

myGrid: An Open Platform for Data-intensive Post-Genomic Functional Analysis

Matthew Addis

IT Innovation Centre

mja@it-innovation.soton.ac.uk

myGrid in a Nutshell

- “The key to bioinformatics is integration, integration, integration,”

bioinformatics expert Jim Golden, Curagen spin-off 454 Corporation

Bioinformatics: Bringing it all together, Nature 17 October 2002

- myGrid is all about integrating bioinformatics tools and data to support the scientific processes of *in-silico* experimentation

Contents

- Project timescales, funding and partners
- What problems is myGrid trying to solve?
- The myGrid approach
- Canned demo
- Web Services, Grid and the Semantic Grid
- Next Steps
- Further information

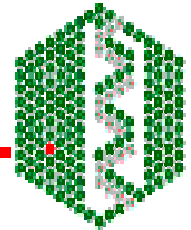
myGrid timescales and funding

- EPSRC eScience pilot project
 - Official start October 2001
 - Actual start January 2002
 - Planned end March 2005
- £3M EPSRC funding + contribution from industrial partners
- 16 RAs, 9 studentships
- Early days – still two thirds of the project to go...

myGrid partners



EMBL
European Bioinformatics Institute



Problems: heterogeneity everywhere

- Seamless access to bioinformatics data sources and tools is not easy
 - Data formats
 - Data access mechanisms
 - Data annotations and interpretations
 - Analysis techniques and implementations
 - Service Providers
- Relatively few standards
 - GO
 - DAS
 - BioMOBY, I3C
- EBI hosted tools 50+
 - Homology & Similarity
 - Prot. Function. Analysis
 - Structural Analysis
 - Sequence Analysis
 - Miscellaneous Tools
- EBI hosted databases 30+
 - Nucleotide Databases
 - Protein Databases
 - Proteome Analysis
 - Structure Databases
 - Microarray Database
 - Literature Databases

Problems: access mechanisms

The collage illustrates various biological databases and their access mechanisms:

- NCBI Blast - Netscape**: A web browser window showing the NCBI Blast search interface.
- NCBI**: The National Center for Biotechnology Information homepage, featuring search options for Nucleotide, Protein, Translations, and more.
- PRINTS**: Protein Fingerprint Database, described as a compendium of protein fingerprints.
- Stanford Microarray Database**: A database for microarray data, with sections for Data Selection for Analysis and Gene Selection & Annotation.
- The Arabidopsis Information Resource (TAIR)**: A comprehensive resource for Arabidopsis thaliana, including search options for Allele, Probe, Clone, Author, Colleague, and Other_Germplasm.
- Antirrhinum majus Genome Database**: A database for the Antirrhinum majus genome, with search options for Simple Search, Text Search, Class Browser, Expression Search, Ace Query, and BLAST Search.
- ExPASy**: The European Bioinformatics Institute's Protein Analysis Pipeline, with search options for Site Map, Search ExPASy, and Contact us.
- PROSITE**: Database of protein families and domains, providing access to protein families and domains through various search methods.

PROSITE Database of protein families and domains

PROSITE is a database of protein families and domains. It consists of biologically significant sites, patterns and profiles that help to reliably identify to which known protein family (if any) a new sequence belongs. [More details / References / Disclaimer](#).

Release 16.46, of 27-Sep-2001 (contains 1098 documentation entries that describe 1486 different patterns, rules and profiles/matrices).

Access to PROSITE

- by description
- by entry name or accession number (PSxxxxx or PDOCxxxxx number)
- by author
- by citation
- by full text search
- SRS - Sequence Retrieval System

Documents

- PROSITE user manual
- List of PROSITE documentation entries
- How to obtain PROSITE
- Document describing the syntax of profiles in PROSITE
- List of programs that make use of PROSITE
- List of abbreviations for journals cited
- List of on-line experts
- The optimal way to develop patterns

Tools for PROSITE

- ScanProsite - Scan a sequence against PROSITE or a pattern against SWISS-PROT
- ProfileScan - Scan a sequence against the profile entries in PROSITE
- [search tools](#)

Services

- by FTP

Courtesy of Mark Wilkinson (BioMOBY)

Problems: data isn't just numbers

InterPro Entry IPR000025 - Netscape

File Edit View Go Communicator Help

Database	InterPro
Accession	IPR000025; Melatonin_receptor (matches 22 proteins)
Name	Melatonin receptor
Type	Family
Dates	08-OCT-1999 (created) 27-MAR-2000 (last modified)
Signatures	PRO0857; MELATONINR (22 proteins)
Parent	PRO00275; Rhodopsin-like GPCR superfamily (3990 proteins)
Children	PRO02278; Melatonin 1A receptor (12 proteins) PRO02279; Melatonin 1C receptor (5 proteins) PRO02280; Melatonin-related 1X receptor (3 proteins)
Function	melatonin receptor (GO:0008502)
Component	membrane (GO:0016020)

Abstract

G-protein-coupled receptors (GPCRs) constitute a vast protein family that encompasses a wide range of functions (including various autocrine, paracrine and endocrine processes). They show considerable diversity at the sequence level, on the basis of which they can be separated into distinct groups. We use the term clan to describe the GPCRs, as they embrace a group of families for which there are indications of evolutionary relationship, but between which there is no statistically significant similarity in sequence [1]. The currently known clan members include the rhodopsin-like GPCRs, the secretin-like GPCRs, the cAMP receptors, the fungal mating pheromone receptors, and the metabotropic glutamate receptor family.

The rhodopsin-like GPCRs themselves represent a widespread protein family that includes hormone, neurotransmitter and light receptors, all of which transduce extracellular signals through interaction with guanine nucleotide-binding (G) proteins. Although their activating ligands vary widely in structure and character, the amino acid sequences of the receptors are very similar and are believed to adopt a common structural framework comprising 7 transmembrane (TM) helices [2, 3, 4].

Melatonin is secreted by the pineal gland during darkness [5]. It regulates a variety of neuroendocrine functions and is thought to play an essential role in circadian rhythms. Drugs that modify the action of melatonin, and hence influence circadian cycles, are of clinical interest (for example, in the treatment of jet-lag). Melatonin receptors are found in the retina, in the pars tuberalis of the pituitary, and in discrete areas of the brain. The receptor inhibits adenylyl cyclase via a pertussis toxin-sensitive G-protein, probably of the Gi/Gi_o class [6].

