

eProtein Scientific Meeting and Workshop

24<sup>th</sup> – 26<sup>th</sup> April 2006

eScience and Protein Annotation  
@ Imperial College London

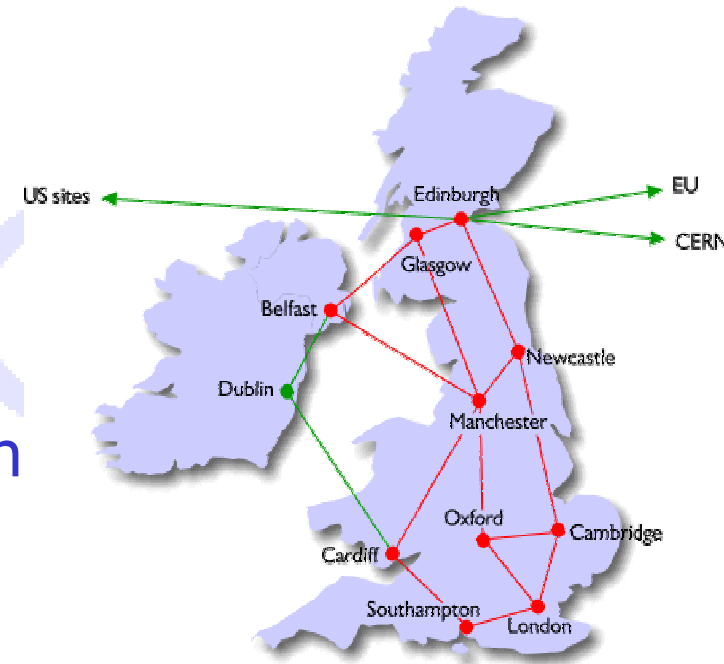
John Darlington

London e-Science Centre

Imperial College London

# London e-Science Centre

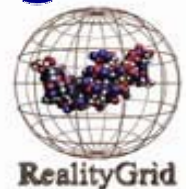
- Enabling e-Science activity
- Multi-disciplinary research
  - Bioinformatics
  - High Energy Physics
  - Computational Engineering
- Development of Next-generation Middleware
  - ICENI and GridSAM



**EPIC**



**ImmunologyGrid.org**



# e-Science & Grid Computing

## e-Science

High-level Support for Computational  
Science

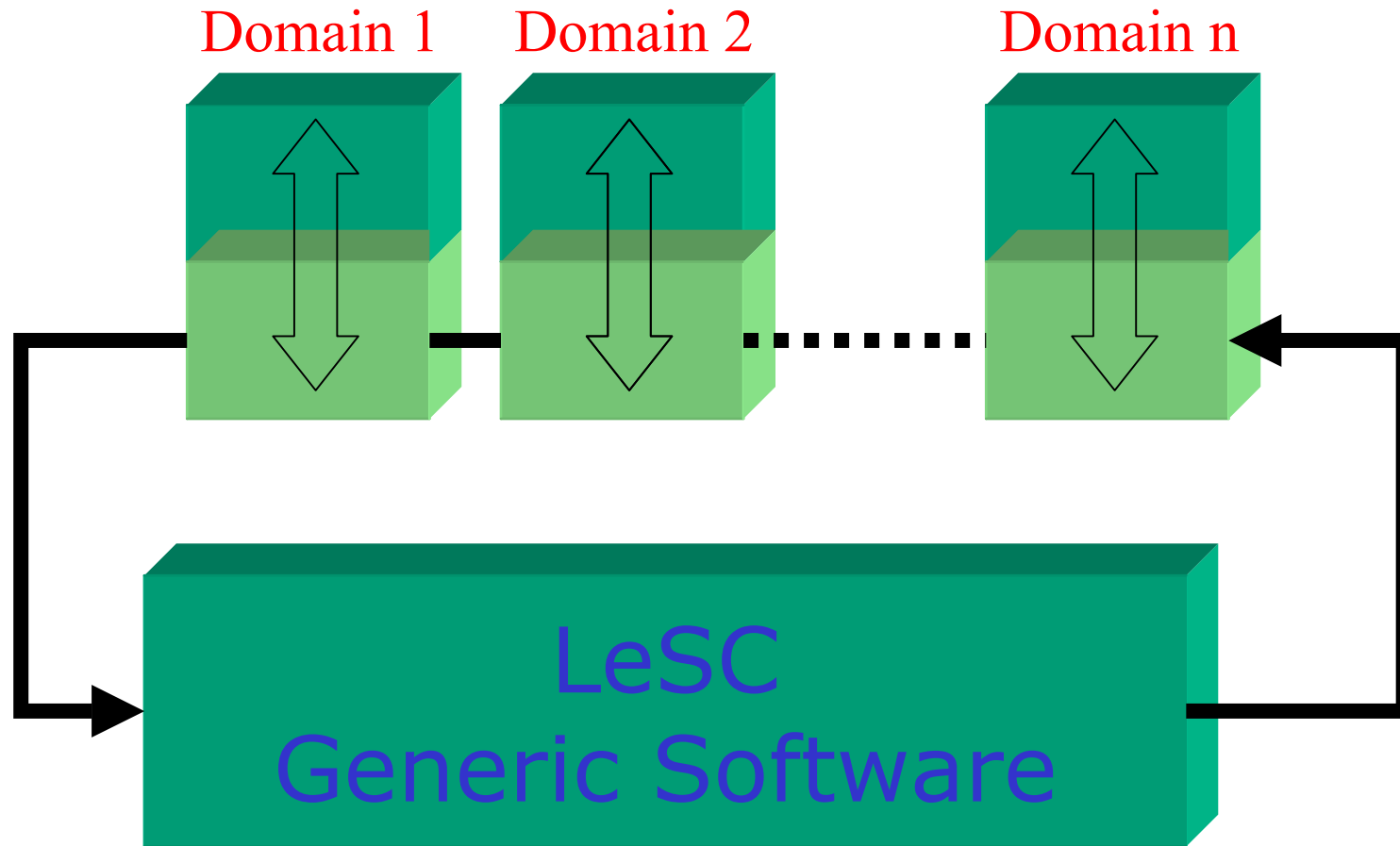
Abstraction/Encapsulation of Methods  
Support for Communities & Collaboration

