Taverna: A Workbench for the Design and Execution of Scientific Workflows

Dr Katy Wolstencroft

myGrid

University of Manchester

Taverna and myGrid

- myGrid a suite of components designed to support *in silico* experiments in biology
- Taverna workbench – main user interface
- Semantic service discovery components
- myGrid Ontology for bioinformatics services
- Provenance components
- myGrid provenance ontology

Client Applications

Provenance Ontology

Workflow Warehouse

myExperiment Web Portal

Taverna Workbench GUI

Provenance Warehouse

Service / Component Catalogue

Feta Information Services

Taverna Workflow Enactor

LogBook Provenance Management

Default Results

Service Ontology

3rd Party Resources (Web Services, Grid Services)

Custom Datasets

Service Management

Resources

Is Taverna Just for Biologists?

- Nothing in the code is specific to biology
- The default list of services ARE bio services, but Taverna doesn't care what they are
- Services from other science disciplines can simply be slotted in

Taverna Workbench v1.5.1.6

File Tools Workflows Advanced Help

Design Results Discover LogBook test2

Fetch today's Dilbert comic 09:58

Save as XML | Save to disk | Save to disk as website | Excel | Close

Status Results Process report

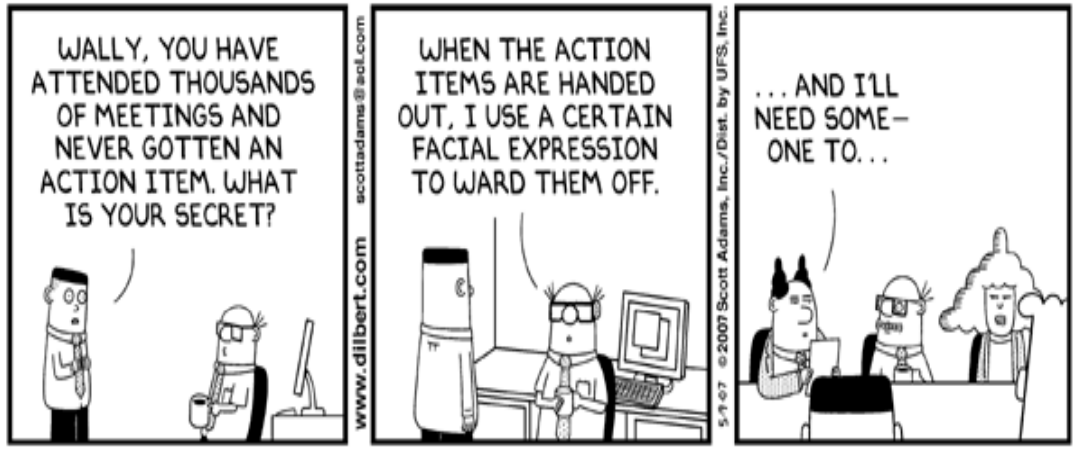
Processor statii

...	Name	Last event	Event timestamp	Event detail	Breakpoint
	getImageLinks	ProcessComplete	09-May-2007 09:58:11		
	findComicURL	ProcessComplete	09-May-2007 09:58:11		
	getPage	ProcessComplete	09-May-2007 09:58:11		

Graph Intermediate inputs Intermediate outputs

image

List
urn:lsid:net.sf.taverna:dataCollection:b2c7b8
application/octet-stream,image/*
urn:lsid:net.sf.taverna:dataItem:3264



© Scott Adams, Inc./Dist. by UFS, Inc.

start taverna-1.5.0 Taverna_astrogrid GgfIvoaMay07W... Medical and Heal... C:\WINDOWS\sy... Taverna Workbe... 10:01

What Taverna Gives you

- Automation
- Implicit iteration
- Implicit parallelisation
- Support for nested workflow construction
- Error handling
  - Retry, failover and automatic substitution of alternates

Extensibility

- Accepts many types of services:

  - web services (WSDL), beanshell scripts, local java scripts, R-processor, Biomart, BioMoby, Soaplab, Workflows

- Easy to add your own services
- Plug-in architecture
  - Easy to build new processor types
  - Easy to extend to include alternative results viewers

Who Provides the Services?

