed, Velocity, Turbine**
 - Apache Java servlet container
 - enterprise information portal
 - Java-based template engine
 - web application framework



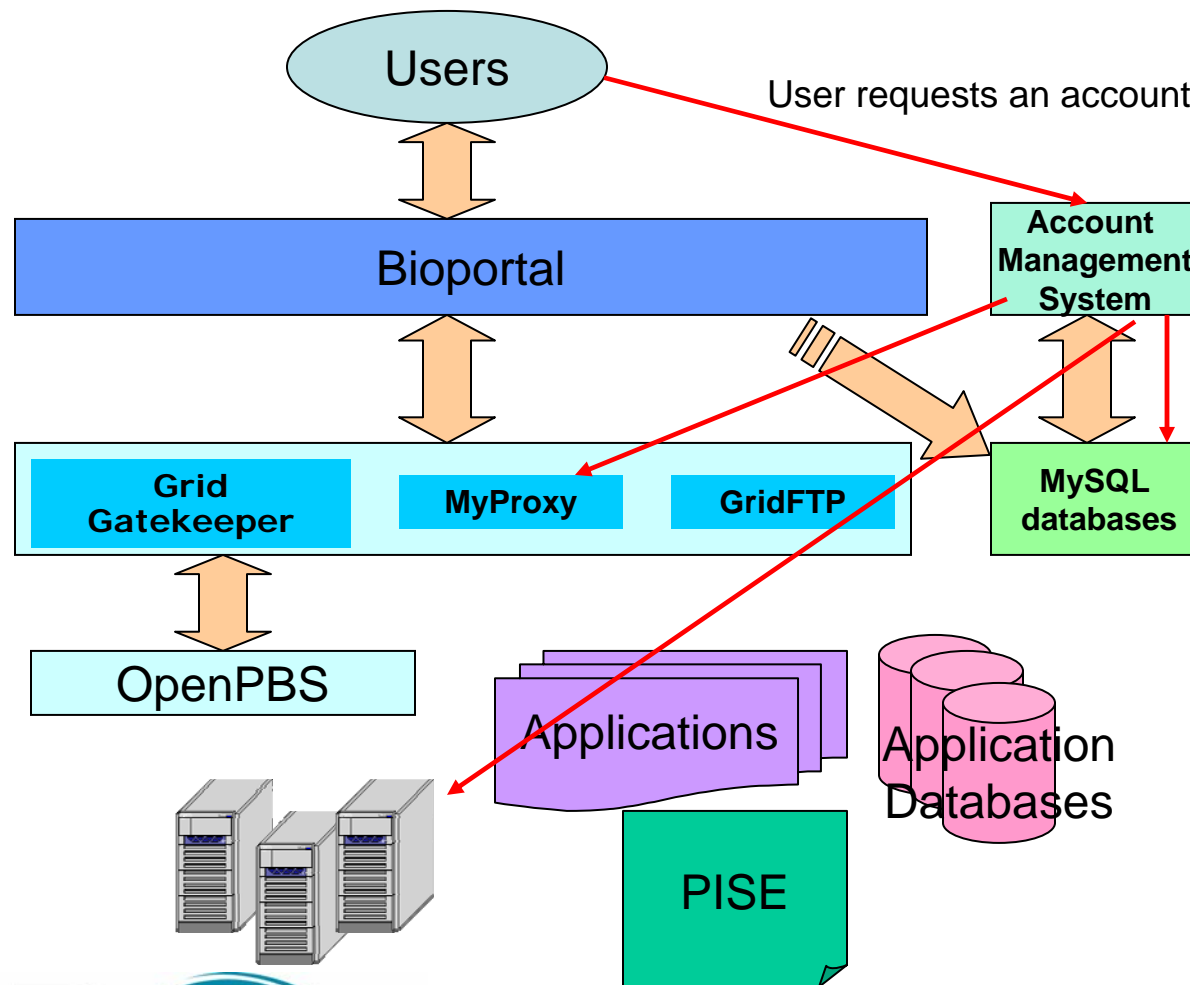
The **Apache Jakarta Project**

<http://jakarta.apache.org/>

Architecture



Account Management



Behind the scenes

1. Create Unix account
2. Create a certificate request
3. Sign the certificate request
4. Update MyProxy
5. Add entry to gridmap file
6. Create a portal account

Community Accounts

- Single account for all jobs,
- Per user on portal
- Maintain audit tracks
- Admin portlet access