Examples

- P49288 ML1C_CHICK
- P49285 ML1A_CHICK
- P49219 ML1C_XENLA
- P49217 ML1A_PHOSU

[View examples](#)

References

- Attwood T.K., Findlay J.B.C. *Fingerprinting G-protein-coupled receptors*. Protein Eng. 7: 195-203(1994). [MEDLINE:94224751] [PUB00004961]

Document Done

review

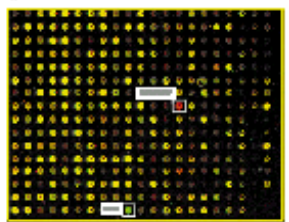
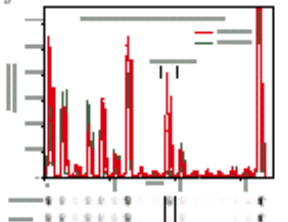




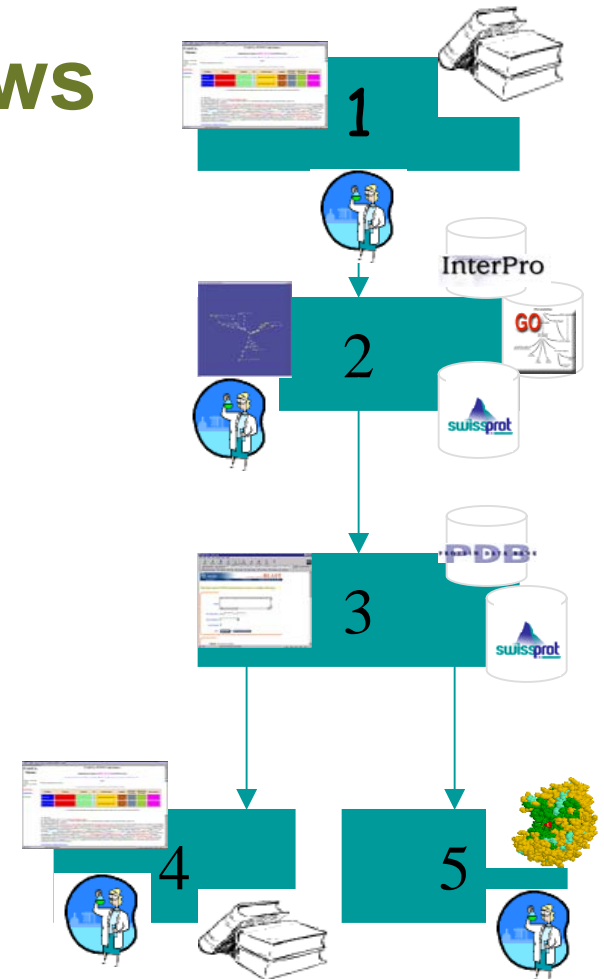
Image analysis and data extraction



Problems: in-silico experiments are workflows

Circadian Rhythms

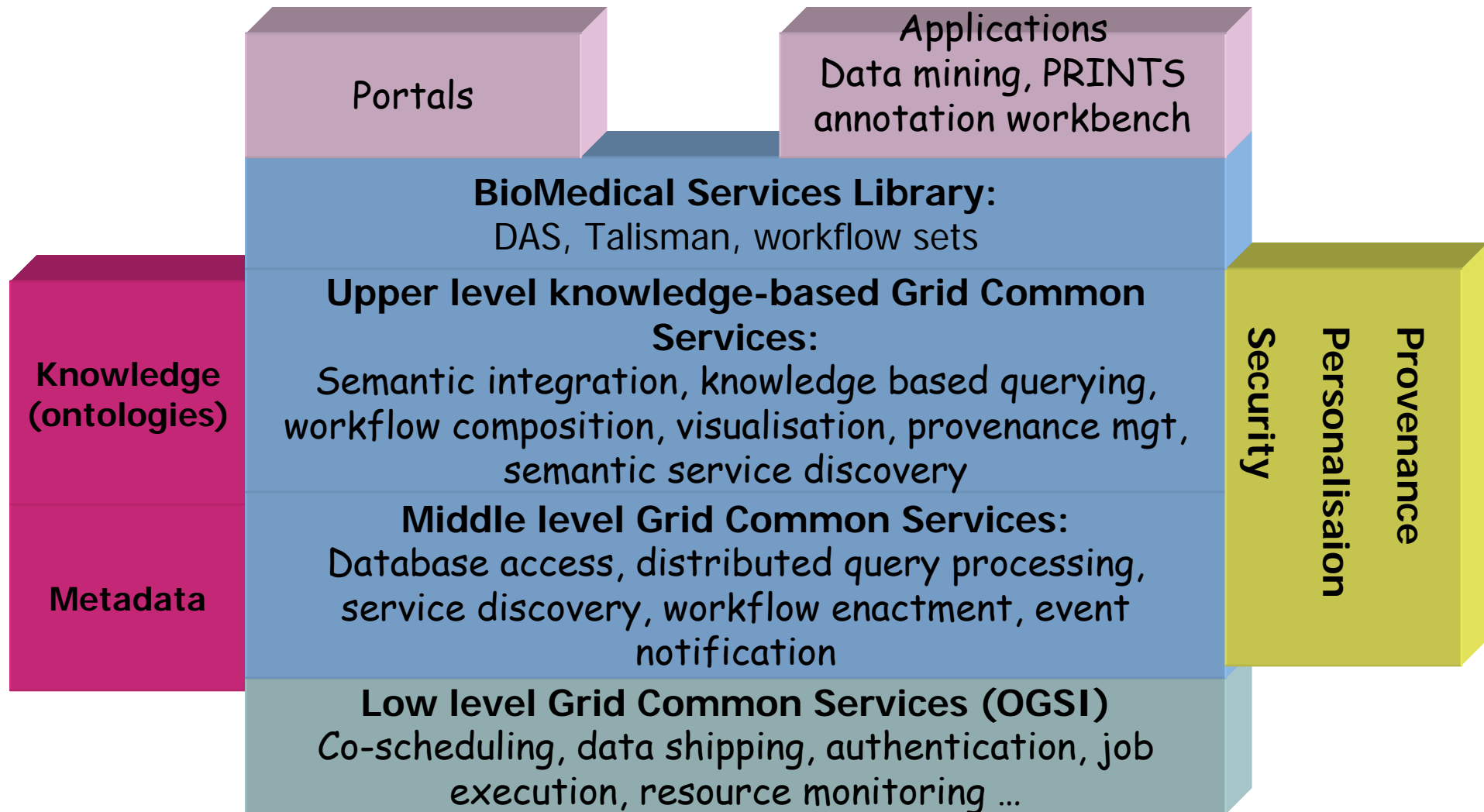
- Has anyone else studied the effect of neurotransmitters on the circadian rhythms in *Drosophila*?
- I've got a cluster of proteins from my experiment. How do their functions interrelate? And what are the proteins with a particular function?
- Is a structure known for my protein? What other proteins have a similar structure?
- What is known about the homologous protein?
- Can I build a homology 3D model?



