CISBIC Data Management, Integration and Bioinformatics

Imperial College London

Sarah Butcher
Centre for Bioinformatics
s.butcher@imperial.ac.uk

Centre for Integrative Systems Biology Imperial College

- **www.imperial.ac.uk/cisbic**
- Faculties of Natural Sciences, Medicine and Engineering
- Exemplars look at host pathogen interactions at the level of molecules and individual cells
- cycle of hypothesis formulation, experimentation, predictive modelling, model validation and formulation of the next hypothesis to deliver major biological insights into complex systems
- develop a core of multidisciplinary expertise which can champion this
- introduce new technologies that enhance the quality and quantity of relevant data

Exemplar Projects

- Glycomics of *Mycobacterium bovis* and *Campylobacter jejuni*
- Spatio-temporal control of phagocytic signalling during uptake of attenuated *Salmonella typhimurium*
- Innate Immune Signalling

CISBIC Core Facilities

- Imaging
- Mass Spectrometry and glycomics
- Mass Spectrometry and proteomics (phosphoproteomics)
- Metabonomics
- Microarray (2 colour)
- Microarray (affymetrix)
- Bioinformatics

Centre for Bioinformatics

- **Opened 2001 with mission to:**
- promote and co-ordinate world-class research and training in Bioinformatics within Imperial
- provide state-of-the-art Bioinformatics support to members of Imperial for their research

- Centralised bioinformatics resources
- Network of affiliates
- An entry point to Bioinformatics within Imperial
- Outreach - London Bioinformatics Forum
- High quality teaching – e.g. Wellcome 4yr PhD, MSc

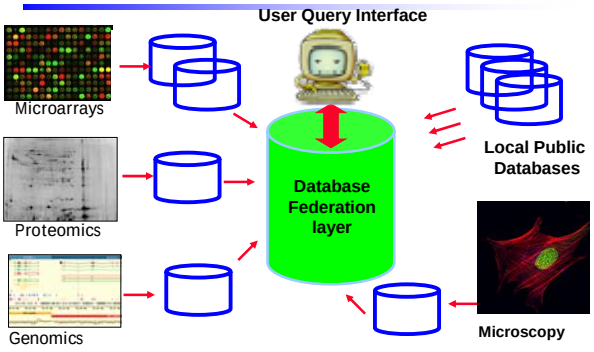
Bioinformatics Support Service

- **4 full-time staff** - wide range of expertise
- **Hardware resources –** dedicated central servers
- **Local copies of public biological databases**
- **Help-desk via email**
- **Web-site** - help documentation/tutorials
  - web-based programs
- **One-to-one tutorial sessions and site visits**
- **Formal teaching** - practical-based training courses
- **Development** - bespoke scripts and interfaces
- **Project-based consultation and collaboration**

Resources For CISBIC

- 2.5 new posts
- Additional bioinformatics support by BSS
- New dedicated database servers (development & production)
- Use of existing core bioinformatics databases
- Use of HPC in LeSC
- Close access to wet-lab researchers and modellers

System Biology Data Management



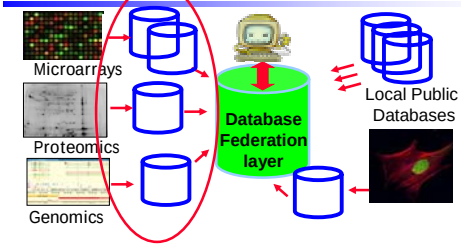
Current Scoping

- Survey existing data management practices within core groups – existing databases/systems?
- **Discuss relevant minimal biological data to store (common to all experimental methods)**
- Survey current 'useable' status of relevant standards e.g. MIAME, PSI and markup languages e.g. SBML
- To barcode or not to barcode?
- *Then design of primary central data management module to store biological data and tag to unique primary accession number*

Ideally....

- Ensure everyone using any of the core facilities **has** to enter crucial biological/experimental data centrally so experiments spanning different technologies times & groups on the **same** biological material can be identified and later cross-mapped
- Tied into equipment booking systems?
- Will electronic lab-books help?
- Temptation to ask for absolutely **every** parameter to be recorded 'just in case' but this affects compliance!
- Has to be easy to use

Experimental Data Repositories



- What is already in use?
- (How) can it be adapted/extended to tie to central modules?
- Where are new databases required?

Complications

- Legacy - multiple microarray databases already in use, some groups have no central repository
- Many data types arising from different groups, methodologies
- Many raw data files too large to exchange e.g. images
- Some complex data very difficult to interpret in isolation e.g. metabonomic
- Some data standards more mature (stable, useable) than others – trying to hit a moving target
- Requirement to manage data while data management/integration system being built

