TECHNISCHE UNIVERSITÄT DRESDEN

ZIH
Zentrum für Informationsdienste und Hochleistungsrechnen

Distributed Job, Workflow and Data Management - Science Gateways as Solution to Rule Them All

Sandra Gesing
Center for Research Computing
sandra.gesing@nd.edu
23 October 2014

UNIVERSITY OF NOTRE DAME

University of Notre Dame

- In the middle of nowhere of northern Indiana (1.5 h to Chicago)
- 4 undergraduate colleges
- ~35 research institutes and centers
- ~12,000 students

Center for Research Computing

UNIVERSITY OF NOTRE DAME

- Software development and profiling
- Cyberinfrastructure/science gateway development
- Geographical Information Systems
- Visualization Support
- Computational Scientist support
- Collaborative research/ grant development
- System administration/ design and acquisition
- ~40 researchers, research programmers, HPC specialists

CRC and OIT building

Center for Research Computing

UNIVERSITY OF NOTRE DAME

- Computational resources: 20,000 cores+
- Storage resources: 2 PB
- Visualization systems
- Systems for virtual hosting
- Prototype architectures e.g., OpenStack
- Access and interface to
  - XSEDE
  - Open Science Grid
  - Blue Waters

CRC HPC Center (old Union Station)

Distributed Infrastructres

- Definition Grid (Ian Foster, 1998)

„A computational grid is a hardware and software infrastructure that provides dependable, consistent, pervasive, and inexpensive access to high-end computational capabilities.“

- Definition Cloud (Sam Johnston, 2008)

„The Cloud is what The Grid could have been.“

- Virtualization
- Services
  - Infrastructure as a Service (IaaS)
  - Platform as a Service (PaaS)
  - Software as a Service (SaaS)

Distributed Job Infrastructres

UNIVERSITY OF NOTRE DAME

- National and international infrastructures (EGI, PRACE, XSEDE)
- UNICORE, HTCondor, Globus Toolkit, gLite
- OpenStack, Amazon EC2, Windows Azure

- Mixed computing paradigms
  - Grid of federated clusters (NGI-DE)
  - Grid of federated clouds (EGI, CERN-openlab)
  - Grid over cloud (Stratuslab)
  - Cloud over grid (WNoDeS (Worker Nodes on Demand Services))

Workflows

A sequence of connected steps in a defined order based on their control and data dependencies

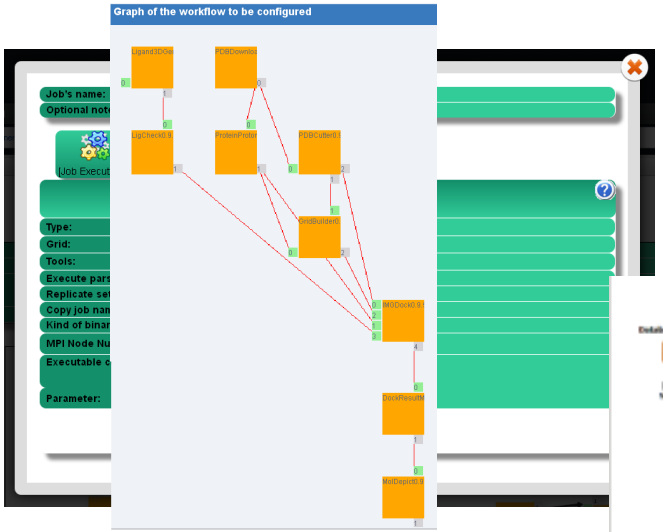
Slide copied from: Stuart Owen „Workflows with Taverna“

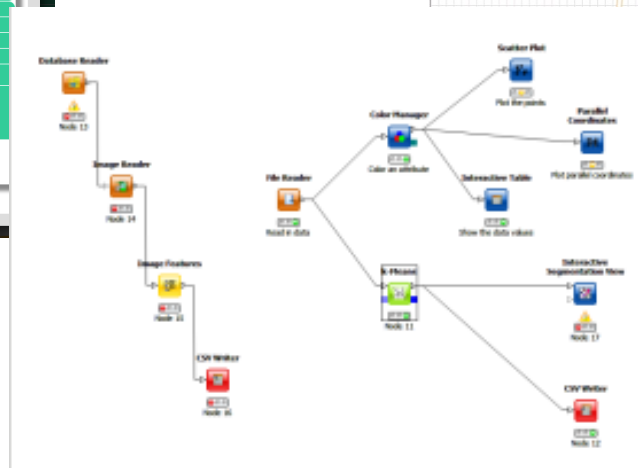
Workflow Systems

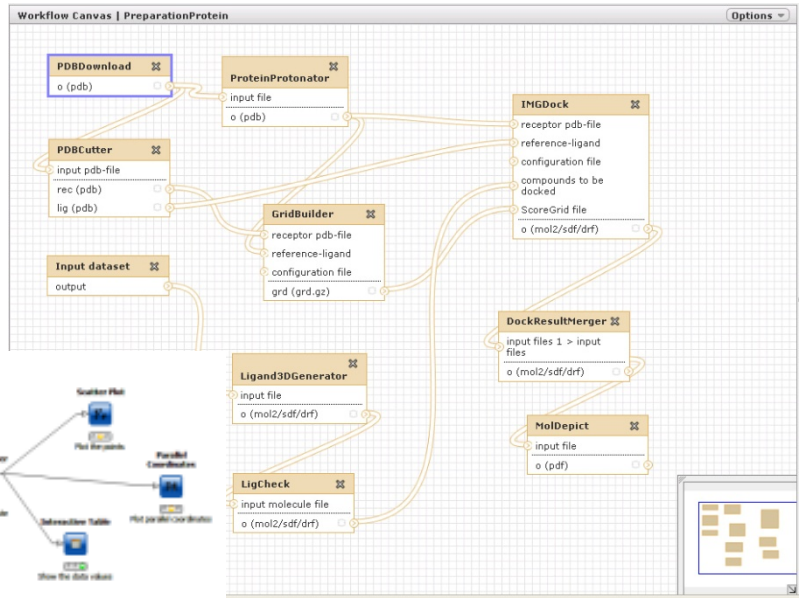
- Different workflow concepts
- Different workflow languages
- Different workflow constructs

Workflow Editors

- Different technologies (workbenches, web-based)
- Different look-and-feel







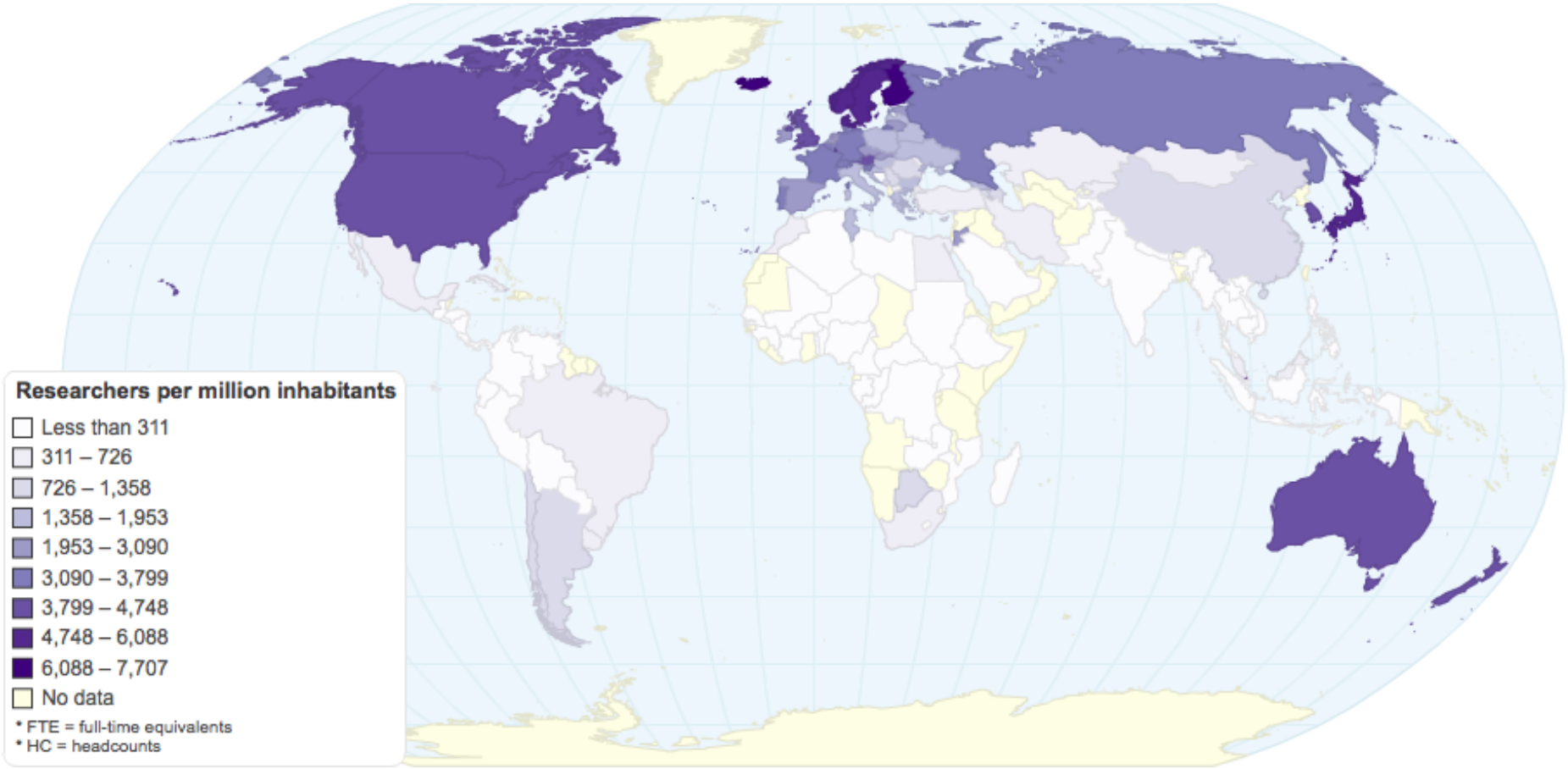
Distributed Data Management

- Amazon Simple Storage Service (Amazon S3)
- Google File System (GFS)
- XtreemFS
- dCache
- iRODS

Figure copied from „iRODS Overview“

Researchers World Wide

- ~7 million researchers world wide



http://chartsbin.com/view/1124
http://ec.europa.eu/euraxess/pdf/research_policies/130122_Researchers%20Report_2012_FINAL%20REPORT_translation_DE_FINAL.pdf

State-of-the-art

- Data-intensive and compute-intensive problems
- Sophisticated tools and methods available
- DCIs (Distributed Computing Infrastructures) available
- Workflow systems available
- Distributed data management available

How do researchers use the tools and distributed environments on a large scale?

Open Issues

- Usability of tools often limited
- Complexity of methods
- Lack of graphical user interfaces

Open Issues

- Usability of tools often limited
- Complexity of methods

```
|=========================================|
|                                         |
| Version: 1.1                            |
| build date: Jan 10 2012                 |
| execution host: vomitoxin               |
| execution time: 2012-09-09, 16:39:43 (MST) |
 ----------------------------------------

Available parameters are ('*' indicates mandatory parameters):
   *  -i <in.file>            input molecule file
   *  -o <out.file>           output file
      -ef <double>            error fraction; print error if fraction of invalid mols is larger
      -write_par <out.file>   write xml parameter file for this tool
      -par <in.file>          read parameters from parameter-xml-file

Available flags are:
      -ri      remove invalid molecules.
      -ut      check for unique topologies
      -nc      no not check for unique conformations
      -rm      remove input file when finished
      -help    show help about parameters and flags of this program


This tool checks all molecules of the given input file for errors. Supported formats are mol2, sdf or drf (DockResultFile, xml-based).

The following checks are done for each molecule:

    * bond-lengths may not be completely senseless (i.e. <0.7 or >2.5 Angstroem)
    * each 'molecule' in the input file may only contain one actual molecule, i.e. there may be no unconnected atoms or fragments.
    * each atom must have a valid assigned element
    * the molecule must be protonated (since this is necessary for docking/(re-)scoring).
    * 3D coordinates must be present (instead of 2D coordinates; also necessary for docking/(re-)scoring)
    * partial charges may not contain completely senseless values (>5 or <-5).
    * each conformation should appear only once within the given file, otherwise it is rejected and not written to the output file. However, if option '-ut' is used,
 molecules will instead be checked for unique topologies.

If option '-ri' is used, only those molecules that pass all those tests are written to the output file. If this option is not used, all molecules are written to outp
ut containing a property 'score_ligcheck' with a value of 1 if the molecule passed all tests or with a value of 0 if it did not pass them.


sshgw-bs[13]
```

