Efficient Mass Spectra Prediction through Container Orchestration with a Scientific Workflow

Maximilian Hanussek
High Performance and Cloud Computing Group
Zentrum für Datenverarbeitung
Eberhard Karls Universität Tübingen

20.06.17
Poznań
• Introduction
• Key-technologies
• UNICORE workflow for QCEIMS
• Results & Discussion
• Conclusion/Outlook
Introduction

- Many useful tools are available
  - Hard to install
  - Difficult to use
  - Small user community

- Use existing technologies to simplify installation and usage
  - Broaden user community

Goal: Evaluation of Docker and UNICORE
Use case

- Quantum Chemical Electron Ionization Mass Spectrometry (QCEIMS)
  - Mass spectra prediction tool (Stephan Grimme et. al)\(^1\)
  - Complex execution procedure
  - Many required additional tools and software packages
  - HPC
- Nonetheless well working tool with good results

• Quantum Chemical Electron Ionization Mass Spectrometry (QCEIMS)
• Docker (container virtualization technology)
• Uniform Interface to Computing Resources (UNICORE)
Step 1

Step 2

Step 3

Step 4

neutral loss (H atom)

Key-technologies (Docker)

- Built Docker Image from Dockerfile
- Most tools included (MNDO99, DFTB+, ORCA, PubChemPy, ...)
- Stable computing environment
- Installation with root permissions
Key-technologies (UNICORE)

- Middleware software for distributed computing systems
- Developed at the research center Jülich and by further partners
- Offering different components:

<table>
<thead>
<tr>
<th>UNICORE Server</th>
<th>UNICORE Workflows</th>
<th>UNICORE Portal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Job submission</td>
<td>Workflow engine</td>
<td>Web access</td>
</tr>
<tr>
<td>Data handling</td>
<td>Connected to UNICORE Server</td>
<td>Use existing workflows</td>
</tr>
<tr>
<td>Resource allocation</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

• Combined usage of Docker and UNICORE
• Automating the QCEIMS calculations
• Most software already preinstalled in Docker image
UNICORE Workflow for QCEIMS

Start

For each molecule

- File splitting
  - Format conversion and annotation
  - Open shell check
    - Open shell molecule
      - End
      - Database insertion
      - Report generation
      - Output merging
    - No open shell molecule
      - Configfile generation
      - Ground state calculation
      - Fragment generation
      - Fragmentation calculation
      - Fragmentation loop
Simple Benchmark Set

- 11 molecules
- 2 QC tools (MNDO99, DFTB+)
- OM2-D3 (MNDO99)
- DFTB3-D3 (DFTB+)
- Absolute value distance$^2$:

\[ D_A(I_E, I_S) = \left( 1 + \frac{I_E - I_S}{I_S} \right)^{1} \]

Glycine (DFTB3-D3, 300 trj., $D_A$: 0.752)
Nicotinamide (DFTB3-D3, 200 trj., $D_A$: 0.660)
Mean molecule score

- **GABA**
- **Glycine**
- **Glycolic acid**
- **Nicotinamide**
- **Acetonitrile**
- **Ethanethiole**
- 2−Aminopropane
- 2−Nitrotoluene
- Vanillic acid
- Benzyl Cyanide
- Umbelliferone

**MNDO99**

**DFTB+**
Runtimes (QCEIMS)

Accumulated walltimes

OM2-D3
DFTB3-D3

CPU time [h]

Number of Trajectories
Results & Discussion (Docker)

- Successful integration as wrapping tool
- Occurred problems
  - Docker group = root
  - User mapping is mandatory
  - Manual garbage collection of exited and dead containers
  - Accumulation of Docker metadata (/var/lib/docker)
Challenges to overcome

- Encapsulation of multiple molecules
- Encapsulation of a molecule for fragmentation
- Merging generated results by unique identifier
- Explicit data staging
- QC tool monitoring
- Unicore-portal workflow adaption
  - Export workflow from UNICORE Rich Client (URC)
  - Integrate input field
  - Set independent path to workflow node scripts
UNICORE portal authentication

TLS login

You will be logged as: CN=Maximilian Hanussek,OU=Universitaet Tuebingen,O=Grid

Login with certificate

Switch to user registration
Job Computation

New UNICORE Job

Application

Job name: New UNICORE Job
Tag: 
Select application: Workflow Template
Select version: any version
Command line arguments: 

Workflow template parameters

Select a template

Editing workflow template file: UNICORE_QCEiMS_workflow.xml

splitSdf_INPUT: glycine.sdf

Submit
Conclusion/Outlook

- Successful simplification of QCEIMS
  - Default parameter change, QC tools (ORCA)
- Docker: Great potential, serious problems
- UNICORE: Generally applicable to complex tools
  - Login/Registration procedure

Relatively good in silico spectra is better than no spectrum.
Acknowledgment

Applied Bioinformatics group Tübingen
• Prof. Dr. Oliver Kohlbacher

High performance and cloud computing group Tübingen
• Dr. Jens Krüger

UNICORE-Support
• Dr. Bernd Schuller

QCEIMS
• Christoph Bauer, Prof. Dr. Stefan Grimme