- Open domain services and resources.
- Taverna accesses 3000+ services
- Third party – we don't own them – we didn't build them
- All the major providers
  - NCBI, DDBJ, EBI …
- Enforce NO common data model.

- Quality Web Services considered desirable

National Center for Biotechnology Information (USA)

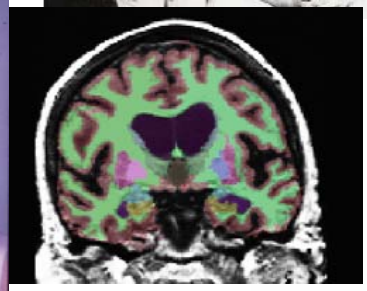
Tokyo, Japan

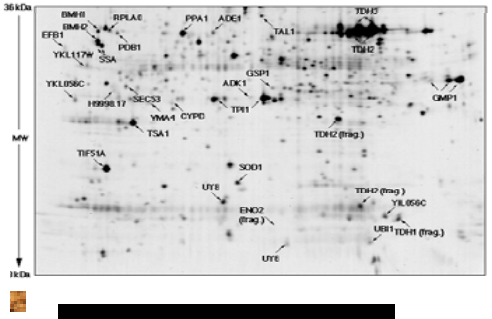
Cambridge, UK

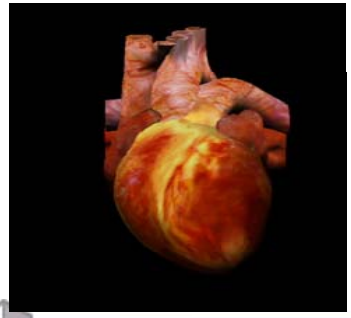
SRS

SeqHound

Who uses Taverna?







~42800 downloads

Users worldwide

- Systems biology
- Proteomics
- Gene/protein annotation
- Microarray data analysis
- Medical image analysis
- Heart simulations
- High throughput screening
- Genotype/Phenotype studies
- Health Informatics
- Astronomy
- Chemoinformatics
- Data integration

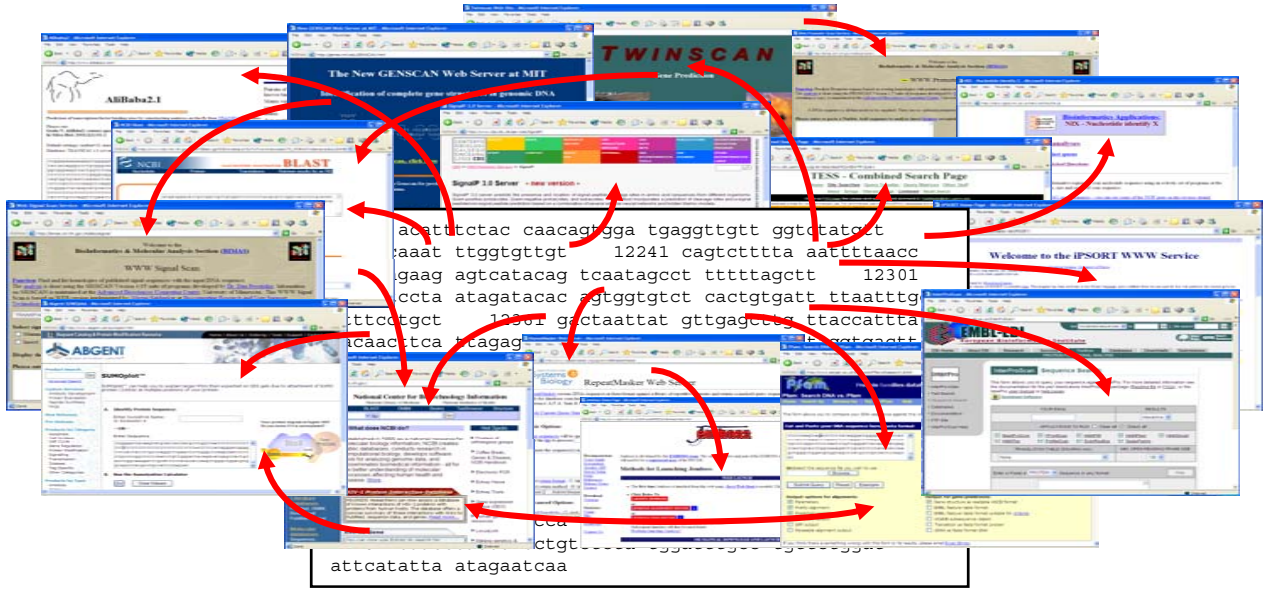
- ISMB07 – 6 posters, 2 demos, 1 BOF, 1 tutorial

What do Scientists use Taverna for?

