MoSGrid – Molecular Simulations in a Distributed Environment

Sandra Gesing
sandra.gesing@uni-tuebingen.de
18 June 2013
Molecular Simulations and Docking

- Prediction and analysis of molecular structures
- Support by sophisticated tools and methods
- Numerous applications, e.g.
  - Materials science
  - Drug design
Molecular Simulations and Docking

• Prediction and analysis of molecular structure
• Support by sophisticated tools and methods
• Numerous applications, e.g.
  • Materials science
  • Drug design
Molecular Simulations and Docking

- Data intensive and compute intensive
- Structure databases, e.g., ZINC with ~20 mio structures
- Sensitive and „expensive“ data
- Distributed data management available
- DCIs (Distributed Computing Infrastructures) available

Why do researchers not use the 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

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
 -read_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 ±6).
* 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.
Open Issues

• 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
Open Issues

- 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

Slide copied from: Stuart Owen „Workflows with Taverna“
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

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
• Users are generally not IT specialists

⇒ User interfaces need to be intuitive and self-explanatory

⇒ 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
Goal of Science Gateways

Usability of software and data

"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)
MoSGrid

Molecular Simulation Grid
• Science gateway integrated with underlying compute and data management infrastructure
• Distributed workflow management
• Data repository
• Open source
MoSGrid

Molecular Simulation Grid
- Science gateway integrated with underlying compute and data management infrastructure
- Distributed workflow management
- Data repository
- Open source

Survey of willingness to share knowledge in the community
⇒ 90% share workflows
⇒ 70% share results after publication
MoSGrid in a Nutshell

Portal
WS-PGRADE
Workflow
High-level middleware service level
gUSE
DCIs
UNICORE 6

Structure
Recipe
Result
Distributed file system
XtreemFS
Data repositories
Result

User-Input
Portal
Grid
Ressource
Recipe Structure
Result

Input: input.gjf
Software: g09, gaussian.log

Data repositories

Sandra Gesing
MoSGrid – Molecular Simulations in a Distributed Environment
16
Job and Workflow Management

grid User Support Environment

User Interface
WS-PGRADE

High-Level Middleware Service Layer
gUSE

DCI Resources
Middleware Layer
UNICORE 6
Distributed Data Management

- XtreemFS is an object-based grid and cloud filesystem
- Replication for availability, locality, bandwidth, and latency
- Easy integration in heterogeneous environments
- UNICORE extension with URL scheme xtreemfs:// available
Data Repository

- Repository consists of data and metadata storage
- MSML (Molecular Simulation Mark-up Language)
- Subset and extension of CML (Chemical Mark-up Language)
- Unified data representation
- Used for storing structures of molecules and macromolecules, simulation descriptions, and results
- Parsers and adapters used for conversions to and from MSML
Repository – Data Management

- First step in workflows: MSML to application specific input
- Results are computed
- Output converted to MSML
- Last step in workflows: MSML to JSON metadata (metadata extractor in UNICORE)
- JSON indexed for searchable results by UNICORE and LUCENE (text search engine API)
Job Configuration

Job's name: ParserProtein
Optional note: Description of Job

Type: unicores
Grid: flavus.informatik.uni-tuebingen.de:8090
Tools: Bash shell 3.1.16
Execute parser: 
Replicate settings in all Jobs: 
Copy job names to tools: 
Kind of binary: Sequential
MPI Node Number: 
Executable code of binary: Recently stored: Durchsuchen...
Parameter: genparser.sh ProteinPro
### Job Configuration

**Job's name:** PDBCutter

**Optional note:** Description of Job

<table>
<thead>
<tr>
<th>Type</th>
<th>unicore</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grid</td>
<td>flavus.informatik.uni-tuebingen.de:8090</td>
</tr>
</tbody>
</table>
| Tools              | PDBCutter 1.0.0, 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, ...
| Execute parser     |           |
| Replicate settings in all Jobs |           |
| Copy job names to tools |           |
| Kind of binary     |           |
| MPI Node Number    |           |
| Executable code of binary |           |
| Parameter          |           |
Remote File Configuration
Domain-specific Workflows

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)
MD Portlet

Molecular Dynamics Simulation Portlet

Welcome Workflow Submission Workflow Status

Get workflow status

Workflows list

NAME
birke/2011-14-1/11:59:42/EQ_Comp
birke/2011-14-1/12:15:31/SIN
birke/2011-14-1/12:19:28/EQ_Comp
birke/2011-14-1/12:52:56/EQ_Comp
jk/2011-14-1/13:41:38/EQ_Comp

https://unicore-bisgrid.uni-paderborn.de

list files Abort workflow Describe

Files to get:

Jmol
MoSGrid – Molecular Simulations in a Distributed Environment
The MoSGrid Science Gateway

Built by 17 institutes and companies

[Logos of the institutions and companies involved in MoSGrid]
The MoSGrid Science Gateway

Used by 125 user groups in 16 countries
The MoSGrid Science Gateway

Currently supports
- 65 workflows in repositories
- 90 applications
- 85 GB data in central repository
- ~50,000 structures (1 – 5 GB)
in user repositories

AutoDock Vina
Gaussian
FlexX
CADDSuite
MoSGrid ended 31.12.2012 but partners participate in

SCI-BUS (SClentific gateway Based User Support)
• EU project 01.10.2011 – 30.09.2014
• Extension of the MoSGrid portal with an interactive molecule editor based on WebGL and a semantic search

ER-flow (Building an European Research Community through Interoperable Workflows and Data)
• EU project 01.10.2012 – 30.09.2014
• Integration of applications in SHIWA simulation platform
• Study of data exchange between workflow systems
• Community management
Thank you!

sandra.gesing@uni-tuebingen.de