Laurent Gatto
Department of Biochemistry, University of Cambridge
Tennis Court Road, Cambridge, CB2 1GA, UK
Email: lg390@cam.ac.uk
Tel: +44 (0) 1223 760253

After a PhD in Evolutionary Genetics at the Free University of Brussels, and complementary studies in Computer Science at the University of Namur, Belgium, I worked for 3 years in industry with a focus on computational pipelines R&D. In 2010, I moved to the University of Cambridge, where I concentrated on data and software infrastructure for proteomics, first as a Post-doctoral Research Associate and later as a Senior Research Associate. In 2014, I was awarded an Sustainability Institute fellowship and recently have become a Software Carpentry instructor.

Since the completion of my PhD, the emphasis of my work has focused on delivering software solutions to tackle specific data-driven scientific questions, that have culminated in several state-of-the-art software, data and educational packages. The extend to which I can deliver research software is constraint by the limited opportunities for research software engineering in academia. The fellowship would be jointly hosted by the Department of Computer Science and the Research IT Services at the UCL. It will give me a unique opportunity to pursue my inter-disciplinary career focusing on research software engineering.

Project summary Modern science relies on ever increasing quantities of data (some have coined it a 'data deluge') and on trustworthy and sustainable software to analyse and interpret them. Each of these pillars of modern scientific activities are generally developed and maintained independently. On one hand, a typical data pipeline starts with the submission of annotated data to an official repository that, after some basic curation, serves them to the wider community. This flow is unidirectional and, once released, becomes inert (updates and further annotations are very rare) and, most often, dormant. On the other hand, the exposure of research software is often limited by the application (or applicability) to a reduced number of data sets. Surprisingly, while each data/software information and development streams can't be conceived in isolation, and their true potential emerges from their interoperation, little efforts exist to facilitate streamlined interactions and discoverability and enable new data/software meta-analyses.

A recent project involved developing, implementing and applying a transfer learning algorithm that learns from different data sources to classify features. In our specific question, the classification is aimed at inferring the spatial sub-cellular localisation of proteins using experimental data and third-party gene ontology annotation. Our algorithm was initially inspired from a leaves image classification problem and was improved and adapted for our use cases. We would like to apply it to another domain, genomics data being an obvious candidate; however, manual exploration and reliance on personal contacts are not effective and substantially slow cross-fertilisation between domains. The transition from the first publication of the algorithm in the International Conference on Machine Learning to a functional implementation and application in spatial proteomics initially relied on a fortuitous discovery. What further transitions and improvements could be anticipated? If these are prohibitively slow or difficult, the dissemination, improvement and generalisation of the software is dramatically limited. The project's goal is to offer systematic ways to enable such transitions.

Concretely, the first major milestone of the project is a set of interconnected databases of experimental data, software and publications, all spanning different disciplines, created from mining relevant repositories. The relations between records within and between these databases will be browseable and searchable through an online portal and programmatically accessible through specific APIs. Given specific properties of existing records, or new, user-defined properties describing software, data or publications, new matches can then be identified. Properties related to quality and sustainability of software and data (metadata, documentation, testing, . . .) will be used to favour the promotion of more sustainable code. Known cross-disciplinary data/software outputs will be used to optimise and assess the relevance of the infrastructure. The second major milestone aims at facilitating the application of the relations identified above by enabling transparent integration of different data source (local or remote, public or private), data sets (transparent to file formats) through data APIs and software architectures (local or remote) through software APIs/pipelines. Milestone one enables systematic exploration of RSE outputs and milestone two supports their automated and standardised meta-analysis. Each of these deliverables will be complemented by dedicated open communication channels to openly publicise and disseminate the new software/data relations and use cases.

Laurent Gatto, PhD

Senior Research Associate
Computational Proteomics Unit
Cambridge Systems Biology Centre
Department of Biochemistry
University of Cambridge

Tel: +44 (0)7747 802315
lg390@cam.ac.uk
http://cpu.sysbiol.cam.ac.uk
https://github.com/lgatto

Appointments

Senior Research Associate University of Cambridge
Cambridge, UK **October 2013 –**
Group Leader – Computational Proteomics Unit

Visiting Scientist EMBL/EBI
Cambridge, UK **January 2010 –**
Member of the Proteomics identifications database (PRIDE) team.

Software Sustainability Institute Fellow 2014.
Software Carpentry Instructor.
Bioconductor project, associated member.

Previous appointments

Post-Doctoral Research Associate Jan 2010 – Sept 2013
University of Cambridge, Cambridge, UK

Bioinformatician, Project Leader Aug 2006 – Dec 2009
DNAVision, Gosselies, Belgium

Education

2000 – 2006 (viva 2006-07-27)
PhD in Science (Molecular Biology Department)
Free University of Brussels (ULB), Belgium

2000 – 2001
Master of Advanced Studies (DEA) in Sciences
Free University of Brussels (ULB), Belgium

1997 – 2000
Masters in Biology (highest honors)
Free University of Brussels (ULB), Belgium

Selected Publications (Out of 26)

Huber W. *et al.*, *Orchestrating high-throughput genomic analysis with Bioconductor*, **Nat Methods** 2015 Jan 29;12(2):115-21.

Gatto L., Breckels LM, Burger T, Wieczorek S and Lilley KS *Mass-spectrometry based spatial proteomics data analysis using* pRoloc, **Bioinformatics**, 2014 May 1;30(9):1322-4.