PISE

- **Pasteur Institute Software Environment (PISE)**
 - generates web interfaces for molecular biology tools
 - XML specification for command line interface
- **Rationale and objectives**
 - simplify specification of program interfaces
 - homogeneous specification mechanisms
 - reuse of existing software interfaces
 - independent development and integration
 - extension for integration with graphical interfaces
 - complexity hiding and commonality
- **Bioportal program described in PISE**
 - semi-automated GUI synthesis from XML via Perl
- **Output is a generated command line**
 - Example: `blastall -p blastp -d env_nr -i query.dat.blast2.1116248106513 -a 2`



An Example PISE XML

Report Application Panel
CHIPS: Codon usage statistics (EMBOSS)

Reset Submit

(● = required, ● = conditionally required)

[Simple chips form](#)

[Input section](#)

[Advanced section](#)

[Output section](#)

Input section

● seqall -- DNA [sequences] (-seqall) : please enter [either](#) :

1. the name of a file: Browse

2. or the actual data here:

(sequence format)

[Return to the main part with your favorite browser's Back function]

Advanced section

cfile [codon usage table name] (-cfile)

☒ Sum codons over all sequences (-sum)

[Return to the main part with your favorite browser's Back function]

Output section

● outfile.out outfile (-outfile)

[Return to the main part with your favorite browser's Back function]

Some explanations about the options

Input section
enter either the name of a file or the actual data
if you are using Netscape 2.x or later, you can select a file by typing its name, or better, by selecting it with the Netscape file browser (Browse button)
OR you can type your data in the next area, or cut and paste it from another application.
(but not both)

Sequence format
The sequence will be automatically converted in the format needed for the program providing you enter a sequence either:
in plain (raw) sequence format or in one of the following known formats:
IG, GenBank, NBRF, EMBL, GCG, DNASTrider, Fitch, fasta, Phylip, PIR, MSF, ASN, PAUP, CLUSTALW.
You may enter in the text area a database entry code, or an accession number, in this form:

database:entry_name
or:
database:accession.