## Grid Computing

**Transparent** use of Distributed Resources

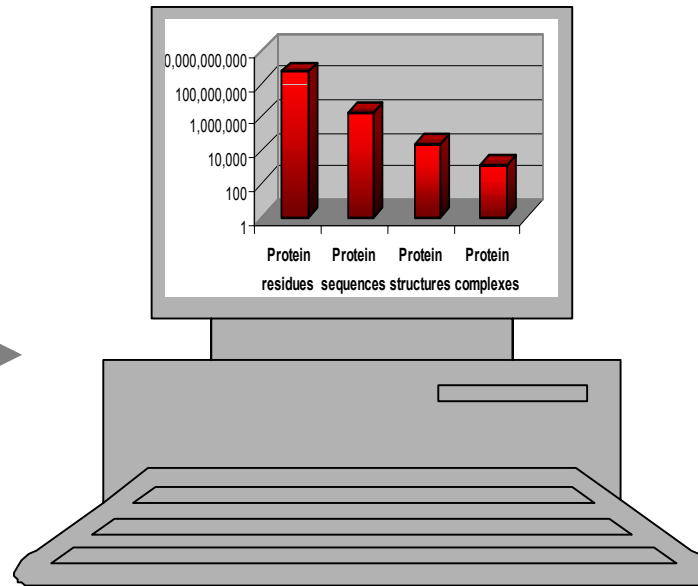
c.f. Bioinformatics!

