Interoperable and Scalable Metabolomics Data Analysis with Microservices

Christoph Steinbeck
and the PhenoMeNal consortium

European Bioinformatics Institute, Hinxton
Friedrich-Schiller-University, Jena
Nuclear fission
Five-dimensional energy landscapes
Seafloor spreading
The view from under the Arctic ice
Career prospects
Sequence creates new opportunities

naturejobs
genomics special
Traditional Medicine  

Personalised Medicine

@genomicsedu  
www.genomicseducation.hee.nhs.uk

FRIEDRICH-SCHILLER-UNIVERSITAT JENA
Phenome - Exposome
Reaction times following external change

- **Metabolism** (Seconds)
- **Genetics** (Decades, Centuries...)
- **Epigenetics** (Days, Months, Years...)
- **Gene Expression** (Hours)
The Metabolome is an easily accessible and dynamically changing Molecular Phenotype
Phenome - Exposome
Phenome - Exposome

Big, well-annotated metabolomics data required to statistically link individual components of the exposome to effects in the molecular phenotype.
Metabolomics

Measures **occurrence** and **concentrations** of many small molecules (**metabolites**) in an organism at once.

**Metabolites:** (Endogenous) small molecules in biological organisms

Wide range of phys/chem properties
Diagnostic fluids

- Urine (time-averaged data)
- Plasma (snap-shot data)

Other accessible analytical compartments

- Specialized fluids and biopsies (selected fluids)
- Pathological fluids
- Artificial fluids

Nicholson et al., Nature, 491(7424), 384–392
Metabolomics uses a wide-range of analytical techniques

Nuclear Magnetic Resonance (NMR)

Mass Spectrometry (MS)
Typical 500 MHz $^1$H-NMR spectrum of human cerebrospinal fluid. Numbers indicate the following metabolites:

1. DSS, 2. imidazole, 3. 2-hydroxybutyric acid, 4. 2-hydroxyisovaleric acid, 5. 2-oxoisovaleric acid, 6. 3-hydroxybutyric acid, 7. 3-hydroxyisobutyric acid, 8. 3-hydroxyisovaleric acid, 9. acetic acid, 10. acetoacetic acid, 11. acetone, 12. L-glutamine, 13. pyruvic acid, 14. L-glutamic acid, 15. citric acid, 16. creatinine, 17. creatine, 18. D-glucose, 19. lactic acid, 20. myo-inositol, 21. D-fructose, 22. formic acid, 23. L-histidine, 24. L-tyrosine, 25. methanol, 26. glycerol

source: http://www.csfmetabolome.ca
Example: Untargeted Metabolomics of Cerebrospinal Fluid

Alzheimer’s Disease (AD), Mild Cognitive Impairment (MCI), Cognitive Normal (CN)

Trushina, et al. PLOS ONE 8, e63644 (2013).
Large Scale Computing with Medical Metabolomics Data

- H2020 e-infra
- 3 Years
- 13 Partners
- 8 Mio €
- 830 PM
- Kick-off 9/15
An e-infrastructure for Large Scale Computing with Medical Metabolomics Data
Microservices

- Architectural design pattern
- Independent, potentially distributed processes
- Language-independent interfaces.
- Services decoupled. Perform a small task (“Do one thing and do it well”).
- Individual service should be easy replaceable.
MS1/MS2 – XCMS-OpenMS

By P. Emami, S. Hermann, C. Ruttkies, O. Spjut, S. Neumman, K. Kultima

FRIEDRICH-SCHILLER-UNIVERSITAT JENA

PhenoMeNaL
Large-Scale Computing for Medical Metabolomics
Workflow Management with
• Galaxy
• Jupiter
• Luigi
• Pachyderm
100s of common metabolomics tools ready to run as Galaxy workflows in the cloud of your choice
Large Scale Computing with Medical Metabolomics Data