- Data gathering and annotating
  - Distributed data and knowledge
- Data analysis
  - Distributed analysis tools and high throughput
- Data mining and knowledge management
  - Hypothesis generation and modelling

Data Gathering

- Collecting evidence from lots of places
- Accessing local and remote databases, extracting info and displaying a unified view to the user



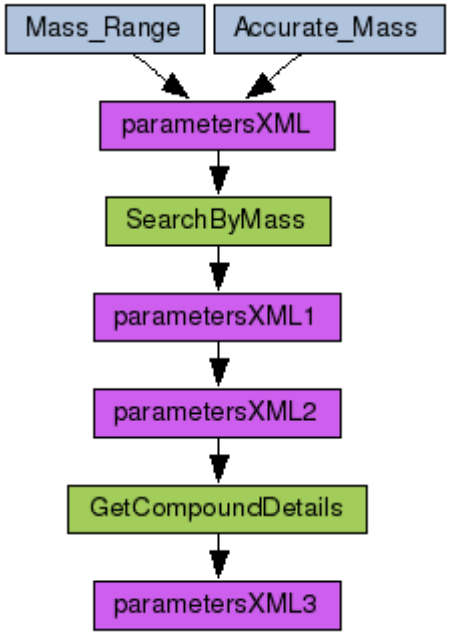
Annotation Pipelines

- Genome annotation pipelines
  - Bergen Center for Computational Science – Gene Prediction in Algal Viruses, a case study.
    - Workflow assembles evidence for predicted genes / potential functions
    - Human expert can ‘review’ this evidence before submission to the genome database
- Data warehouse pipelines
  - e-Fungi – model organism warehouse
  - ISPIDER – proteomics warehouse

Chemoinformatics

- CDK Taverna – a chemoinformatics plugin

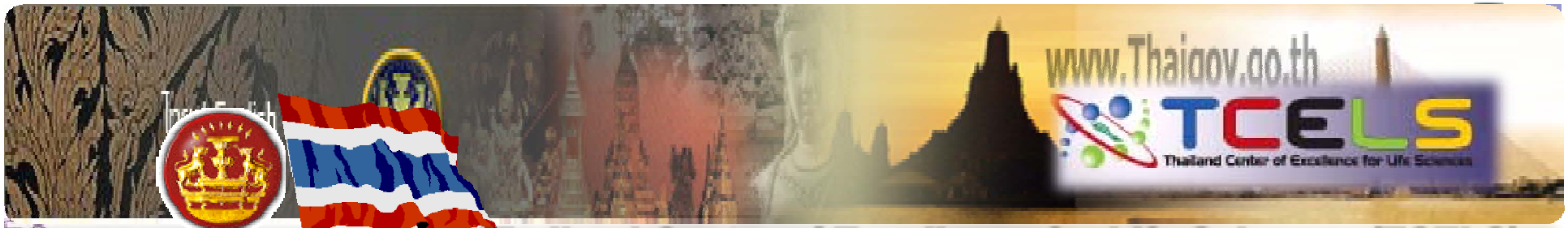
Chemical information retrieval workflows



- Structures
- reactions
- object relational data

Data Analysis

- Access to local and remote analysis tool
- You start with your own data / public data of interest
- You need to analyse it to extract biological knowledge



Prime Minister's Office

Thailand Center of Excellence for Life Sciences (TCELS)