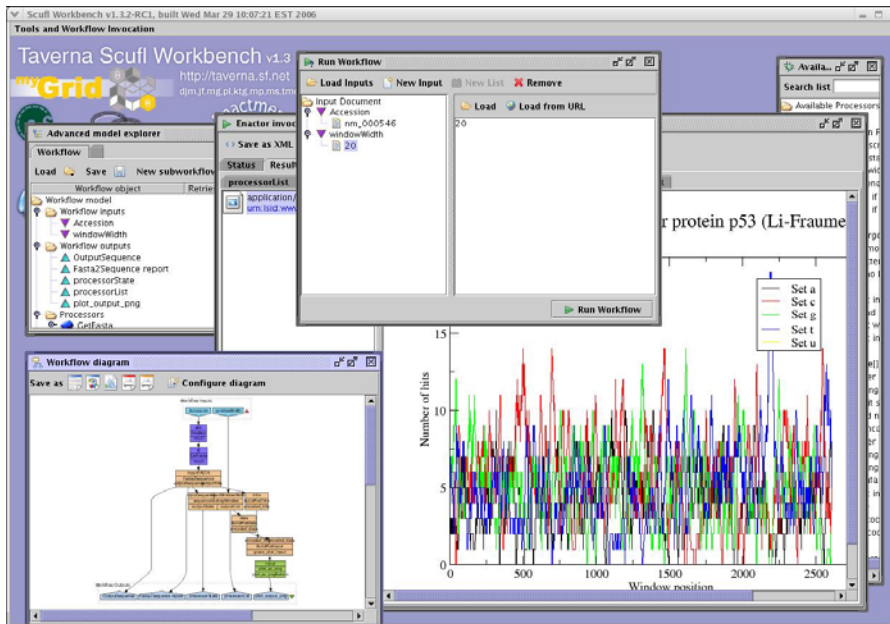
Input Section

```
<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>

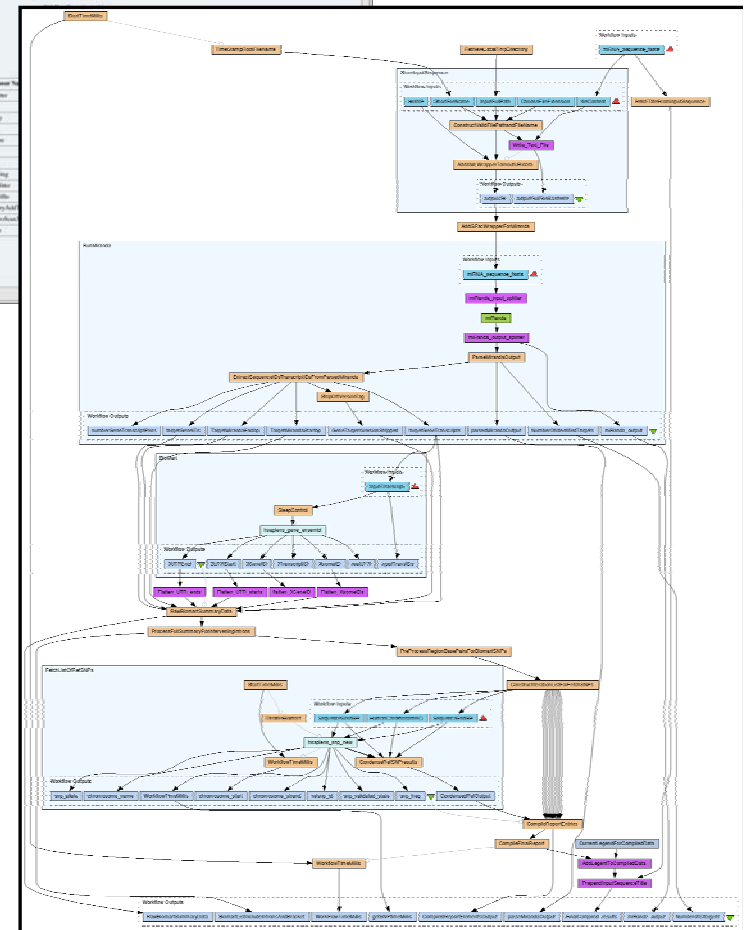
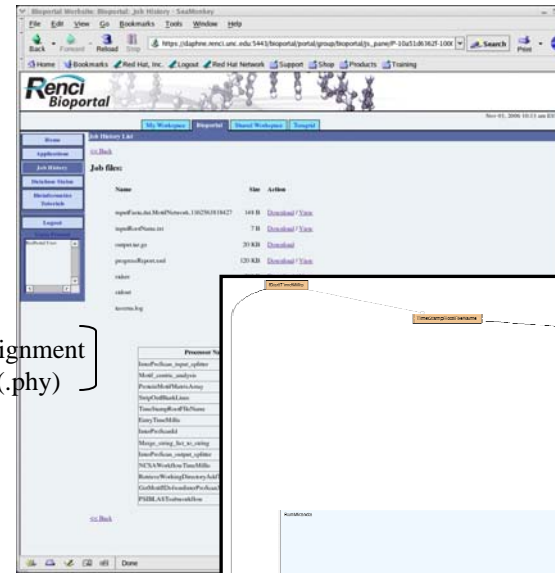
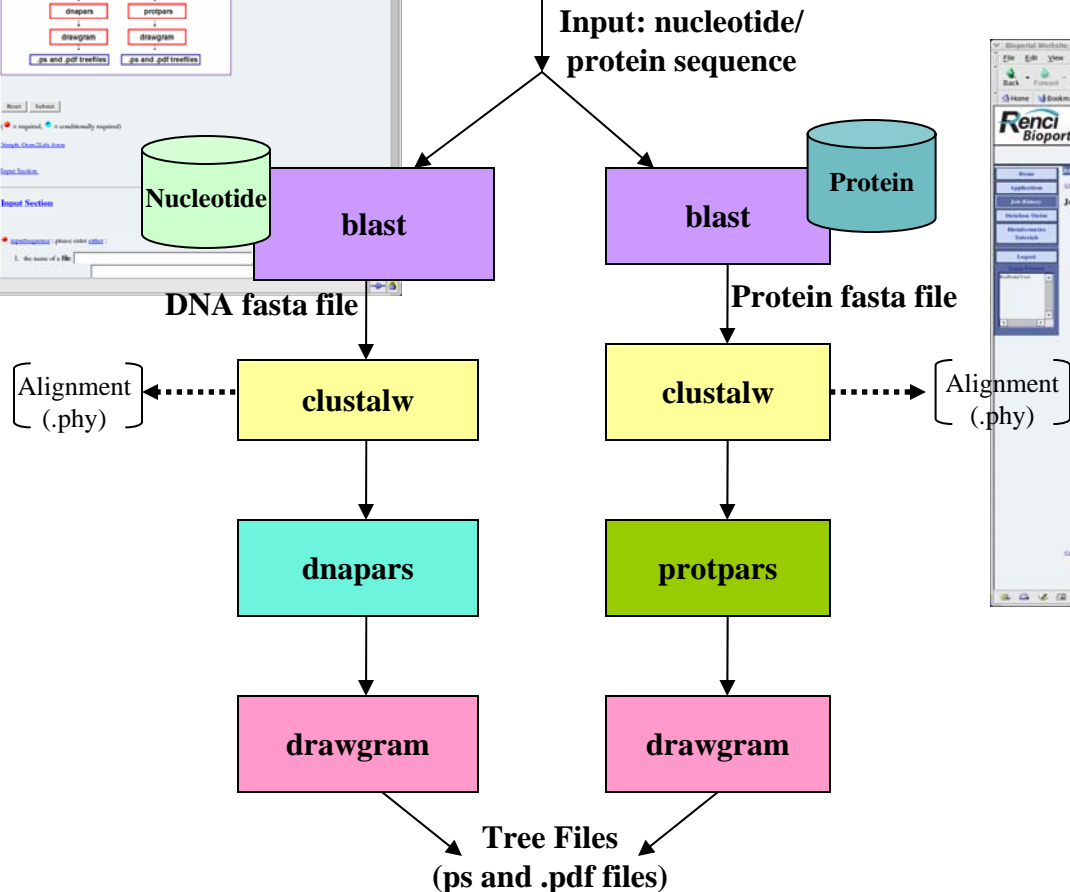
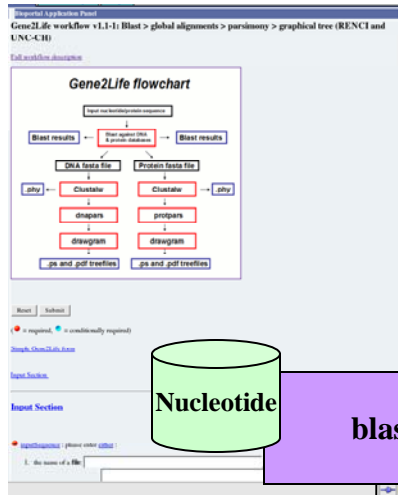
        <language>perl</language> <code>1</code>
      </pipe>
    </attributes>
  </parameter></parameters></paragraph>
</parameter>
```