Problems: e-Science

- Personalisation
 - Who else has asked this question & can I use/adapt their approach?
 - I want to annotate and publish my process for use by others
 - I want to store and access my personal datasets
- Provenance
 - Which type, version and provider of BLAST did I use?
 - What was the workflow and the results at each stage?
 - I want to publish my findings as an protein annotation
 - Ownership, credit, trust, immutable and auditable data
- Change management and notification
 - When was P12345 last updated?
 - Has PDB changed since I last ran this workflow?
 - Has the data provenance changed?
 - Are there new or alternative services that I can use?

myGrid approach: marketecture diagram



myGrid approach: modular set of services

- Ontologies and knowledge-based services
 - Description and interpretation of data, tools and services
- Workflow authoring, publication and enactment
 - Service orchestration and process automation
- Directories and repositories
 - Services, workflows, data and results
- Portal
 - Easy access to all myGrid services
- Notification and event propagation
 - Change management
- Knowledge Extraction
- Security and trust

Example: describe the workflow

The image displays three overlapping screenshots of the myGrid web application, illustrating the workflow for describing a service and creating a workflow.

myGrid Service

A description of a myGrid service:

- requires input
- uses method
- produces result
- performs task
- uses resource
- is function of

Files
Jobs
Current Users
History
Import
Logout

Match

myGrid Service

A description of a myGrid service:

- requires input
- uses method
- produces result
- performs task
- uses resource
- is function of

Files
Jobs
Current Users
History
Import
Logout

Match

myGrid Service

A description of a myGrid service:

- requires input
- uses method
- produces result
- performs task
- uses resource
- is function of

Files
Jobs
Current Users
History
Import
Logout

Match

myGrid Workflow Description

The workflow consists of these services:

```

graph TD
    A[mygrid retrieve GO term using SwissProt accession number] --> B[mygrid display 3D plot of Gene Ontology lattice]
    B --> C[Back]
    B --> D[Finish]
  
```

Back Finish

myGrid Service Matching

The above description matches these services:

Example: execute the workflow

The screenshot displays the myGrid web application interface within a Microsoft Internet Explorer browser window. The address bar shows the URL: `http://128.243.22.166:8080/myGridPortal/index.html`. The main page features the "myGrid" logo and a sidebar with navigation buttons: Files, Jobs, Current Users, History, Import, and Logout. A "Select Service" dialog box is open, listing three options: "BioFetch Service (Cambridge)", "BioFetch Service (Milan)", and "BioFetch Service (London)". The "BioFetch Service (Cambridge)" option is selected. The main content area displays the status of a workflow execution, showing the ID "1019051503123" and the status "Running". Below this, a "Description" section lists the workflow steps: "Display workflow provenance in native format", "Display selected object(s) in HTML", "Stop the workflow without waiting for results", and "Can't view the workflow output". The bottom of the browser window shows a status bar with "Applet started." and "Local intranet".

Example: view results

http://128.243.22.166:8080/myGridPortal/index.html - Microsoft Internet Explorer

File Edit View Favorites Tools Help

uk.ac.ebi.escience.tmo.tviewer.TViewer

GO:5748 - mitochondrial

Files

Jobs

Current Users

History

Import

Logout

myGrid

Applet started.

7601

http://128.243.22.166:8080/myGridPortal/index.html - Microsoft Internet Explorer

File Edit View Favorites Tools Help

myGrid

Workflow ID	Status
workflow.king:1018605180834	Could not retrieve the status, sorry :(

Operation	Description
Display XML	Display workflow provenance in native format
Display HTML	Display selected object(s) in HTML
Stop Job	Stop the workflow without waiting for results
View Output	View the workflow output

Workflow Provenance Data

Workflow ID	workflow.king:1018605180834
User	mygrid

Inputs

Name	Type	Value
SwissprotAccessions	string[]	<div> Q9VVKQ5 Q9VWM4 P91657 Q9V720 Q9VXM4 Q9VBS4 Q9XZ14 Q9VUY1 Q9V6R4 Q9VV15 Q9VMB6 Q9VXZ8 Q9VKR8 Q9VGC4 Q9VC43 Q95078 Q26307 Q9VIC4 Q24511 Q9Y092 Q9VLZ7 Q95SV2 Q9VLP1 Q9VVD2 </div>

Outputs

Name	Type	Value
null	string[]	<div> GO:0008020 GO:0016789 GO:0016787 GO:0006821 GO:0007268 GO:0005247 GO:0005554 GO:0016301 GO:0006869 </div>

Applet started.

