Drs Tim Griffin and Pratik Jagtap along with the Galaxy-P team from the University of Minnesota are working to develop workflows on an open source platform for the analysis of multi-omic data. They are currently focusing on using a Galaxy-based framework to investigate the integration of genomic datasets with mass spectrometry-based ‘omics’ data. But in the long term, they aim to expand the platform to cope with many other ‘Big Data’ domains.

Currently, a major limitation to what we can discover from complex datasets derived from next-generation technologies is our ability to analyse them. This is where the work of Dr Tim Griffin, Dr Pratik Jagtap and their research team will play an important role.

THE ‘BIG DATA’ ERA

Moore’s Law predicts that computing power will double approximately every two years, and with this, the cost of high-powered machines will also decrease. However, this cannot continue indefinitely and 2017 may be the crunch point at which physical limitations intervene, with the rate of progress becoming ever more saturated.

But what influence has this increase in computer power had on science? One of the major advances has been the ability to generate data using next generation, high throughput techniques, resulting in ‘Big Data’. Although ‘Big Data’ has been used to define many datasets, the term often corresponds to what are now commonly known as ‘omics’ datasets – genomics, metabolomics, proteomics, transcriptomics and epigenomics to name but a few. For example, in biomedical science, we see large scale, system-wide approaches being used more and more commonly. These include the 1000 Genomes Project, the emergence of personalised medicine – tailored to an individual’s needs – and systems biology, examining multiple, interacting pathways concurrently as one giant network.

However, the analysis of these large and complex datasets requires an analytical platform which can cope with the intense informatics requirements, as well as the ability to access disparate software from different ‘omics’ domains. Many wet-bench researchers will not have access to this level of compute-power or expertise locally, and therefore there is an increase in remote, or cloud, open-access platforms being used to access the necessary bioinformatic tools needed to cope with the complex results that researchers are obtaining.

ONE SOLUTION FOR ALL

At the University of Minnesota, Drs Tim Griffin, Pratik Jagtap and team are working on solutions to analyse these complicated datasets. This is a multi-disciplinary, collaborative project between Dr Griffin’s lab and the Minnesota Supercomputing Institute, which involves software developers, data scientists and wet-bench biological researchers. Specifically, the team are focusing on mass spectrometry (MS)-based ‘omics’ data (metabolomics and proteomics) and how they can harness an existing open-source framework, called Galaxy.

Put simply, mass spectrometry represents a high throughput technique that sorts ions based on their mass to charge ratio. Once certain signatures have been recorded for individual ions, this information can, for example, be extrapolated to identify peptides, the building blocks of proteins. Tandem mass spectrometry (MS/MS) further expands on this by using at least two stages of mass analysis.

GALACTIC PLATFORM

Galaxy was originally developed over a decade ago to solve problems in genomic}

One of the major advancements is the ability to generate data using next generation, high throughput techniques, resulting in ‘Big Data’
The analysis can also be extended to sequence variants, which may previously transcriptomic data translated in silico that result from sequence variations at expression of protein sequence variants data. From here, they aim to verify the (Galaxy-P) team investigates ways in which sharing and reproducibility. The latter is across the globe. Supported by a team of

Existing studies have already used the An example of where Galaxy-P (galaxyp. samples.

metaproteomics gateway). This platform provide the necessary tools for future discovery.

If your research were awarded a considerable amount of money and granted access to the world’s most powerful computer – which informatics tool would you develop? A tool that integrates outputs from all ‘omics platforms and provides a ‘Google earth’ like interactive visual data. Such a tool would be extremely useful to a biological researcher in both providing an overview of datasets and identifying potential interpretation while providing opportunities to dive into regions of interest for validation and actionable interventions/ follow-ups. We continue to be amazed and fascinated by the depth of analyses that the Galaxy platform offers in challenging fields of research. Another avenue might be to use such a powerful compute platform to re-analyse existing publically available proteomic and transcriptomic datasets using newer multi-omics tools, and develop tools to mine for new discoveries.

What was the biggest challenge you had to overcome when developing Galaxy-P? The development of tools and workflows for the goal of analysis of mass spectrometry data provided challenges at many levels. Be it at the conceptualisation stage, or at grant seeking stage, or at tool selection or workflow stage, we looked at all the challenges as opportunities. Deciding which of the many effective software tools to implement in Galaxy has been a challenge, as well as understanding the many different ‘omics sub-fields and how different software tools work, as well as which are at the forefront in terms of functionalities. However, the biggest challenge and priority in efforts has been to maintain the relevance of workflows in a constantly emerging environment where the inputs are diverse and outputs differ and newer interpretations.

What is the most niche/unexpected dataset that you’ve been asked to analyse? The breadth of biological research and flexibility of the Galaxy-P workflows has exposed us to many interesting datasets. These range from human salivary datasets for metaproteomics and proteogenomics, to dental plaque metaproteomes in presence of sugar to the study of metaproteomes from the North Pacific Oceans. But the most unexpected dataset has been the study of cardiacmuscular protein expression in hibernation of ground squirrels. Human hearts lose the ability to function at temperatures of 20°C and below. The study tried to shed light on how the heart of hibernating animals can withstand these low temperatures. We are certain that we will continue to see more of these interesting datasets as we continue our research work.

In the future, do you see Galaxy-P becoming a desk-based tool that can easily and universally be used by anyone, anywhere in the world? The research community has been using Galaxy platform for genomics studies for quite some time now and there is a stable ecosystem of developers and users, which makes this sustainable. We have seen a gradual increase in interest in using Galaxy-P amongst researchers as we have promoted it via research publications, workshops and presentations worldwide. Along with the Galaxy community of developers and researchers, we have been working on making the workflows available via downloadable tool containers or by making public instances available so that researchers can access pre-installed tools and workflows for the research areas of their interest. The vision for the future is that researchers will access these software tools remotely, where they are housed on powerful cloud-based hardware.

Leading on from this, do you think that younger students and early career researchers should be given compulsory bioinformatic training as part of their studies? Absolutely! Bioinformatics has become a necessary research skillset for experimental researchers. Programming skills enable young researchers to perform novel analyses of previously acquired data. For users, analytical and data interpretation skills expand their ability to seek newer avenues in their research fields. We strongly believe that bioinformatics training will help in introducing and honing skills in programming and data processing, and helps in continuing to expand the breadth and depth of questions that can be sought by the future generation of scientists. ‘Big Data’ will only continue to be generated in biological research, and having the ability to speak both languages in terms of biology and computational science will be a critical skill, and one that is very much in demand in years to come.