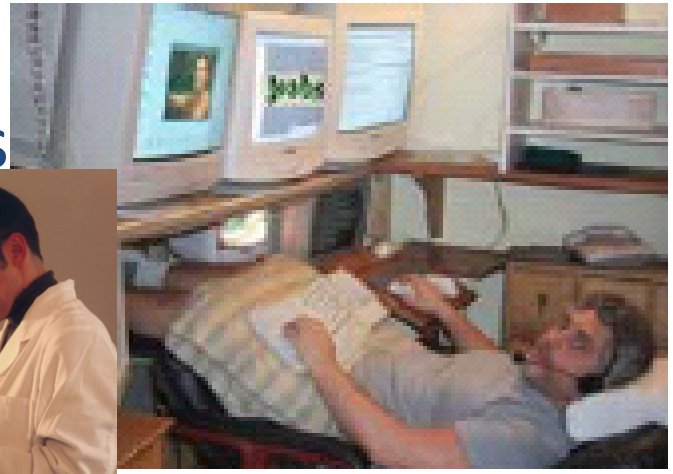
Open Issues

- Usability of tools often limited
- Complexity of methods
- Lack of graphical user interfaces
- Workflows
- Complexity of infrastructures
- Users are generally not IT specialists

Open Issues

- Usability of tools often limited
- Complexity of methods
- Lack of graphical user interfaces
- Workflows
- Complexity of infrastructures

not IT specialis



Open Issues

- Usability of tools often limited
- Complexity of methods
- Lack of graphical user interfaces
- Workflows
- Complexity of infrastructures
- Users are generally not IT specialists

⇒ **User interfaces need to be intuitive and self-explanatory**

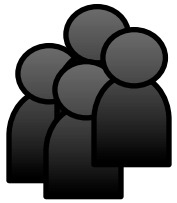
⇒ **Science gateways**

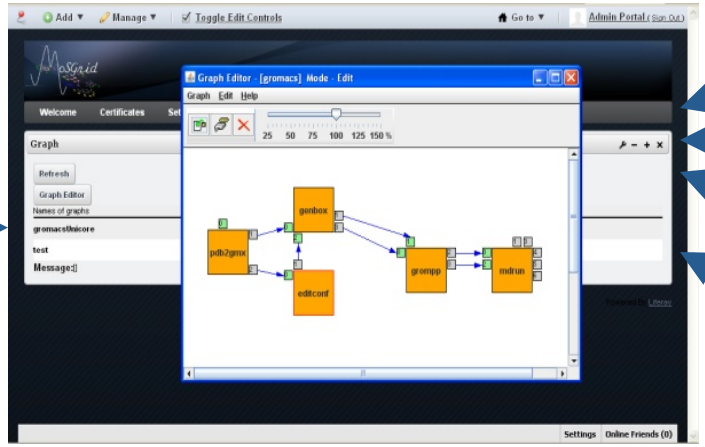
Science Gateways

"A Science Gateway is a community-developed set of tools, applications, and data that is integrated via a portal or a suite of applications, usually in a graphical user interface, that is further customized to meet the needs of a specific community."

TeraGrid/XSEDE

Community





GROMACS FAST. FLEXIBLE. FREE.

XTREEMFS

Web-based Science Gateways

- Single point of entry
- Possibility to customize views and tools
- Store user preferences
- No installation of software on the user's side
- No firewall issues

Slartibartfast: "I must warn you, we're going to pass through, well, a sort of gateway thing."
Arthur Dent: „What?"
Slartibartfast: "It may disturb you. It scares the willies out of me."
(Douglas Adams in "The Hitchhiker's Guide to the Galaxy")

Goal of Science Gateways

UNIVERSITY OF NOTRE DAME

Usability of software

"After all, usability really just means that making sure that something works well: that a person … can use the thing - whether it's a Web site, a fighter jet, or a revolving door - for its intended purpose without getting hopelessly frustrated."

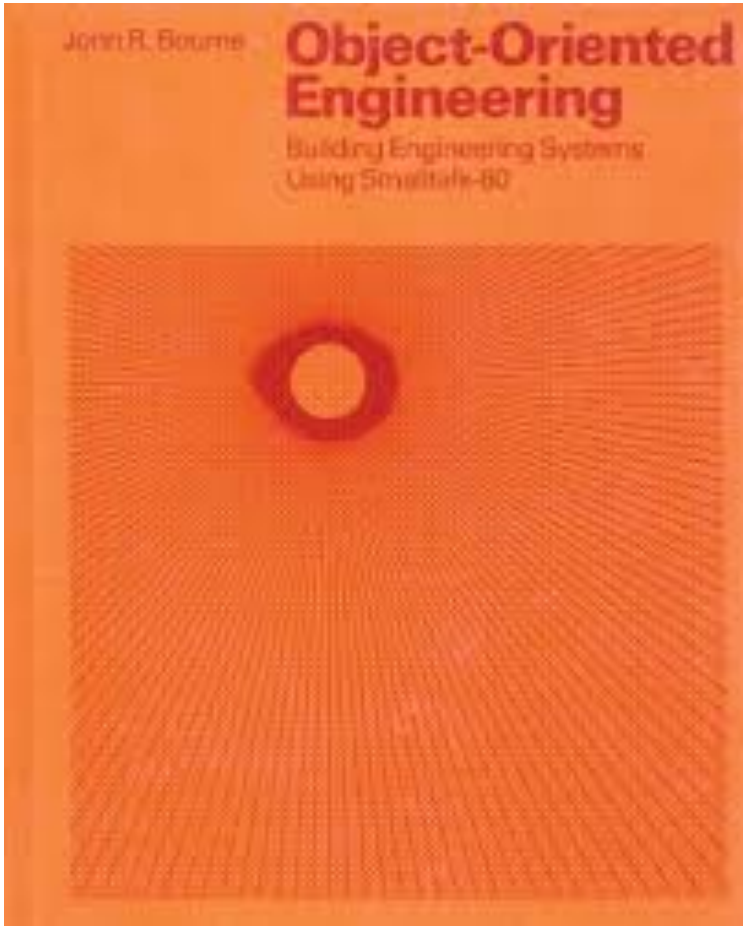
(Steve Krug in *"Don't make me think!: A Common Sense Approach to Web Usability"*, 2005)

Re-Usability

- Sharing of knowledge and data
- Re-Using of „recipes“ and workflows
- Re-Usability of software

“The key to productivity is reusability. The easiest way to produce code is obviously to have it already!"

(John R. Bourne in *“Object-oriented Engineering: Building Engineering Systems Using Smalltalk-80”*, 1992)



Re-Usability

Re-inventing is not always necessary...

Re-Usability

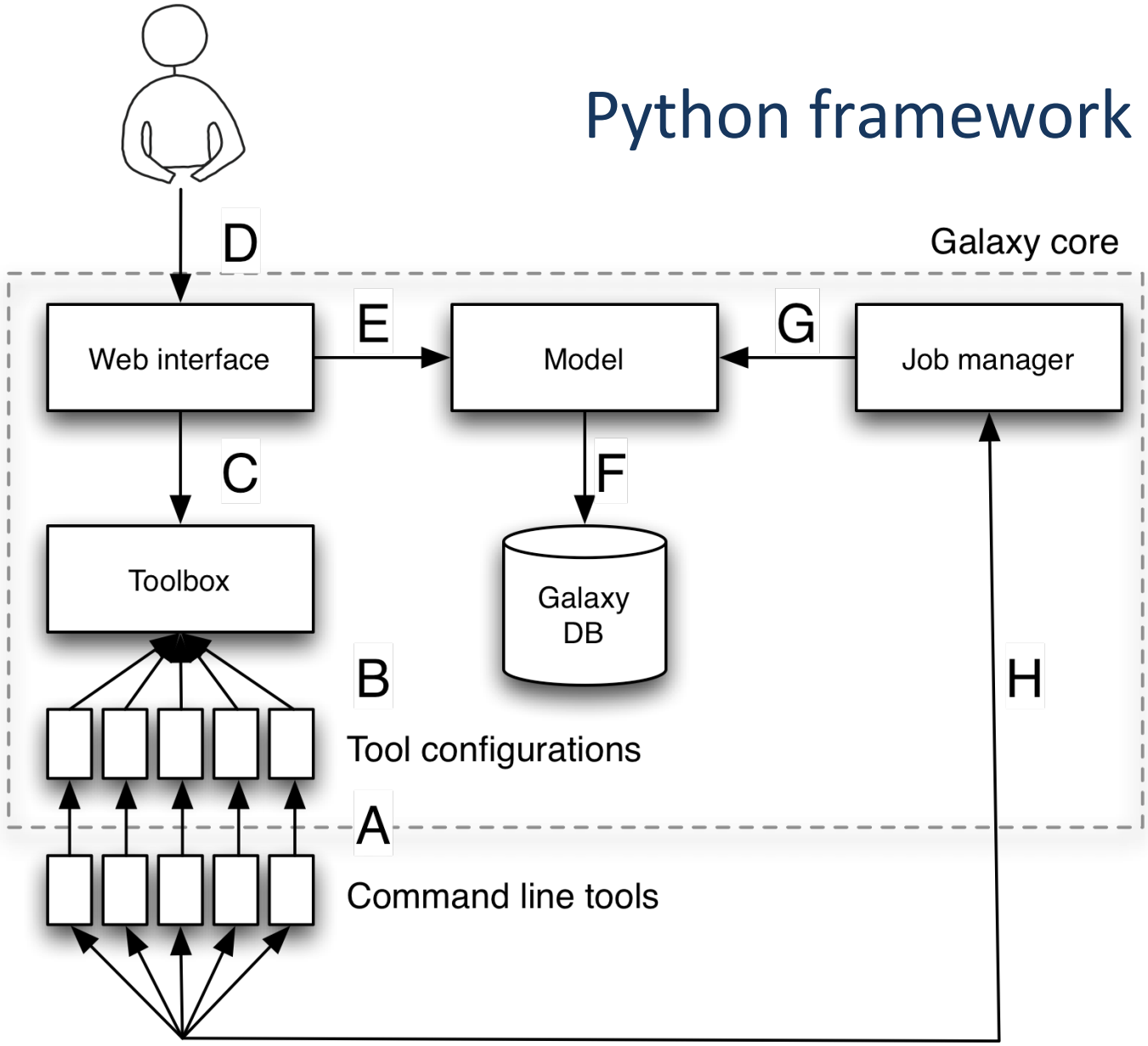
... but the model should fit to the demands of the community

Diverse Approaches

- Science gateway frameworks (Galaxy, WS-PGRADE)
  - Static layout
  - Layout extendable
  - Workflow-enabled
- Portal frameworks (Liferay)
- Content management systems (Drupal)
- Libraries for implementation (Django)
- APIs for implementation (Apache Airavata, Agave)

Galaxy

Python framework



Galaxy

UNIVERSITY OF NOTRE DAME

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Admin | Help | User | Using 9.4 MB

Tools

search tools

- Get Data
- Send Data
- ENCODE Tools
- Lift-Over
- Text Manipulation
- Filter and Sort
- Join, Subtract and Group
- Convert Formats
- Extract Features
- Fetch Sequences
- Fetch Alignments
- Get Genomic Scores
- Operate on Genomic Intervals
- Statistics
- Wavelet Analysis
- Graph/Display Data
- Regional Variation
- Multiple regression
- Multivariate Analysis
- Evolution
- Motif Tools
- Multiple Alignments
- Metagenomic analyses
- FASTA manipulation

Hello world! It's running...

To customize this page edit `static/welcome.html`

Galaxy is an open, web-based platform for data intensive biomedical research. The Galaxy team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University. The Galaxy Project is supported in part by NHGRI, NSF, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.

History

test history
8.8 MB

44: Main output for creating files (Test3.txt)
1 line
format: text, database: ?
`Test3.txt`

43: Main output for creating files (Test2.fasta)
1 line
format: text, database: ?
`Test2.fasta`

42: Main output for creating files (Test1.txt)
1 line
format: text, database: ?
`Test1.txt`

41: Main output for

Parametrization

UNIVERSITY OF NOTRE DAME

Analyze Data | Workflow | Shared Data | Visualization | Admin | Help | User

Compute sequence length (version 1.0.0)

Compute length for these sequences:

2: http://bx.psu.edu/~clements/Events/GMOD2013/m.vannielli.sequence.fasta

How many title characters to keep?:

0

'0' = keep the whole thing

Execute

What it does

This tool counts the length of each fasta sequence in the file. The output file has two columns per line (separated by tab): fasta titles and lengths of the sequences. The option *How many characters to keep?* allows to select a specified number of letters from the beginning of each FASTA entry.