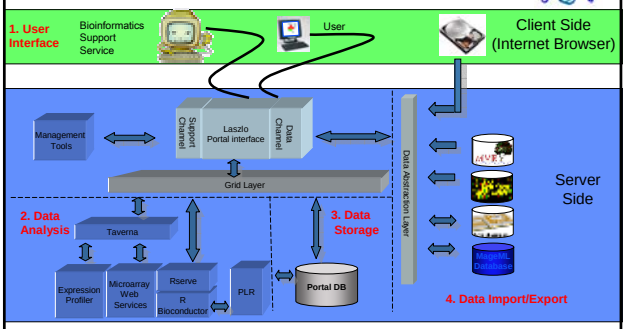
Integration Stages

- Data federation layers to associate data within type-specific repositories and allow centralised inter-database queries
- Integration with existing public databases for added-value
- Improved data input and query interfaces
- Updating whole in line with emerging standards and user feedback
- Data sharing policies (external)

Challenges

- **Underpin Integrative Systems Biology at Imperial**
- **Provide functional data management that:**
- Is readily accessible to biologists
- Supports data exploration and mining by the modelling community
- Is flexible – allows adaptation for ongoing changes in standards (e.g. PSI), remit (new methods, new data types)

A Distributed System for E-Support of Microarray Data Analysis and Management - EMAAS



Complications of Supporting Microarray Analyses

- **Multiplicity of software** with different interfaces, running on different platforms
- **Steep learning curve** for complex software and analyses
- Need to continually assess, integrate, maintain, support new software and tools in **fast-changing field**
- **No single common data format** - data integration & transfer often difficult
- **Data storage** – large data, often not centralised
- **Efficient support difficult and time-consuming** when remote staff don't have ready access to experimentalist's data

EMAAS

- **E**xtensible **M**icro**A**rray **A**nalysis **S**ystem
- The development of a **portal** providing:
- Simple, robust **access to up-to-date resources** for **microarray data storage** and **analysis**
- **Distributed** availability of the portal allowing access to large compute power facilities required for microarray analysis and storage
- An integrated system to optimise distance **user support** and **training** using these amenities

Project Components

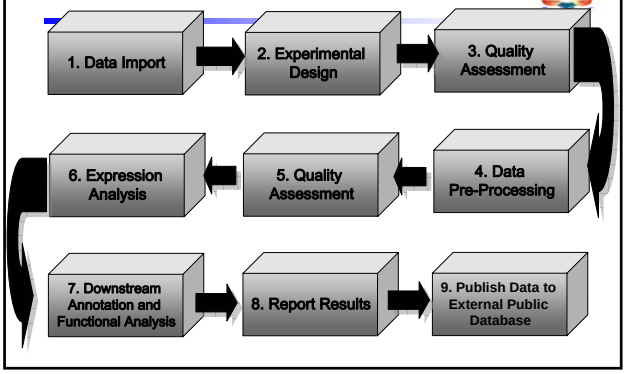
3 integrated components:

- **LeSC** – portal infrastructure, web and grid services
- **MAC** – access to data repositories and intermediate database for analysis tracking
- **BSS** – Implementation of portal interface and integration of data analysis tools

Architecture Summary

- Underlying database POSTGRES
- Portal interface developed in OpenLazlo
- Job/resource scheduling using GRIDSAM
- Main microarray QA and analyses using R/Bioconductor
- GO annotations/KEGG pathway retrievals using Taverna SCUFL workbench

Overview of a Typical Microarray Data Analysis Workflow



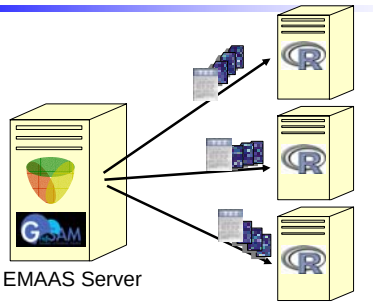
Microarray Data Mining Resource (MiMiR)

MiMiR

MicroarrayCentre

- Comprehensive solution for storage, annotation and exchange of Affymetrix microarray data
- Based on MAGE format and MIAME-supportive
- Currently being extended to capture array-associated clinical information

Analysis - GridSAM - R

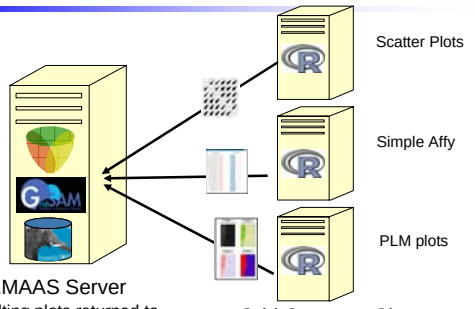


EMAAS Server

data, experimental description and the QA function name passed

Grid Compute cluster nodes

Analysis -GridSAM - R



EMAAS Server

Resulting plots returned to EMAAS server for portal display

Grid Compute Cluster

Reporting Results

- All analysis steps, methods and parameters are captured
- A .pdf report can be generated from the analysis steps to describe how the analysis was performed
- Facilitate users data export experience through direct links to MIAMExpress (The ArrayExpress web submission tool)
- Animated tutorials via OpenLaszlo Media tools
- Users who have data in MiMiR can export data automatically to Array Express

Acknowledgements

- BBSRC for funding CISBIC
- Douglas Young and Jaroslav Stark
- Other 15 CISBIC co-applicants
- Mike Sternberg
- Geraint Barton
- James Abbott, Derek Huntley