# Applied Computing

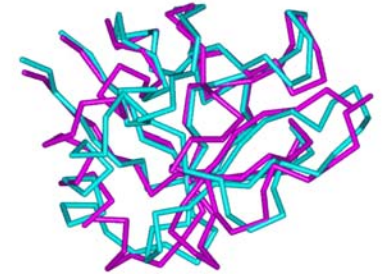


# The Scope of Problem

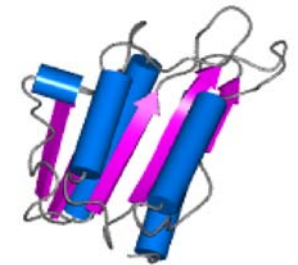
Query  
sequence



Matching  
homologs



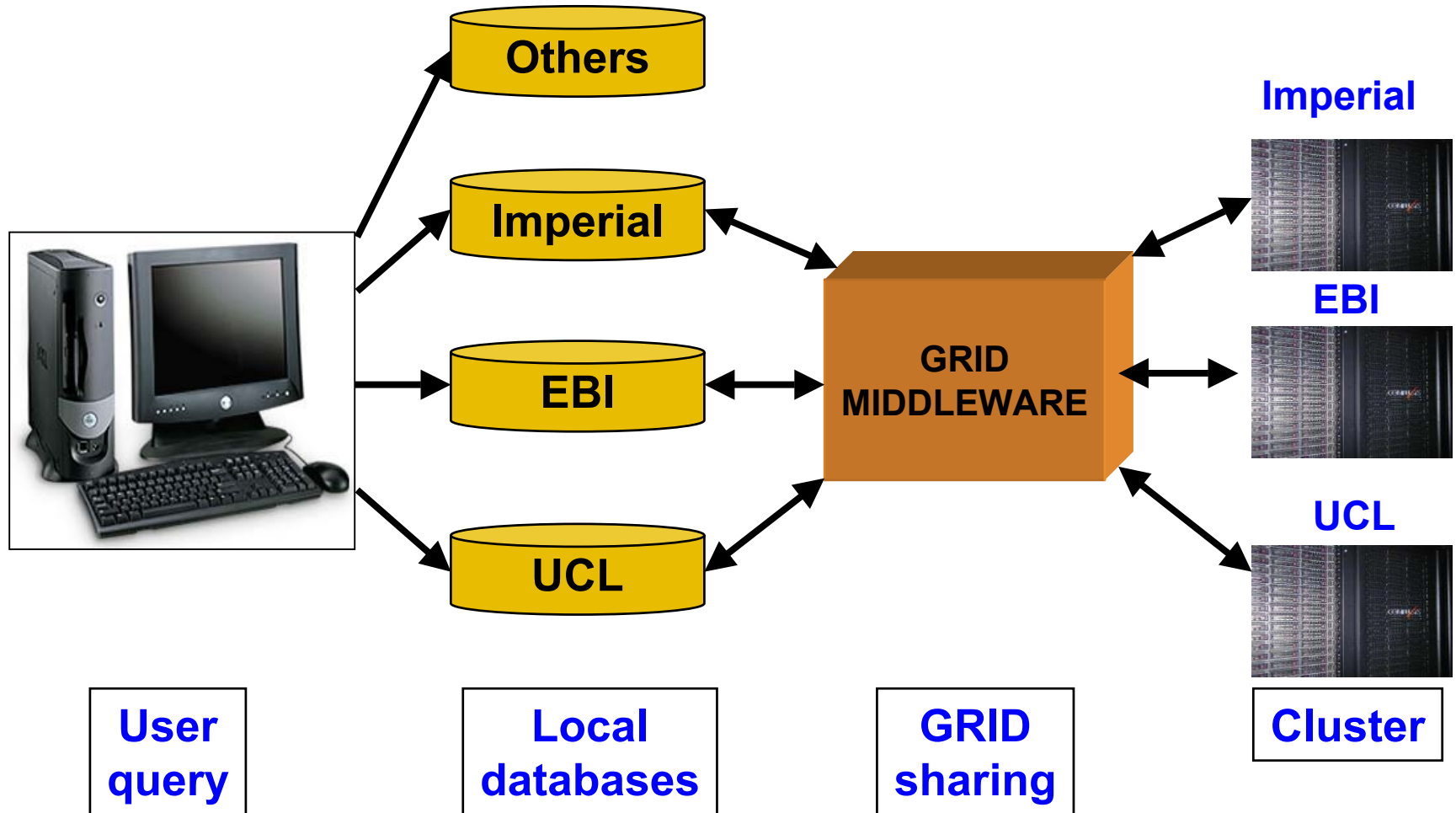
Matching  
fold



2,737,104 Sequences in UniProtKB

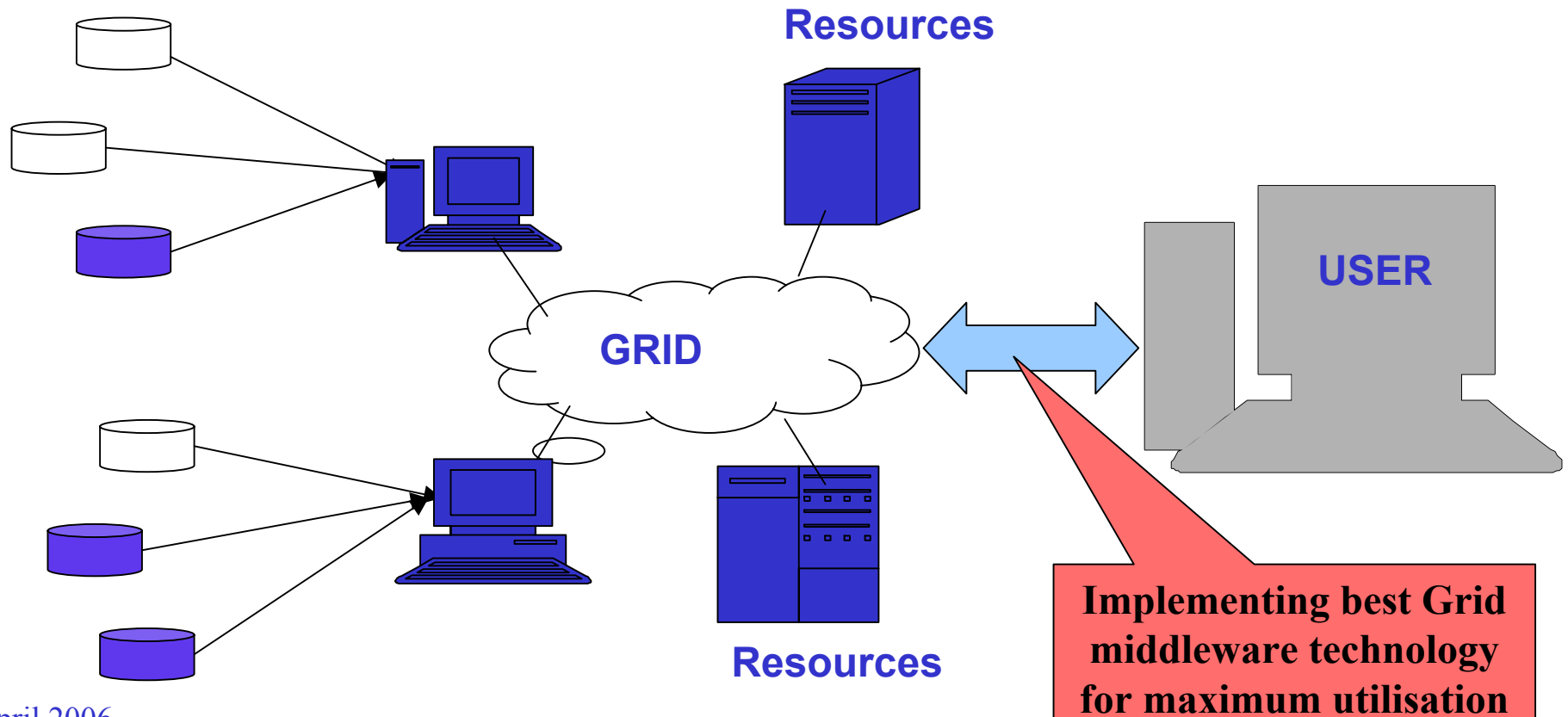
**Computationally expensive!**

# The eProtein Project



# The Key Challenges

- Develop best strategy to provide transparent high performance computing from distributed resources



# IC e-Science Networked Infrastructure

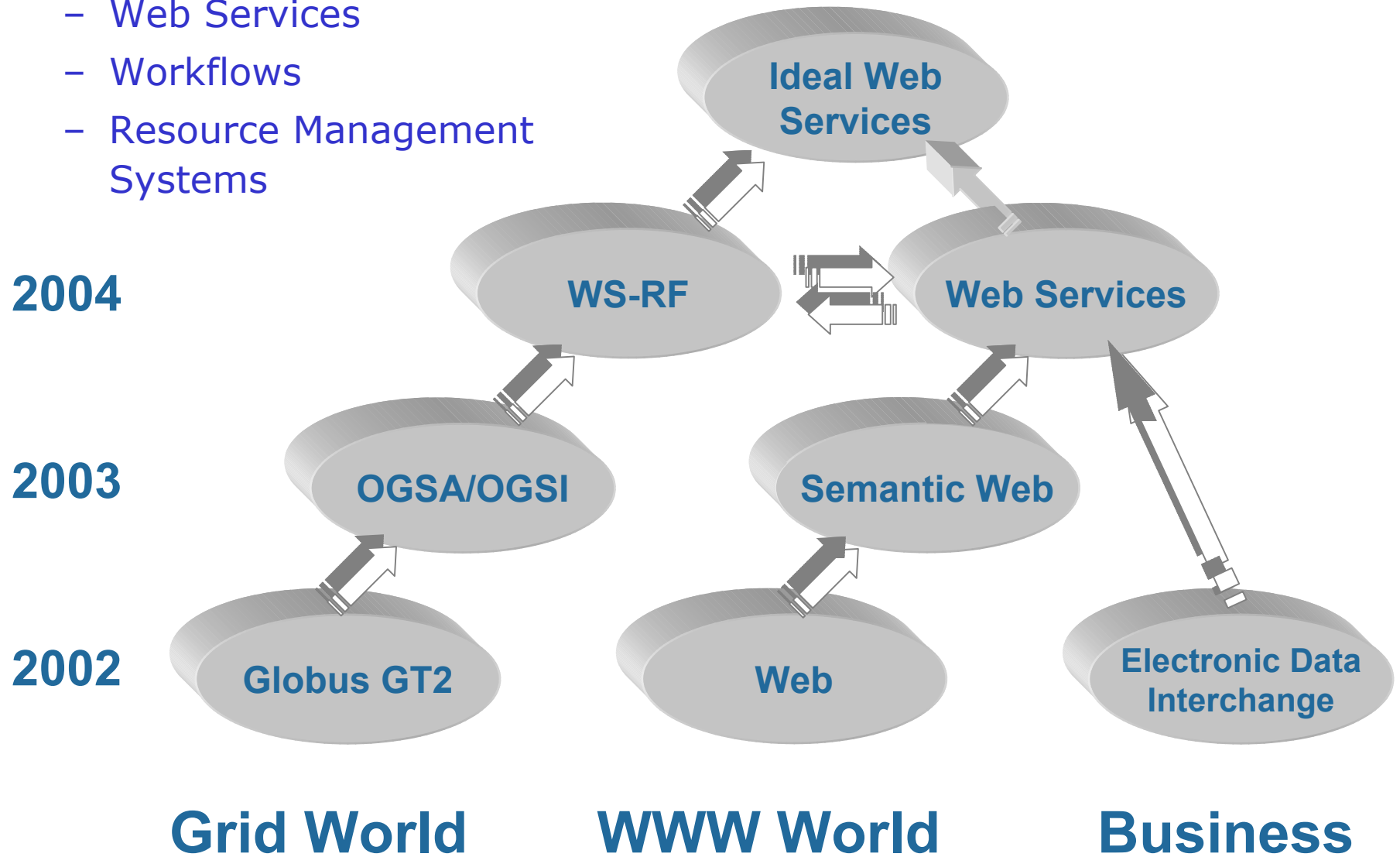
- A Grid middleware that abstracts the user away from the complexity of Grid
- Represents compute, storage and software resources as services which communicates using standard protocols
- An end-to-end middleware consisting of Grid service infrastructure and a service management framework



