^{my}Grid : Workflow based *in silico* experiments in biology

Katy Wolstencroft, Tom Oinn, Phillip Lord, Jun Zhao Carole Goble and the ^{my}Grid team & May Tassabehji / Hannah Tipney Medical Genetics, St Mary’s Hospital Manchester

What is $^{\text{my}}$Grid?

- e-Science pilot research project funded by EPSRC http://www.mygrid.org.uk
- Manchester, Newcastle, Sheffield, Southampton, Nottingham, EBI and RFCGR, also industrial partners.
- ‘targeted to develop open source software to support personalised *in silico* experiments in biology on a grid.’
- Now - platform grant for $^{\text{my}}$Grid 2 and $^{\text{my}}$Grid is an OMII-UK (Open Middleware Infrastructure Institute) node

What is ^{my}Grid?

“A comprehensive loosely-coupled suite of middleware components specifically to support data intensive *in silico* experiments in biology”

- Distributed computing
- Workflow Design and Enactment
- Provenance and Data management
- Semantic Discovery

Collaborators

LION
AstraZeneca
The Wellcome Trust
SOURCEFORGE.net
EMBL-EBI
European Bioinformatics Institute
MERCK
gsk GlaxoSmithKline
MANCHESTER 1824
The University of Manchester
University of Southampton
iT innovation
Sun microsystems
IBM
network inference
geneticXchange
EPISTEMICS LTD
knowledge is our business
THE UNIVERSITY OF NOTTINGHAM
UNIVERSITY OF NEWCASTLE UPON TYNE

Thanks to the other members of the Taverna project, http://taverna.sf.net

Motivation

Bioinformatics is an open Community

- Open access to data
- Open access to resources
- Open access to tools
- Open access to applications

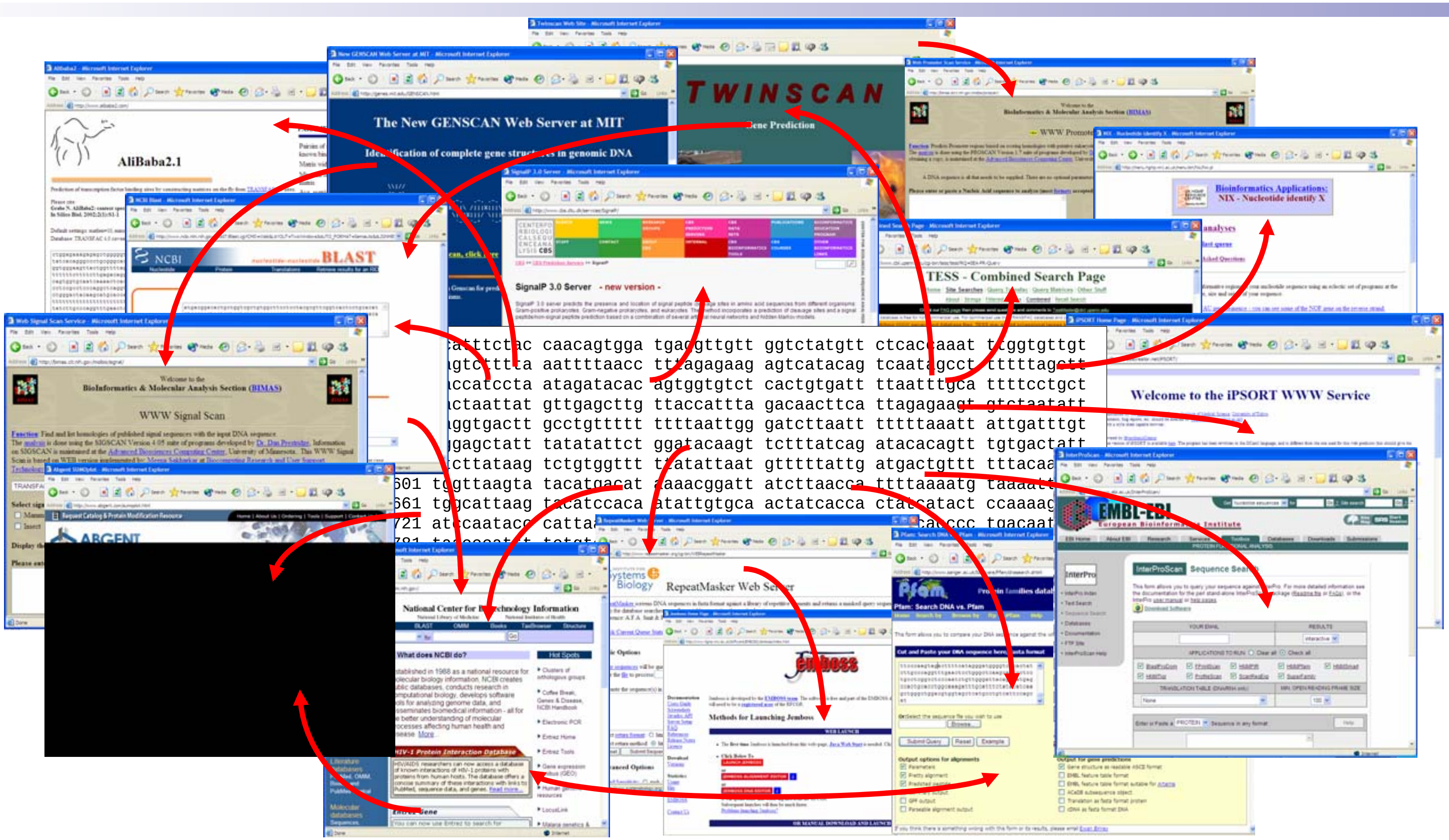
Global Bioinformatics

Problems

- Heterogeneous data
- Distributed resources
- Potentially requires supercomputing power
- Very few standards – I/O formats, data representation, annotation

Integration and interoperability between resources is difficult

Traditional Approach


AliBaba2.1
The New GENSCAN Web Server at MIT
TWINSCAN
Gene Prediction
SignalP 3.0 Server - new version -
NCBI BLAST
TESS - Combined Search Page
Bioinformatics Applications: NIX - Nucleotide identify X
WWW Signal Scan
BioInformatics & Molecular Analysis Section (BIMAS)
ARGENT
Welcome to the iPSORT WWW Service
EMBL-EBI European Bioinformatics Institute
InterProScan Sequence Search
National Center for Biotechnology Information
RepeatMasker Web Server
EMBOSS
Pfam
Plam: Search DNA vs. Pfam

```
atttctac caacagtgga tgaggttgtt ggtctatgtt ctcaccaaat ttggtgttgt
tgtctttta aattttaacc tttagagaag agtcatacag tcaatagcct tttttagctt
ccatccta atagatacac agtggtgtct cactgtgatt ttaatttgca ttttcctgct
ctaattat gttgagcttg ttaccattta gacaacttca ttagagaagt gtctaatatt
ggtgactt gcctgttttt ttttaattgg gatcttaatt tttttaaatt attgatttgt
gagctatt tatatattct ggatacaagt tctttatcag atacacagtt tgtgactatt
cttataag tctgtggttt ttatattaat gtttttattg atgactgttt tttacaa
601 tggttaagta tacatgacat aaaacggatt atcttaacca ttttaaaatg taaaatt
661 tggcattaag tatatcccaca atattgtgca actatcacca ctatcatact ccaaaag
721 atccaatacc cattat
```

[redacted]

Cutting and Pasting

- Advantages:
  - Low technology on both server and client side
  - Very robust: Hard to break
  - Data integration happens along the way
- Disadvantages:
  - Time consuming (and painful!)
    - Can be repeated rarely
    - Limited to small data sets
  - Error prone:
    - Poor repeatability

Pipeline Programming

- Advantages
  - Repeatable
  - Allows automation
  - Quick, reliable, efficient
- Disadvantages
  - Requires programming skills
  - Difficult to modify
  - Requires local tool and database installation
  - Requires tool and database maintenance!!!

