

London Bioinformatics Service

Sarah Butcher and David Colling

Outline

- Why do we want to have a bioinformatics service running over the Grid?
- What we plan to do
- Why LondonGrid?
- Who are we?
- Conclusions

Note that implementation of this service is at an early stage

A Brief History Of Genome Sequencing

1977 Fred Sanger *et al* sequence all 5,375bp phage Φ -X174 first complete sequencing of an organism

1980 ~56 DNA gene sequences in public domain, ~180 by 1983

1995 Venter *et al* first complete genome sequence of a microorganism - *Haemophilus influenzae*

1996 *Saccharomyces cerevisiae* - first complete eukaryotic genome finished

1998 *Caenorhabditis elegans* - first multicellular eukaryote genome (97Mb) is completed

1999 Chromosome 22 - first complete human chromosome is published

2000 Celera publishes the *Drosophila melanogaster* (fly) Genome

2000 Celera announces 'completion' of the human genome, public at 85% coverage

2001 Rough draft of human genome published in Nature and Science

2003 'Full' coverage of human genome – final assembly published

Why do we need a bioinformatics service over the Grid?

- As we move through the post-genomic era, the sheer size and complexity of data we are able to generate is still continuing to grow rapidly. As new leaps in technology allow us to sequence whole genomes in a matter of days or months, with a commensurate orders of magnitude drop in cost.
- The initial publically-funded human genome took over 11 years to get to first draft and cost tens of millions of dollars. Recently, 454 Technologies and Baylor College of medicine announced the completed genome of Jim Watson in 2 months at a cost of less than \$1 million.
- The drive of technological advance on data size is mirrored throughout the biological domain, with microarrays now regularly reaching 5 million data points per 'chip', and up to one hundred 'chips' to analyse per experiment
- Compounding the issue of sheer data volume, the drive towards an integrative systems biology approach requires the ability to inter-relate and mine high throughput data arising from genomics, transcriptomics, metabonomics, proteomics experiments in order to form complex models of natural processes and regulatory networks.
This means a growing requirement for even non-computing specialist labs to look to distributed computing to solve their computational problems

What we plan to do

- Within EGEE there has been some very effective work on running bioinformatics over the Grid. However, this has been done (largely) as part of specific challenges. We are trying to build a prototype service available to all bioinformatics groups within London
- A service implies a level of reliability and ease of use
 - User interfaces are very important providing straight forward and easy to use environments to non-specialist users.
 - Initially deployed across an environment that we can monitor and control.
- This will be run over the resources within LondonGrid (more about LondonGrid later).

What we plan to do

- We will (as far as possible) deploy non-commercial software to avoid licensing issues.
- While the list of software that we are planning to deploy is not yet fixed an initial list looks something like:
 - General use cross compatibility software (EMBOSS)
 - High throughput sequencing assembler for short fragments (e.g. MOSAIK)
 - Programs broadly useful for large-scale genome annotation:
 - Blast suite (for sequence database searching)
 - PSI-Blast (for protein database searching for remote homologues)
 - Interproscan (multiple algorithms for searching Interpro database – used for detection of known protein domains, functional patterns and motifs)
 - Infernal (non-protein encoding RNA detection)
 - Phylogenetics (e.g. MrBayes)

Note that not all bioinformatics software is suitable for EGEE style grids, but a lot is!

What we plan to do

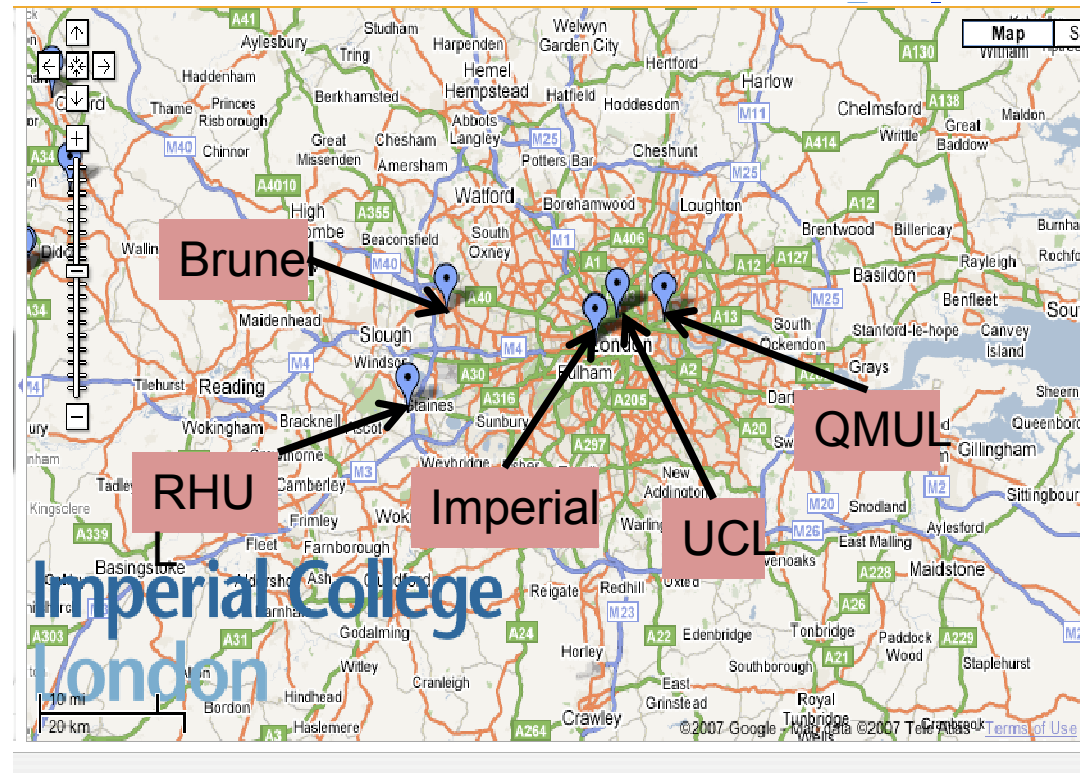
- Many of the problems that we will see have already been tackled in other Grid projects and so there is much that we can learn.

These will include:

- Reliable distribution and versioning of software
- Reliable distribution and versioning of data(bases)
- The ability to process workflows
- Tuning how a specific sort of application runs in a given environment.
- Appropriate user interfaces

So why LondonGrid?

- The five institutes (over 9 sites) that form LondonGrid have been collaborating on e-Science projects since 2001
- There is a support structure covering e-Science infrastructure across the sites with weekly technical meetings
- There is a management structure to determine usage priorities.
- The e-Science activities at the institutions are close – politically, geographically, and in network terms
- Bioinformatics communities at these (and other London sites) are already closely linked



All this makes this an ideal Grid on which to build a prototype service

Who are we?

- Currently this effort is being led by the Bioinformatics Centre and e-Science groups at Imperial College.
 - Experience in porting/running/service provision of bioinformatics applications to many platforms and general experience of grid porting and operations.
 - Leading role in both London Bioinformatics forum and LondonGrid.
- In discussions with possible industrial partners (GSK, Constellation Technologies, Equinox, etc) with who we are applying for funding.
- Expect other bioinformatics groups from within London to join this activity

Conclusions

- We are trying to build a prototype bioinformatics service within London where reliability and usability will be the key to its use
- We are doing this in London because of the need for a controlled and reliable environment when prototyping
- We will build on much of the work performed elsewhere, both within grid projects and outside of them.