# Sensible Convergence

## Solutions

- Web Services
- Workflows
- Resource Management Systems



# Web Services and Service Oriented-Architecture (SOA)

- Web Service: A service which is available through a standardized interface (SOAP) and described through a standardized interface description language (WSDL)
- Service-oriented architecture is essentially a collection of services which can communicate with each other in a client-service manner

# Grid Job Submission and Monitoring Service (GridSAM)

- A job submission and monitoring web service
- One of the first system to support
  - Job Submission Description Language (JSDL)
- Works as a wrapper around the following job execution systems:
  - SSH, Condor Pool, Grid Engine 6, Globus 2.4.3
- The main functionality utilised:
  - Submit and Start Jobs, Monitor Jobs, Transfer Files



# GridSAM Overview

- For resource owners GridSAM works as a web service for execution on heterogeneous resources uniformly through *Forking* or *SSH*, *Condor Pool*, *Grid Engine 6*, *Globus 2.4.3*
- For end-users GridSAM works as a set of end-user tools and client-side APIs to interact with a GridSAM web service
- The main functionality is :
  - Submit and Start Jobs
  - Monitor Jobs
  - Terminate Jobs
  - File transfer
  - Client-side submission scripting
  - Client-side Java API



# Job Submission Description Language (JSDL)

- XML template language for describing core aspects of a job

Job Definition

Job Description

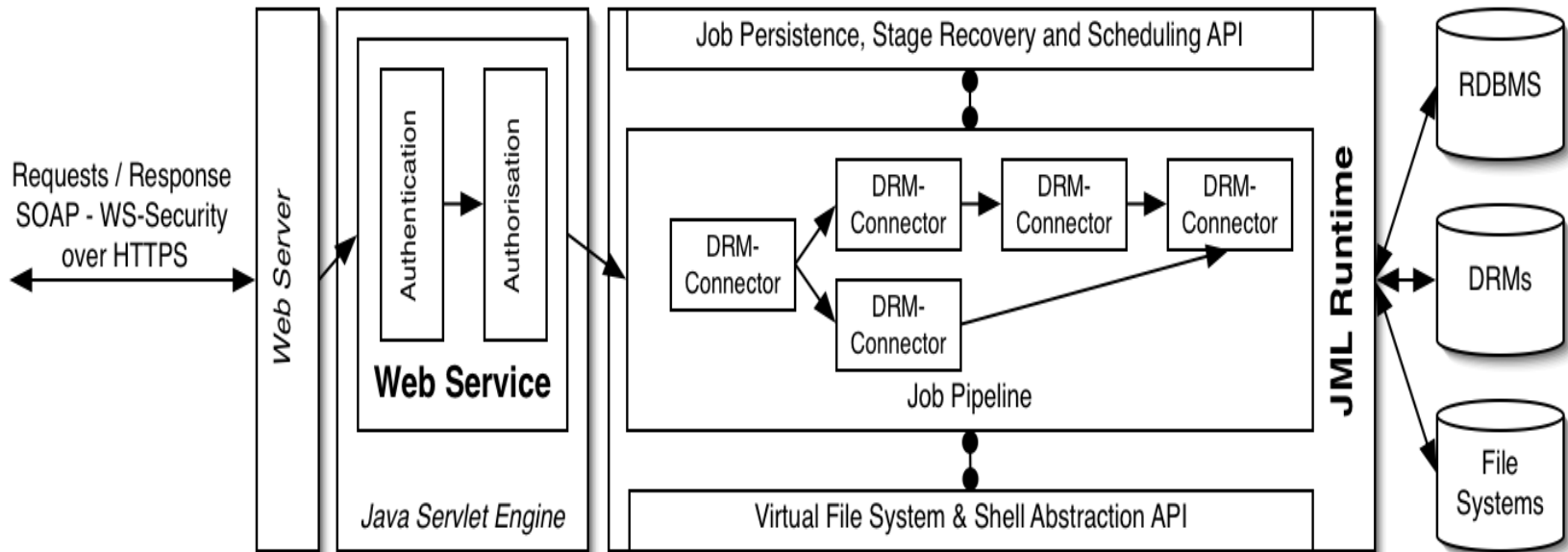
Application

For example, a BLAST application is described as:

For example, a BLAST application is described as:

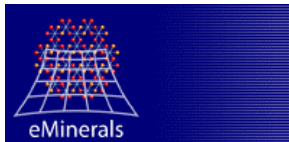
```
<DataStaging> <FileName>input-blastrun.in</FileName>  
<CreationFlag>overwrite</CreationFlag>  
<Source>  
<URI>ftp://anonymous@gridsam.lesc.doc.ic.ac.uk:55521/public/input-files/input-blastrun.in</URI>  
</Source>  
</DataStaging>
```