^{my}Grid Requirements

- Automation
- Reliability
- Repeatability
- Distributed resources
- Few programming skill required

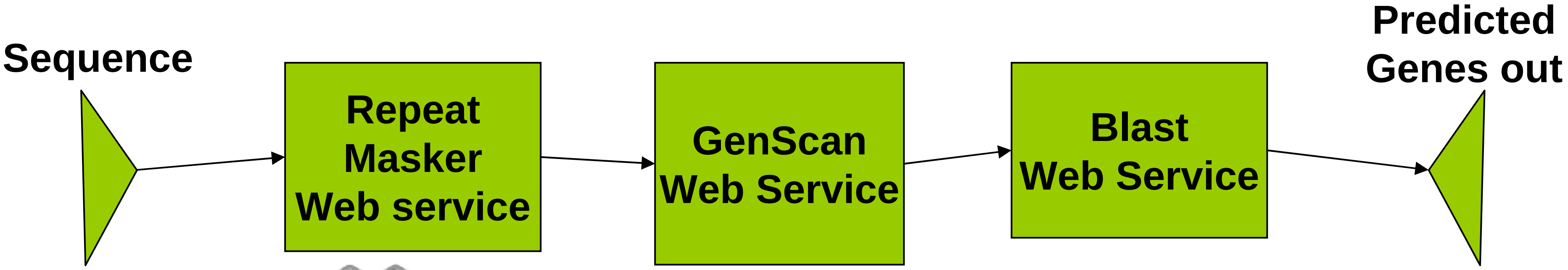
myGrid Approach - Workflows

General technique for describing and enacting a process

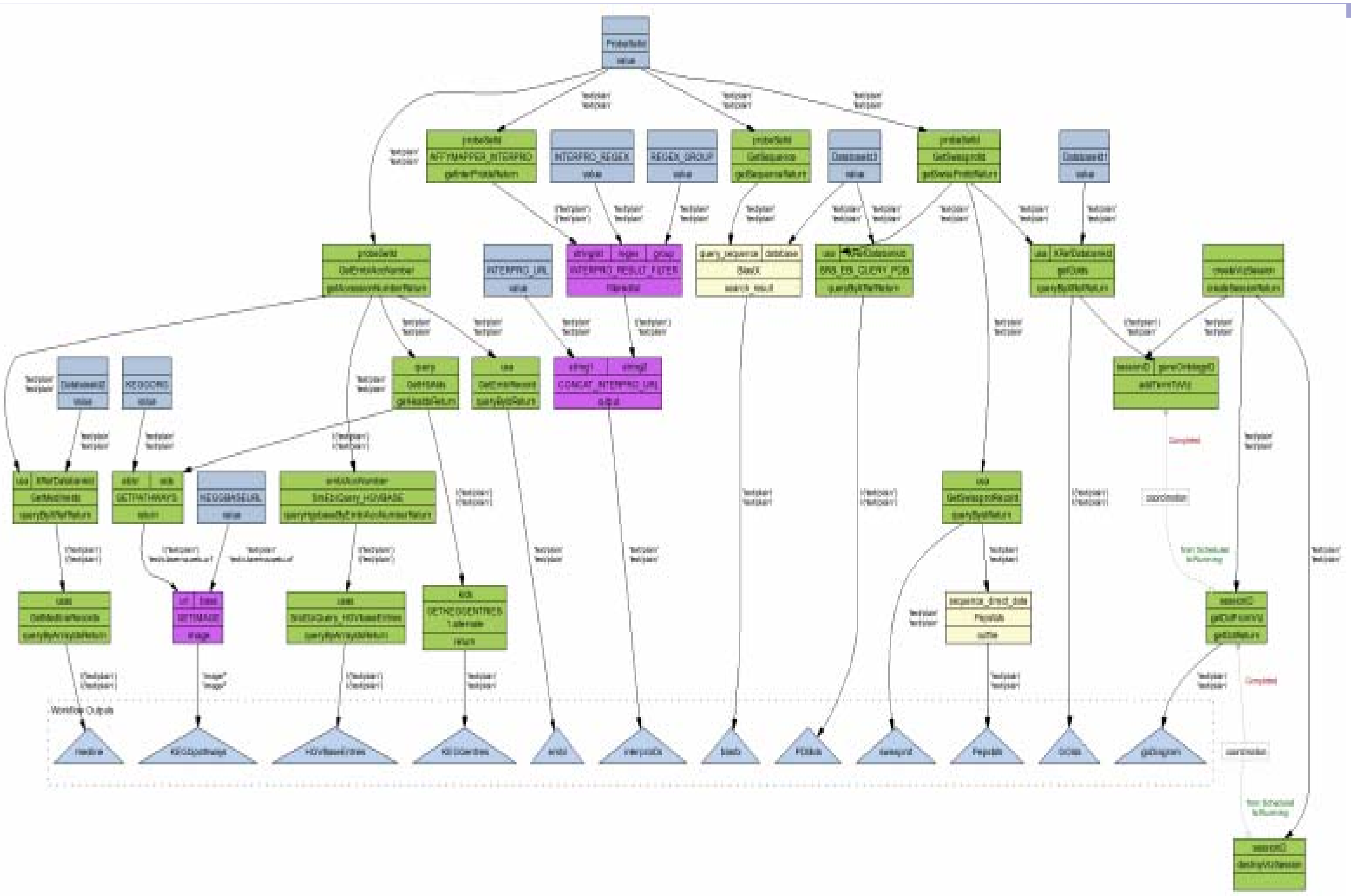
describes *what* you want to do, not *how* you want to do it

Simple language specifies how bioinformatics processes fit together – processes are web services

- High level workflow diagram separated from any lower level coding – therefore, you don't have to be a coder to build workflows