A ready-made, well-tested, best-practice, Virtual Research Environment

Literally a click of a button
Managing large sets of machines in the cloud

Container Orchestration

Infrastructure Provisioning

CRE

Containers

C. Orchestrator

Container Engine

Multi-VMs OS

Hardware/IaaS

Software Provisioning

kubernetes

TERRAFORM

docker

CRE

Google Cloud Platform

AWS

openstack

FRIEDRICH-SCHILLER-UNIVERSITAT JENA

PhenoMeNal

Large-Scale Computing for Medical Metabolomics
Generic Cloudification

From tool

Select tool

Document

Containerise

Portable

Galaxy PROJECT

Workflow Wrapping

Register in Galaxy

Interoperable

Test

Deploy

Test

Tag for release

Integration test

Reproducible

Review

Package

User test

To release

Full audit trail: Git version control, build, and test reports
Economies of Scale
Economies of Scale

- Assume 24,000 samples. Each takes 1 minute to process. That is 400 CPU core hours, or 16 days, or **2 weeks**.
Economies of Scale

- Assume 24,000 samples. Each takes 1 minute to process. That is 400 CPU core hours, or 16 days, or 2 weeks.
- At $0.04 per core hour, plus 1TB storage, that’s $16 in the cloud.
Economies of Scale

- Assume 24,000 samples. Each takes 1 minute to process. That is 400 CPU core hours, or 16 days, or 2 weeks.
- At $0.04 per core hour, plus 1TB storage, that’s $16 in the cloud.
- It is also $16 if you rent 400 cloud servers and you’d be done in 1 hour.
Assume 24,000 samples. Each takes 1 minute to process. That is 400 CPU core hours, or 16 days, or 2 weeks.

At $0.04 per core hour, plus 1TB storage, that’s $16 in the cloud.

It is also $16 if you rent 400 cloud servers and you’d be done in 1 hour.

It is also $16 if you rent 24,000 cloud servers, and be done in 1 minute (ignoring overhead …)
Economies of Scale

- Assume 24,000 samples. Each takes 1 minute to process. That is 400 CPU core hours, or 16 days, or 2 weeks.
- At $0.04 per core hour, plus 1TB storage, that’s $16 in the cloud.
- It is also $16 if you rent 400 cloud servers and you’d be done in 1 hour.
- It is also $16 if you rent 24,000 cloud servers, and be done in 1 minute (ignoring overhead …)
Testing scalability

Large data set: 2000 NMR spectra

Processing tool: BATMAN

Local cloud: MedBio 50 vCPUs

Medium cloud: Microsoft Azure 300 vCPUs

Large scale cloud: EMBASSY 1000 vCPUs

Desk top system 8 cores

Measure runtime, speedup, efficiency...

Large data set: 2000 NMR spectra

Processing tool: BATMAN

Local cloud: MedBio 50 vCPUs

Medium cloud: Microsoft Azure 300 vCPUs

Large scale cloud: EMBASSY 1000 vCPUs

Measure runtime, speedup, efficiency...
PhenoMeNal scalability – BATMAN NMR processing

- 4 systems:
  - Desktop (8 cores)
  - MedBio (50 vCPUs)
  - Azure (300 vCPUs)
  - EMBASSY (1000 vCPUs)

- Considerable speed up possible
  - Running time for 2000 spectra down from 3 days (1 core) to 10 mins (1000 vCPUs)
Summary

• Large analytical chemistry data produced in medical metabolomics

• Computational analysis can take days or weeks on a single node

• **Industry standard tools** such as Kubernetes, Terraform and Ansible allow us to
  
  • seamlessly **deploy analysis pipelines**
  
  • composed of **microservices**
  
  • on **1000s of cloud machines**
  
  • at **low costs** and **no upfront investment**

• The free and open PhenoMeNal infrastructure encapsulates those **orchestration tools** together with **hundreds of tools in computational metabolomics**
Funding and Collaborators

European Commission
Grant # 654241

PhenoMeNal Consortium

Netherlands Metabolomics Centre

IMPERIAL COLLEGE
LONDON

EMBL-EBI

ISAsoftwaresuite

MRC Human Nutrition Research

UNIVERSITY OF CAMBRIDGE