Taverna workflows: provenance and reproducibility

Aleksandra Pawlik
The University of Manchester

Workflow approaches to investigation of biological complexity
STFC/NERC Workshop 15-16 October 2013
Workflows for improvement

Workflows are more than just pipelines...

- Scaling up automated execution
- Bringing together distributed and continually changing resources
- Dealing with different standards, interfaces and implementation
- Support for repeatable analysis
Taverna Engine Execution

- Workflows in Scufl2
- Functional dataflow, simple control flows, implicit iteration
- Linking services and tools
- Different data resources and formats
- “In Workflow Programming” (eg. Beanshell scripting)
- Provenance collection: W3C PROV-O, OPM
- Plug-in Framework
  - Infrastructures: Web Services (SOAP, REST), Grid, HPC
  - Common Tools: Excel Spreadsheets, Google Refine, R
- OAuth security plug-in
Taverna Workbench

- Customizable for domains (eg. expose services only for biodiversity)
- Desktop application
- Intermediate results views
- Plug-in framework

![List of services](image)

![Workflow engine to run workflows](image)

![Construct and visualise workflows](image)
Taverna User Spectrum

Taverna

Workflow Engineer
Computational Scientist
Domain Scientist (Workflow User)

Concept Knowledge

Domain

Workbench
Workbench Components
Lite

Domain-Specific Website / Tool / Portal

High Workflow Visibility

Low

myGrid
- Right apps, right users
- Commodity apps:
  - Web. Spreadsheets. R.
- Customisation
- Mixed workflow / scripting
- Deployment / Portability
  - Web based / desktop
  - Virtualised deployments
  - Cloud hosted service
  - A cloud-enabled local host
- Local ownership
- Capability building

Apps
- Domain/task specific apps that incorporate (an ecosystem of) workflows. Integrate

Workflow
- Parameterised, integrative, multi-step (data) pipelines, analytics, computational protocols. Can be repetitively reused.

WFMS middleware
- Support design, config. and execution of workflows. manage utility actions for data, logging, security, compute, error. Shield incompatibilities & complexity.

Infrastructure
- Legacy, others and your own software, datasets, services, codes, and platforms. Optimise and manage use of computing infrastructure.
Reuse and Reproducibility

Research Article

Genetic analysis of the *Trichuris muris*-induced model of colitis reveals QTL overlap and a novel gene cluster for establishing colonic inflammation


Abstract

Background: Wide Association Inflammatory Bowel Disease susceptibility loci for colitis in pigs to date. We have now started to perform an experiment.

Methods: We analyzed the pigs in details and found that the pigs had improved susceptibility to colitis. The pigs were then analyzed and found to be susceptible to colitis.

Results: 7 autosomal QTL regions were associated with the establishment of chronic inflammation in pigs. The QTLs on chromosomes 3 (A2, A3), 7 (A24, A25), and 8 (A27) were found to be associated with susceptibility to colitis. The QTLs on chromosomes 3 (A24, A25) and 8 (A27) were found to be associated with susceptibility to colitis.

Conclusion: Genetic susceptibility to induced colonic mucosal inflammation in the pigs revealed different biological properties. Combined methodology in transcripational and pathway data allowed identification of biologically relevant candidates implicated as a colitis susceptibility gene in functional relevance.

Keywords: Trichuris muris, Colitis, Genetic susceptibility, Gastrointestinal tract.

Workflow

- **Trichuriasis Induced Colitis (Paul Fisher)**
  - Added by Paul Fisher ... more than 1 year ago (16/02/11 @ 15:26:36)
- **KEGG pathways common to both QTL and microarray based investigations (Paul Fisher)**
  - Added by Paul Fisher ... more than 1 year ago (16/02/11 @ 13:08:18)
- **Pathways and Gene annotations for RefSeq ids (Paul Fisher)**
  - Added by Paul Fisher ... more than 1 year ago (16/02/11 @ 13:07:54)
- **Pathway and Gene to Pubmed (Paul Fisher)**
  - Added by Paul Fisher ... more than 1 year ago (16/02/11 @ 13:07:10)
- **Pathways and Gene annotations for QTL region (Paul Fisher)**
  - Added by Paul Fisher ... more than 1 year ago (16/02/11 @ 13:00:39)
~6,000 members
over 300 groups,
over 3,000 workflows
Taverna Components

Workflow Blocks made of a workflow

- Well described
- Well behaved
- Well looked after
- Agreed fail
- Agreed formats in and out
- Agreed provenance

Deposited in myExperiment
Grouped into families
Provenance: how did you do it?

- The link between computation and results
- Collecting -> Using Provenance
- Reporting at different scales/levels

PDIFF: comparing provenance traces to diagnose divergence across experimental results [Woodman et al, 2011]
Enabling **reproducible**, transparent research.

http://www.w3.org/community/rosc/

Research Objects

http://www.researchobject.org/

**bundles and relates digital resources of a scientific experiment or investigation using standard mechanisms**
Taverna in Galaxy

Tools

Wrap as Tool

Workflow in my experiment

Upload

Galaxy execution

Taverna server
The Taverna Suite of Tools

Workflow

User Interfaces

Workbench

Taverna Lite

Workflow Engine

Workflow Provenance

Workflow Server

Interaction Server

Workflow Components

Activity and Service
Plug-in Manager

Service Catalogue

Third Party Tools

Web Portals / Gateways

Client User Interfaces

Third Party Tools

Virtual Machine

Interaction Server

Command Line

Player

APLs
Sustainability and user support

Freely available
Open source
Current version 2.4

80,000+ downloads across version

Windows/Mac OS X/Linux/Unix

Tutorials and Workshops

Active user forum & support

www.taverna.org.uk
Taverna in other projects

BioDiversity Virtual e-Laboratory
www.biovel.eu

SCAPE
www.scape-project.eu

Wf4Ever
www wf4ever-project.org

VPH-Share
www.vph-share.eu

HELIO
www.helio-vo.eu

RightField
HELIO
www.helio-vo.eu

Pacific Northwest National Laboratory
www.pnnl.gov

Scientific Workflows and Provenance Working Group
www.dataone.org

SHIWA
www.shiwa-workflow.eu

iPlant Collaborative
www.iplantcollaborative.org
**Data-centric Computation**
Scientific workflows over Distributed Cyber-Infrastructure.

**Data sharing**
libraries and catalogues for all types of scientific artefacts and all types of scientists.

**Knowledge Management**
Metadata, semantics digital exchange, preservation, publishing

**Software Engineering**
Software sustainability, software and data policy, training

**Applications**
- Astronomy
- Library
- Digital Preservation
- Biodiversity
- Biology
- Systems Biology
- Chemistry
- Public Health
- Astro-Physics
- Social Science
For more information

- Taverna
  - http://www.taverna.org.uk

- myExperiment
  - http://www.myexperiment.org

- myGrid
  - http://www.mygrid.org.uk