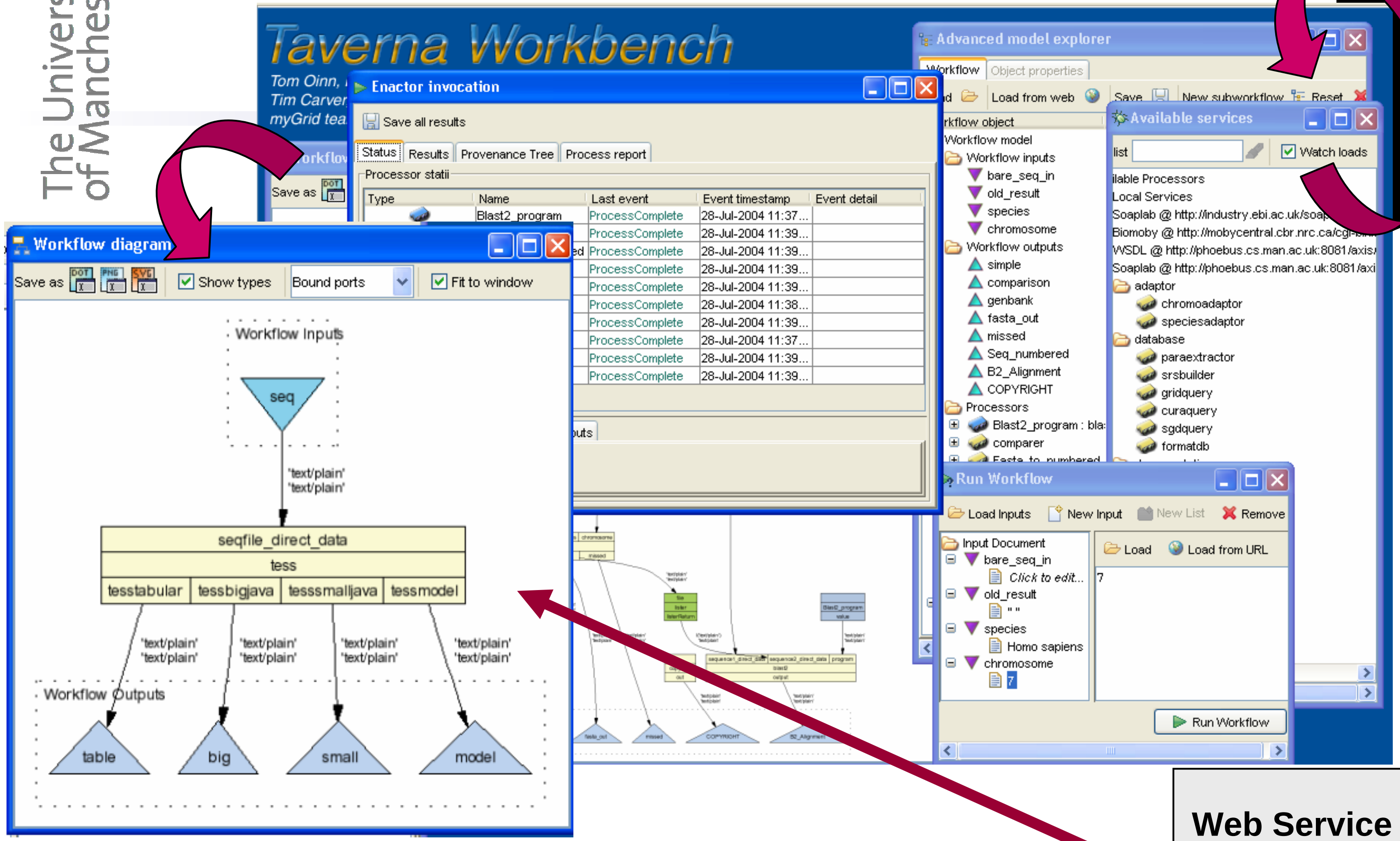
Workflow Diagram



Workflow Advantages

- Automation
  - Capturing processes in an explicit manner
  - Tedium! Computers don’t get bored/distracted/hungry/impatient!
  - Saves repeated time and effort
- Modification, maintenance, substitution and personalisation
- Easy to share, explain, relocate, reuse and build
- Releases scientists/bioinformaticians to do other work
- Record
  - Provenance: what the data is like, where it came from, its quality
  - Management of data (LSID - Life Science Identifiers)

Workflow Components



Enactor invocation

Save all results

Status | Results | Provenance Tree | Process report

Processor statii

Type	Name	Last event	Event timestamp	Event detail
	Blast2_program	ProcessComplete	28-Jul-2004 11:37...	
		ProcessComplete	28-Jul-2004 11:39...	
	ed	ProcessComplete	28-Jul-2004 11:39...	
		ProcessComplete	28-Jul-2004 11:39...	
		ProcessComplete	28-Jul-2004 11:39...	
		ProcessComplete	28-Jul-2004 11:38...	
		ProcessComplete	28-Jul-2004 11:39...	
		ProcessComplete	28-Jul-2004 11:37...	
		ProcessComplete	28-Jul-2004 11:39...	
		ProcessComplete	28-Jul-2004 11:39...	

Run Workflow

Input Document
- bare_seq_in: Click to edit...
- old_result: " "
- species: Homo sapiens
- chromosome: 7

Freefluo

Freefluo
Workflow
engine to run
workflows

Web Service | **e.g. DDBJ BLAST**

SOAPLAB Web Service | **Any Application**

Scufl Simple Conceptual Unified Flow Language
Taverna Writing, running workflows & examining results
SOAPLAB Makes applications available

Workflow Services – Web Services

- automated programmatic internet access to applications
- Many bioinformatics resources provide web service versions of popular tools - e.g. NCBI BLAST

myGrid > 1000 web services for bioinformatics

SeqHound – Database of biological sequences and tools

BioMart – Federated query system

EMBOSS – Sequence analysis tools

BioMoby – Collection of web services

EBI SOAPLAB – Collection of supported services

Web Service Creation

Many service providers provide web service implementations of applications

SoapLab

- For wrapping ‘legacy’ command-line applications for use in Taverna

GowLab

- For wrapping html resources

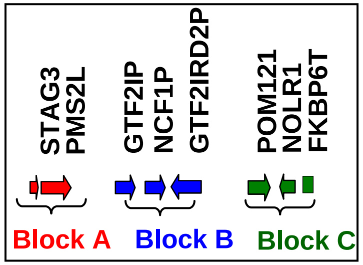
potentially ANY bioinformatics tool / resource can be wrapped and used in ^{my}Grid workflows

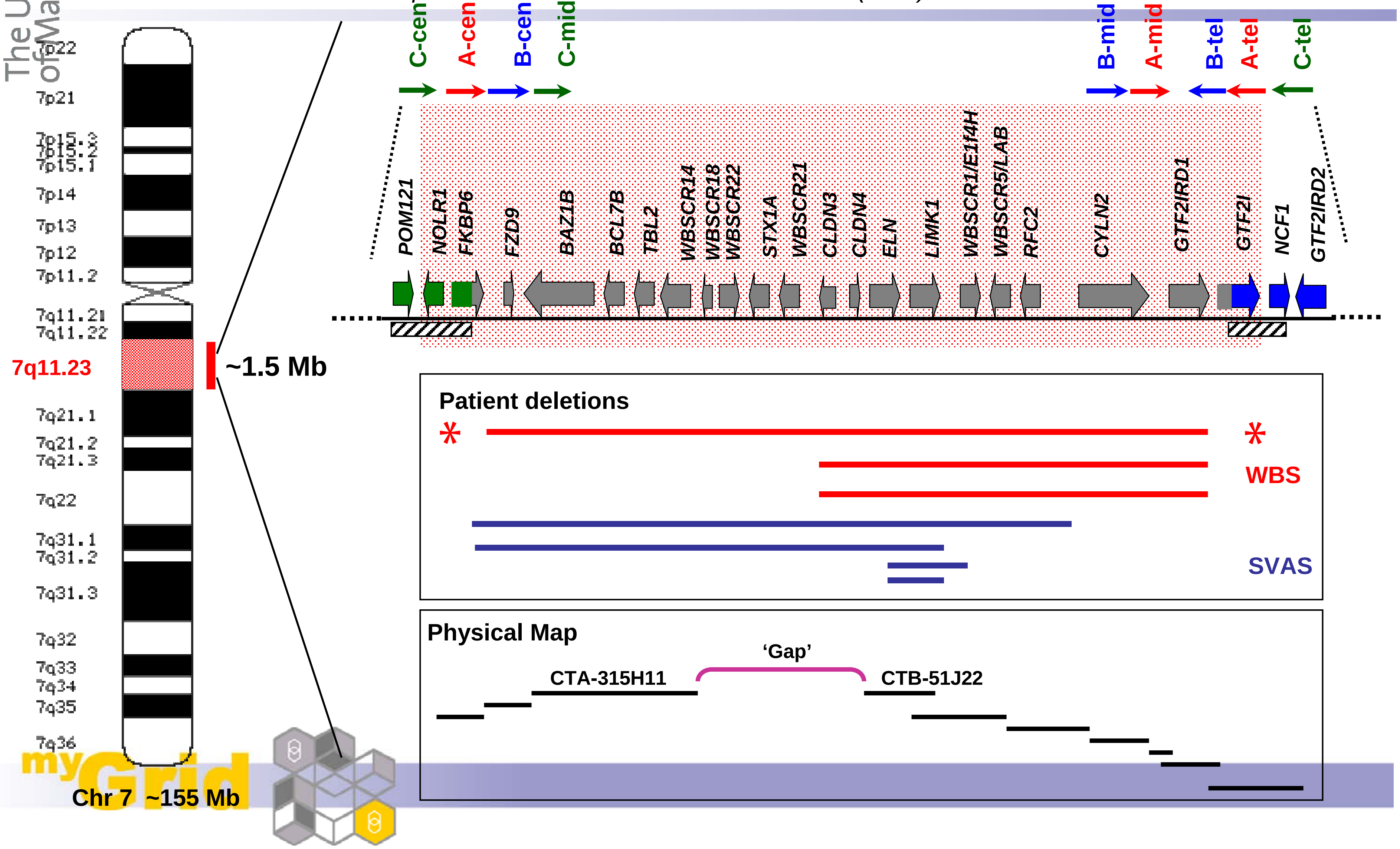
Use Cases:
Williams-Beuren Syndrome (WBS)