Example

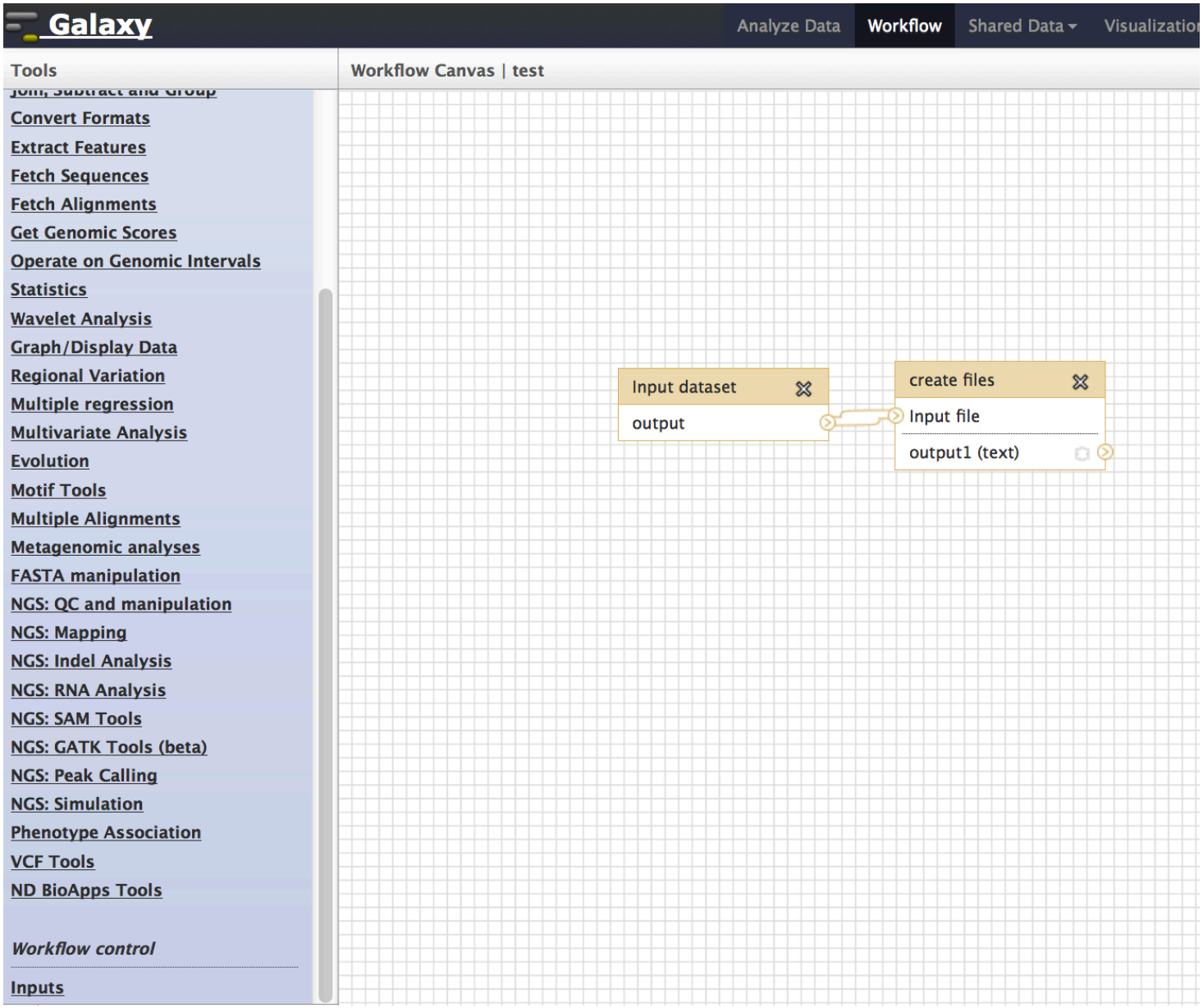
Suppose you have the following FASTA formatted sequences from a Roche (454) FLX sequencing run:

```
>EYKX4VC02EQLO5 length=108 xy=1826_0455 region=2 run=R_2007_11_07_16_15_57_
TCCGCGCCGAGCATGCCCATCTTGGATTCCGGCGCGATGACCATCGCCCGCTCCACCACG
TTCGGCCGGCCCTTCTCGTCGAGGAATGACACCAGCGCTTCGCCCACG
>EYKX4VC02D4GS2 length=60 xy=1573_3972 region=2 run=R_2007_11_07_16_15_57_
AATAAAACTAAATCAGCAAAGACTGGCAAATACTCACAGGCTTATACAATACAAATGTAAfa
```

Running this tool while setting **How many characters to keep?** to **14** will produce this:

```
EYKX4VC02EQLO5  108
EYKX4VC02D4GS2   60
```

Workflows


Galaxy
Analyze Data
Workflow
Shared Data
Visualizatio
Tools
Workflow Canvas | test
Convert Formats
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Wavelet Analysis
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Motif Tools
Multiple Alignments
Metagenomic analyses
FASTA manipulation
NGS: QC and manipulation
NGS: Mapping
NGS: Indel Analysis
NGS: RNA Analysis
NGS: SAM Tools
NGS: GATK Tools (beta)
NGS: Peak Calling
NGS: Simulation
Phenotype Association
VCF Tools
ND BioApps Tools
Workflow control
Inputs
Input dataset
output
create files
Input file
output1 (text)

Administration

UNIVERSITY OF NOTRE DAME

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Admin | Help | User

Administration

Security
- Manage users
- Manage groups
- Manage roles

Data
- Manage quotas
- Manage data libraries
- Manage local data (beta)

Server
- View data types registry
- View data tables registry
- View tool lineage
- Reload a tool's configuration
- Profile memory usage
- Manage jobs
- Review tool migration stages

Tool sheds
- Search and browse tool sheds

Form Definitions
- Manage form definitions

Sample Tracking
- Manage sequencers and external services
- Manage request types
- Sequencing requests
- Find samples

Administration

The menu on the left provides the following features

- **Security** – see the **Data Security and Data Libraries** section below for details
  - **Manage users** – provides a view of the registered users and all groups and non-private roles associated with each user.
  - **Manage groups** – provides a view of all groups along with the members of the group and the roles associated with each group (both private and non-private r
    you to manage the users and roles that are associated with the group.
  - **Manage roles** – provides a view of all non-private roles along with the role type, and the users and groups that are associated with the role. The role names in
    and groups that are associated with the role. The page also includes a view of the data library datasets that are associated with the role and the permissions ap
- **Data**
  - **Manage data libraries** – Data libraries enable a Galaxy administrator to upload datasets into a data library. Currently, only administrators can create data libra
    When a data library is first created, it is considered "public" since it will be displayed in the "Data Libraries" view for any user, even those that are not logged in
    library by associating roles with the data library's "access library" permission. This permission will conservatively override the [dataset] "access" permission for

    For example, if a data library's "access library" permission is associated with Role1 and the data library contains "public" datasets, the data library will still only
    data library's "access library" permission is associated with both Role1 and Role2 and the data library contains datasets whose [dataset] "access" permission is
    be able to access the library, but will not see those contained datasets whose [dataset] "access" permission is associated with only Role1.

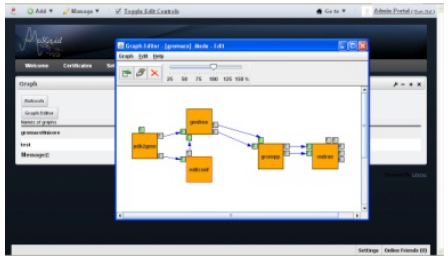
    In addition to the "access library" permission, permission to perform the following functions on the data library (and its contents) can be granted to users (a lib
    dataset).

    - **add library item** – Users that have the role can add library items to this data library or folder
    - **modify library item** – Users that have the role can modify this library item
    - **manage library permissions** – Users that have the role can manage permissions applied to this library item

    The default behavior is for no permissions to be applied to a data library item, but applied permissions are inherited downward (with the exception of the "acc
    library itself). Because of this, it is important to set desired permissions on a new data library when it is created. When this is done, new folders and datasets a
    permissions. In the same way, permissions can be applied to a folder, which will be automatically inherited by all contained datasets and sub-folders.

    The "Data Libraries" menu item allows users to access the datasets in a data library as long as they are not restricted from accessing them. Importing a library
    but will be a "pointer" to the dataset on disk. This approach allows for multiple users to use a single (possibly very large) dataset file.
- **Server**
  - **Reload a tool's configuration** – allows a new version of a tool to be loaded while the server is running
  - **Profile memory usage** – measures system memory used for certain Galaxy functions
  - **Manage jobs** – displays all jobs that are currently not finished (i.e., their state is new, waiting, queued, or running). Administrators are able to cleanly stop lon
- **Forms**
  To be completed
- **Sequencing Requests**
  To be completed
- **Cloud**
  To be completed

