Facilitating synthetic biology literature mining and searching for the plant community

Robert Davey
Email Address: robert.davey@tgac.ac.uk

The Idea

We propose to set up and run a hackathon at TGAC to extend the ContentMine literature mining system in order to allow plant-focused synthetic biology papers to be mined for facts and subsequently searched through the Grassroots Genomics web portal.

Who We Are

Robert Davey (robert.davey@tgac.ac.uk) – Robert joined TGAC in February 2010 as the lead software engineer on the MISO LIMS project, which was released as an open source framework in June 2012. He went on to become the Core Bioinformatics Project Leader, and was then appointed as Data Infrastructure and Algorithms (DIA) Group Leader in late 2012. The DIA science faculty group is involved in researching innovative data-driven informatics, developing aspects of underlying analysis and management infrastructure and processes (Galaxy, iPlant UK), as well as research projects to investigate new infrastructures to facilitate the sharing and publication of data and related metadata (Collaborative Open Plant Omics, Field Pathogenomics, Grassroots Genomics). Robert's main interests are in enterprise-grade software development, data management and associated HPC infrastructure, metadata and ontologies, sequence analysis and quality control pipelines, novel visualisation strategies for sequencing and biological data, and the open source ethos.

Ksenia Krasileva (ksenia.krasileva@tgac.ac.uk) – Ksenia is a Group Leader with a joint appointment at The Genome Analysis Centre and The Sainsbury Laboratory. Ksenia joined Norwich Research park in December 2014 moving from University of California Davis where she held Fellowship from National Institute of Food and Agriculture (NIFA) to develop functional genomic tools for wheat working with Jorge Dubcovsky. Ksenia's group continues to advance functional genomic toolbox for wheat and related species and to apply this tools to study plant innate immunity. The projects in the group span comparative genomic analyses, working with wheat TILLING populations and CRISPR technologies. Ksenia's main research goal is integration of new technologies for rapid pathway dissection and engineering favourable traits, such as disease resistance. An important philosophy of the group is to conduct science in a way that bridges the gap between bioinformaticians, biologists and breeders and to provide these communities with communication streams, such as the Grassroots Genomics platform, for integrative data evaluation and data analyses.

Nicola Patron (nicola.patron@tsl.ac.uk) – Nicola is a molecular and synthetic biologist at The Sainsbury Laboratory (TSL), a world-leading research institute working on the science of plant-microbe interactions. At TSL Nicola designs and develops molecular tools to engineer plant genomes and produce novel functions in plant cells and supervises the Norwich Research Park's International Genetically Engineered Machine (iGEM) team. As a bioengineer, Nicola is also interested the societal impacts of synthetic biology and the

complex intellectual property issues that surround genetic sequences, DNA and natural products. Nicola is a current fellow of SynBioLEAP, a program that aims to catalyse a community of emerging leaders in synthetic biology to create bold new visions and strategies for developing biotechnology in the public interest.

Richard Smith-Unna (rds45@cam.ac.uk) – PhD student, Plant Sciences Cambridge, already with several papers on plant genomics. RSU is very adept with modern software tools and design and has built genome assembly systems for plants. He has also worked for over a year with ContentMine in developing a declarative system of scrapers (quickscrape) which retrieves all the published material for a paper (text, images, captions, abstract, supplemental data, CSV files, etc.). Quickscrape is easy to use, and can be extended to new journals and publishers (last weekend a chemist wrote 4 scrapers in an afternoon). Very recently he has developed getpapers which lets researchers submit general queries to a range of repositories (especially EuropePMC, CORE (UK repositories)) and retrieve the URLs, which are then passed to quickscrape.

Peter Murray-Rust (pm286@cam.ac.uk) – is a (retired but highly active) chemist in Cambridge University. He is on the editorial board of several open journals (e.g. BMC J. Cheminformatics), and has been a member of the Project Advisory Board of EuropePMC for nearly 10 years. He is also a highly active member of the Advisory Board of the Open Knowledge Foundation, creating initiatives such as Open Access lists, and the Panton Declaration for Open Scientific Data, followed by Panton Fellowships. He has fought for legal reform of Copyright in the UK especially for Content Mining (formulating the slogan “The Right to Read is the Right to Mine”) and supported the Hargreaves legislation in 2014. In 2014 he was awarded a Shuttleworth Fellowship to “change the world” through setting up ContentMine. ContentMine’s intention is to liberate all scientific facts from the literature which is now legal in the UK.

Implementation

We propose to fund an open hackathon for technologists and biologists to come together and produce concrete digital outputs that facilitate the indexing and searching of synthetic biology texts. Utilising the open technologies of the Grassroots Genomics (**http://www.tgac.ac.uk/grassroots-genomics**) project at TGAC and the ContentMine (**http://contentmine.org/**) platform from the University of Cambridge, the hackathon will mutually improve both platforms to enable better access to research literature in plant synthetic biology.