- Contiguous sporadic gene deletion disorder
- 1/20,000 live births, caused by unequal crossover (homologous recombination) during meiosis
- Haploinsufficiency of the region results in the phenotype
- Multisystem phenotype – muscular, nervous, circulatory systems
- Characteristic facial features
- Unique cognitive profile
- Mental retardation (IQ 40-100, mean~60, ‘normal’ mean ~ 100)
- Outgoing personality, friendly nature, ‘charming’

Williams-Beuren Syndrome Microdeletion

Eicher E, Clark R & She, X An Assessment of the Sequence Gaps: Unfinished Business in a Finished Human Genome. Nature Genetics Reviews (2004) 5:345-354
Hillier L et al. The DNA Sequence of Human Chromosome 7. Nature (2003) 424:157-164



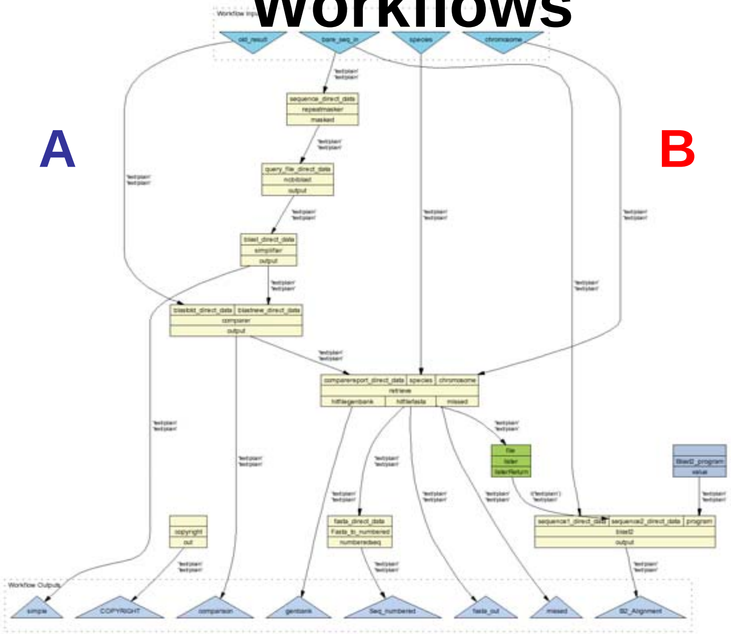


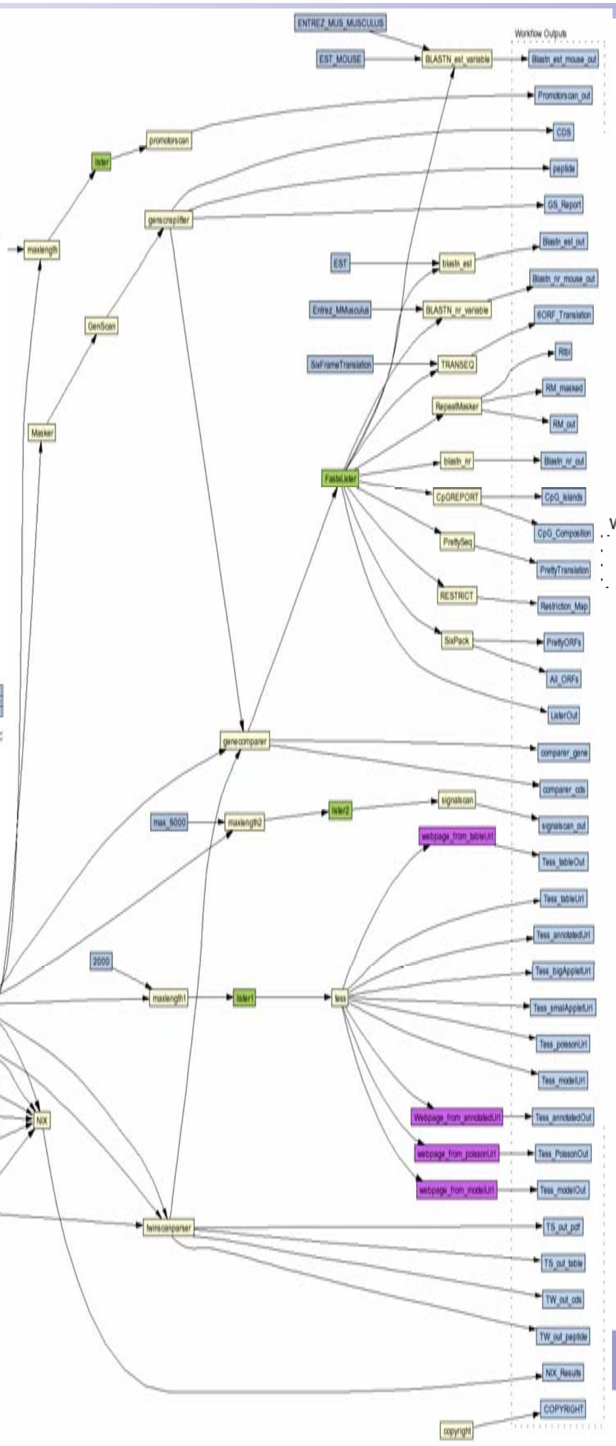
Filling a genomic gap *in silico*

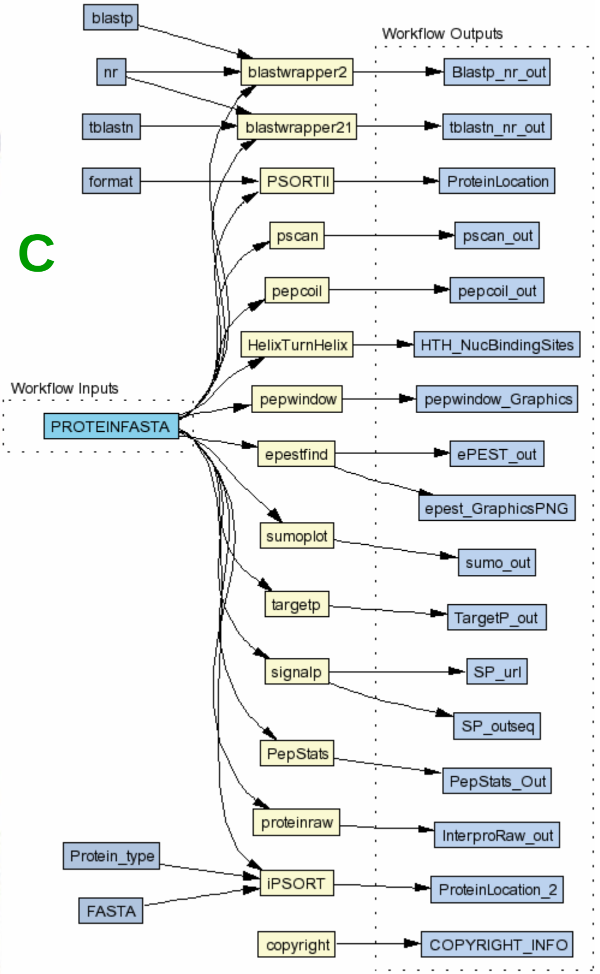
- Identify new, overlapping sequence of interest
- Characterise the new sequence at nucleotide and amino acid level

- Frequently repeated – info rapidly added to public databases
- Time consuming and mundane
- Don’t always get results
- Huge amount of interrelated data is produced

The Williams Workflows







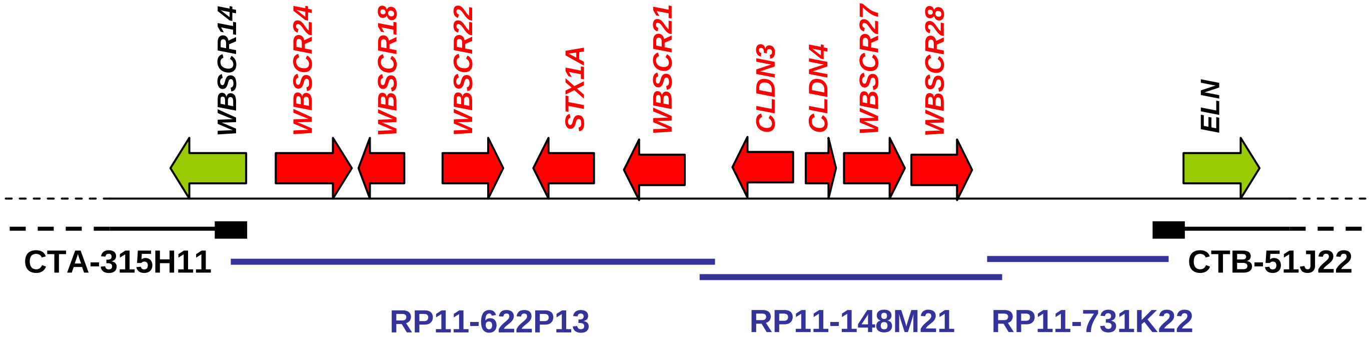