WS-PGRADE



User Interface
WS-PGRADE
Liferay

Workflow storage

Application repository

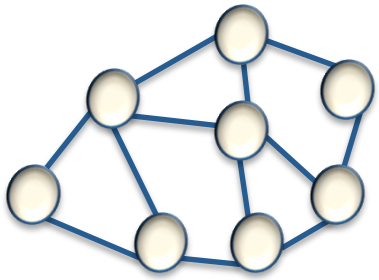
Information system

Workflow engine

Submitters

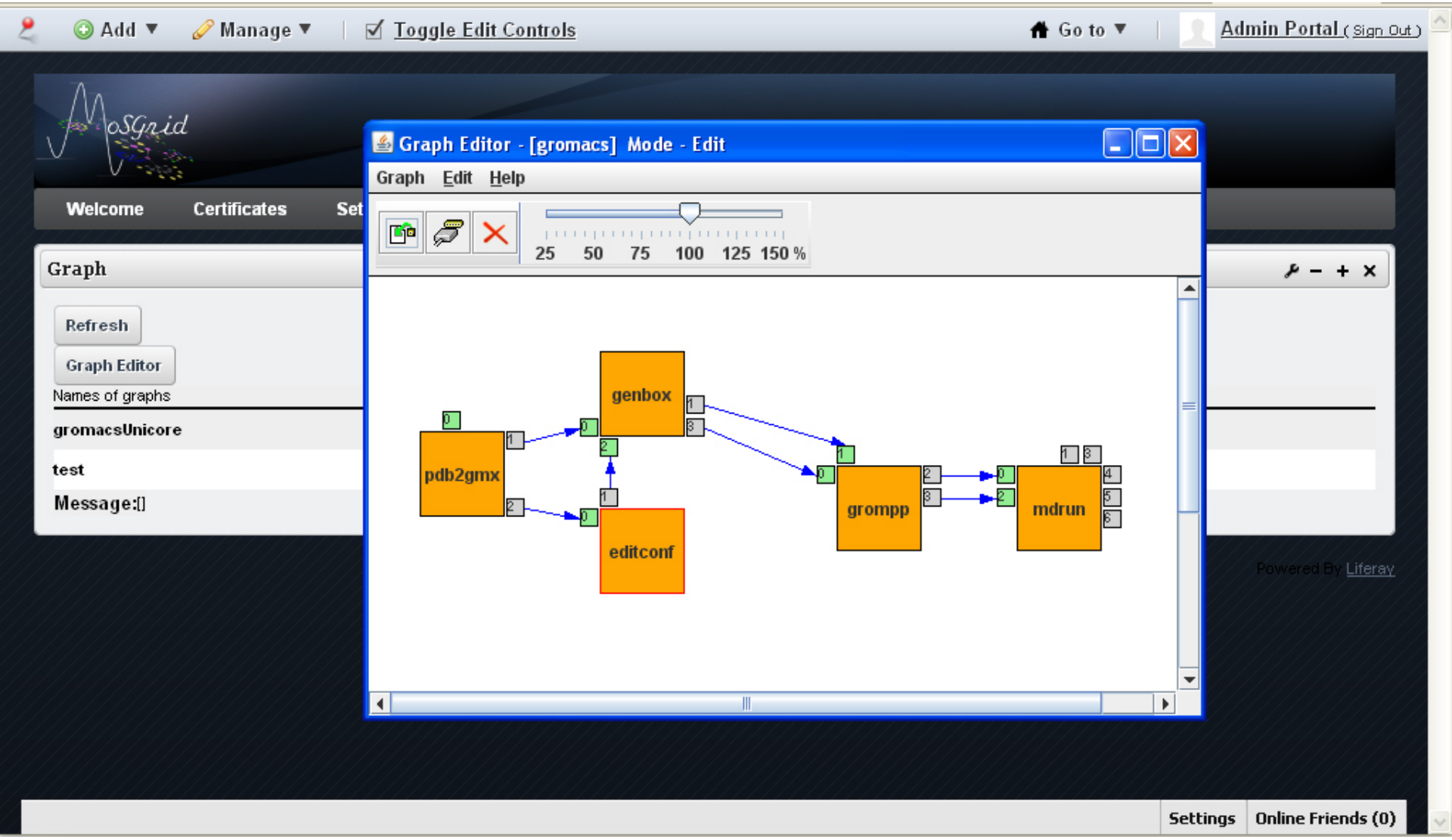
Logging

High-Level Middleware Service Layer
gUSE

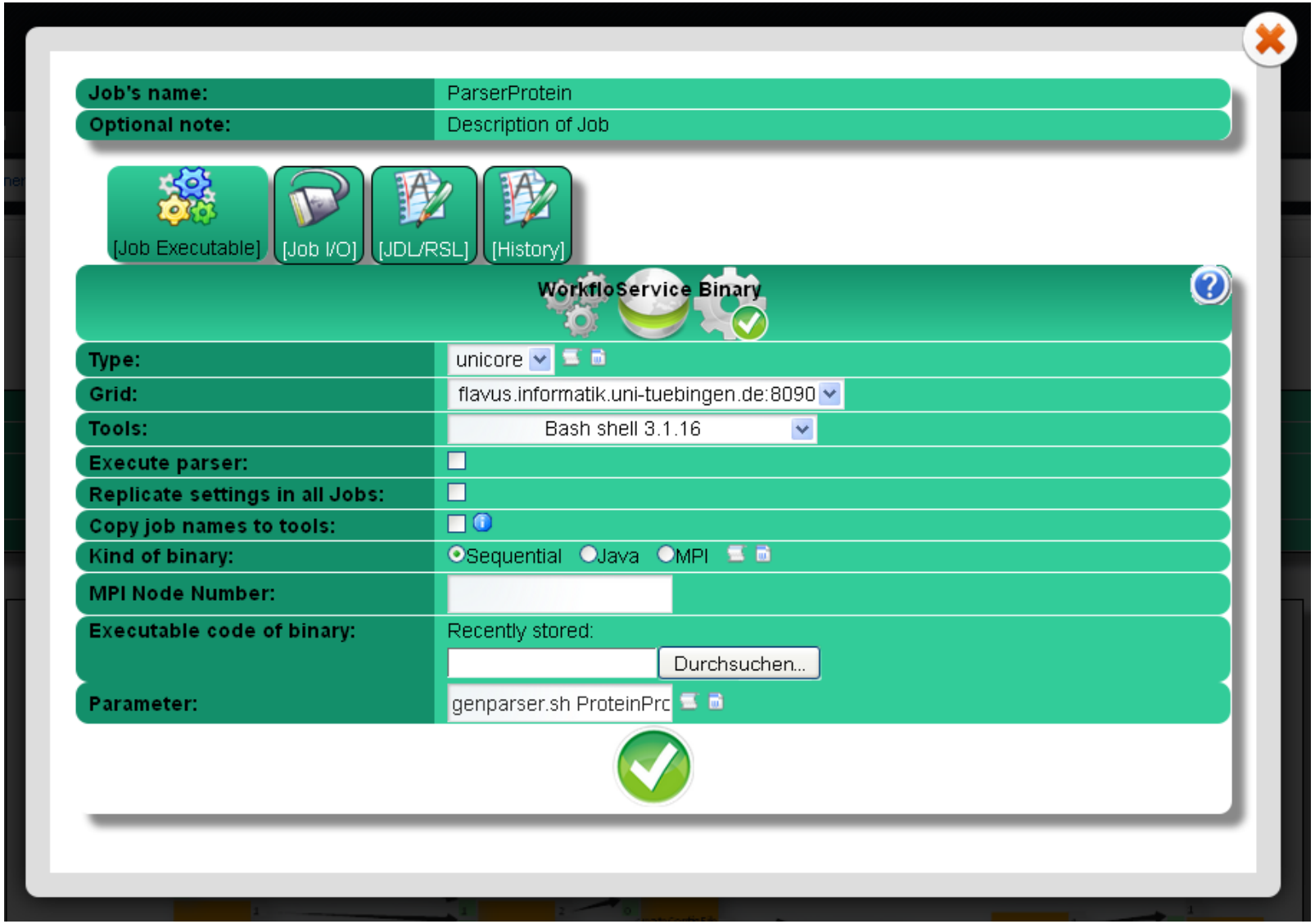


DCI Resources Middleware Layer

WS-PGRADE


Add
Manage
Toggle Edit Controls
Go to
Admin Portal (Sign Out)
MoSGrid
Welcome
Certificates
Graph
Refresh
Graph Editor
Names of graphs
gromacsUnicore
test
Message:[]
Graph Editor - [gromacs] Mode - Edit
Graph
Edit
Help
25
50
75
100
125
150 %
genbox
pdb2gmx
editconf
grompp
mdrun
Powered By Liferay
Settings
Online Friends (0)

Job Configuration



Monitoring

MoSGrid Portal > Workflows > Concrete

2011-1-13 14:53	finished	Details	Delete
2011-1-17 12:0	finished	Details	Delete
2011-2-9 9:34	finished	Details	Delete
2011-1-18 9:40	finished	Details	Delete
2011-2-1 14:44	finished	Details	Delete
2011-2-7 18:55	finished	Details	Delete
2011-2-15 9:21	finished	Details	Delete
2011-1-14 10:38	finished	Details	Delete
2011-1-18 10:13	finished	Details	Delete
2011-2-10 12:56	finished	Details	Delete

Selected WF Instance:
2011-2-10 12:56

Job	Status	Instances	[Actions]	
cell	finished	1	View finished	Hide

Sorting method: Method 1 Range: 20 From: Refresh

PID	Resource	Status	View info			
0	https://unicore6-bisgrid.uni-paderborn.de:8080/bisgrid/services/JobManagement?res=0106fb75-d006-488b-a97a-6f8c60c25daa	finished	Logbook	std. Output	std. Error	Dowload file output

Powered By Liferay

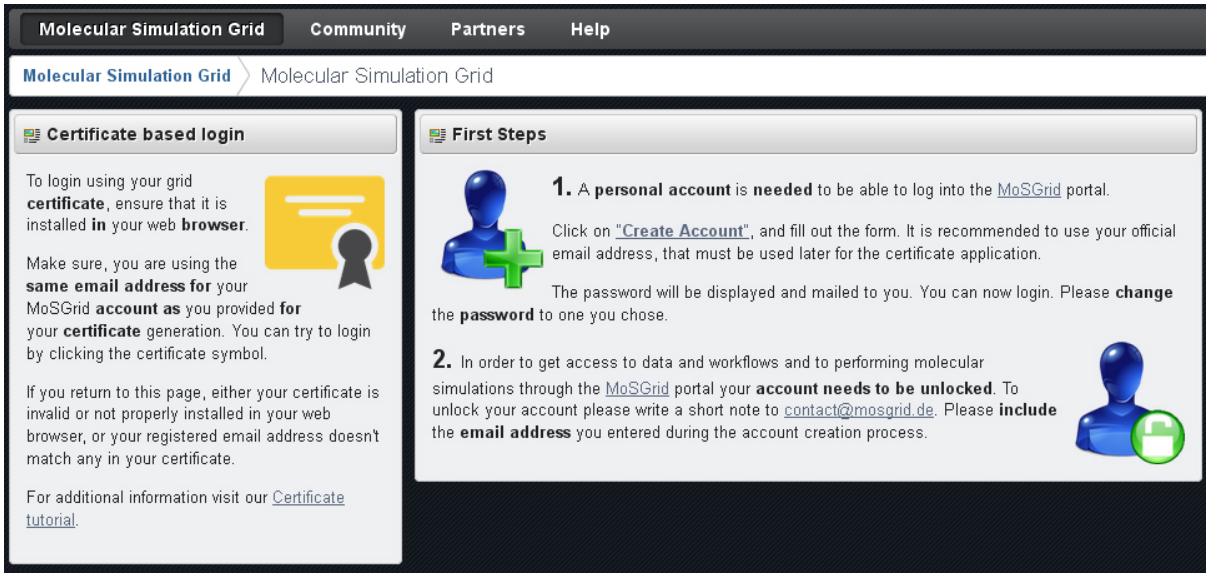
Settings | Online Friends (0)

MoSGrid – WS-PGRADE Example

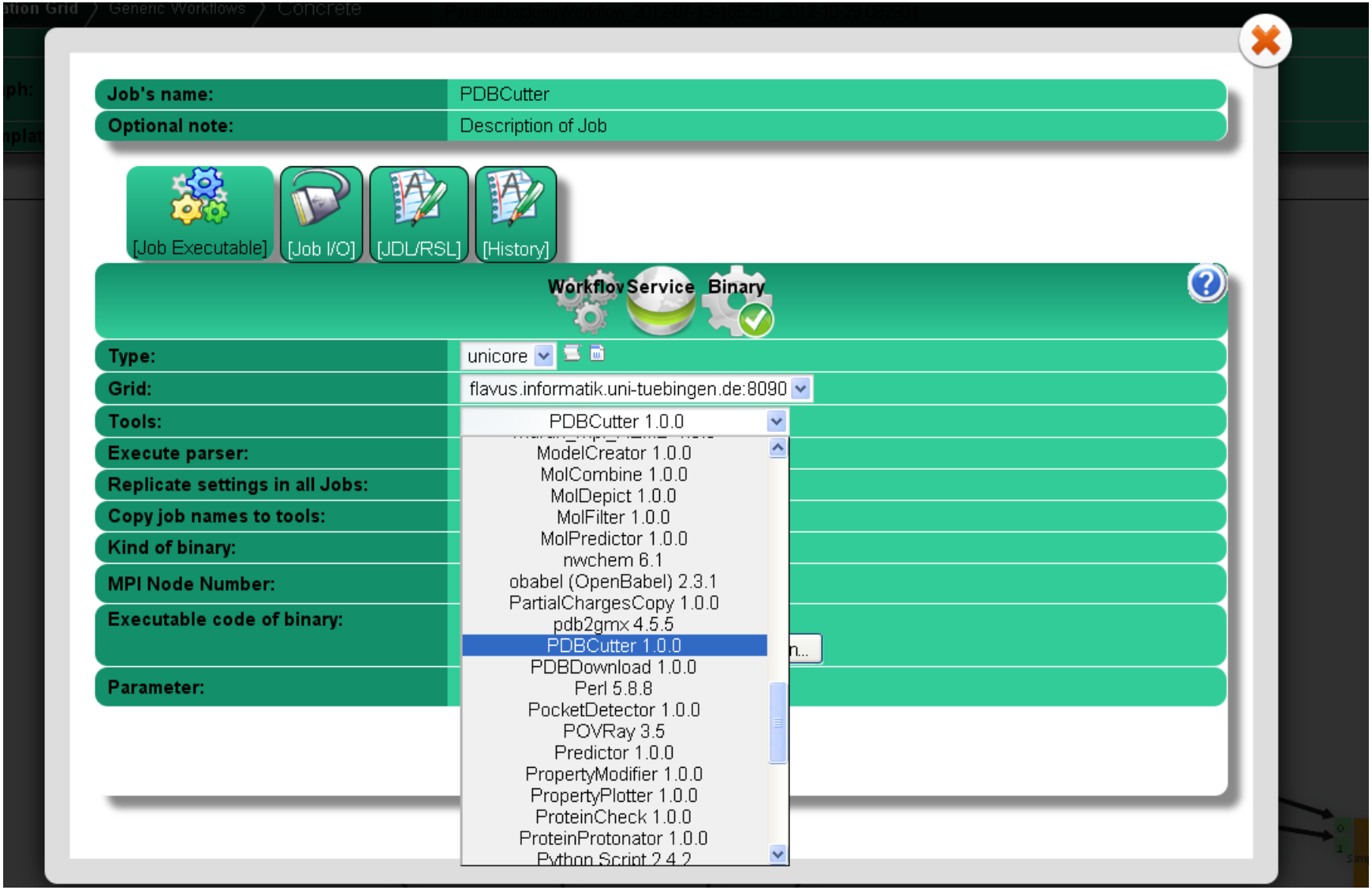
UNIVERSITY OF NOTRE DAME

Molecular Simulation Grid

- Science gateway integrated with underlying compute and data management infrastructure
- Distributed workflow management
- Data repository
- Metadata management



Tools


Job's name:
PDBCutter
Optional note:
Description of Job
[Job Executable]
[Job I/O]
[JDL/RSL]
[History]
Workflow Service Binary
Type:
unicore
Grid:
flavus.informatik.uni-tuebingen.de:8090
Tools:
PDBCutter 1.0.0
Execute parser:
Replicate settings in all Jobs:
Copy job names to tools:
Kind of binary:
MPI Node Number:
Executable code of binary:
Parameter:
ModelCreator 1.0.0
MolCombine 1.0.0
MolDepict 1.0.0
MolFilter 1.0.0
MolPredictor 1.0.0
nwchem 6.1
obabel (OpenBabel) 2.3.1
PartialChargesCopy 1.0.0
pdb2gmx 4.5.5
PDBCutter 1.0.0
PDBDownload 1.0.0
Perl 5.8.8
PocketDetector 1.0.0
POVRay 3.5
Predictor 1.0.0
PropertyModifier 1.0.0
PropertyPlotter 1.0.0
ProteinCheck 1.0.0
ProteinProtonator 1.0.0
Python Script 2.4.2

File Management

MoSGrid – Application Areas

UNIVERSITY OF NOTRE DAME

Molecular Dynamics

- Study and simulation of molecular motion

Quantum Chemistry

- Study and simulation of molecular electronic behavior relative to their chemical reactivity

Docking

- Main focus on evaluation of ligand-receptor interactions (e.g., for drug design)

Docking Portlet

UNIVERSITY OF NOTRE DAME

Docking Portlet

Import | Submission | Monitoring | About | ?

Select an imported instance

Import
StandardDockingWorkflow_2012-03-30-125439_29.(

Please fill the input mask to submit your workflow

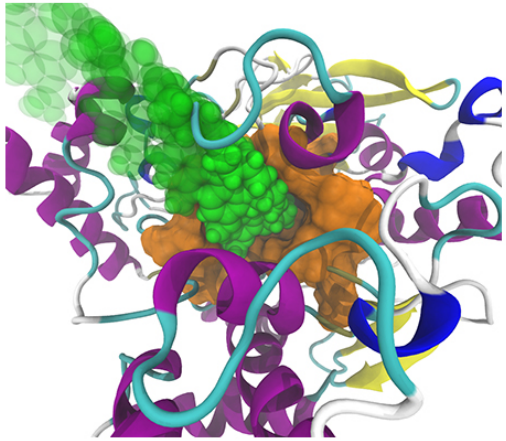
PDBCutter

Filename *
1DX6.pdb
Upload PDB

PDB Model *
Model 0

☑ Chain A
☐ Chain S

Select a protein chain from your PDB input file to act as receptor (secondary structure) including the binding pocket (orange).