Local intranet

Extensions to example

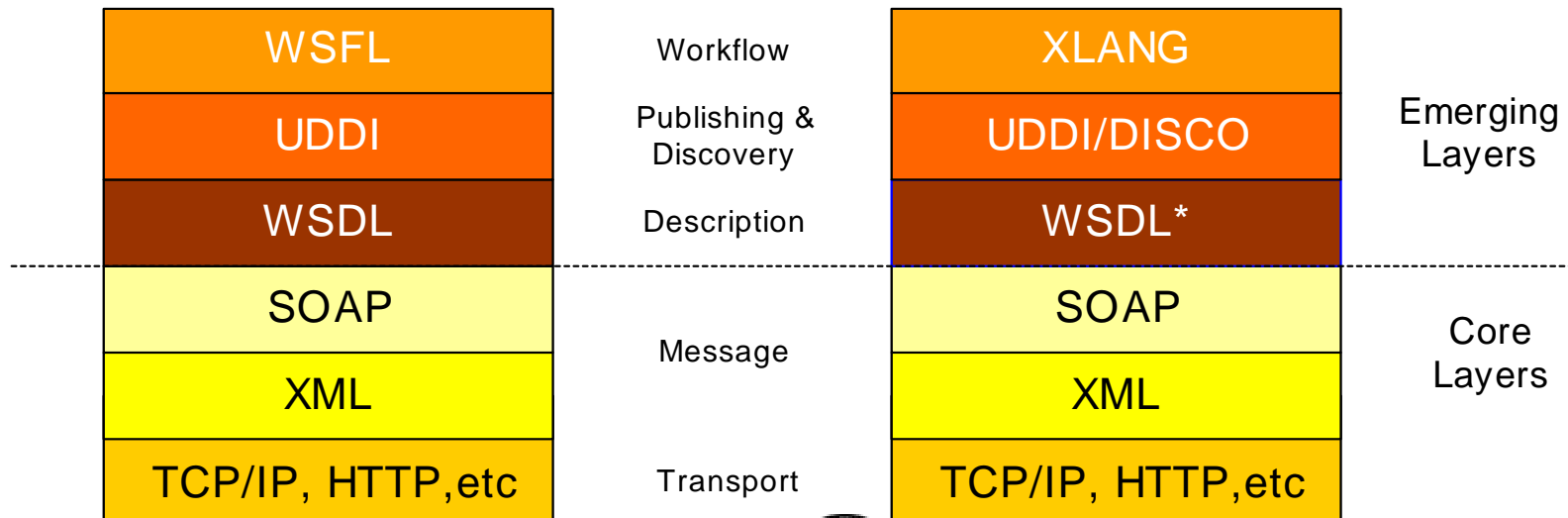
- Knowledge extraction from literature
 - Extract medline references from SWISSPROT annotations
 - Retrieve journal abstracts
 - Analyse abstracts using NLP and ontologies
 - Automatically hypertext link literature
- Change management
 - User notified by EBI of changes to services
 - User (or their software agent) re-enacts workflow
 - New results compared with previous results
- More complexity: EMBOSS workflow
 - Combined two concurrent application flows
 - Executing 7 applications using 45 web service invocations

Underpinning technologies: Web Services

- Platform, language and object model neutral
- Industry drive with a wide range of tools
- Rapid standardisation

IBM

Microsoft

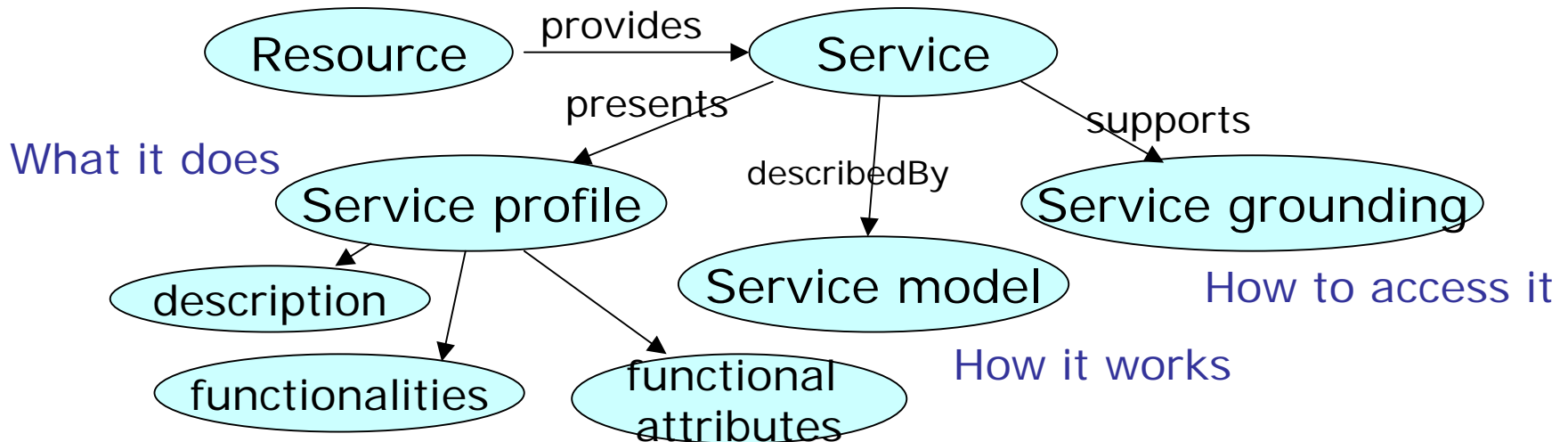


Semantics is the key to interoperability

- Web Services and XML only provide syntax and a framework for communication
 - Still need human readable specifications
 - Still need software developers to write scripts
- Semantics are needed to allow software to ‘understand’ the data and the function of services
 - Improves service discovery by guided searches and inference
 - Substitution of alternative services that have the same function
 - Automatic generation of clients that use and orchestrate services
 - Automatic selection and application of data transformations

Ontologies

- Ontologies: the shared and common understanding of a domain often in the form of a taxonomy
- Applies to services and people as well as biology
- Can be reasoned over by software



myGrid layered service ontology

1. Class of service

protein sequence alignment,
protein sequence database.