A: Identification of overlapping sequence

B: Characterisation of nucleotide sequence

C: Characterisation of protein sequence

The Biological Results

Four workflow cycles totalling ~ 10 hours
The gap was correctly closed and all known features identified



314,004bp extension

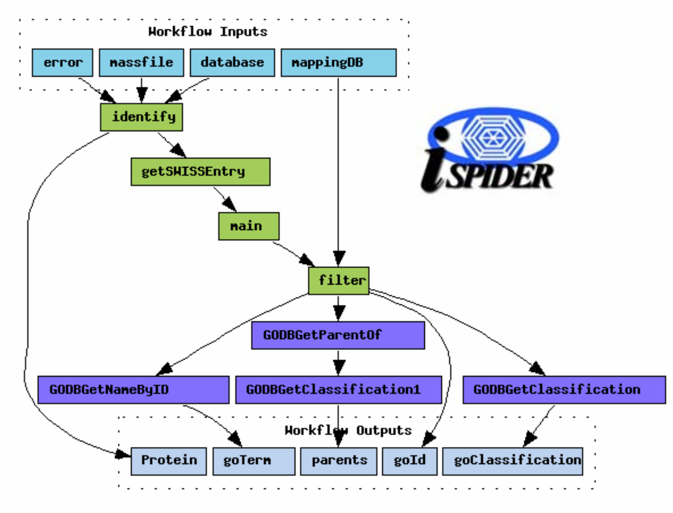
All nine known genes identified
(40/45 exons identified)

The Workflow Experience

Have workflows delivered on their promise? **YES!**

- Correct and biologically meaningful results
- Automation
  - Saved time, increased productivity
  - Process split into three, you still require humans!
- Sharing
  - Other people have used and want to develop the workflows
- Change of work practises
  - *Post hoc* analysis. Don’t analyse data piece by piece receive all data all at once
  - Data stored and collected in a more standardised manner
  - Results amplification

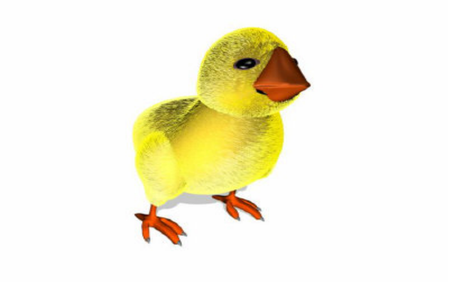
Workflow Reuse



Mouse genome

Trypanosomiasis in cattle

Chicken genome



Chris Wroe, Carole Goble, Antoon Goderis, Phillip Lord, Simon Miles, Juri Papay, Pinar Alper, Luc Moreau *Recycling workflows and services through discovery and reuse* Concurrency and Computation: Practice and Engineering *accepted for publication*