Gatto L., Christoforou A. *Using R and Bioconductor for proteomics data analysis.* **Biochim Biophys Acta** 2014 Jan;1844(1 Pt A):42-51.

Chambers M. *et al.* *A Cross-platform Toolkit for Mass Spectrometry and Proteomics*, **Nature Biotechnology**, 30, 918 – 920, 2012.

Gatto L. and Lilley K.S. *MSnbase – an R/Bioconductor package for isobaric tagged mass spectrometry data visualisation, processing and quantitation*, Bioinformatics, 28(2), 288-289, 2012.

Selected Software Development

Manipulating and exploring protein and proteomics data, 2014.

pRolocGUI: Interactive visualisation of organelle (spatial) proteomics data, 2014.

rpx: An R interface to the ProteomeXchange repository, 2014.

pRoloc: A unifying bioinformatics framework for organelle proteomics, 2012.

hpar: A simple interface to and data from the Human Protein Atlas project, 2012.

rols: An R interface to the Ontology Lookup Service, 2012.

MSnbase: Base Functions and Classes for MS-based Proteomics, 2011.

mzR: Raw mass-spectrometry data parsig library.

Selected oral communications (invited talks)

Gatto L. *Spatial proteomics: Combining experimental and annotation data to predict protein sub-cellular localisation.* 13 Jan 2015, European Bioconductor Developer meeting, EMBL, Heidelberg.

Gatto L. *An overview of the (growing) R/Bioc ecosystem for mass spectrometry and proteomics*, MRC Clinical Sciences Centre, Imperial College London, 5 December 2014, London.

Gatto L. *Computational Challenges in Mass Spectrometry-Based Spatial Proteomics*, HUPO meeting, Computational Mass Spectrometry Initiative, 8 Oct 2014, Madrid.

Current Grants

May 2014 – Oct. 2015 BBSRC Tools and Resources Development Fund (£144,112). *Automated identification of optimal data-specific organelle clusters using freely available protein annotations.* **PI Laurent Gatto**, University of Cambridge.

March 2014 – February 2018 BBSRC Strategic Longer and Larger grant. *A spatio-temporal map of the developmental fly interactome.* **Research co-investigator**, with PI Professor Simon Hubbard, University of Manchester.

March 2013 – November 2015 Addenbrookes Charitable Trust. *Protein biomarkers for vascular calcification in end-stage renal disease.* **Co-applicant**, with PI Dr Thomas Hiemstra, University of Cambridge.

Teaching

I am regularly teaching at national and international workshops and University courses (i.e. MPhil in Computation Biology, Cambridge in 2012-2014) on computational biology, scientific programming and Software Carpentry Bootcamps. These teaching activities are on a voluntary basis.

Reviewing activities

Journals: Proteomics, Bioinformatics, BMC Genomics, BMC Research Notes, Expert Reviews in Proteomics, Journal of Proteomics, F1000Research, PeerJ, Journal of Open Research Software.

Conferences: ISMB/ECCB. **Funding agencies**: Medical Research Council.

Host departments

Prof. David Jones, Computer Science and Dr. James Hetherington, RITS.
Also support from Prof. Christine Orengo, Structural and Molecular Biology Department.

Outline of the aims and activities of the fellowship

There are a wide range of excellent outputs related to RSE at large: software, algorithms and data sets of great scientific value. These are however too often produced, developed and exploited in isolation, limiting their dissemination and full potential across disciplines. The aim of the project is to enable individual experimental, computational scientists and RSEs to broaden up the relevance of their research outputs to different fields of applications. The project centres around the following activities: (1) data collection of relevant resources across scientific disciplines; (2) data mining and modelling to effectively inform users of most relevant interlinked data, software and publication records; (3) efforts in standardisation and automation, to facilitate more rigorous and systematic meta-analysis; and (4) the project will foster scientific and RSE community involvement to assure that most relevant resources and results are served for the communities.

Justification of the choice of host department and fit to UCL's wider initiatives in RSE

The UCL is at the forefront of research software engineering, recognising the unique set of skills and experience stemming from computational sciences and software engineer. The Research IT Services and the Department of Computer Science constitute the ideal environment for such as applied projects, that aims at making best use of existing data sources and infrastructure skills to produce a usable and useful service to the global RSE and research communities. UCL's efforts in new technologies and analytics as applied to scholarly content, data and software in the frame of the Big Data Institute are particularly relevant too. This combination of departments at the UCL is the ideal environment for my professional development to reinforce my experience is research software engineering and improve my skills in large scale infrastructure serving the science community.

Importance of fellowship in terms of science and engineering research enabled, including contributions to EPSRC's priorities

The EPSRC recognises the importance of software and data sets as primary research outputs in their own rights. It has invested millions of pounds per year in software over the last 5 years to promote RSE activities. The project offers a way to maximise these investments by promoting discoverability, re-use and new application of existing software.

The project enables and promotes collaboration by identifying relevant and reusable data and software, both for software/data users and producers, in particular with respect to more effective dissemination and collaboration across communities. Software and data re-use, and in particular across disciplines, is a major hallmark of success and impact. The action of the project lies at the interface of experimental and computational research outputs and can support researchers to cross these fields and thus promote opportunities and innovation in software provision and maximise the academic and economic impact of RSE outputs. By means of software re-use promotion and discovery, the project will promote open dissemination of sustainable and trustworthy software and facilitate the development of communities of users and developers across disciplines.