# GridSAM Implementation



# Users and Evaluators

- Users



VO TechBroker

MANGO



CCS  
UCL



- Evaluators



- Contributors

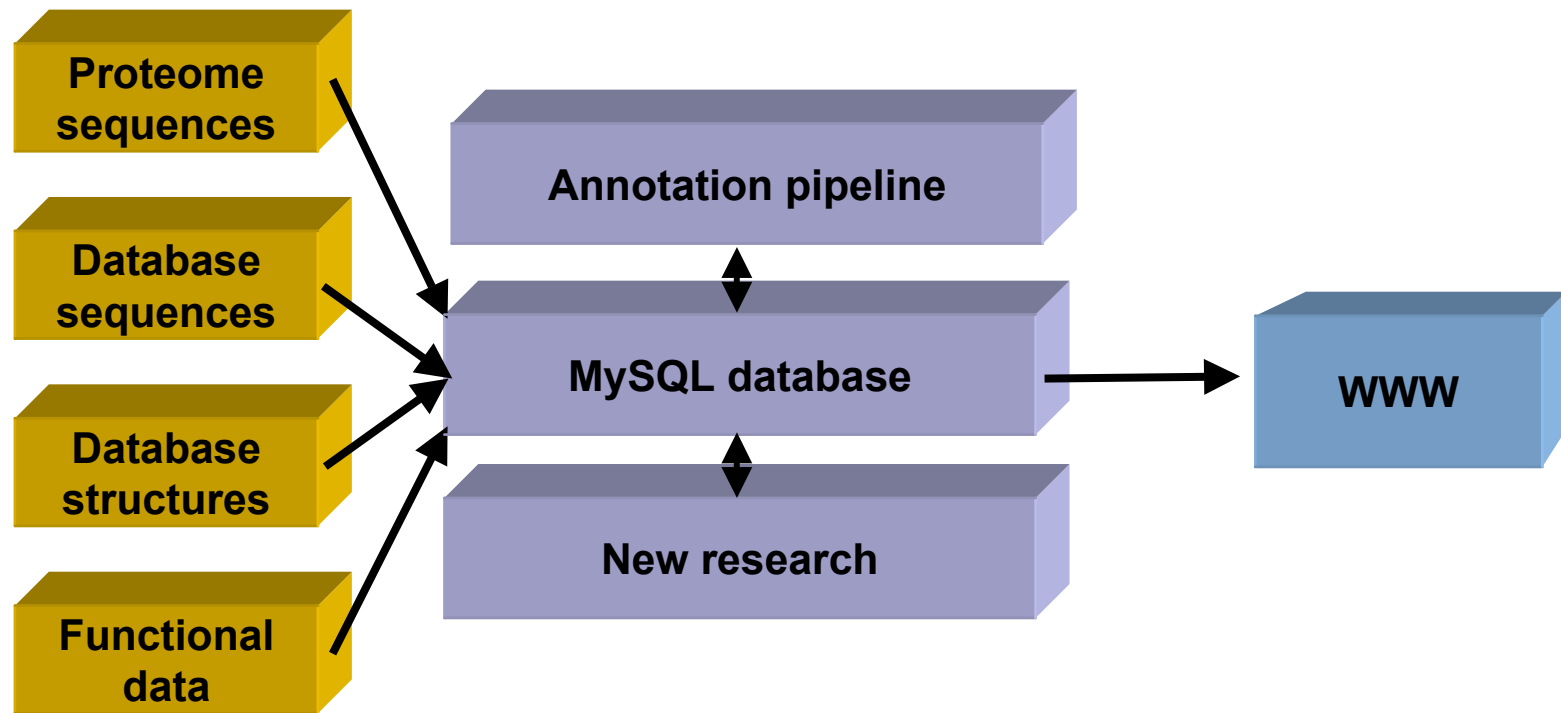


# Grid-based Resource Management

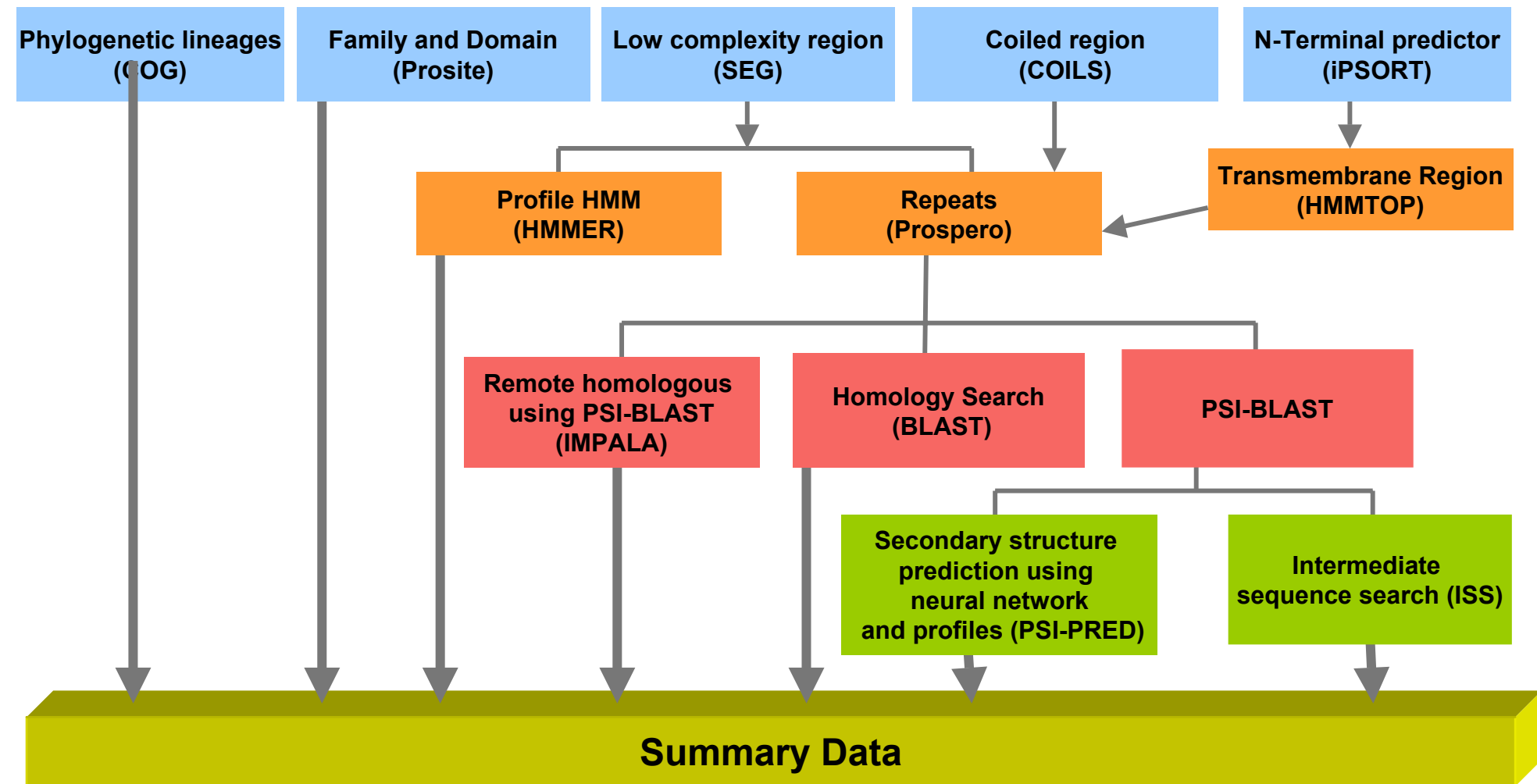
- 3D-GENOMICS Database
  - 3,132,369 sequences from 261 genomes
  - The annotation pipeline is based on homology and fold recognition methods
- Workflow Management System
  - Grid technology implemented for the workflow