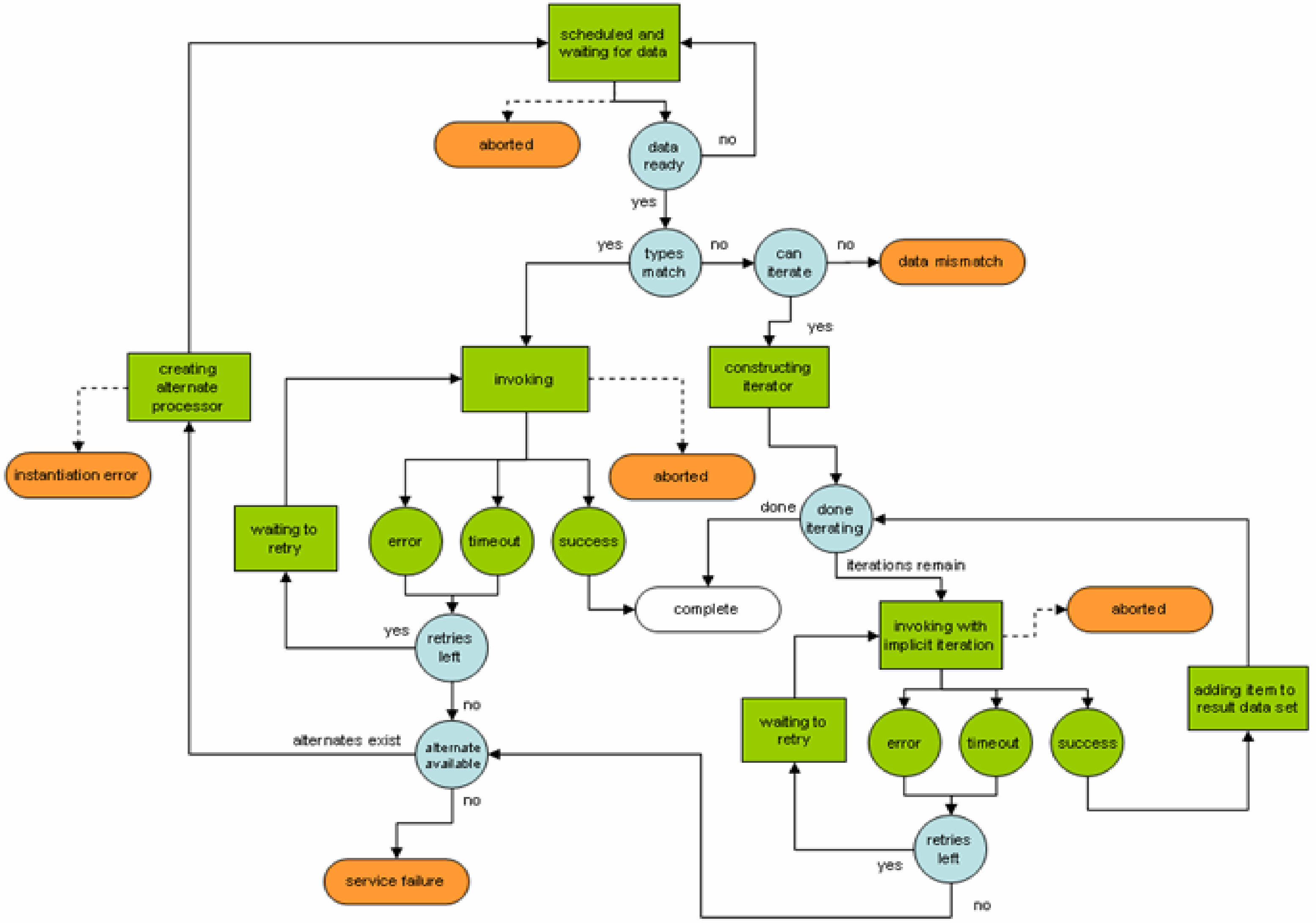
Web Service Issues

- Most owned by other people
- No control over service failure
- Some are research level

However - Taverna can

- Notify users of service failures
- Instigate retries
- Substitute services from other sites

Fault Tolerance

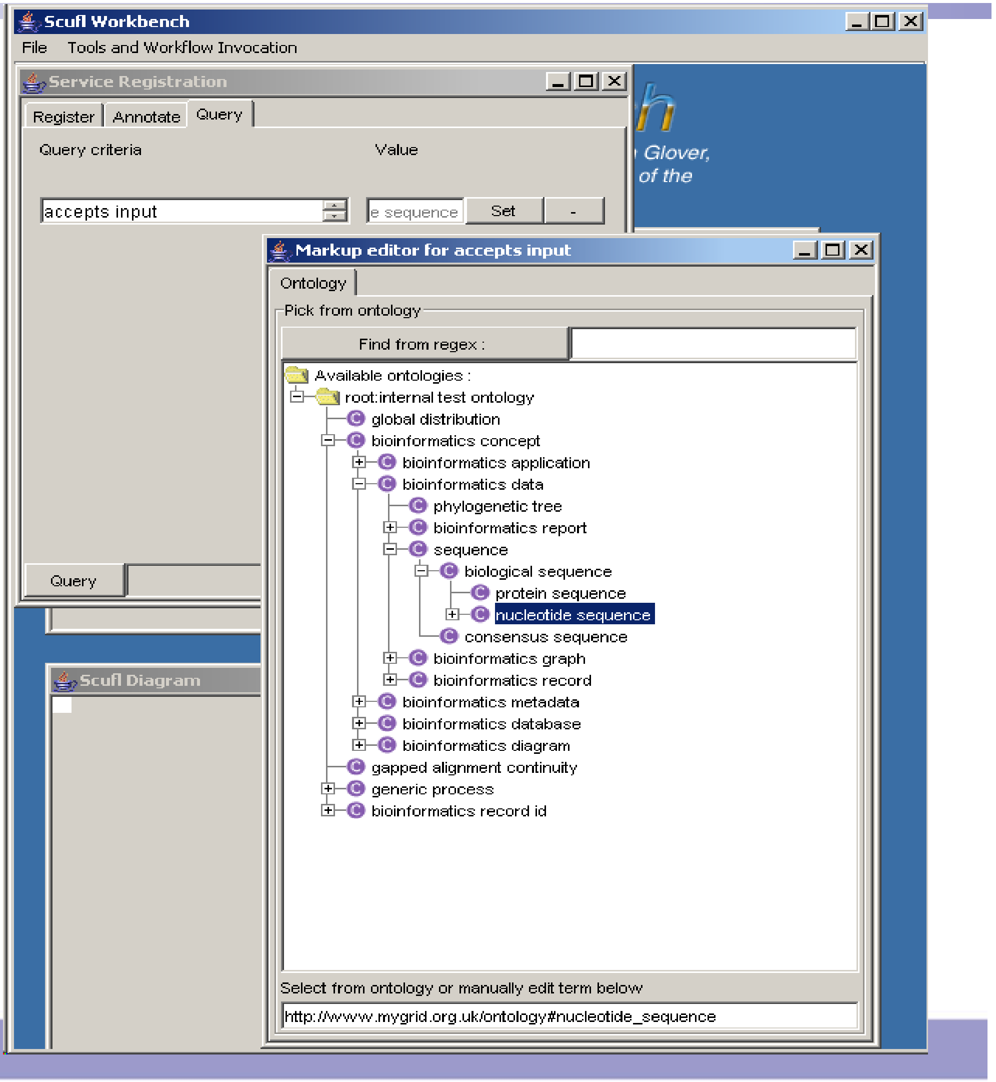


Service Discovery – Feta

Services – only useful if the user can find them!