2. Specific services

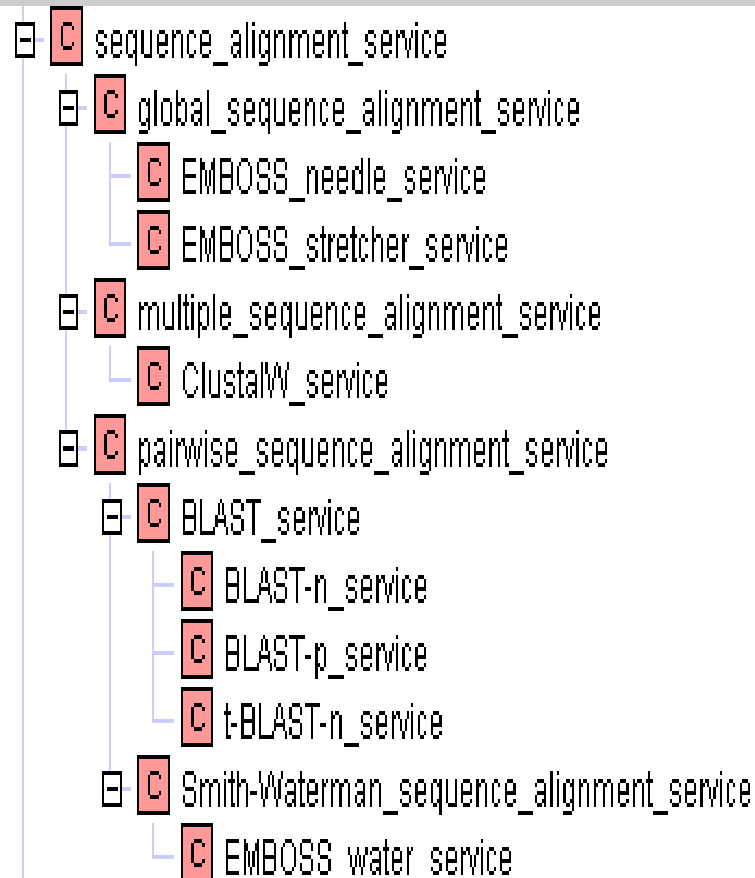
BLASTn is a tool for computing
sequence homology that uses the
BLAST algorithm over nucleotides

3. Instance description of specific services

BLASTn service is provided by EBI

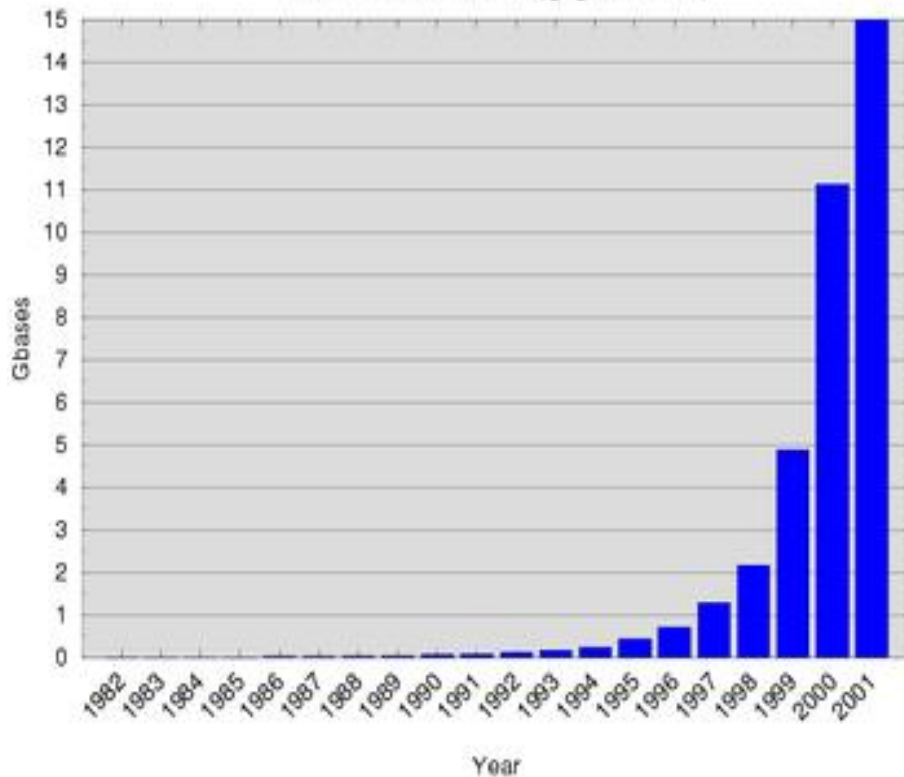
4. Description of invoked instance of a service