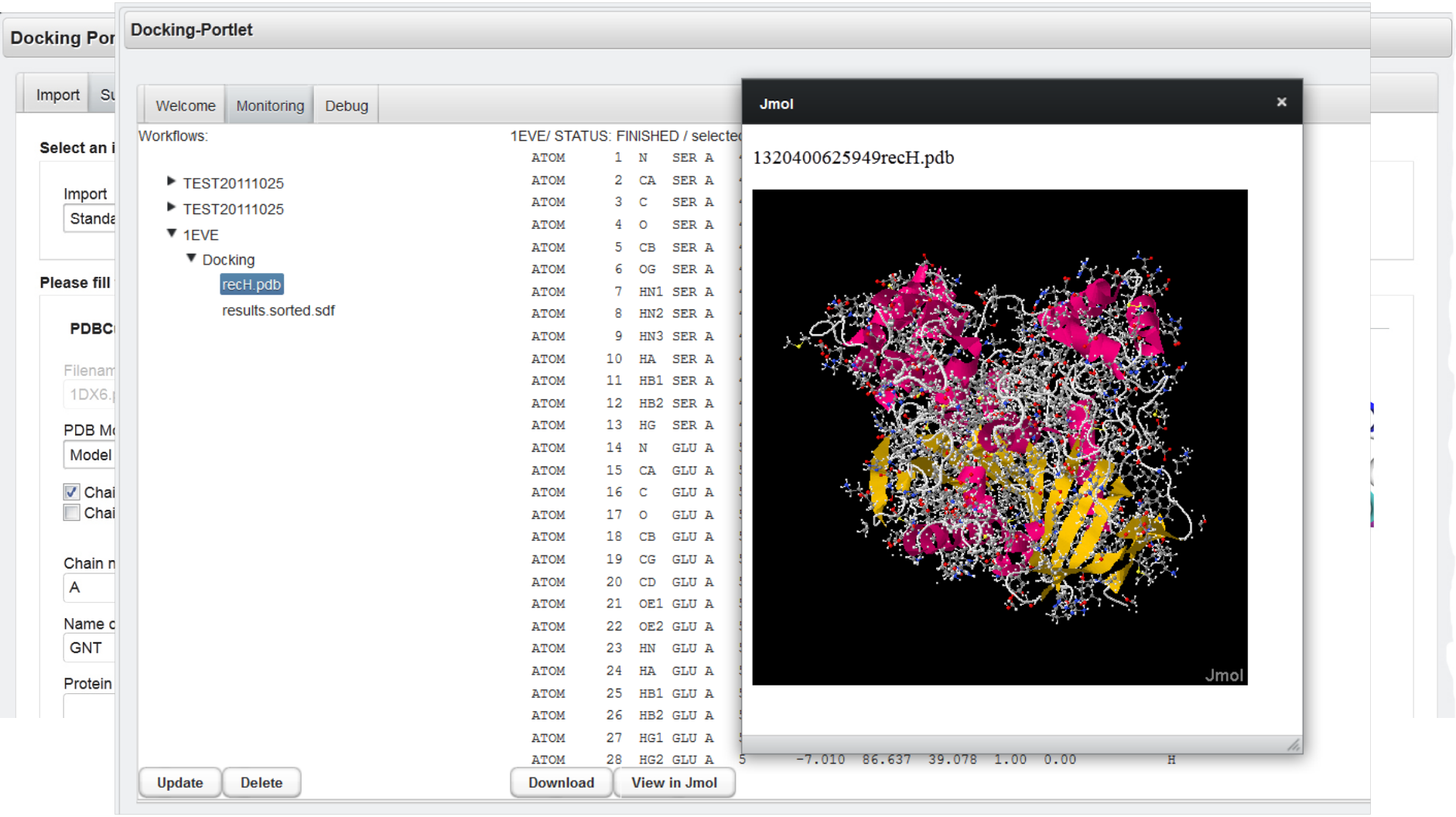
Chain name of ligand *
A

Name of ligand as stated in pdb file *
GNT

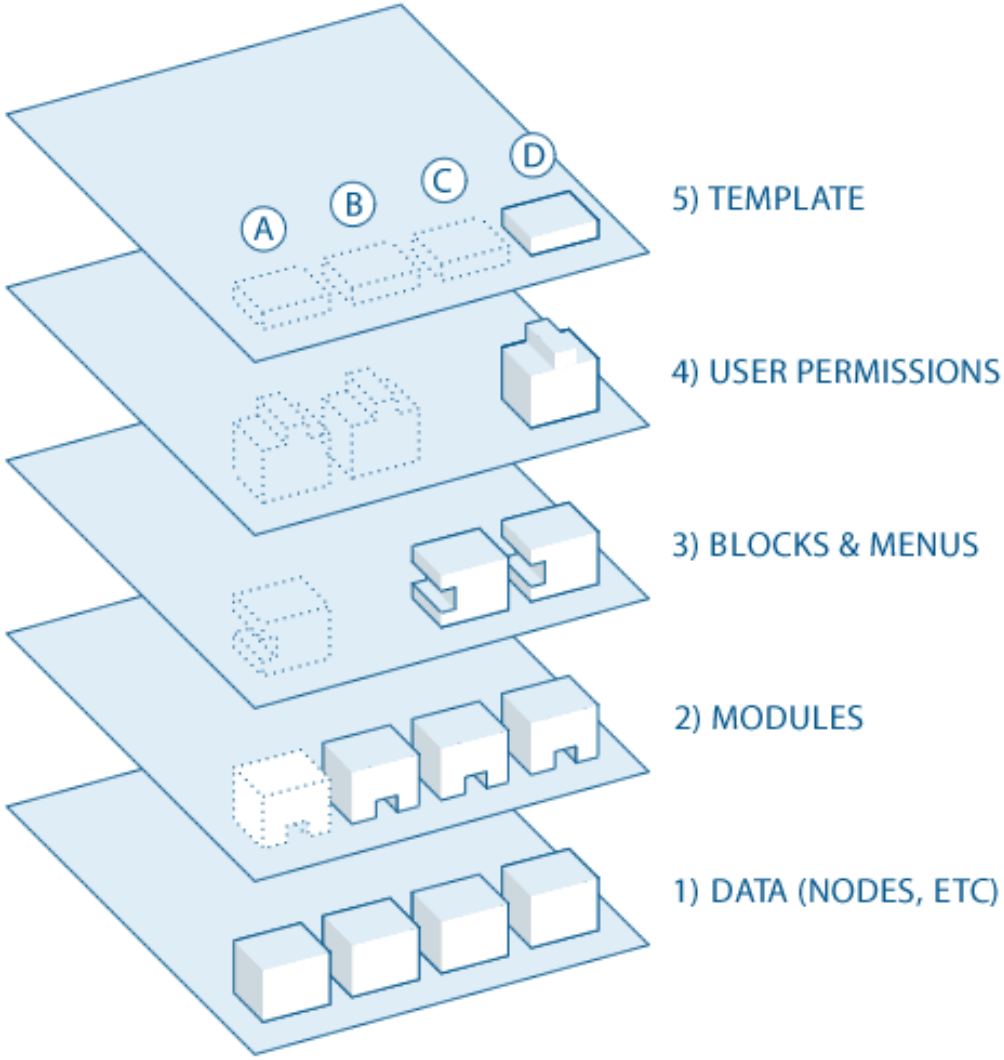
Specifiy a reference ligand (green) by it's three letter code including the corresponding chain. It might be necessary to open the input PDB file with an editor. This information is required for the indentification of the binding site and the calculation of RMSD values.

Protein Chains that are to be deleted

Docking Portlet



Drupal


A
B
C
D
5) TEMPLATE
4) USER PERMISSIONS
3) BLOCKS & MENUS
2) MODULES
1) DATA (NODES, ETC)

VectorBase – Example for Drupal

UNIVERSITY OF NOTRE DAME

Enter search terms GO

LOGIN

VectorBase

Bioinformatics Resource for Invertebrate Vectors of Human Pathogens

ABOUT | ORGANISMS | DOWNLOADS | TOOLS | DATA | HELP | COMMUNITY | CONTACT US

Welcome to VectorBase!

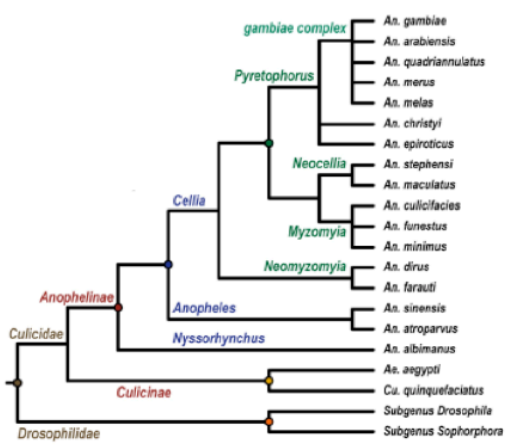
VectorBase is an NIAID Bioinformatics Resource Center dedicated to providing data to the scientific community for Invertebrate Vectors of Human Pathogens. We aim to provide a forum for the discussion and distribution of news and information relevant to invertebrate vectors, as well as access to tools to facilitate the querying and analysis of the data sets presented on this site.

DATA

GENOMES | TRANSCRIPTS & TRANSCRIPTOMES | PROTEINS & PROTEOMES | MITOCHONDRIAL SEQUENCES | POPULATION BIOLOGY

TOOLS & RESOURCES

Pause



First pass annotation for nine Anopheline species available

VectorBase and The Anopheles Genomes Cluster announce the first pass annotation of nine Anopheline genomes. The predictions were generated using *ab initio* and similarity approaches utilising transcriptome data and taxonomically informative proteomes. Gene models were aggregated using the MAKER system. These gene sets are available for browsing, searching via BLAST and download.

An. albimanus	*An. arabiensis*
An. christyi	*An. dirus*
An. epiroticus	*An. funestus*
An. minimus	*An. quadriannulatus*
An. stephensi	

Want to see your BLAST, ClustalW and HMMer jobs?

Login above or Register here.

POPULAR ORGANISMS

Anopheles gambiae | *Aedes aegypti* | *Culex quinquefasciatus*

RECENT ADDITIONS

Anopheles funestus | *Phlebotomus papatasi* | *Biomphalaria glabrata*

All organisms

LATEST NEWS

August 14, 2013
VectorBase Release VB-2013-08
June 28, 2013
VectorBase Release VB-2013-06

More news

DID YOU KNOW?

New search engine at VectorBase

Searching via the box at the top of all pages now lets you find more than just genes! Most site content is now searchable. ... From Newsletter 13 (Sep

Tools

UNIVERSITY OF NOTRE DAME

Enter search terms GO

VectorBase
Bioinformatics Resource for Invertebrate Vectors of Human Pathogens

LOGIN

ABOUT | ORGANISMS | DOWNLOADS | TOOLS | DATA | HELP | COMMUNITY | CONTACT US

Tool

Home » Tools

BLAST

Due to browser compatibility problems in Safari and Firefox, we recommend using **Google Chrome** when using blast. We are working on these issues and apologize for any inconvenience this may cause you.

Paste your sequences here

Upload FASTA File

Browse... No file selected.

Program

blastn | tblastn | tblastx | blastp | blastx

blastn - Nucleotide vs. Nucleotide

Job Control

Load results

Job ID

RESET

SUBMIT

Datasets

- All Datasets
- *Aedes aegypti*
- *Anopheles albimanus*
- *Anopheles arabiensis*
- *Anopheles christyi*
- *Anopheles darlingi*
- *Anopheles dirus A*
- *Anopheles epiroticus*