# 3D-GENOMICS Architecture

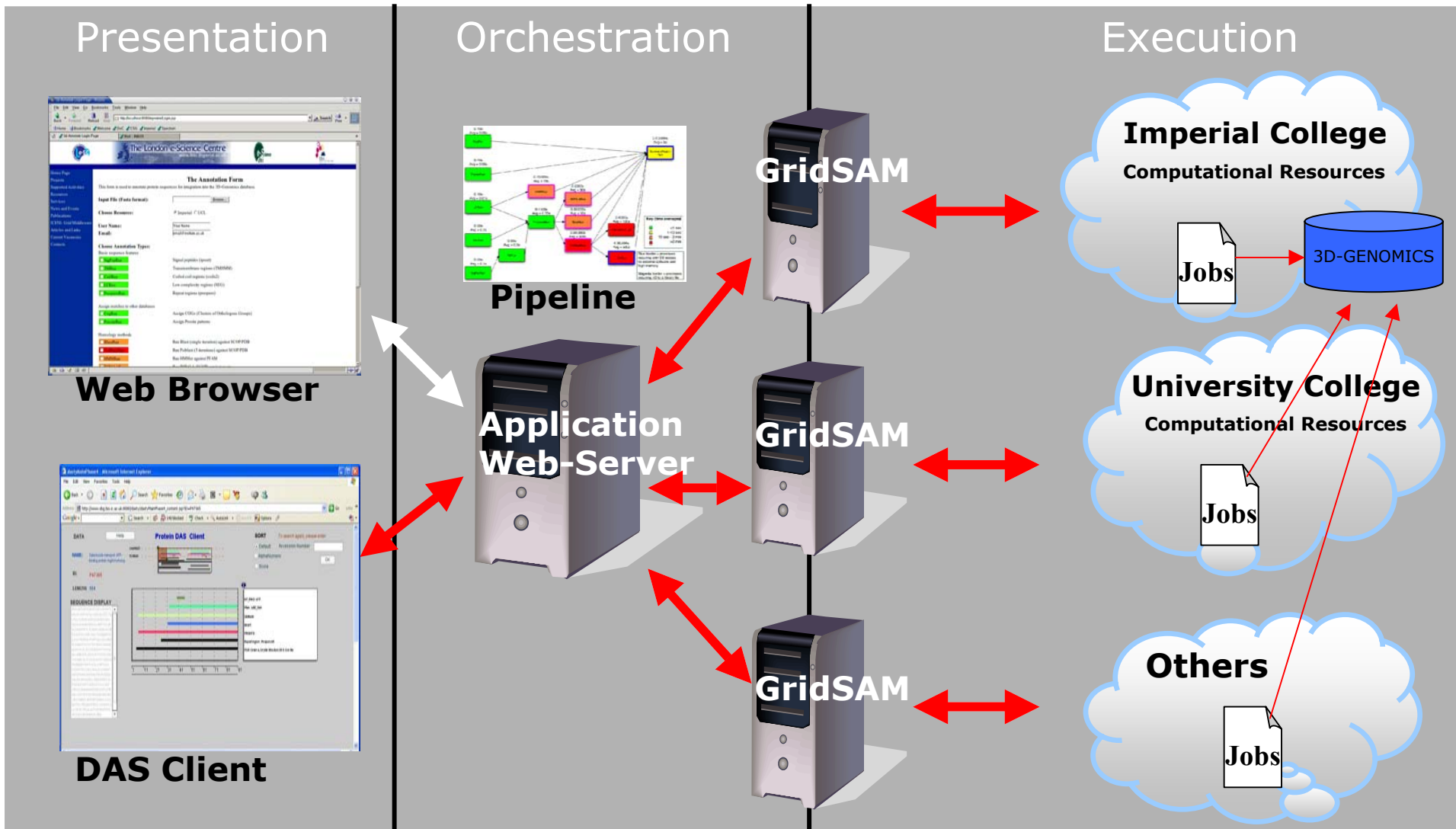
- A MySQL database which provides the structural and functional information for the protein annotation



# The Annotation Pipeline



# Workflow Management System

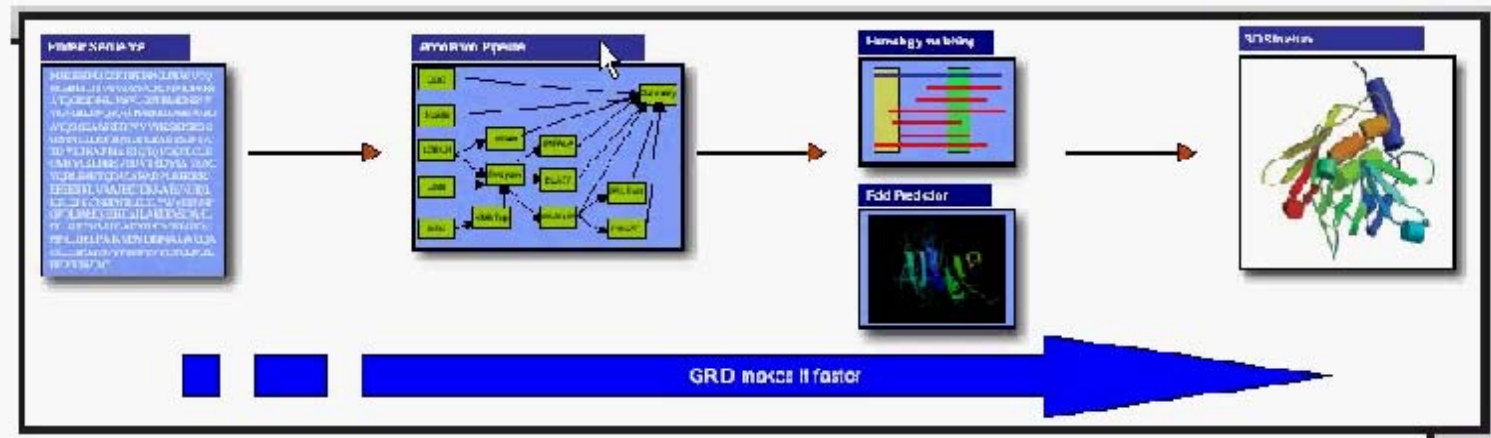




- Home Page
- Projects
- Supported Activities
- Resources
- Services
- News and Events
- Publications
- ICENT Grid Middleware
- Articles and Links
- Current Vacancies
- Contacts

