Designing challenging 'dry' bioformatics projects: exploiting public databases of genetic and post-genomic plant science data

Carol Wagstaff, Department of Food Biosciences, University of Reading, Reading RG6 6AP
E-mail c.wagstaff@reading.ac.uk

Background and rationale

Like many institutions we have experienced the pressure of many final year students wishing to do projects coupled with extremely limited financial resources that are not sufficient to support a laboratory based project without additional funding from existing research grants. Finding 'dry' projects that still provide a challenge to the student that goes beyond a literature review or dissertation is not easy, but the requirements are that the student can find out something novel, follow scientific method and have the scope to achieve the maximum grade if the project goes well. I have recently started to offer projects that make use of the wealth of post-genomic information and free databases that now exist, together with sequence information for many organisms, to design projects where information from one species can be used to inform a programme of research on another species.

Food Biosciences is small compared to most Biological Sciences departments (about 55 students per year of undergraduate study and the same on MSc programmes) and about 20 staff offer approximately three project titles each. At present, I am the only member of staff offering bioinformatics projects, although some of my colleagues do offer alternative dry projects in the form of conducting food choice surveys or accessing results from large-scale diet and health studies. We have no restrictions on project choice, other than to limit the number of students per staff member to around four. Both 'wet' and 'dry' projects carry the same amount of credit (40 out of 120 credits) in the final year and run over two terms. This is the second year of offering informatics projects in my present job, but I also ran them in my previous position when I was a post-doc in a Biological Sciences department. I would say that they were more popular amongst the Biologists than those studying Food Science or Nutrition, probably because the former have a better background in genetics and plant science and are more aware of the growth of bioinformatics within their discipline.

How to do it

Informatics projects involve 3 elements:

1) A biological problem;
2) Molecular genetics; and
3) Database interrogation.

The project can start at many different points depending on the prior knowledge of the student. Some will have a good understanding of the biological problem being investigated e.g. antioxidants in plants, but not of molecular genetics or database interrogation, whereas others will have knowledge and experience of different elements. Essentially the informatics project brings together components of all three areas by the end of the study.

I would always advise starting with the biological principles behind the project, explaining to the student the real-world relevance of the investigation. For example, my students are all studying some aspect of food biosciences and are wary of a project that looks too much like pure plant science. Once they appreciate that considerable breeding efforts go into our plant-based food crops and that plants donot make antioxidants (or any other secondary product) for our benefit, they begin to see the relevance. I generally have to do a lot of explaining about what Arabidopsis is and why it is so useful, but a student of plant sciences would have less need of this, and perhaps more need of an introduction to which plant products are of dietary significance to humans. Thankfully, with food issues having such a high profile in the media there is a high level of general awareness of dietary goods and evils amongst students.

These projects require a fairly heavy input of time from the supervisor at the beginning of the project in order to familiarise the student with the relevant databases, but once the student is equipped with the relevant tools the project requires much less effort to supervise. The important thing is that students feel confident to go and try things for themselves.

The student will collect a variety of data in the form of gene/protein sequences, descriptions of gene/protein function and expression values. The projects are written in the form of a research paper — our department has just taken the decision to use Biosciences Horizons (http://biohorizons.oxfordjournals.org/) as the guiding format – and these projects are therefore assessed

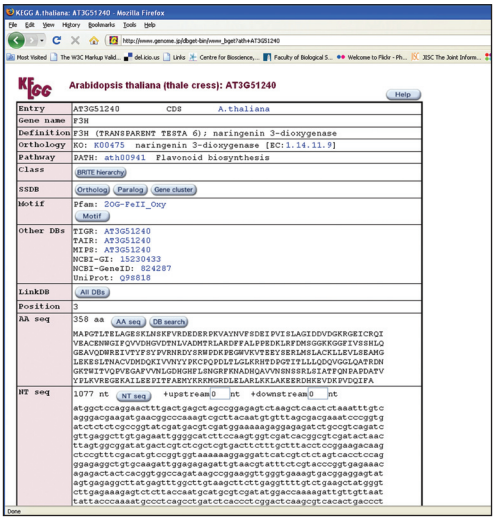
against the same criteria as laboratory based studies.

The following case study illustrates the resources I directed students to for a particular project: 'Using Arabidopsis to identify targets for future research to manipulate the flavonoid content of lettuce.'

a) Direct the student towards some reading designed to familiarise them with the different types of flavonoids, under what situations (e.g. stress) the plant produces them, and the importance of different flavonoid groups in the diet. Hopefully with a bit of guidance they will then decide to focus on one major pathway — for example anthocyanin biosynthesis.

b) Show the student how to use the Kegg metabolic pathway maps for Arabidopsis www.genome.jp/kegg/pathway.html. The flavonoid biosynthesis pathway can be selected from the list of secondary metabolites and the reference pathway changed to the organism Arabidopsis thaliana. You should now be at www.genome.jp/kegg/pathway/ath/ath00941.html. On this page, the rounded boxes link to other pages from the same metabolic map (in this case specific flavonoid groups such as anthocyanins, phenylpropanoids). It can be useful to follow such links if you are looking for genes that regulate large chunks of the pathway. Each square box represents a gene that is thought to regulate that step of the reaction. If it is shaded green the information comes from Arabidopsis. If you select a square box by clicking on it you will be taken to the details of that gene, including AGI code, and genomic and cDNA sequence. Save the information in a separate document.