Both ContentMine and the Grassroots Genomics portal use a search infrastructure based on Lucene (**https://lucene.apache.org/**), a text-based search engine, to store data taken from academic papers and supplementary data. Currently they are both using the default implementation. However, all of the indexing and scoring functionality used within Lucene for its searches is open to customisation so can be tailored specifically for plant-focused and synthetic biology terms and journals. The sections of academic papers where terms appear, *e.g.* abstract, materials and methods, and so on, can be taken into account when determining the weighting of particular terms, and thus the priority of the results returned to the user. Furthermore, custom parsers can be written for common types of supplemental

data, for example experimental results and log files,to extract metadata that would otherwise be unavailable to be searched. We will first extend ContentMine's ability to scrape and search relevant and important synthetic biology literature resources, and subsequently extend the Grassroots Genomics infrastructure to take advantage of this functionality.

We will use the funds to pay for attendees' travel and accommodation where possible, giving preference to early career researchers. We are also committed to promoting gender and ethnicity balance in our events and will ensure a productive and respectful environment for all participants. TGAC has excellent training facilities and interconnectivity to the Internet that would be able to be used for free. Davey and Smith-Unna will organise and run the hackathon, and due to their strong developer backgrounds, can assist with and promote all software outputs. Software engineers in the Davey group at TGAC will donate their time as helpers and fellow hackers. All outputs will be openly created, updated and shared in real-time using the GitHub collaborative software development platform, and will be integrated into the ContentMine and Grassroots Genomics platforms during the hackathon.

Benefits and outcomes

Our proposal will establish a new connection between TGAC and the University of Cambridge, enabling technological advances to be made that will benefit the community not only on the Norwich Research Park and at the University of Cambridge, but also to all users of the ContentMine platform. Through the synthetic biology expertise of our partners in The Sainsbury Laboratory, we will be able to tailor both the ContentMine and Grassroots Genomics infrastructures to allow plant researchers to access a larger corpus of open knowledge more quickly and easily than ever before. We will take advantage of TGAC's extensive National Capability in high-performance computing hardware to uplift the technological aspects of the proposal, supporting the hackathon to produce usable and fast outputs in a very short time.

Previous mapping exercises of the synthetic biology landscape include mining data from scientific literature in order to identify trends, networks, funders and leaders to provide information to agencies, policy makers and individuals (https://dx.doi.org/10.1371/journal.pone.0034368). The Synthetic Biology Project at the Woodrow Wilson Centre began mapping the emerging field of synthetic biology in 2009. They produced maps that showed cities and world regions (http://www.synbioproject.org/inventories/maps-inventory/map-analysis-2013) in which synthetic biology was experiencing growth. Their maps can be used to examine the locations of companies, universities, research institutions, government and military laboratories and policy centers that are active in this field. They also sought to investigate the number of products that reached the market that used synthetic biology technologies or contain ingredients that were the product of synthetic biology (http://www.synbioproject.org/cpi). They mined media to produce a report that looked at press coverage of synthetic biology and examined aspects of synthetic biology that may be cause for either potential public acceptance or rejection of the technology.

Hackathons are a very effective way of incentivising and formalising the quick development of novel functionality. Previous hackathons in the science space, such as Open Data Day (http://opendataday.org/), the Bioinformatics Open Source Conference CodeFest (http://www.open-bio.org/wiki/Codefest_2015), and the Collaborations Workshop HackDay (http://www.software.ac.uk/cw15) have produced tangible and beneficial outputs within a short hackathon timeframe. In the same way, we will produce the following outputs:
– Scrapers and stylesheets for selected journals with a synthetic biology, or associated, audience, e.g.
ACS Synthetic Biology
PLOS
BioEngineering Bugs
Nature
Nature Methods
Nature Biotechnology
Nature Plants
New Phytologist
Plant Cell
Plant Physiology
The Plant Journal
Plant Methods
– Scrapers for iGEM wiki pages that have use plant chassis or parts from plants
– List of keywords and filtering methodologies for plant synthetic biology to enable effective and targeted searching of the resources listed above

ContentMine has run several workshop/hackathons and now the software can be used productively within minutes of starting the hack. Previous events include Open Science Brasil, New Delhi, OpenCon 2014, EBI, Wellcome Trust (4 times oversubscribed), Cochrane Collaboration (clinical trials), Edinburgh/ARRIVE (animal testing), and Cambridge Chemistry. We expect that some of the attendees will become expert enough to run future events using the software.

We will involve a group of potential platform users from The Sainsbury Laboratory, The Genome Analysis Centre and The John Innes Centre to evaluate the outputs from the hackathon and rate their value to the community of plant researchers that are already applying, or looking to adapt, synthetic biology methodologies in their research programmes.

Budget

The exact number will depend on applications to the hackathon, but we aim to fully fund 6-8 MSc/PhD students/post-doctoral researchers to attend the hackathon (costed at approximately £500 per head), with their travel and accommodation costs being covered by the grant. ContentMine leaders' (Murray-Rust, Smith-Unna and 1 other) expenses (travel to TGAC and accommodation) will be covered by the grant. TGAC staffing costs to help run and organise the hackathon will be absorbed by TGAC.

Coffee, lunches, and an evening dinner will be provided for the attendees by TGAC as a sponsorship in kind.

We have dedicated access to TGAC's excellent new training facilities and will be supported by the 361 Division (Training and Outreach) staff in order to assist with the hackathon logistics (technical details, travel arrangements, lunches, and so on). This will greatly reduce the overhead cost of the hackathon so we will be able to support a larger number of attendees.