Feta semantic discovery – can find services by:

- **Bioinformatics task**
- **Resource used**
- **Bioinformatics method**
- **Input format**
- **Output format**



Query criteria: accepts input

Value: e sequence

Pick from ontology

```
Available ontologies :
  root:internal test ontology
    global distribution
    bioinformatics concept
      bioinformatics application
      bioinformatics data
        phylogenetic tree
        bioinformatics report
        sequence
          biological sequence
            protein sequence
            nucleotide sequence
          consensus sequence
        bioinformatics graph
        bioinformatics record
      bioinformatics metadata
      bioinformatics database
      bioinformatics diagram
    gapped alignment continuity
    generic process
    bioinformatics record id
```

Select from ontology or manually edit term below

http://www.mygrid.org.uk/ontology#nucleotide_sequence

Data Management

- Workflows can generate vast amount of data - How can we manage and track it?
- Data
- AND metadata
- AND experiment provenance

  - LSIDs - to identify objects

  - Semantic Web technologies (RDF, Ontologies)

    - To store knowledge provenance

  - Taverna workflow workbench & plugins

    - Ensure automated recording

Provenance

- Stored in RDF–Resource Description Framework
- < Subject, Predicate, Object>

e.g. <*urn:actualInputParameter0, isAbout: my:provenance*>

Identified by a URI and typed by an ontology

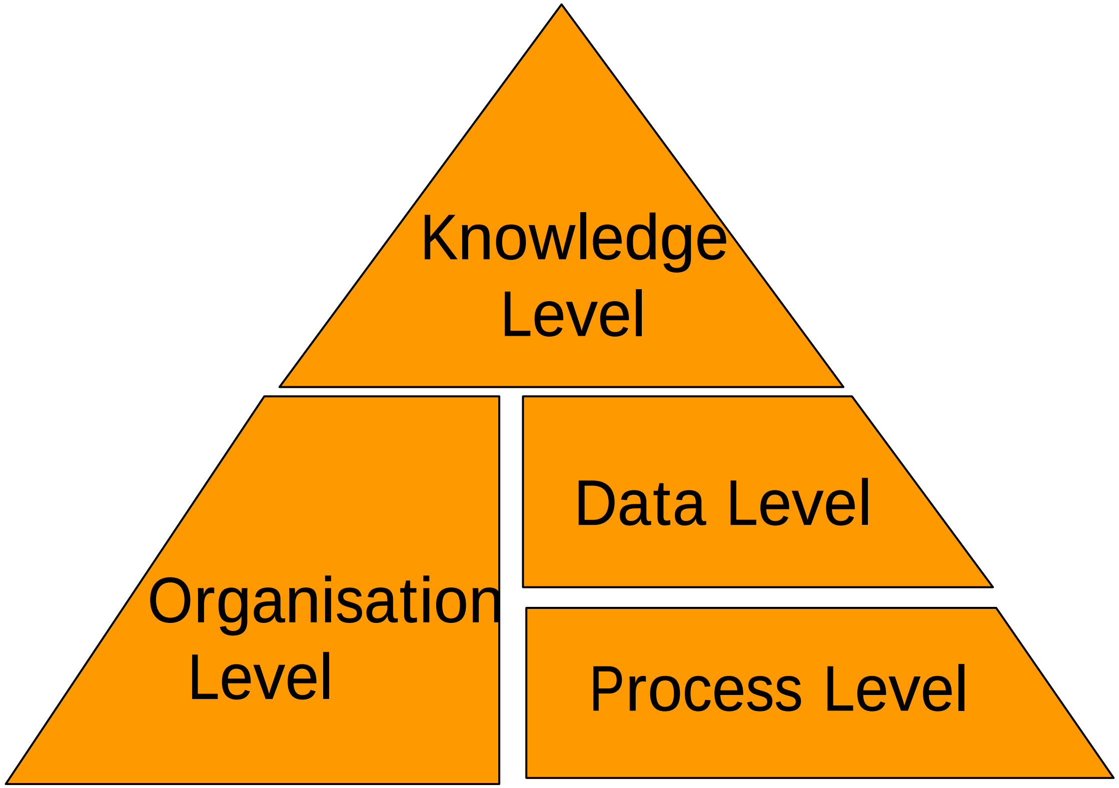
- Inputs and outputs and ^{my}Grid services can be defined using ^{my}Grid Service Ontology – describes bioinformatics processes

Life Science Identifiers

- Life Science Identifiers (LSIDs) are the standard adopted by the Object Management Group (OMG) for the identification of life science data objects

  urn:lsid:ncbi.nlm.nlh.gov.lsid.biopathways.org:genbank_gi:7717376
- LSIDs used throughout ^{my}Grid to ID data objects from external sources as well as internally created data.
- Using a standard mechanism for identification allows for more efficient and cohesive exchanges between ^{my}Grid components

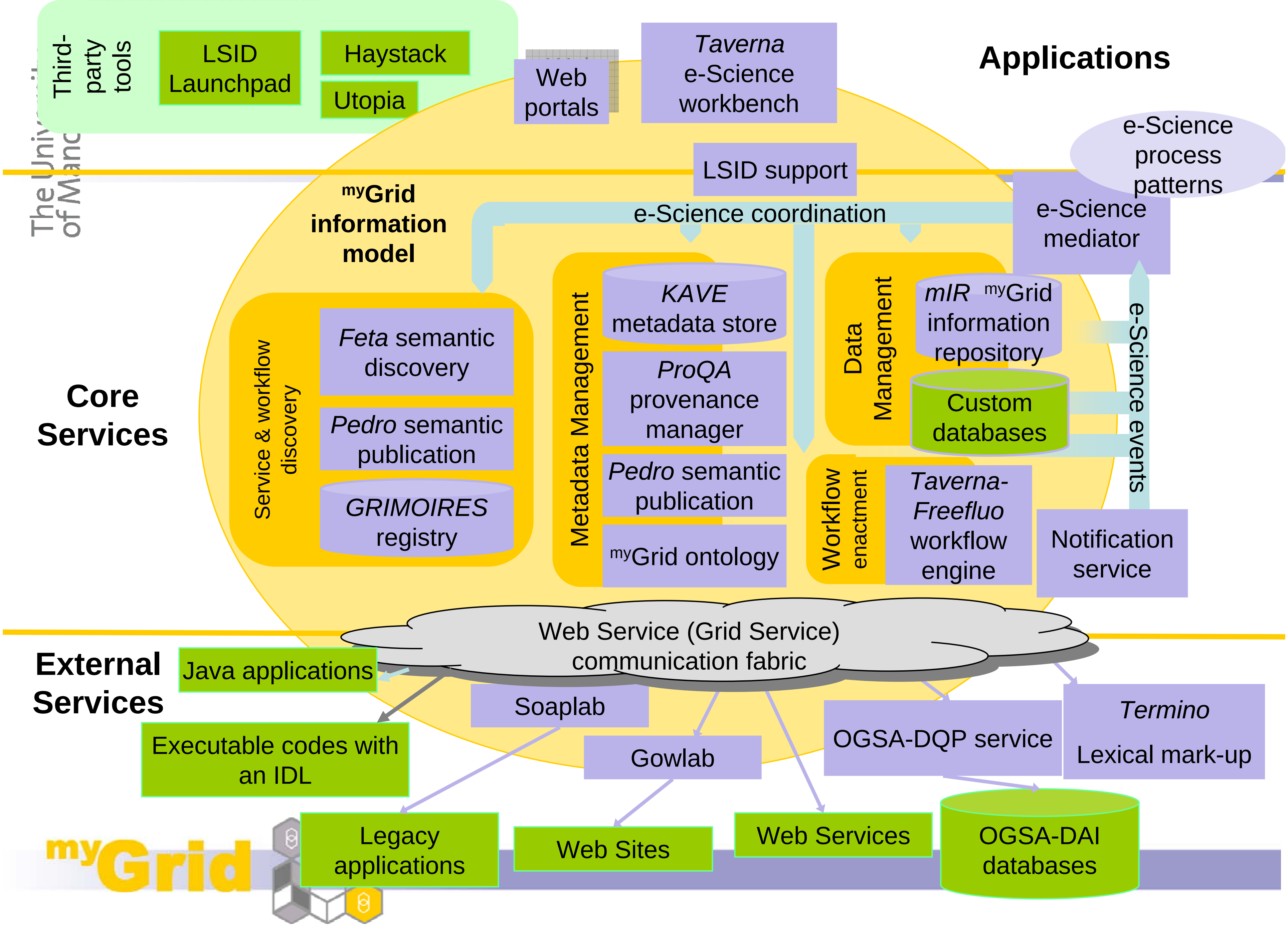
Provenance Pyramid



^{my}Grid Architecture

In keeping with the bioinformatics community

- Open architecture
  - Service Oriented Architecture
  - Loosely coupled
  - Web services based
  - Assemble your own components
  - Designed to work together