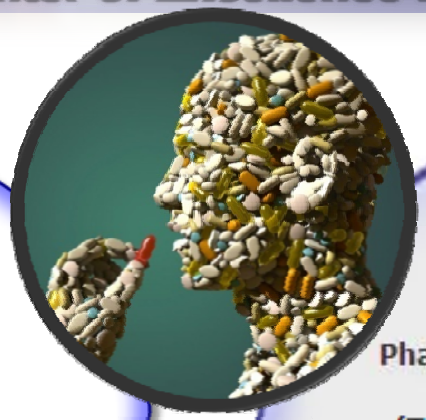
HIV-1 Pharmacogenomics (Nevirapine-rash)
อ. ศศิโสภิณ เกียรติบูรณกุล

Drug allergy: Allopurinol, Carbamazepine
อ.ทิชา ลิ้มสุวรรณ, อ. **วสุ ก่าชัยเสถียร**

Pharmacogenomics in Childhood acute lymphoblastic leukemia
อ สุรเดช หงส์อิง

Pharmacogenomics in oncology-chemotherapy
อ. เอกภพ สิระชัยนันท์

Pharmacogenomics in Thalassemia
อ. ธันยชัย สุระ



Pharmacogenomics in Psychiatric Diseases
(Tsunami victims and relatives)
Chunlalongkorn Hospital
Rajanukul Institute

Wasun Chantratita
Project director

Pharmacogenomics project

myGrid

Individualized medicine

Pharmacogenomics

- Heavy use of R-Statistics for clinical data analysis
- Association study of Nevirapine-induced skin rash in Thai Population
- A systemic (bodywide) allergic reaction with a characteristic rash
  - 100 Cases: rash – 100 Cases: no rash controls
  - 10,000 SNP significantly associated with rash
  - Pathway analysis and systems biology
  - Prioritising SNPs
  - Functional studies
  - Diagnostic tools

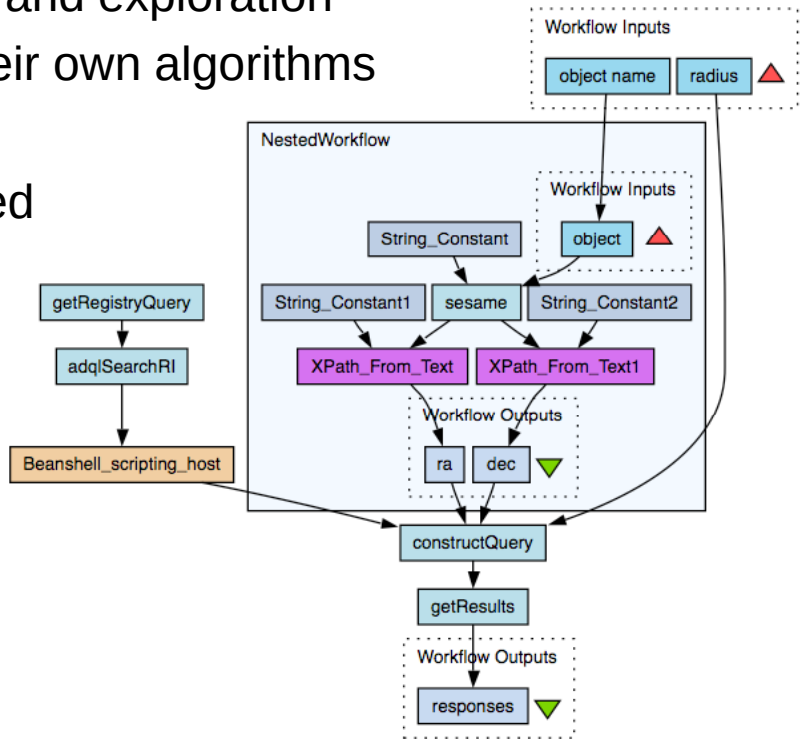
Health Informatics

- Medical imaging – MIAS Grid
  - Investigating cartilage thickness during drug trials
  - 2D and 3D brain image registration

- PsyGrid
  - Investigating first episode psychosis over 2 years and study epidemiology of schizophrenia
  - support studies and clinical trials in mental health

AstroGrid – A working Virtual Observatory (VO)

- High throughput datamining facilities for interrogating those databases
- A uniform archive query and data-mining software interface
- The ability to browse simultaneously multiple datasets
- A set of tools for integrated on-line analysis of extracted data
- A set of tools for on-line database analysis and exploration
- A facility for users to upload code to run their own algorithms on the datamining machines
- An exploration of techniques for open-ended resource discovery