## 3D-Annotate

The e-Protein project is using emerging Grid technologies to combine heterogeneous resources at multiple sites (Imperial College London - IC, European Bioinformatics Institute - EBI, University College London - UCL) collaborating in the execution of these proteome annotation pipelines. The pipeline used within the project is utilising homology and fold recognition methods to assign structures to the proteomes and generate three-dimensional models. The London e-Science Centre is developing ICENT a service-oriented middleware framework which is used extensively within the e-Protein project to capture the workflow of this pipeline, and map it to resources on the Grid.



**Click for  
Login Page**

Login Page

**Input File (Fasta format):**

**Input Fasta File**

**Choose Resource:**

Imperia  UCL

**User Name:**

**Email:**

**Notification via e-mail**

**Choose Annotation Types:**

Basic sequence features

SigPepRun

TMRun

CoilRun

LCRun

ProsperoRun

- Signal peptides (psort)
- Transmembrane regions (TMHMM)
- Coiled coil regions (coils2)
- Low complexity regions (SEG)
- Repeat regions (prospero)

Assign matches to other databases

CogRun

PrositeRun

- Assign COGs (Clusters of Orthologous Groups)
- Assign Prosite patterns

Homology methods

BlastRun

PsiBlastRun

HMMRun

- Run Blast (single iteration) against SCOP/PDB
- Run Psiblast (5 iterations) against SCOP/PDB
- Run HMMer against PFAM

**Select application for the pipeline**

<b>Submission</b>	Done
<b>Monitoring</b>	Done

Results at [3D-GENOMICS HOME PAGE](#)

Please check the output from the following link. If your browser does not automatically redirect you in page

**Jobs execution  
summary and  
link to the  
result page**

[Back to top](#)

Comments to [shd@doc.ic.ac.uk](mailto:shd@doc.ic.ac.uk). Copyright 2005 The London e-Science Centre.

This page was last modified on Wed Sep 14th 09:51:36 BST 2005



DATA

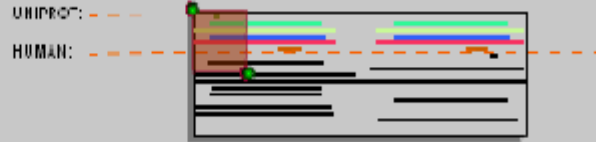
Help

# Protein DAS Client

**SORT**

*To search again, please enter*

**NAME:** Galactoside transport ATP-binding protein mglA homolog.



**ID:** P47365

**LENGTH:** 564

Default: Accession Number

AlphaNumeric

Score

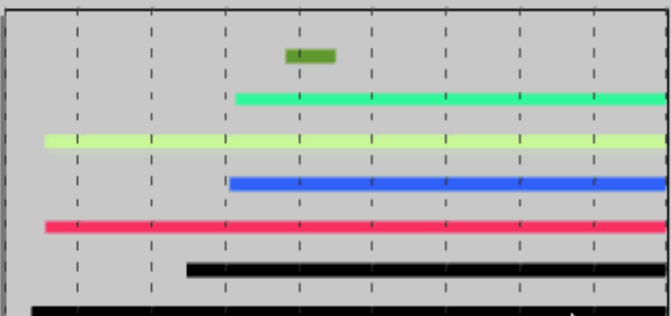
**Accession Number Search**

## SEQUENCE DISPLAY

```

MEHVAFMMEHSHSFCNGCIANVDV
SIVAVENTVHTLGENGAGHSITLSIF
G.LYIPDSGIFG EIQVNFSSICAVK
HIGMVHQHFLIENYVLDNHLGNE
SRFGFLP LIRIVCEAIKIMIKGIF
VQLIQVNSILTVGQQQRYEILIV.FF
DSNLIPIEFTAVLSDLEQNFILIRNF
KILGKTVLQSHLNEIQVADTATVLR
LGMVGSFDVITTTVEKALLRMGVE
LHQTKHTTDFVAKDEFVLQCNLNL
FLNISLAYIFLVPCNNIHAQDIIIN
PIFDUNISFLNLTTSWYTPK.VIGL
N.LGLSYQENTD EIPFAHQEIFAP
GVGNGQSQD.VNLCQIEIASN-LIF
NVIDSFAQIRKPIWAGIDVLEDRHI
YGLDULVYFHTVWQIENFPFS
VRF.IPIREPLSN.IIKFDVPGRE
GSRVYKLSGNGQFLIKKED.HQ
NQLLVAGVTPGLNGAIVHNSLLPI
ARNRILLVSYELDQIALACTVAVNI
G*NGMGIAD.MRQSSIP.LIRG

```



IP\_RIND: ATP  
 Pfam: ABC\_tran  
 DOMAIN:  
 Smart:  
 PROSITE:  
 Repeatregator: Prospero\_H1:  
 SCOP: Superfamily 1.1.1.1 E.Coll.Ma

Type: PDB  
 Name: 1q1b  
 Start: 4 End: 230  
 Description: Chain A, Crystal Structure Of E. Co  
 Score: -

**Annotation summary**

**Clickable Annotation**

# Big Crunch



Mars  
204 Dual  
Processor  
Opteron

Viking  
260 Dual  
Processor  
Pentium 4



Condor Pool  
>270 Workstations



# Experimental Results

## General Summary

Experiment wall-clock time: 43:58:48 (hh:mm:ss)  
No. of sequences in the experiment: 38931  
No. of completed annotations: 36669 (94.19%)  
No. of failed annotations: 2262 (5.81%)  
No. of annotations per hour: 833.76  
Average wall-clock time per annotation: 0:14:55 (hh:mm:ss)  
Expected duration of serial annotation: 0:1:50 (hh:mm:ss)  
Total number of jobs: 305781  
Total wall-clock time of jobs: 1167:39:0 (hh:mm:ss)  
Number of jobs completed per hour: 6952.72

