Data Exploitation in the CLF OCTOPUS Facility

Dan Rolfe

Central Laser Facility, Research Complex at Harwell
Overview

- OCTOPUS
- Challenges
- Vision
- Approaches
- Examples
- Conclusion
• National imaging facility with peer-reviewed, funded access
• Located in Research Complex at Harwell
• Cluster of microscopes and lasers and expert end-to-end multidisciplinary support
• Operations and some development funded by STFC
• Key developments funded through external grant – BBSRC, MRC
Challenge

• Under-exploitation of potential to solve grand biomedical challenges
  – Rapidly improving measurement methods giving unprecedented detail
  – Studies involve combining multiple techniques
  – Huge computational and data challenges are the bottleneck

• Moving towards multi-facility studies
  – Diamond, ISIS, OCTOPUS, SCD

• OCTOPUS as an example and our focus
  – Large and growing variety of imaging and analysis modalities and tools
  – Increasing quantities and varieties of studies/samples
  – Computation increasingly critical but challenging/alien to most users
Challenges

- Challenging image analysis – often very manual
- User expertise in numerical/computational work
- Every study is different
- Developing/applying new algorithms all the time – agility
- Metadata – often limits ability to exploit large studies
- Complex combination of sample, instrumental and analyses techniques, platforms, licensing models...
- Numbers/varieties of datasets, not just size
- Different computing architectures
- Delivering this in a way effective for user and facility
- Integration across facility
Example

BBRSC LoLa project looking at EGFR cell signalling in cancer
- Driven OCTOPUS single molecule developments
- User in plant cell imaging now catching up in scale of challenge

1 part of a PhD project
- 1 experimental technique
- 50 experimental conditions
- 30 datasets for each condition
- 1000 single molecule tracks for each condition
- Multiple properties & events of interest in each track
- Comparison of just one property…
Large scale comparisons
Vision

Multi-technique bioimaging solutions centre of excellence
*with data exploitation to match advanced instrumentation and lead new developments*

• Flexible vertical and horizontal integration of data and computation in OCTOPUS
  – From start to end of studies, across all methods in study
  – Robust, quantitative, automated approach to data analysis, visualisation and management
  – Scalable, for complex, large studies
  – For expert and non-expert users
  – Secure and convenient access to capabilities
  – *Enable step change in capability through development and integration of varied data and algorithms – make current tour-de-force experiments routine*
Approaches

- Multidisciplinary collaboration
- Approaches from other field (astronomy algorithm, PX automation)
- Robust, reproducible and quantitative
- Bayesian analysis – well defined questions, explicit assumptions, errors – objective basis for conclusions
- Remove manual tuning parameters
- Loose coupling to maximise chance of integration and scaling
Multidimensional single molecule tracking

- Automated registration & tracking in multiple channels
  - Computer vision
  - Bayesian feature detection from astronomical galaxy detection

- Instrumental metadata from acquisition
  - Flexible specification of many instrument configurations

Core suite

- Advanced GUI for browsing, selection, visualisation, post-analysis
  - Track scores and heuristics to sort and filter
    - “Is this possible?”
    - “Is this reliable?”
  - Tools for selecting subset for pooling and further analysis
  - One dataset at a time
Post analysis

• Some batch jobs, some from GUI, some both
• Various algorithms & quantities
  – Colocalisation & interactions
  – MSD & diffusion
  – …
• Combine and pool multiple datasets
• Compare quantities between pooled groups of datasets
  – Sample metadata a limiting factor
FLImP

- Needham et al, PLOS one, 2013
  - Exploit change in image of diffraction-limited spot on photobleaching to measure sub-diffraction separations
  - Multidimensional fit with bootstrap error estimate for per-spot confidence intervals
  - Count and measure multiple distances in samples
  - Bayesian analysis to interpret combination of many results
    - Fingerprint of molecular interactions in cancer
• Limiting factor – selecting suitable tracks
  • Manual process
  • Need many tracks to build up distribution
  • Takes experienced user weeks to analyse days of data
  • Applying machine learning algorithms to encapsulate expert user knowledge
  • Autoencoder for non-linear dimensionality reduction of data
  • Neural network for non-linear classification
  • Factor \( \sim 3 \) speedup so far
BIGGLES

- Globally optimal Bayesian tracking solution
- Determine probability of all possible tracking solutions
- Honestly reflect ambiguities in result
- No magic tunable parameters – all determined objectively – automation
- Gibbs sampler + Metropolis Hastings
Data and computation management

- Goal – end-to-end integration of OCTOPUS
  - Enable scalable, convenient deployment of heterogeneous software on heterogeneous hardware
  - Hardware
    - STFC/CLF MPI and GPU clusters, OCTOPUS analysis nodes
  - Convenient to users, support staff and developers
  - Integration with STFC user access systems
  - Rapid integration and deployment of new capabilities
  - Integration of
    - in-house developments for SM tracking
    - External tools e.g. OMERO for modalities it supports
Challenges

- Sufficient, reliable metadata – data acquisition & user...
- Wide variety of methods and software
  - Closed/open software – ability to tailor/modify
  - Proprietary software – licensing on scalable/multiuser systems
  - Different assumptions about data organisation
- Different OSs – deploying, integrating, porting
- Speed/reliability of remote desktops
ICAT Job Portal

- Robust, secure and flexible data and job management
- Configurable dataset and job types (options, metadata)
- Loosely coupled
  - Supports deployment of existing software without rewriting
  - Wrapper scripts reconstruct expected filesystem structure, data and parameters for application
- Data and computation colocated on-site
- Batch and interactive jobs

Fisher, Phipps & Rolfe, IWSG, 2013
ICAT Job Portal

- On- and off-site access via web-portal and integrated remote desktop
- Underlying data storage ICAT
  - Common with Diamond and ISIS
- Linux hosting, access from any
- Store raw and processed data
- Complementary richer metadata database

• Status
- WIP – developed with limited resources until recently
- Aim for staged deployment starting 2016
Conclusion

• This was just “one” OCTOPUS imaging modality!
  – Major task – we’ve made a start…

• Ada Lovelace Centre
  – Similar challenges and visions across facilities

• OCTOPUS/IJP work in progress
  – Would form ideal proof-of-principle for a wider campus vision of integration

• Beneficiaries of solving problem
  – OCTOPUS user community
  – STFC user community
  – Other imaging and .. facilities
  – UK biomedical research
Acknowledgements

- CLF
  - Marisa Martin-Fernandez
  - Michael Hirsch
  - Jianguo Rao
  - Teodor Boyadzhiev
  - Chris Duncan
  - Sarah Needham
  - Laura Zanetti-Domingues
  - Selene Roberts
  - Chris Tynan
  - Stephen Webb
  - Benji Coles
  - Dave Clarke
  - Rest of OCTOPUS … in the future…

- SCD
  - Steve Fisher
  - Kevin Phipps
  - Tom Griffin
  - Brian Ritchie
  - Rebecca Fair
  - Brian Matthews
  - Martyn Winn
  - Cheney Ketley
  - Jens Jensen
  - Derek Ross

- King’s College London
  - Peter Parker, Simon Ameer-Begg

- RCaH IT
- Users
- Funding
  - BBSRC
  - MRC
  - STFC