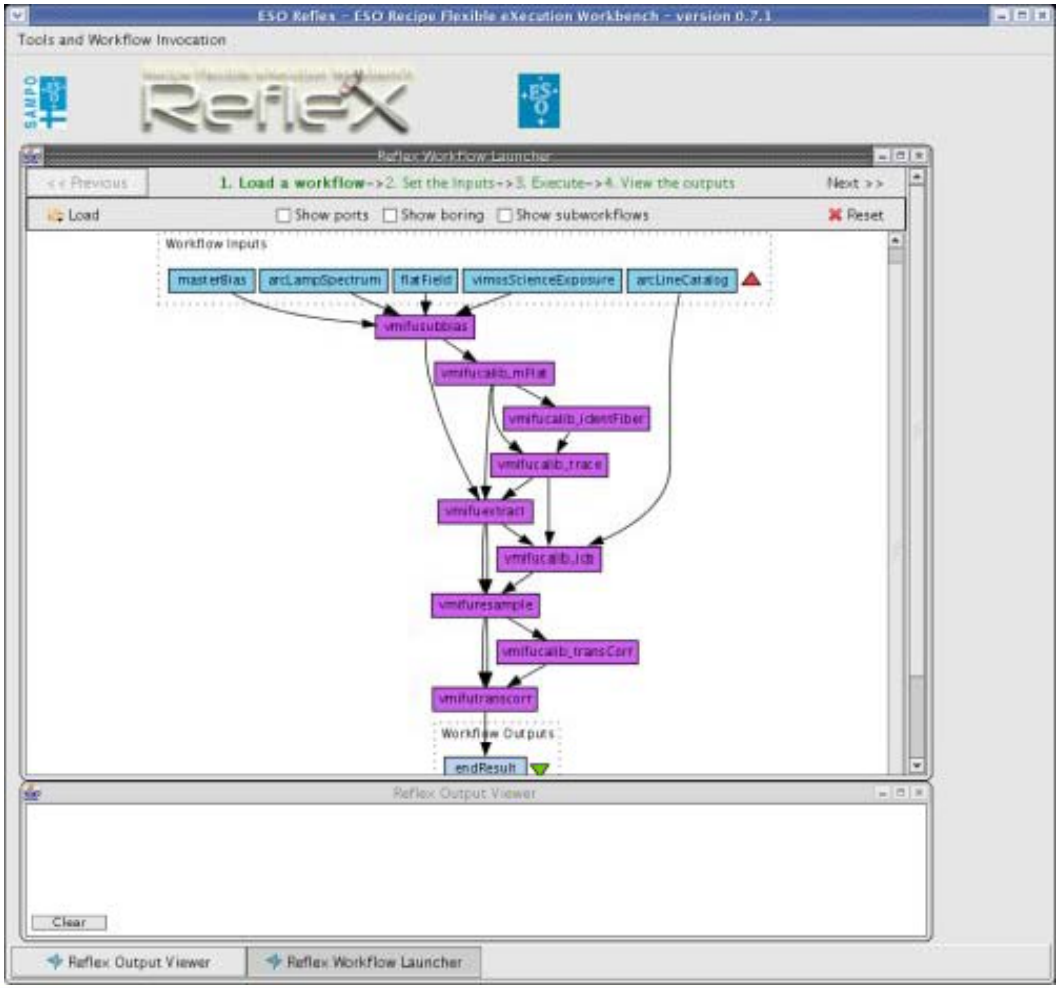


Sampo
European Southern Observatory project

Workflows for data reduction

Reasons for choosing Taverna

- Open source
- Free
- Allows customisation
- Easy to use and adapt
- Designed for science
- Most workflow engines are meant for business applications
- Very robust
- Actively developed
- Good support for web services