Options

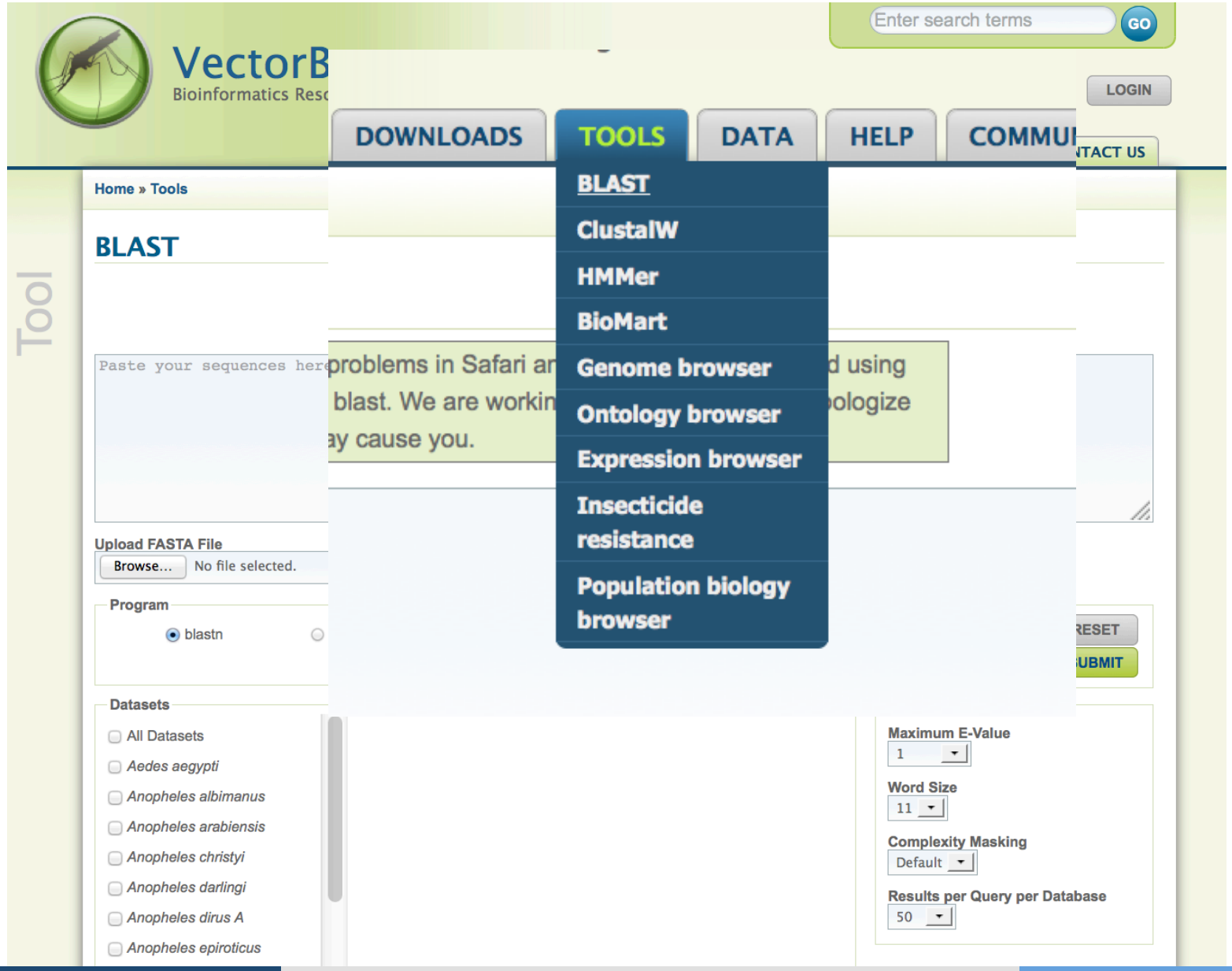
Maximum E-Value: 1

Word Size: 11

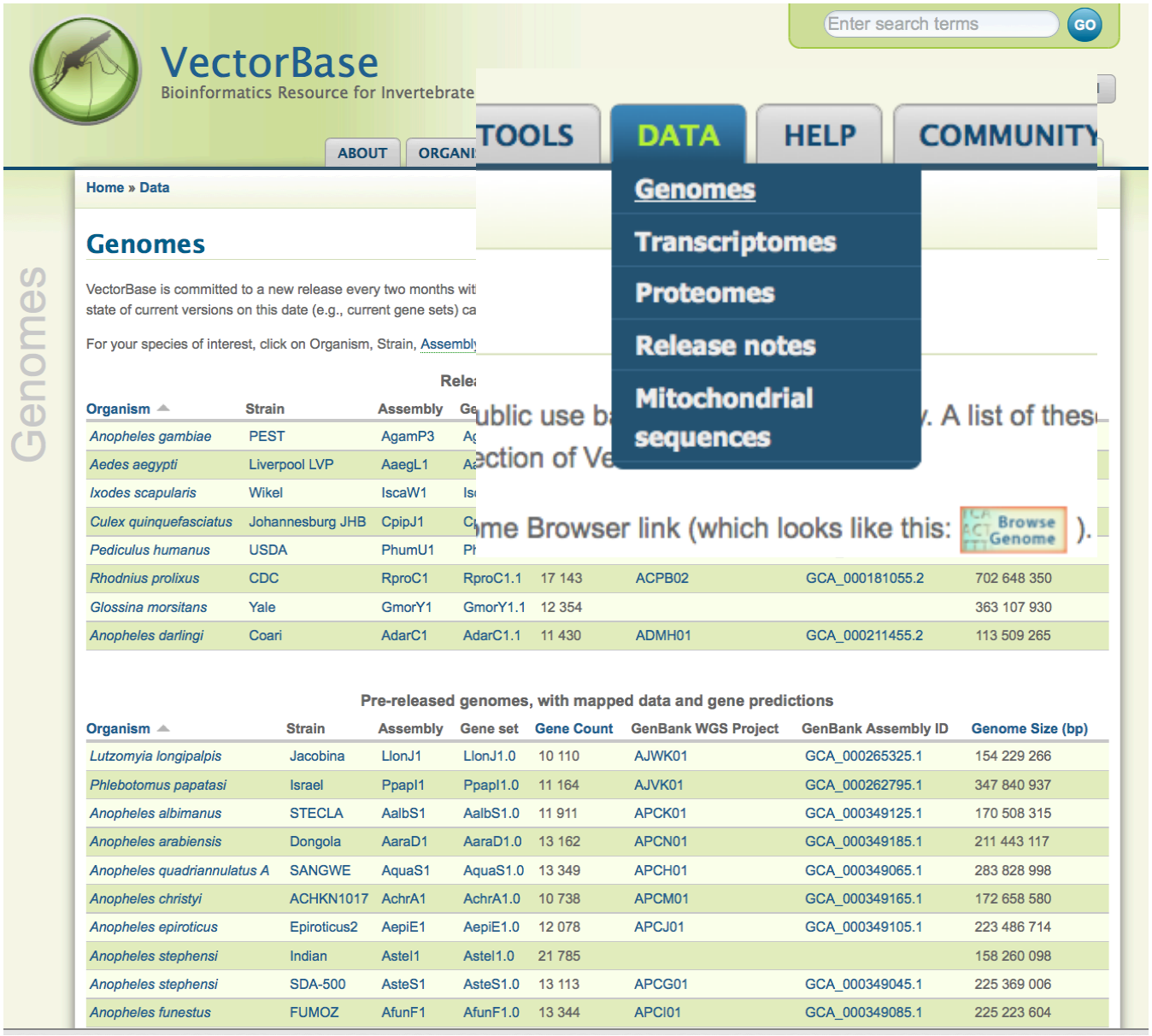
Complexity Masking: Default

Results per Query per Database: 50

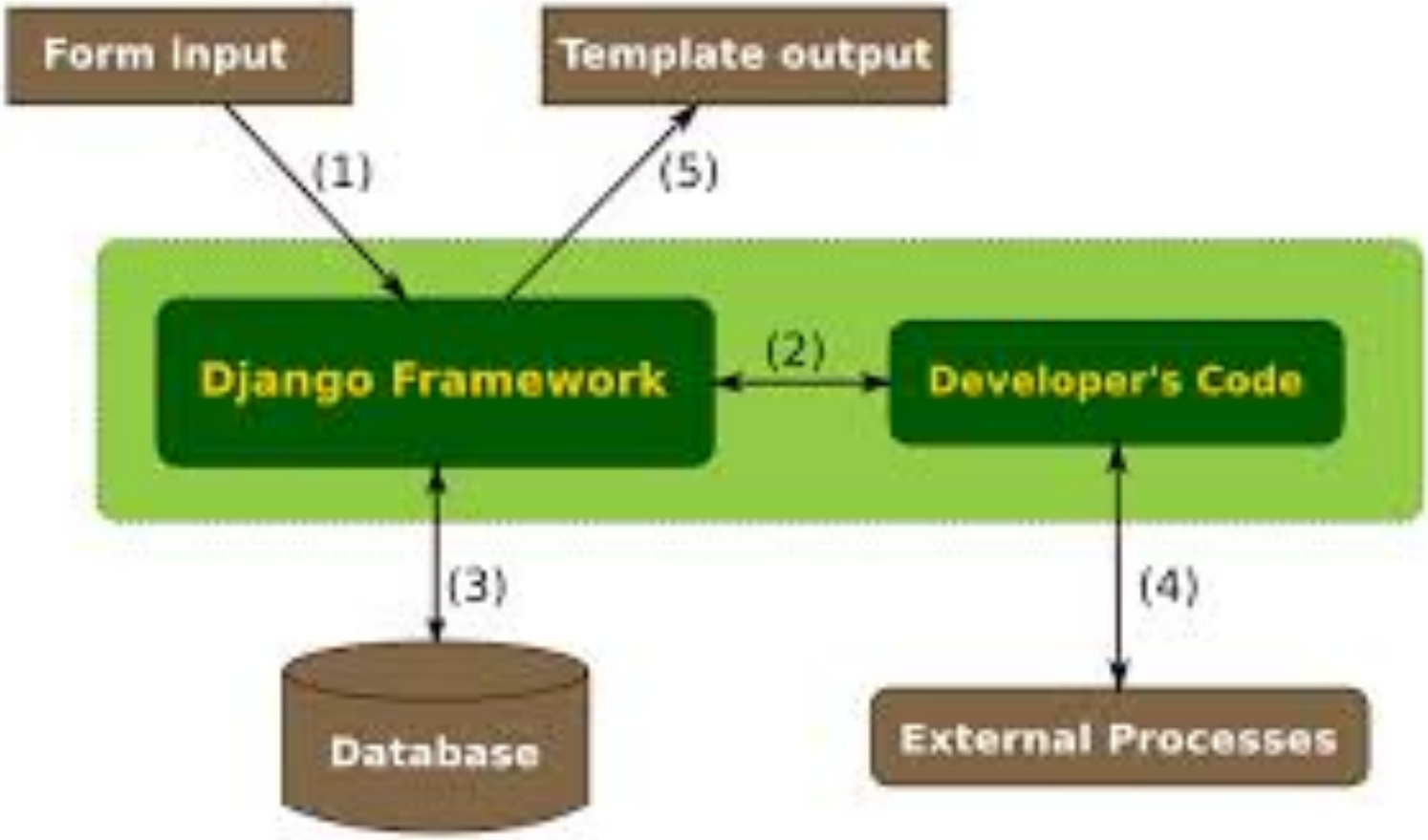
Tools


Enter search terms
GO
LOGIN
DOWNLOADS
TOOLS
DATA
HELP
BLAST
ClustalW
HMMer
BioMart
Genome browser
Ontology browser
Expression browser
Insecticide resistance
Population biology browser
Home » Tools
BLAST
Tool
Paste your sequences here
Upload FASTA File
Browse...
No file selected.
Program
blastn
Datasets
All Datasets
Aedes aegypti
Anopheles albimanus
Anopheles arabiensis
Anopheles christyi
Anopheles darlingi
Anopheles dirus A
Anopheles epiroticus
Maximum E-Value
1
Word Size
11
Complexity Masking
Default
Results per Query per Database
50

Tools


VectorBase
Bioinformatics Resource for Invertebrate
Enter search terms
GO
ABOUT
ORGANI
TOOLS
DATA
HELP
COMMUNITY
Genomes
Transcriptomes
Proteomes
Release notes
Mitochondrial sequences
Home » Data
Genomes
VectorBase is committed to a new release every two months wit
state of current versions on this date (e.g., current gene sets) ca
For your species of interest, click on Organism, Strain, Assembly
ublic use b
. A list of thes
ection of Ve
me Browser link (which looks like this:
Browse Genome
).
Genomes
Organism
Strain
Assembly
Anopheles gambiae PEST AgamP3
Aedes aegypti Liverpool LVP AaegL1
Ixodes scapularis Wikel IscaW1
Culex quinquefasciatus Johannesburg JHB CpipJ1
Pediculus humanus USDA PhumU1
Rhodnius prolixus CDC RproC1 RproC1.1 17 143 ACPB02 GCA_000181055.2 702 648 350
Glossina morsitans Yale GmorY1 GmorY1.1 12 354 363 107 930
Anopheles darlingi Coari AdarC1 AdarC1.1 11 430 ADMH01 GCA_000211455.2 113 509 265
Pre-released genomes, with mapped data and gene predictions
Organism Strain Assembly Gene set Gene Count GenBank WGS Project GenBank Assembly ID Genome Size (bp)
Lutzomyia longipalpis Jacobina LlonJ1 LlonJ1.0 10 110 AJWK01 GCA_000265325.1 154 229 266
Phlebotomus papatasi Israel PpapI1 PpapI1.0 11 164 AJVK01 GCA_000262795.1 347 840 937
Anopheles albimanus STECLA AalbS1 AalbS1.0 11 911 APCK01 GCA_000349125.1 170 508 315
Anopheles arabiensis Dongola AaraD1 AaraD1.0 13 162 APCN01 GCA_000349185.1 211 443 117
Anopheles quadriannulatus A SANGWE AquaS1 AquaS1.0 13 349 APCH01 GCA_000349065.1 283 828 998
Anopheles christyi ACHKN1017 AchrA1 AchrA1.0 10 738 APCM01 GCA_000349165.1 172 658 580
Anopheles epiroticus Epiroticus2 AepiE1 AepiE1.0 12 078 APCJ01 GCA_000349105.1 223 486 714
Anopheles stephensi Indian AstelI1 AstelI1.0 21 785 158 260 098
Anopheles stephensi SDA-500 AsteS1 AsteS1.0 13 113 APCG01 GCA_000349045.1 225 369 006
Anopheles funestus FUMOZ AfunF1 AfunF1.0 13 344 APCI01 GCA_000349085.1 225 223 604

Django


Form input
Template output
(1)
(5)
Django Framework
(2)
Developer's Code
(3)
(4)
Database
External Processes

VECNet – Example for Django

UNIVERSITY OF NOTRE DAME

VECNet

About Tools Login

VECNet

Vector Ecology and Control Network

Our Work

Though malaria remains both treatable and preventable, 350-500 million people worldwide are infected with the disease every year, with up to one million cases ending in death. Nearly 85 percent of the victims who die are younger than five years old.

