Web-based Science Gateways for Structural Bioinformatics

Sandra Gesing
Applied Bioinformatics Group
sandra.gesing@uni-tuebingen.de
19.09.2012
Tübingen
University of Tübingen

Department of Computer Science
(Wilhelm Schickard Institute)
University of Tübingen

• University was founded in 1477
• Since 1863 natural sciences as first German university
• Since 1990 computer science
Applied Bioinformatics Group

Prof. Oliver Kohlbacher – Applied Bioinformatics Group (Director of the Quantitative Biology Center)

Research topics
• Structural Bioinformatics
• Immunoinformatics
• Systems Biology
• Computational Proteomics and Metabolomics
“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
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
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
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
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)
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”)
Structural Bioinformatics

• Numerous applications for structural bioinformatics (prediction and analysis of molecular structure), e.g.
  • Materials science
  • Drug design
• Sensitive and „expensive“ data
• Support by sophisticated tools and methods
• DCIs (Distributed Computing Infrastructures) available
Motivation

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

- Usability of tools often limited
- Complexity of methods

<p>|============================================================================|
| Version: 1.1                                                             |
| build date: Jan 15 2012                                                 |
| execution host: Vomitorin                                               |</p>
<table>
<thead>
<tr>
<th>execution time: 2012-01-09, 16:39:43 (MST)</th>
</tr>
</thead>
</table>

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 mol is larger
-write_par <out.file>  write xml parameter file for this tool
-get <in.file>         read parameters from parameter.xml-file

Available flags are:

-ri                    remove invalid molecules.
-ut                    check for unique topologies
-ne                    no check for unique conformations
-rt                    remove input file once 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 dcf (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 angstroms)
* 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 output 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.

Sandra Gesing
Motivation

• Usability of tools often limited
• Complexity of methods
• Lack of graphical user interfaces
• Workflows
Motivation

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

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

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

- Usability of tools often limited
- Complexity of methods
- Lack of graphical user interfaces
- Workflows
- Complexity of infrastructures
Mo8va8on

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

Web-based Science Gateways for Structural Bioinformatics

NIST Smart Grid Framework 1.0 September 2009
Motivation

- Usability of tools often limited
- Complexity of methods
- Lack of graphical user interfaces
- Workflows
- Complexity of infrastructures
- End users focus on topics diverse to computer science
- End users with different computational background
Motivation

- Usability of tools often limited
- Complexity of methods
- Lack of graphical user interfaces
- Workflows
- Complexity of infrastructures
- End users focus on topics diverse to their backgrounds
- End users with different computational backgrounds
Motivation

- Usability of tools often limited
- Complexity of methods
- Lack of graphical user interfaces
- Workflows
- Complexity of infrastructures
- End users focus on topics diverse to computer science
- End users with different computational background

⇒ User interfaces need to be intuitive and self-explanatory

⇒ Science gateways
Molecular Simulation Grid (D-Grid project)

Goal

Providing users with Grid access to molecular simulation tools and docking tools via a workflow-enabled portal

• Implementation of high-performance computing
• Use of the D-Grid-infrastructure
MoSGrid

Molecular Simulation Grid

Goal

Providing users with Grid access to molecular simulation tools and docking tools via a workflow-enabled portal

• Implementation of high-performance computing
• Use of the D-Grid infrastructure
MoSGrid

Molecular Simulations

Goal

Providing users with Grid access to molecular simulation tools and docking tools via a workflow-enabled portal

- Implementation of high-performance computing
- Use of the D-Grid infrastructure
Molecular Simulation Grid (D-Grid project)

Goal

Providing users with Grid access to molecular simulation tools and docking tools via a workflow-enabled portal

- Implementation of high-performance computing
- Use of the D-Grid-infrastructure
- Annotations of results
- Data mining
- Open-source
- Internationalization
MoSGrid

• survey of willingness to share knowledge in the community
  ⇒ 70% share results after publication, 90% share workflows
• evaluation of portals
  ⇒ Liferay, gUSE/WS-PGRADE
• evaluation of workflow engines
  ⇒ UNICORE, WS-PGRADE
• evaluation of Grid middlewares
  ⇒ UNICORE
    (D-Grid infrastructure GT 4, gLite, UNICORE)
MoSGrid in a Nutshell

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Structure

Recipe

Result

Distributed File System

XtreemFS

Portal

WS-PGRADE

Workflow

High-level middleware service level

gUSE

DCIs

UNICORE 6

Result

User-Input Portal Grid Ressource Recipe Structure Result

Result Coordinates (xyz/pdb) Multiplicity/Charge Job Type #CPU/Memory

Job Creation input.gjf g09 gaussian.log

Web-based Science Gateways for Structural Bioinformatics
Security

User
Security

User

- register
- apply for certificate
- acknowledge

RAs

VOs
Web-based Science Gateways for Structural Bioinformatics

Security

Portal server

User

- authenticate
- provide certificate
- register
- apply for certificate
- acknowledge

RAs

VOs
Security

User

authenticate

provide certificate

register

apply for certificate

acknowledge

Portal server

Account

authorised

assigned to

Roles

authorised

Publicly available content

Role-dependent available content

RAs

VOs
Security

User
- authenticate
- provide certificate
- register
- apply for certificate
- acknowledge

Portal server
- Account
- authorised
- assigned to
- Roles
- authorised
- Role-dependent available content

Middleware services
- provide credential file
- get registered users
- authenticate via credential file
- authorise via VO membership

Publicly available content

RAs

VOs

HPC facilities

MiddleWare services

Roles

Authorised

Account

Roles

Available content

RAs

VOs

HPC facilities
Credentials

• X.509 user certificates
• SAML (Security Assertion Markup Language)
  - Minimize credential data transfers
  - Set of maximum hops for trust delegation
  - Usable for single sign-on infrastructures (e.g., Shibboleth)
Security Infrastructure

User Interface
WS-PGRADE

User Interface
WS-PGRADE

High-Level Middleware
Service Layer

SAML Assertion

SAML Assertion

Grid Middleware Layer
UNICORE

Service Certificate

SAML Assertion

XtreemFS

gUSE Services

BFT

TLS

UNICORE Atomic Services

Batch System

UNICORE Target System Interface

File System

HPC Facilities
D-Grid

User
User Certificate