BLAST as executed for a particular
workflow



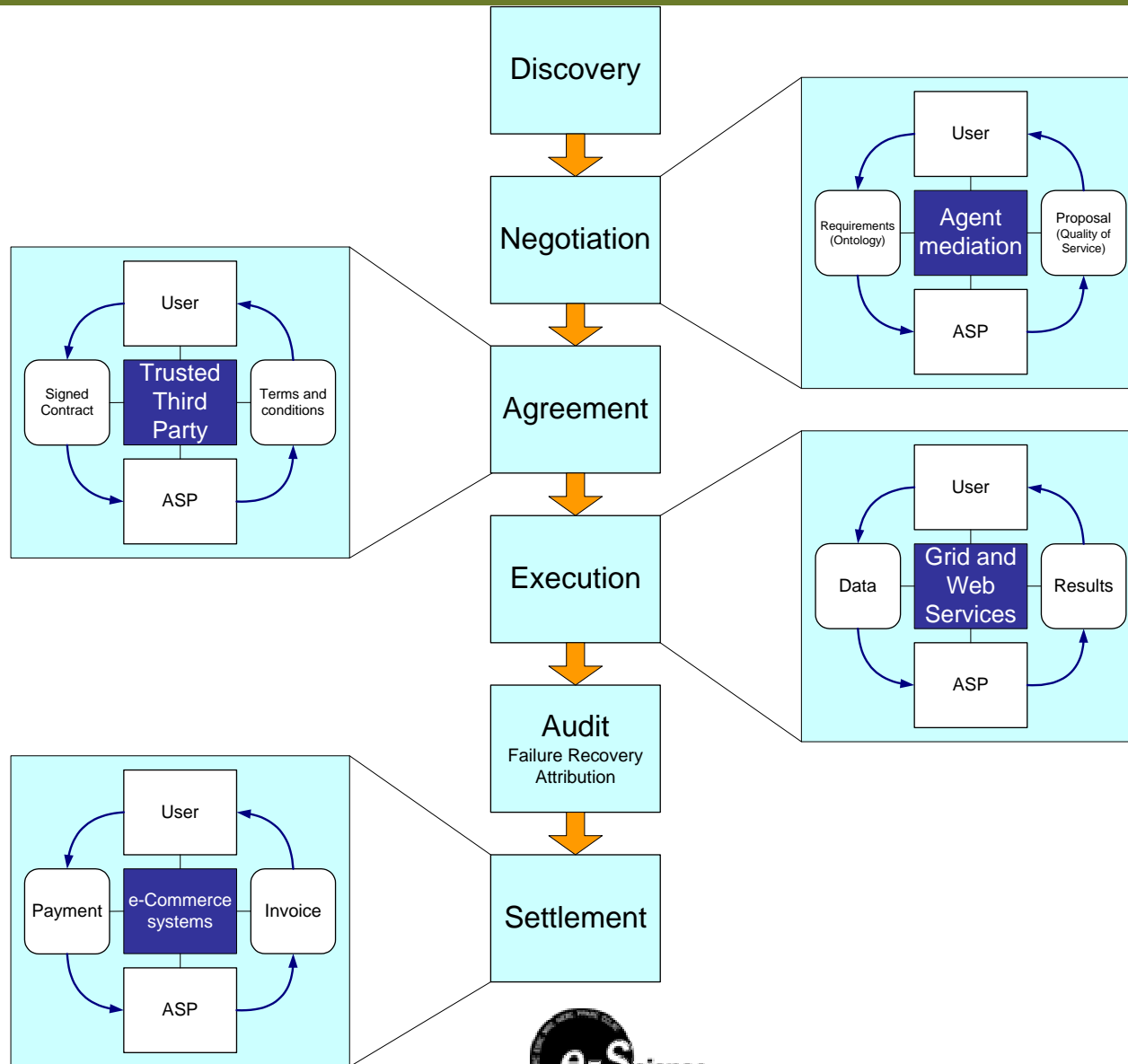
Underpinning technologies: Grid

EMBL Database Growth
total nucleotides (gigabases)



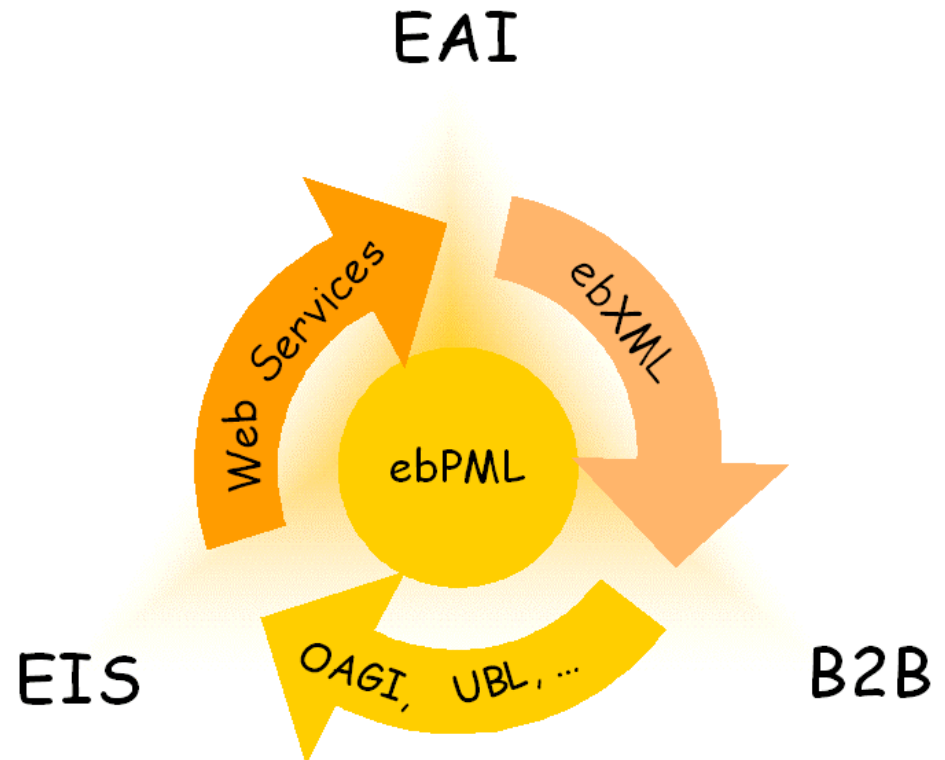
- Can't continue to centralise all resources in one place
 - Public service providers
 - Commercial users
- Large data volumes
 - Transfers and staging
 - Distributed query
- Computationally intensive
 - Resource management