Recent global efforts have contributed to declines in malaria-related sickness and death, but while the present available tools for controlling malaria are effective, they will not by themselves eliminate the disease. There is a need for new strategies to eliminate malaria.

VECNet

VECNet is a consortium of institutions assembled to address the need for new strategies to eliminate malaria, which requires an understanding of how interventions affect the transmission of the disease across different geographic areas where the mosquitoes that transmit malaria differ in their behavior.

Follow @VECNetNews

VECNet Alpha Release August 14, 2013

The VECNet website is currently in "Alpha Release" while the VECNet team tests its design and functionality. If you received an invitation to be an Alpha tester, please request an account here. If you are interested in helping to test the "Beta Release" expected later this year, please register and indicate that you would like to be a Beta tester.

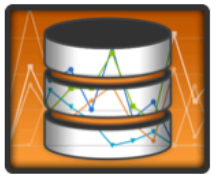
Data

VECNet enables national malaria control managers, researchers, product developers, funding bodies and policy makers to ask questions such as: *'What is the intensity and type of intervention/s required to achieve one malaria death per 100,000 in this particular population?'* and *'What is the impact on malaria transmission of a new tool that potentially kills a mosquito any time it takes a sugar meal, seeks a blood meal or lays eggs?'*

To find answers to these questions, VECNet is developing three resources: the **Digital Library**, the **Data Warehouse Browser** and a **Modeling Platform**. These tools provide users with access both to data on malaria transmission and to modeling software to create simulations of various scenarios. The simulations use geospatially specific data to analyze the potential of different combinations of control interventions to reduce malaria transmissions in a given area.

Digital Library

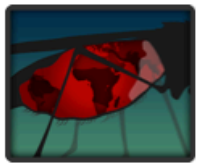
The Digital Library assembles all published and unpublished data on malaria vectors. This extensive database enables the analysis of transmission risk as a function of vector ecology and behavior via the Modeling Platform.



Data Warehouse Browser

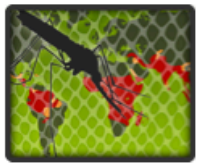
The Data Warehouse Browser is an incentive-based platform for data sharing, and enables easy-to-use, secure access to data for use with any of the Modeling Platform tools. It presents research data in ways that allow model simulations with specific data in geographically defined areas. Users can access all existing data, including their own model input and output data.

Modeling Platform



Transmission Simulator

Researchers use their data as input to model the sensitivity of transmission to changes in the behaviors of vectors resulting from responses to interventions or changes in the environment/ecology.



Risk Mapper

Risk Mapper analyzes the impact of particular interventions on malaria. National malaria control programs can use it to compare the distribution of vector control interventions against the distribution of specific vector species to determine the appropriateness of the intervention.

Product Impact Evaluator (PIE)

Investors use PIE to estimate the value of new control tools. With PIE, product developers estimate the effect of candidate tools on malaria transmission and can then develop and refine Target Product Profiles to achieve a desired level of impact.

Computational Intervention portFolio EvaluatoR (CIFER)

CIFER is an amalgamation of the outputs of Transmission Simulator, Risk Mapper, and PIE, combining the estimates of contributions from individual vector species, individual geographies, and individual interventions to overall malaria transmission. Policymakers can refine the suite of tools needed to achieve malaria eradication by analyzing how interventions affect the transmission of the disease and, as importantly, where interventions will not be able to achieve effective control and elimination.

Risk Mapper

UNIVERSITY OF NOTRE DAME

VECNet

About Tools Logout (sgesing)

Home page

BUILD SCENARIO

- Start page
- Location
- Model Parameters
- Dominant vectors
- Behavior/Habitat
- Baseline Transmission
- Interventions
- Summary

Risk Mapper / Build Scenario / Start Page

Step 1 of 8 : Start Page

Name the scenario and provide a helpful description (optional)

Scenario name: Type in a scenario name

Description: Type in a description of the scenario (optional)

Start Over

Next Step

Risk Mapper

UNIVERSITY OF NOTRE DAME

- Home page

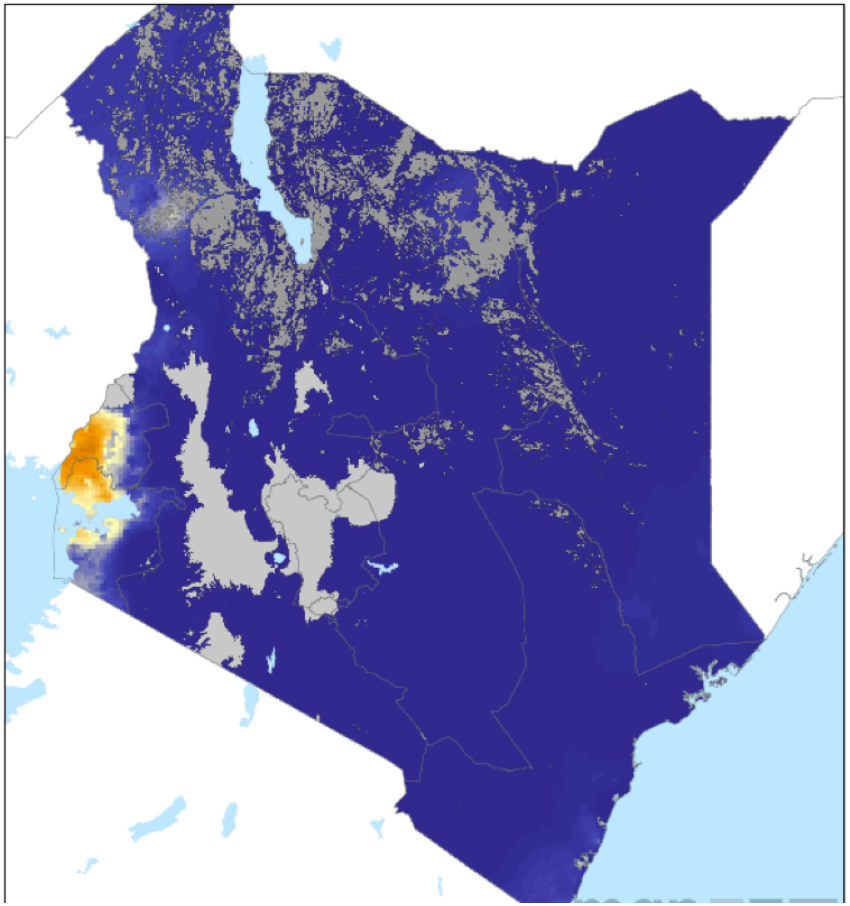
BUILD SCENARIO

- Start page
- Location
- Model Parameters
- Dominant vectors
- Behavior/Habitat
- Baseline Transmission
- Interventions
- Summary

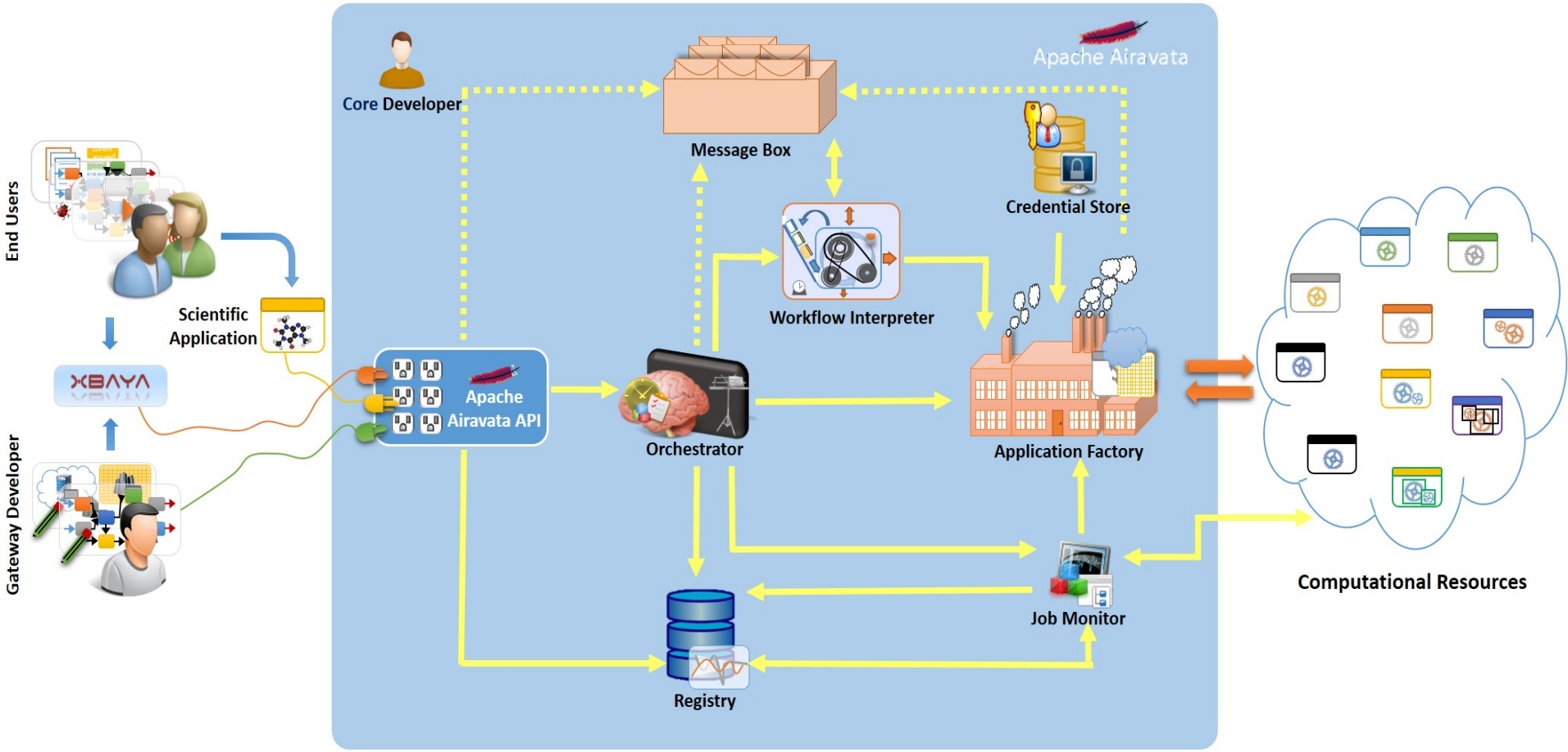
Risk Mapper / Build Scenario (Kenya) / Baseline Transmission

Step 6 of 8 : Baseline Transmission

Annual EIR data, Courtesy: Malaria Altas Project



Apache Airavata - API



Apache Airavata

- XBaya Workflow Suite
  - GUI for workflow composition and monitoring
  - export to various workflow languages BPEL, SCUFL, Condor DAG, Jython and Java
- GFac - an application wrapper service
- WS-execution
- Registry-API: A thick client registry API for Airavata to put and get documents. Current registry implementation is on top of derby or MySQL

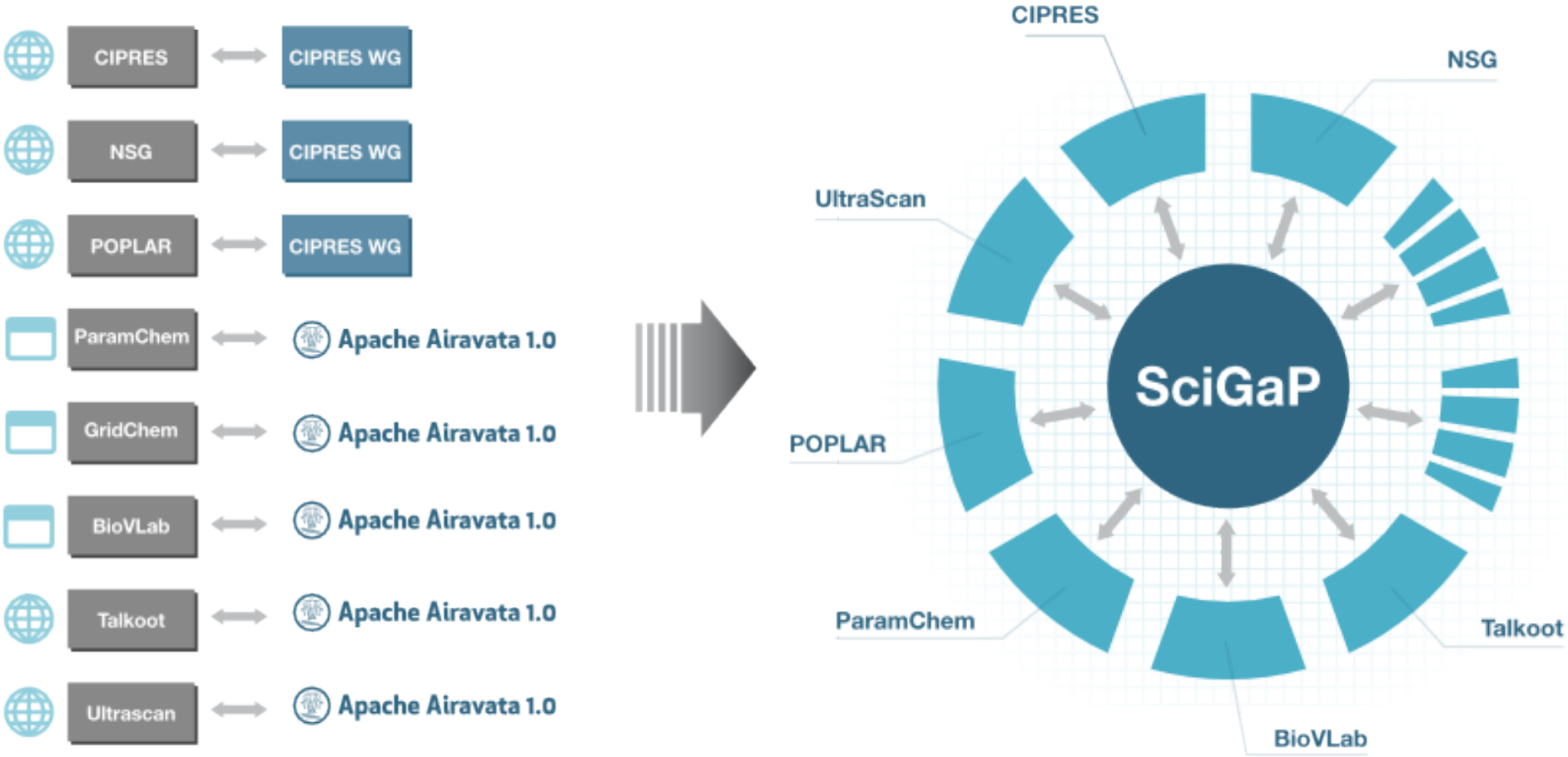
Services on basis of Apache Axis2 services