# Pipeline Processes

## LC Summary

Total wall-clock time: 23:25:0 (hh:mm:ss)  
Average wall-clock time: 0:0:2 (hh:mm:ss)  
No. of jobs: 38993

## PROSPERO Summary

Total wall-clock time: 40:39:4 (hh:mm:ss)  
Average wall-clock time: 0:0:3 (hh:mm:ss)  
No. of jobs: 36744

## TM Summary

Total wall-clock time: 24:20:3 (hh:mm:ss)  
Average wall-clock time: 0:0:2 (hh:mm:ss)  
No. of jobs: 37498

## SIGPEP Summary

Total wall-clock time: 25:6:47 (hh:mm:ss)  
Average wall-clock time: 0:0:2 (hh:mm:ss)  
No. of jobs: 38879

## COIL Summary

Total wall-clock time: 23:57:5 (hh:mm:ss)  
Average wall-clock time: 0:0:2 (hh:mm:ss)  
No. of jobs: 38950

## COG Summary

Total wall-clock time: 24:31:14 (hh:mm:ss)  
Average wall-clock time: 0:0:2 (hh:mm:ss)  
No. of jobs: 39000

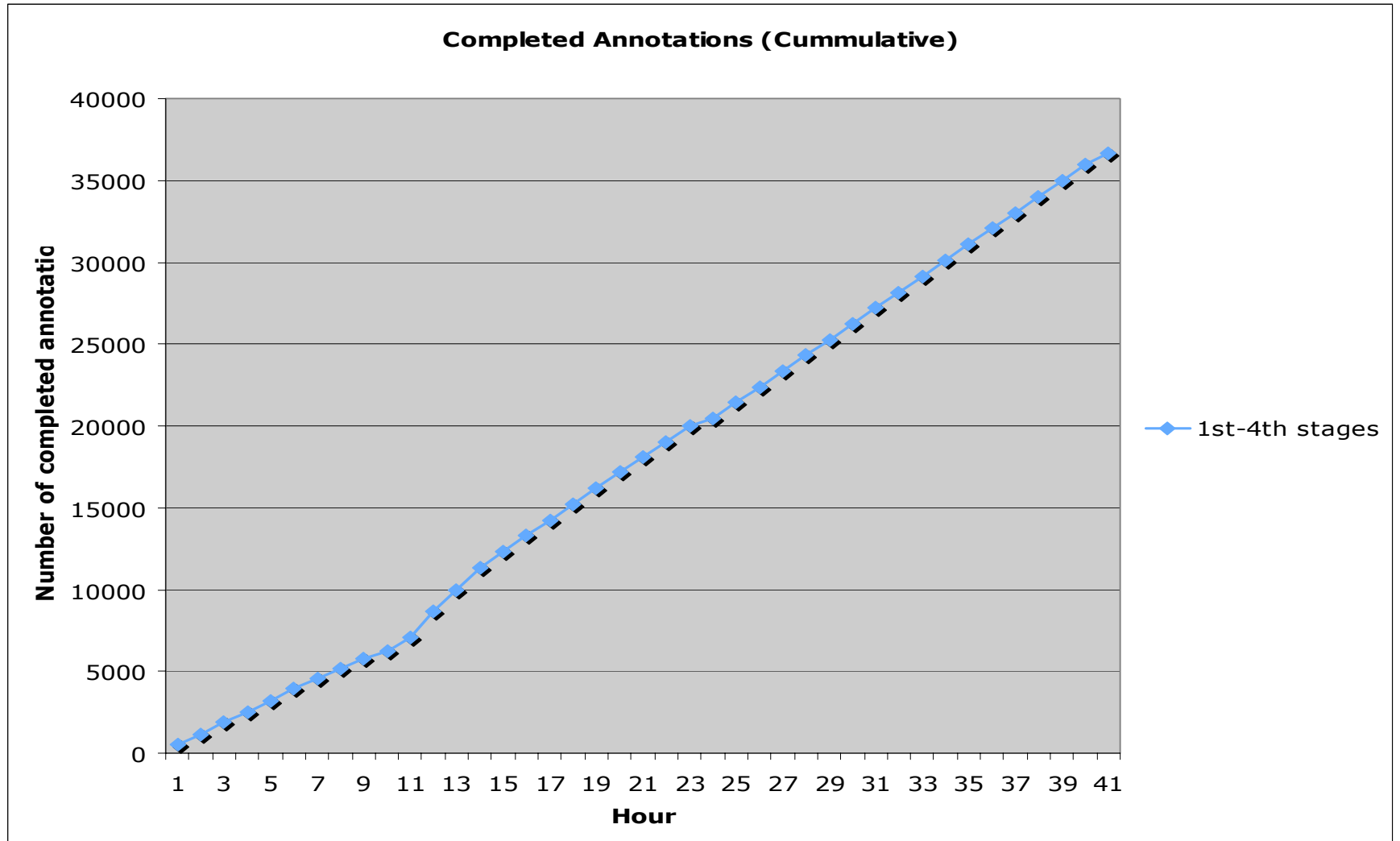
## HMM Summary

Total wall-clock time: 974:53:12 (hh:mm:ss)  
Average wall-clock time: 0:1:35 (hh:mm:ss)  
No. of jobs: 36711

## PROSITE Summary

Total wall-clock time: 30:46:35 (hh:mm:ss)  
Average wall-clock time: 0:0:2 (hh:mm:ss)  
No. of jobs: 39006

# Annotation Throughput



# Achievement

- First UK protein modelling group to provide workflow for annotation of human and other major species.
- Establishing methodology for protein annotation by variety of strategies and comparing results to identify problems.
- Develop best strategy to provide high performance computing from distributed resources making best use of available multi-site research
- Provide end-users with a powerful robust GRID interface

# Acknowledgement

## London e-Science Centre

A Stephen McGough

William Lee

Shikta Das

Murtaza Gulamali

Angela O'Brien

Oliver Jevons

## Structural Bioinformatics Group

Keiran Fleming

Nabil Ali

Arne Muller

Robert MacCallum