Processes for outsourcing

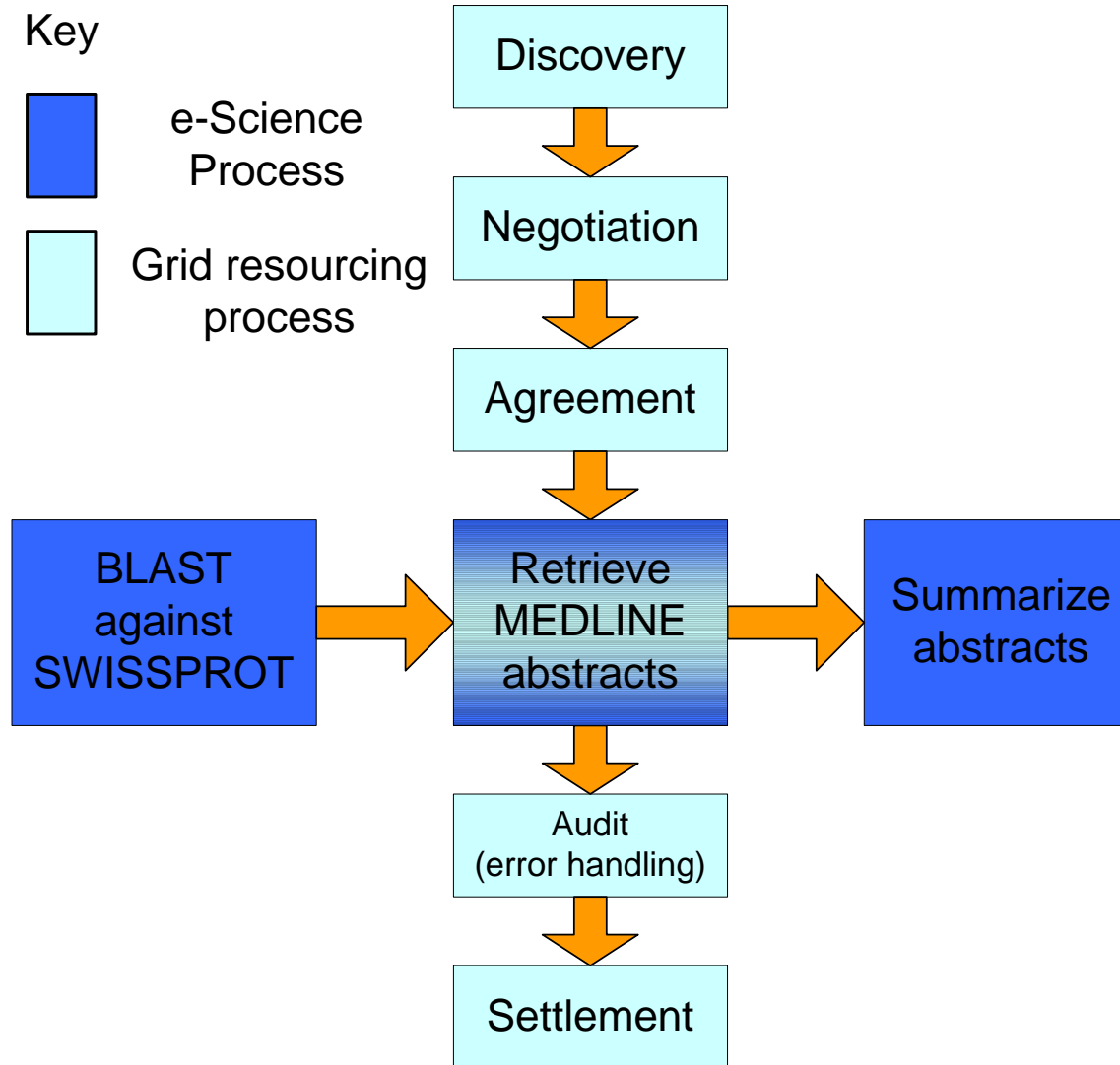


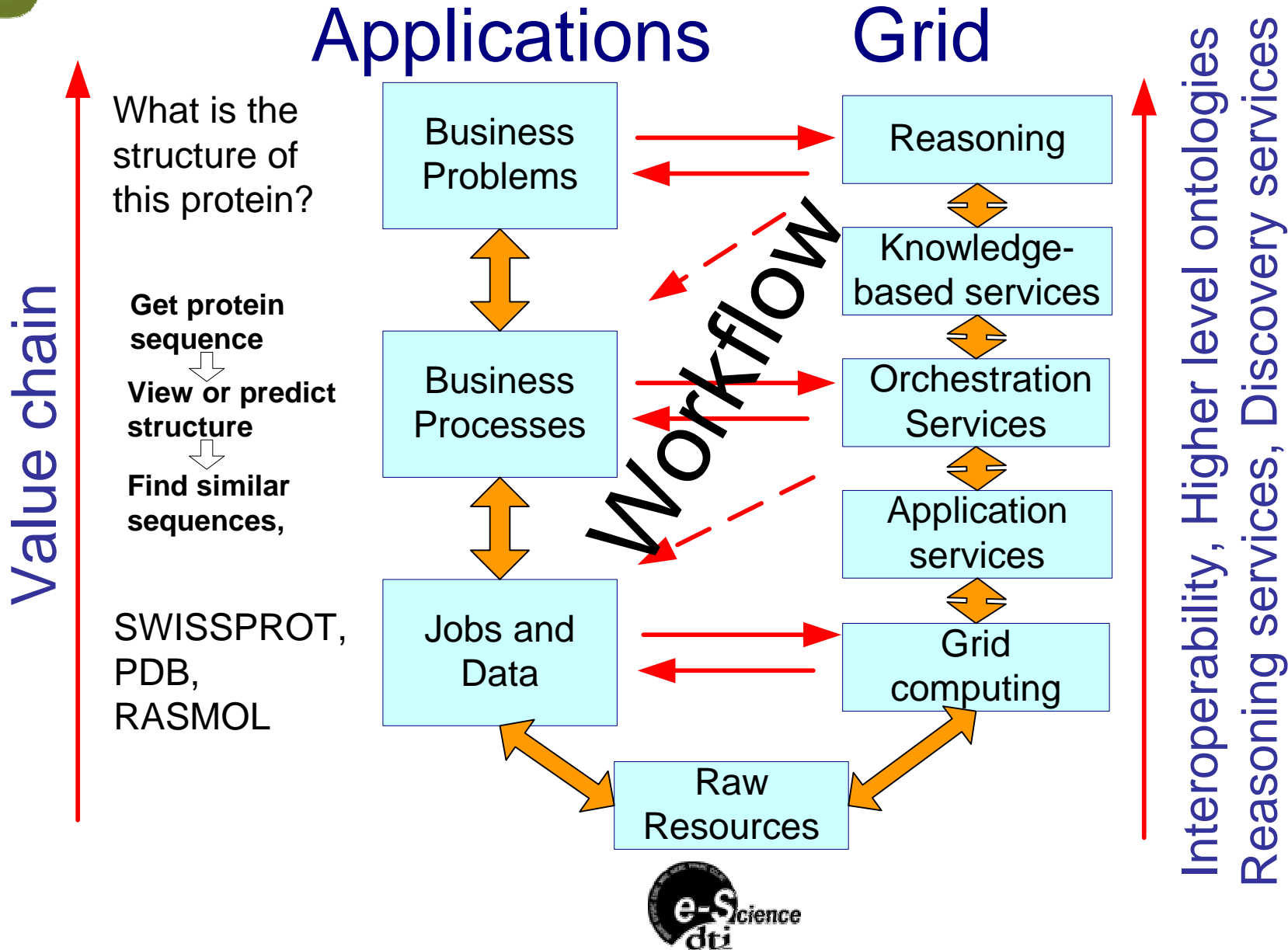
Underpinning technologies: workflow

Workflow Management Coalition	WfMC
OMG	EDOC UML
Web Services	Xlang WSFL BPEL4WS BPML