Apache Airavata Features

- desktop tools and browser-based web interface components for managing applications, workflows and generated data
- sophisticated server-side tools for registering and managing scientific applications on computational resources
- graphical user interfaces to construct, execute, control, manage and reuse of scientific workflows
- interfacing and interoperability with with various external (third party) data, workflow and provenance management tools

SciGap – Example for Airavata

Science Gateway Platform as a Service



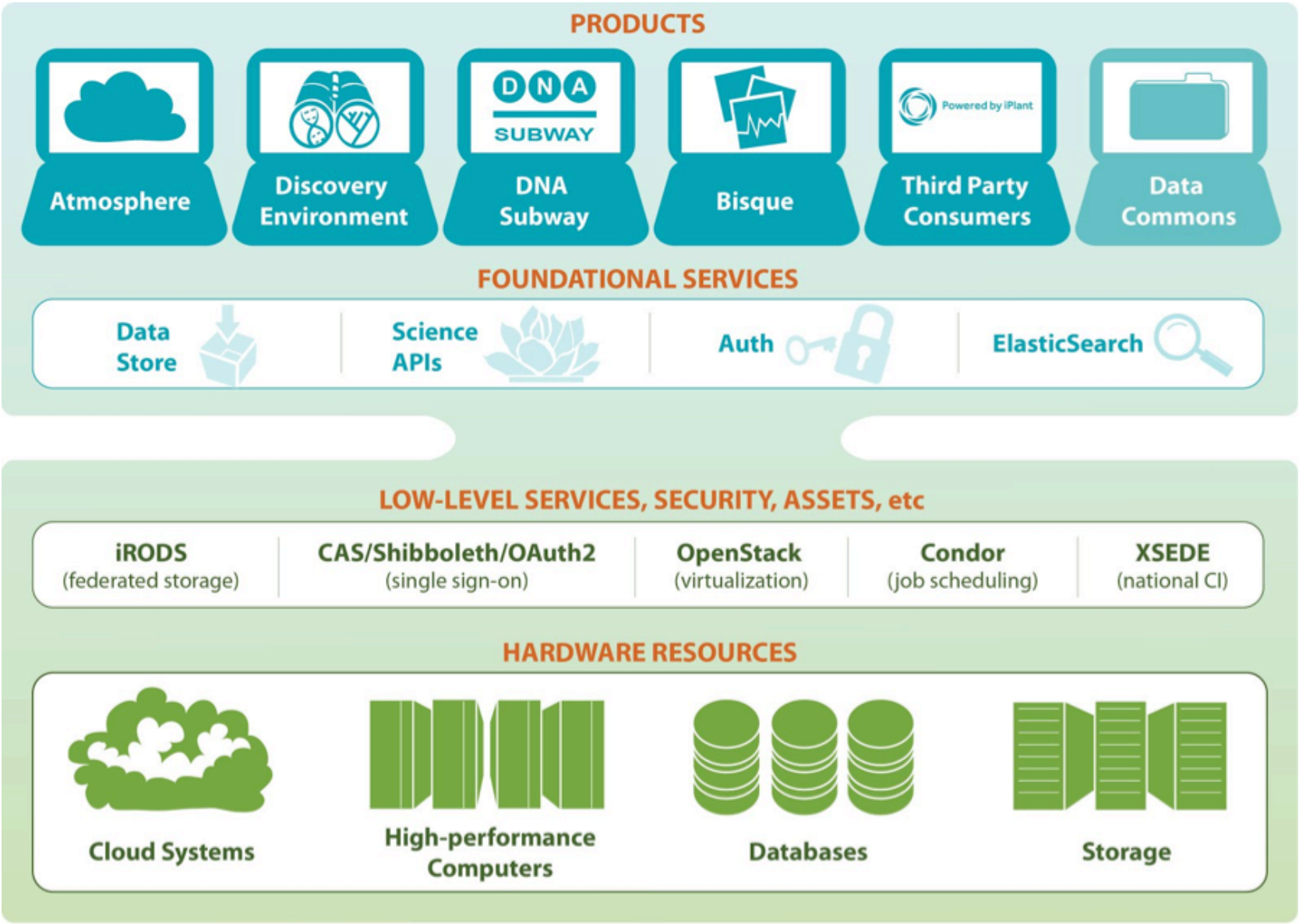
Agave API

UNIVERSITY OF NOTRE DAME

Science-as-a-Service API Platform

- REST API
- Authentication (OAuth2)
- System Management
- Application Management
- Job Management
- Data Management
- Notifications and Events
- SDK (Java, Python, PHP)
- User interfaces with HTML5, JSP

iPlant – Example for Agave API



iPlant – Example for Agave API

UNIVERSITY OF NOTRE DAME

Bisque WebApp | Services | Upload | Download | Images | Find images using tags | Martha Narro

3. Results:

The module ran in 57 seconds

Tracked root tips

Plot | Export

1680.0000 um

IBM229s4.multiroot.tif

Bisque root tip angles.png

Prev/Next | Zoom | Tool | Annotate | Sidebar | Search...

Tip angles

tip-0

Index	Value
0	-0.0066222
1	0.038471
2	0.051409
3	0.00014604
4	-0.00041889
5	-0.0037973
6	0.057078
7	0.084265
8	0.13622
9	0.26852
10	0.23715
11	0.38589
12	0.33344
13	0.50285
14	0.58676
15	0.52631

tip-1

tip-2

tip-3

tip-4

tip-5

tip-0 tip-1 tip-2 tip-3 tip-4 tip-5

Help and workflow

1. Select input image

Select an image by double clicking the image of interest, it will be loaded into the viewer for the step 2. The input time series first image should look something like this:

The 100% magnification on our images lookes like this:

2. Select initial tip positions

Science Gateways Ruling Them All?

UNIVERSITY OF NOTRE DAME

Crucial Topics

- Close collaboration with user communities
- Knowledge about available technical solutions

Sounds easy but…

- Requirements of user communities often not so clear
- Technologies sometimes still under development for certain building blocks

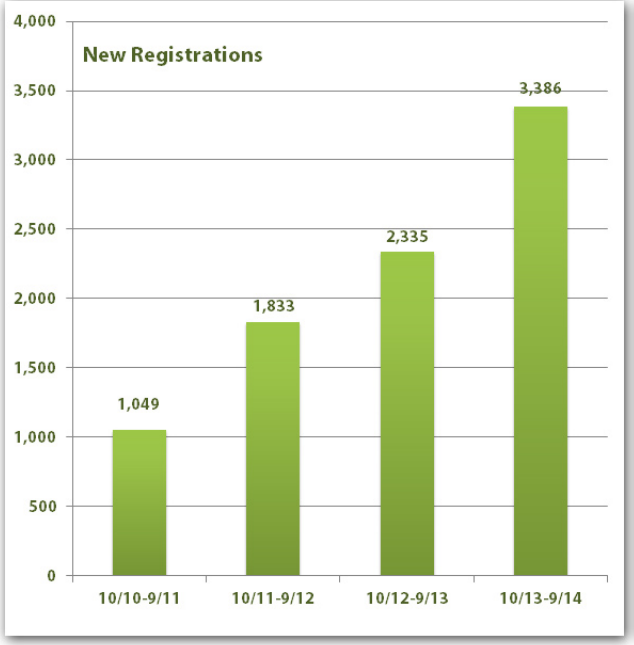
➔ Slow uptake of solutions

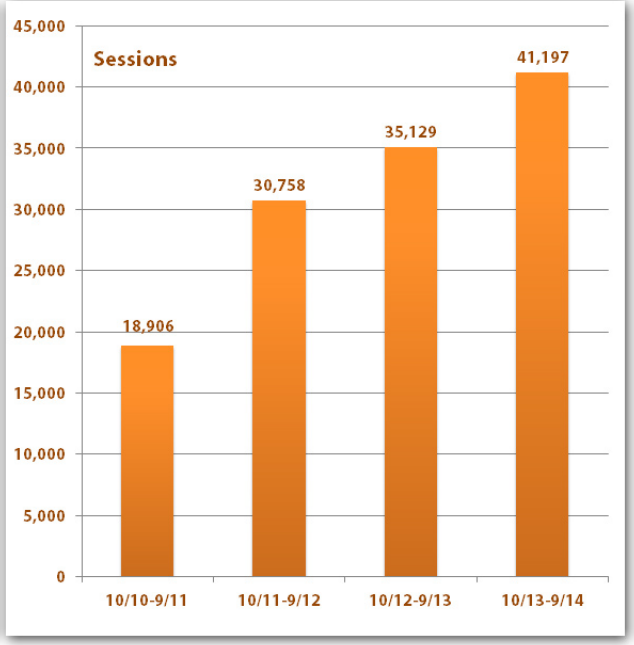
➔ Larger effort for creating science gateways

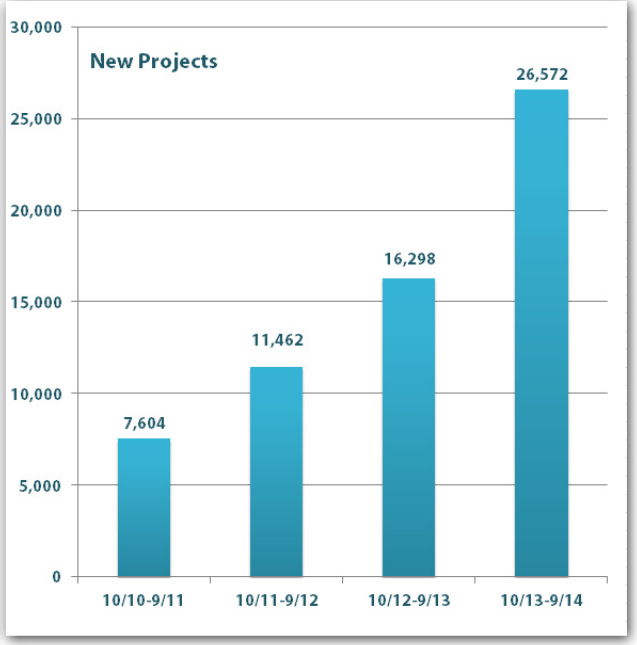
Science Gateways

A new era…

- Novel developments of web-based agile frameworks
- Infrastructure providers report that science gateways are more used than commandlines







http://www.iplantcollaborative.org

Usability vs. Re-Usability

- User side
  - Methods
  - Workflows
  - Data

  ➔ Re-usability increases usability on the user side

- Admin/Developer side
  - Frameworks
  - Libraries
  - Source code
  - Knowledge of programming languages

  ➔ Usability and re-usability depend on support, documentation and scalability

Usability vs. Re-Usability

- User side
  - Layout
  - Visualization
  - Security

  ➔ Re-used parts may be not sufficient, usability depends on the features needed in the community

- Admin/Developer side
  - Integration with computing and data infrastructures
  - Security

  ➔ Usability and re-usability depend on available infrastructures

New Science Gateway - Checklist

UNIVERSITY OF NOTRE DAME

- Demands of the user community on the user interface
- Demands on security and identities
- Demands on computing and data resources
- Workflows
- Performance

- Existing tools and models
- Available underlying infrastructure
- Available documentation and support
- Effort on development and maintenance

Information on Science Gateways

UNIVERSITY OF NOTRE DAME

- IEEE Technical Area on Science Gateways

  http://ieeesciencegateways.org

- Science Gateway Institute

  http://sciencegateways.org

- Science Gateway Workshops (partner workshops)

  Europe: IWSG (International Workshop on Science Gateways)

  USA: GCE (Gateway Computing Environments)

  Australia: In planning stage

- XSEDE Science Gateways

  https://www.xsede.org/gateways-overview

- NERSC Science Gateways

  http://portal.nersc.gov

- Center for Research Computing at Notre Dame

  http://researchcomputing.nd.edu

sandra.gesing@nd.edu