Workflow Support Using Taverna

- Coordination of multiple applications
 - ability to repeat
- Taverna from ^{my}Grid
 - support for Bioinformatics workflows
 - link web services
- Generic Service Toolkit from IU
 - Used to wrap PISE applications
- First iteration: pre-composed workflows
 - Gene2Life, etc



Workflow Examples




- Gene2Life (Bioinformatics)
- MotifNetwork (NCSA)
- SNPsFromMicroArray (CCEGA)

Outline

- ✓ **Overview**
- ✓ **Bioportal Architecture**
 - **Technologies**
 - **Workflow Support**
- **Experiences**
- **Conclusions and Future Work**

Current Bioportal Applications

- Applications (~140 distinct codes)
- Application Suites 
 - **EMBOSS**
 - European Molecular Biology Open Software Suite
 - **GLIMMER**
 - gene identification in microbial DNA
 - **HMMER**
 - Hidden Markov Model program for profile-based sequence analysis
 - **NCBI**
 - diverse set of tools
 - **PHYLIP**
 - PHYLogeny Inference Package for inferring phylogenies
- Others (incomplete list)
 - ClustalW, FASTA
 - mpiBLAST (soon)

Standard bioinformatics databases

- **NCBI Aggregate (95 GB)**
 - three formats: native, BLAST and WUBLAST
- **GenBank (206 GB)**
- **GenPept (3 GB)**
- **PDB (6.3 GB)**
- **Prints (72 MB)**
- **RepBase (8.6 MB)**
- **UniProt (12 GB)**
- **PFam (8.7 GB)**
- **ProSite (16 MB)**
- **TransFac (36 MB)**