© 2002 ebPML.org





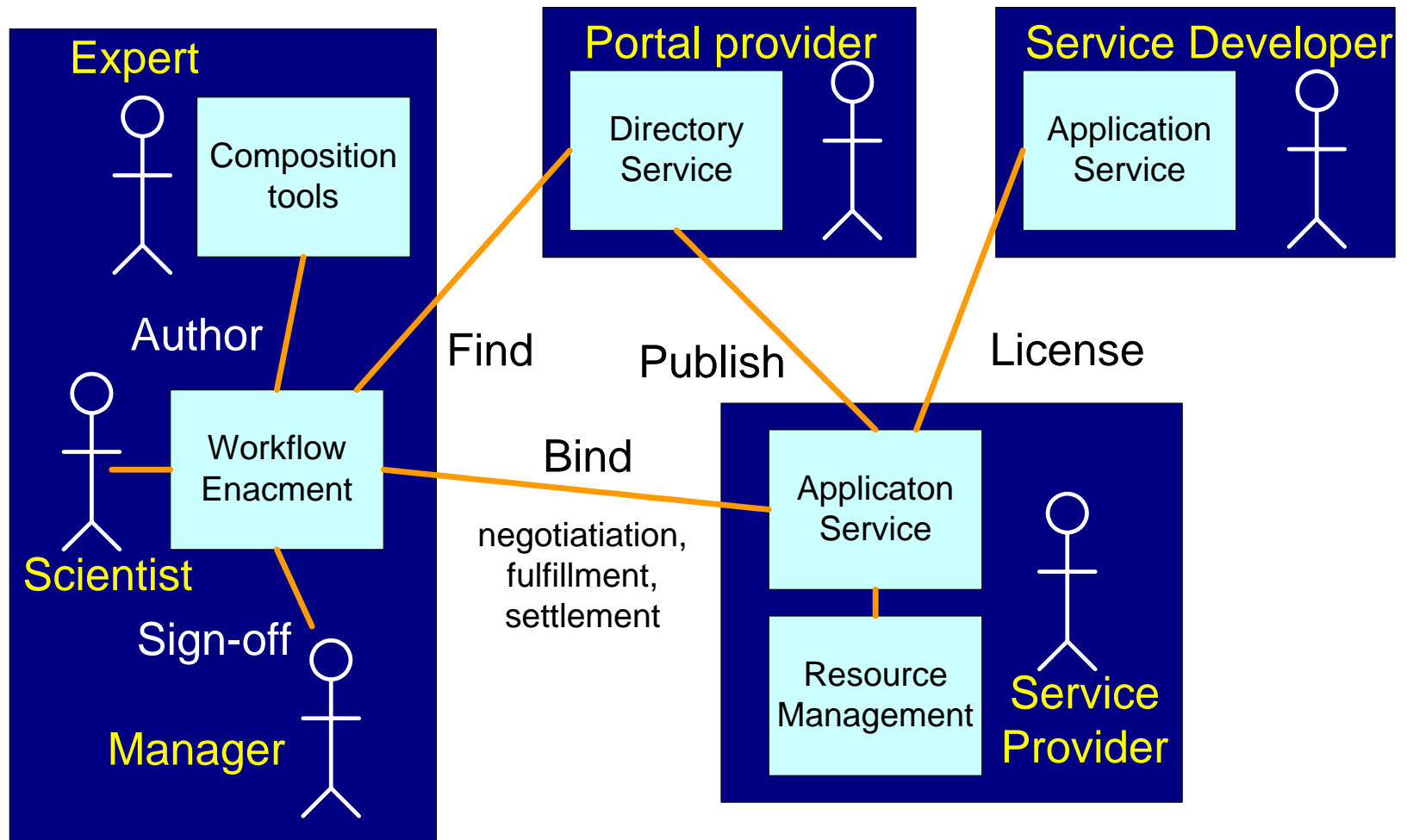
Next steps (1)

<div>Open Source</div> <div>Open Bio Foundation BioJava, BioPerl ...</div>		<div>Other Projects</div> <div>Astrogrid, Geodise, CLEF, Comb-e-chem, BIRN, OGSA-DAI</div>
<div>(DeFacto) Standards</div> <div>OMG LSR, I3C, MGED, Gene Ontology</div>		
<div>Semantic Web</div> <div>RDF, RDFS, DAML+OIL</div>	<div>Bioinformatics integration platforms</div> <div>DAS, OpenBSA, ISYS, OpenMMS, Kleisli, Ensembl, AppLab, SRS, BioNavigator, DiscoveryLink, K1 TAMBIS. BioMOBY ...</div>	
	<div>Web Services</div> <div>XML, SOAP, WSDL, UDDI</div>	<div>Distributed Computing Environments</div> <div>CORBA, RMI, JavaOne</div>
	<div>GRID</div> <div>Globus/SRB/Condor/Sun Grid Engine</div>	

Next steps (2)

- Develop and execute real world use cases
 - More complexity
 - Interaction with the user and their local tools
- Continue to research and develop infrastructure
 - Provenance and security
 - Personalisation
 - Distributed Query
- Release myGrid to the community as open-source
 - Spring 2003

Next steps (3)



Conclusion

- myGrid aims to develop middleware to integrate bioinformatics tools and data in a way that supports the scientist
- The setting is bioinformatics but the results are intended to be generally applicable to e-Science.
- A mix of standard, vanguard and bleeding edge technologies, advanced development and (some) research.
- Semantics and workflow are key to making the most of Web Services and Grid technologies

More information

- myGrid
 - Project Web Site <http://www.mygrid.org.uk>
 - Email myGrid mygrid@cs.man.ac.uk
- IT Innovation
 - Grid Projects <http://www.it-innovation.soton.ac.uk/grid>
 - Matthew Addis mja@it-innovation.soton.ac.uk