Third-party tools

LSID Launchpad

Haystack

Utopia

Web portals

Taverna e-Science workbench

Applications

e-Science process patterns

LSID support

myGrid information model

e-Science coordination

e-Science mediator

Service & workflow discovery

Feta semantic discovery

Pedro semantic publication

GRIMOIRES registry

Metadata Management

KAVE metadata store

ProQA provenance manager

Pedro semantic publication

myGrid ontology

Data Management

mIR myGrid information repository

Custom databases

e-Science events

Core Services

Workflow enactment

Taverna-Freefluo workflow engine

Notification service

Web Service (Grid Service) communication fabric

External Services

Java applications

Executable codes with an IDL

Soaplab

Gowlab

OGSA-DQP service

Termino Lexical mark-up

Legacy applications

Web Sites

Web Services

OGSA-DAI databases

myGrid Users

Widespread uptake

- Bioinformatics
- Systems Biology
- Chemistry
- Medical Physics

Many new e-Science projects using myGrid platform

myGrid Alliance: Application

CLEF

iSPIDER

eFamily

mias GRID

ComparaGRID

e-Fungi

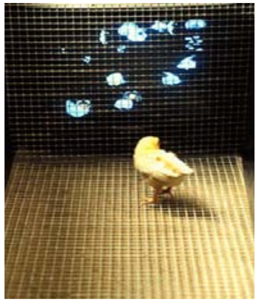
PsyGrid

Integrative Biology
UK - NZ Partnership

Unilever Cambridge
Centre for Molecular Informatics

Small molecules,
Murray-Rust, Cambridge

Chicken genome
Roslin Institute



New projects - New directions for ^{my}Grid

Bioinformatics data – mostly sequence based, workflow processes take several hours

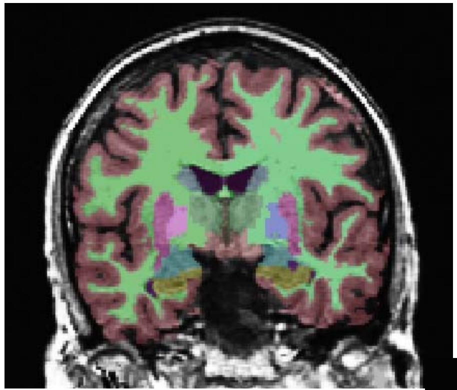
Other Fields

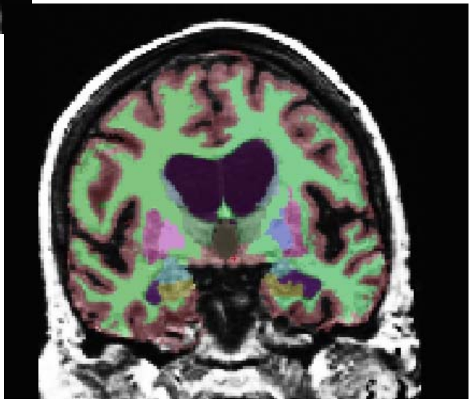
- More data / larger data
- Long running workflows – weeks/months

MANCHESTER
1824

The University
of Manchester

New Directions



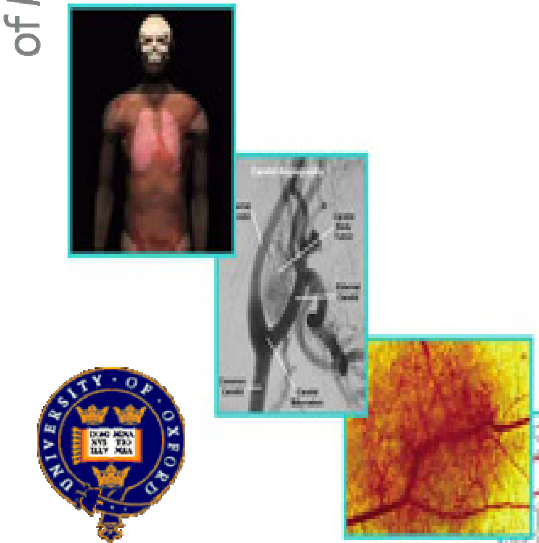


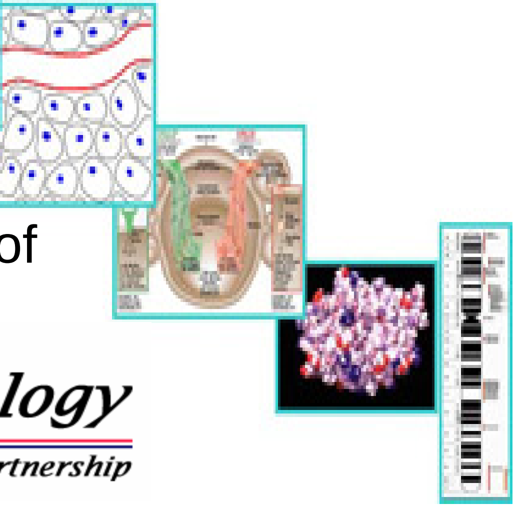
Long Running Workflows

Large Data Sets

MIAS-Grid

PsyGrid





Computational steerage of heart simulation codes

Integrative Biology
UK - NZ Partnership

myGrid

Conclusions

- myGrid enables interoperability and integration of bioinformatics resources
- Legacy applications and new tools can be made available in the workbench
- User driven approach to development
- myGrid offers a whole suite of components to design and enact workflows and trace experiments

^{my}Grid and WBS People

Core

Matthew Addis, Nedim Alpdemir, Pinar Alper, Tim Carver, Rich Cawley, Neil Davis, Alvaro Fernandes, Justin Ferris, Robert Gaizaukaus, Kevin Glover, Carole Goble, Chris Greenhalgh, Mark Greenwood, Yikun Guo, Ananth Krishna, Peter Li, Phillip Lord, Darren Marvin, Simon Miles, Luc Moreau, Arijit Mukherjee, Tom Oinn, Juri Papay, Savas Parastatidis, Norman Paton, Terry Payne, Matthew Pockock Milena Radenkovic, Stefan Rennick-Egglestone, Peter Rice, Martin Senger, Nick Sharman, Robert Stevens, Victor Tan, Daniele Turi, Anil Wipat, Paul Watson, Katy Wolstencroft and Chris Wroe.

Users

Simon Pearce and Claire Jennings, Institute of Human Genetics School of Clinical Medical Sciences, University of Newcastle, UK

Hannah Tipney, May Tassabehji, Andy Brass, St Mary’s Hospital, Manchester, UK

Postgraduates

Martin Szomszor, Duncan Hull, Jun Zhao, John Dickman, Keith Flanagan, Antoon Goderis, Tracy Craddock, Alastair Hampshire

Industrial

Dennis Quan, Sean Martin, Michael Niemi, Syd Chapman (IBM)

Robin McEntire (GSK)