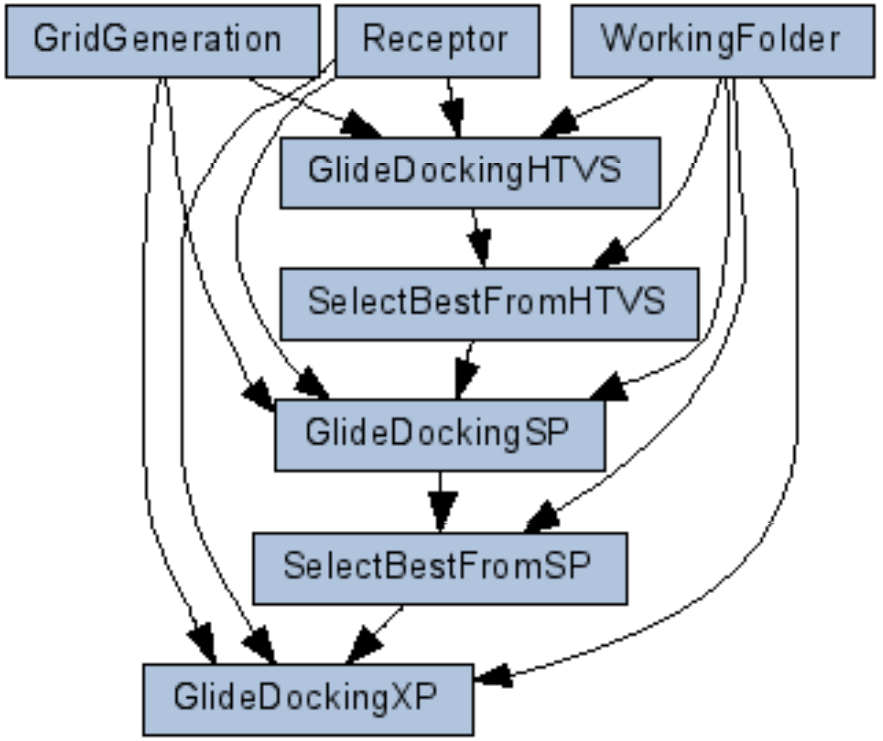
Chemoinformatcs

- CDK Taverna
  - Molecular modelling, spectroscopical support, statistics, clustering etc

- Chimatica Drug Discovery
  - Disease Target Validity
  - Drug Candidate Performance

Building pipelines using two Chemoninformatics toolkits:

  - Schrodinger
  - Accelrys



Summary

Taverna

- Allows access to distributed data and analysis tools
- Allows automation of data flow
- Extensible and flexible for science

Workflows are encapsulations of your experiments

myGrid acknowledgements

Carole Goble, Norman Paton, Robert Stevens, Anil Wipat, David De Roure, Steve Pettifer

- **OMII-UK** Tom Oinn, Katy Wolstencroft, Daniele Turi, June Finch, Stuart Owen, David Withers, Stian Soiland, Franck Tanoh, Matthew Gamble, Alan Williams, Ian Dunlop
- **Research** Martin Szomszor, Duncan Hull, Jun Zhao, Pinar Alper, Antoon Goderis, Alastair Hampshire, Qiuwei Yu, Wang Kaixuan.
- **Current contributors** Matthew Pocock, James Marsh, Khalid Belhajjame, PsyGrid project, Bergen people, EMBRACE people.
- **User Advocates and their bosses** Simon Pearce, Claire Jennings, Hannah Tipney, May Tassabehji, Andy Brass, Paul Fisher, Peter Li, Simon Hubbard, Tracy Craddock, Doug Kell, Marco Roos, Matthew Pocock, Mark Wilkinson
- **Past Contributors** Matthew Addis, Nedim Alpdemir, Tim Carver, Rich Cawley, Neil Davis, Alvaro Fernandes, Justin Ferris, Robert Gaizaukaus, Kevin Glover, Chris Greenhalgh, Mark Greenwood, Yikun Guo, Ananth Krishna, Phillip Lord, Darren Marvin, Simon Miles, Luc Moreau, Arijit Mukherjee, Juri Papay, Savas Parastatidis, Milena Radenkovic, Stefan Rennick-Egglestone, Peter Rice, Martin Senger, Nick Sharman, Victor Tan, Paul Watson, and Chris Wroe.
- **Industrial** Dennis Quan, Sean Martin, Michael Niemi (IBM), Chimatica.
- **Funding** EPSRC, Wellcome Trust.

http://www.mygrid.org.uk

http://www.myexperiment.org