Database update mechanism

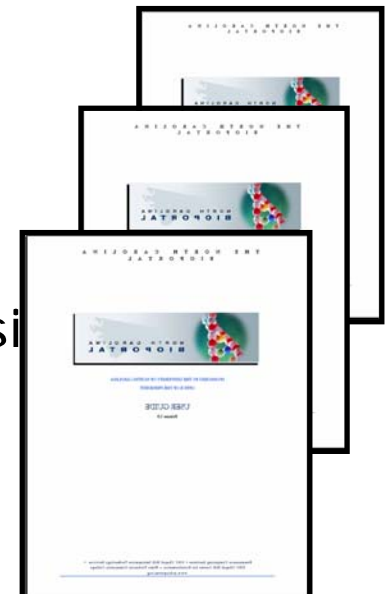
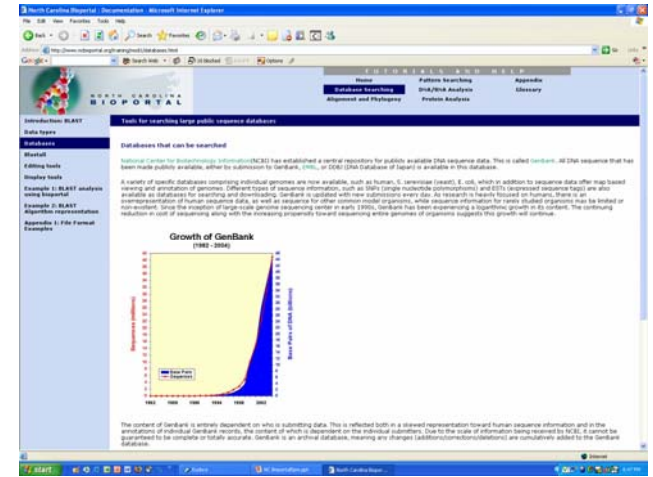
- follows the schedule of the distribution source
- currently NCBI Aggregate is the only one updated nightly

Bioportal Experiences

- **Infrastructure**
 - 34 node Linux cluster/ 1.73 TB storage array
 - TeraGrid sites
- **Enthusiasm from different user communities**
 - simplicity was/is the watchword
- **Account management complexities**
 - simplified account management software created
 - manage community account in the backend
- **Interesting policy issues**
 - balance between encouraging open access and community building and strong authentication for
 - load balancing for complex workflow execution
 - data replication, mirroring and staging for execution efficiency
 - the balance between simplicity of use and expressive power.

Education and Outreach

- **Bioportal Training Materials**
 - database searching
 - alignment and phylogeny
 - pattern searching
 - DNA/RNA analysis
 - protein analysis
- **Packaged and distributed**
 - for local installations
- **Workshops**
 - ongoing at various places
- **Capstone material for courses**
 - **Department of Animal Science, NC A&T**
 - ANSC 665: Techniques in Biotechnology
 - ANSC 771 Bioinformatics & Genome Analysis
- **Future: Bioportal DVD tutorial**



Conclusions and Future Work

- Provide standard tools to access distributed data and resources
 - reduces the learning curve
- Support new technologies, services
 - Globus 4.0, OGCE 2.0, TeraGrid accounting
- Infrastructure
 - closer integration with Taverna
 - allow users to compose workflows
 - mirror essential Bio web services
 - dynamic job scheduling across multiple sites
 - load driven based on community use
 - Open Science Grid
- Portal application suite
 - expand workflows, application and databases based on user feedback



GRIDtoday Names RENCi's Bioportal Top Life Sciences Grid Implementation

CHAPEL HILL, NC, September 12, 2006 – The North Carolina/TeraGrid Bioportal, developed by the Renaissance Computing Institute, was recognized by the readers of *GRIDtoday* in the publication's inaugural Readers' and Editors' Choice Awards.

Related Events at SC|06

- **RENCI Booth (# 1143) Near TACC and CISCO booths**
 - **Tuesday, November 14, 3-4 pm**
 - M Reed. Workflows for Genetics and Melanoma Research
 - **Tuesday, November 14, 4-5 pm**
 - J. McGee. The North Carolina and TeraGrid Bioportal
 - **Wednesday, November 15, 3 pm**
 - E. Jakobsson & G. Rendon. The Motif Network: A Workflow for Molecular Biology
 - **Wednesday, November 15, 4 pm**
 - M. Reed. Workflows for Genetics and Melanoma Research
 - **Thursday, November 16, 11:30 am**
 - J. McGee. The North Carolina and TeraGrid Bioportal
- **Argonne National Laboratory/TeraGrid (#1925)**
 - **Tuesday, November 14, Noon- 1 p.m.**
 - J. McGee. The North Carolina and TeraGrid Bioportal
 - **Wednesday, November 15, Noon- 1 p.m.**
 - J. McGee. The North Carolina and TeraGrid Bioportal