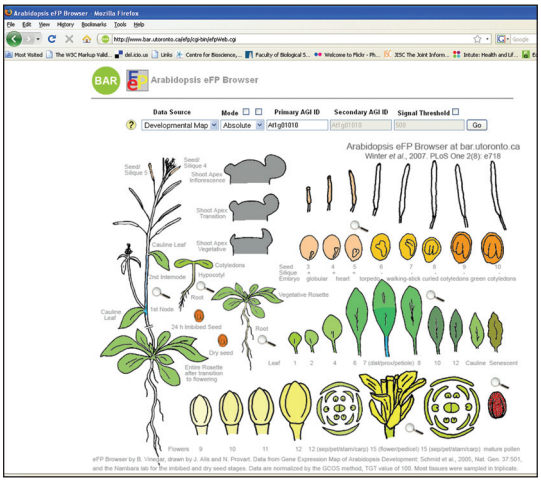
For example, www.genome.jp/dbget-bin/www_bget?ath+AT3G51240 encodes a gene involved in the synthesis of a number of important flavonoids, including anthocyanins. (See screenshot below.)



c) Ask the student to copy the amino acid sequence (ringed in red) from this page and paste it into the lettuce database blast facility at http://cgpdb.ucdavis.edu/database/sms/query.html using fasta format. Check the lettuce EST database and tblastn search boxes before running the blast. A number of sequences are produced with significant alignments. This is a good opportunity to explain to the student what to look for when assessing alignments — good % match over a short region or lower % over the whole sequence. Take care because the lettuce ESTs are not all full length. The lettuce genome is not fully sequenced so the student needs to check they have a homologue (or several) to their gene of interest in the lettuce database before proceeding with the more onerous tasks below.

d) The project can now go in several different directions. At this juncture I usually show the student some software for interrogating expression datasets such as Genevestigator https://www.genevestigator.ethz.ch/gv/index.jsp using their virtual northern tool. This tool can be used to select genes of interest and find out in which tissues they are expressed, at what stage of development and in response to which stress stimuli — all by taking publicly available array data.

Another really nice tool for visualising the spatial expression patterns of genes can be found at www.bar.utoronto.ca/ using the eFP browser tool (see screenshot below), although others on this site are worth a look too.



All the tools mentioned here of course use Arabidopsis so you will need to go back to your original gene IDs — but at this point the students are furnished with the knowledge of which ones show up in lettuce and are worth pursuing. This is particularly important when dealing with gene families and the tools mentioned above come into their own to really answer some questions. For example, the gene identified in part b above encodes a naringenin 3-dioxygenase. This is a family of several related genes and by interrogating a tool like Genevestigator the student can work out

which one is most highly expressed in the desired tissue (leaves) and at what stage of development. The closest homologue to this gene can then be identified from the lettuce database and if the work were moving into the lab this would be the best sequence for any subsequent molecular work to modify flavonoid biosynthesis in lettuce.

e) Some students like to take a more biochemical approach and enjoy finding tools to analyse the predicted protein sequences and the corresponding nucleotide sequences to search for binding sites and recognition motifs using tools such as Prosite, www.expasy.org/prosite/. They could potentially move onto examining folding patterns and identifying docking sites on the mature protein if sufficient staff expertise is available.

f) There are lots of other tools and the students will probably surprise you by finding some of their own! www.arabidopsis.org, www.bar.utoronto.ca/ and www.expasy.org, each host a number of resources, and are good places to start.

Advice on using this approach

a) Be careful the student doesn't lose sight of the aim of the project and collect data without thinking what it means. It is very easy to be overwhelmed by pretty images and gene lists with this project and a bit of direction as to what to focus on is usually helpful.

b) You have to insist that the student is organised and writes as good a lab book for this project as they would for a 'wet' project. Otherwise they will end up with pages of unidentified sequence and very little clarity on what anything means. I encourage them to write down their thought processes as well as a straightforward record of what they have done.

Troubleshooting

It can be frustrating if databases or servers go down. It certainly helps if the student has their own laptop and a fast internet connection rather than being reliant on the average university resource. Some weak students can give into the temptation to stay at home and 'play' with databases without actually achieving anything and in these cases the supervisor will need to impose a more rigid structure of activities.

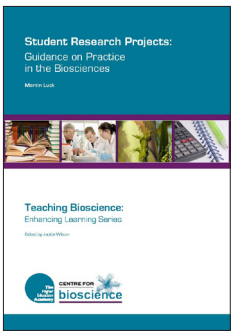
Does it work?

The strengths are that the student can rapidly develop a sense of independence and ownership of their project. Many like the flexibility of working practice it offers and bioinformatics projects can be helpful for students trying to juggle work and family pressures with study. However, it can be quite easy to lose focus and some students can feel as though they are drowning in data. I ran two very similar projects (on aspects of leaf development) at my previous institution with students with very different marks in their second year. To my surprise the 'first class' student did less well than the one who had a 2ii average carried forwards, and the latter really blossomed academically during this project, discovering an ability to synthesis large amounts of information and use the data to understand what was happening at the biological detail in far more detail. This person was able to produce a theory of leaf development and shape determination that challenged the boundaries of what was already known.

Further developments

Recently, new tools have become available to make simple phylogenetic trees using the TreeView function of BLAST as NCBI. This facility allows the development of projects that ask when certain genes evolved and therefore when certain biochemical pathways were in place. Inferences can be made from this to the way plants functioned at the time.

Additional materials



This case study was included in the Teaching Bioscience: Enhancing Learning guide entitled *Student Research Projects: Guidance on Practice in the Biosciences,* written by Martin Luck and published by the Centre for Bioscience. The associated website (www.bioscience.heacademy.ac.uk/resources/TeachingGuides/) contains a downloadable version of this case study

Case Study published October 2008

The Higher Education Academy
CENTRE FOR bioscience

Centre for Bioscience
Room 9.15 Worsley Building
University of Leeds, Leeds, LS2 9JT
Tel / Fax: 0113 343 3001 / 5894
Email: heabioscience@leeds.ac.uk
Web: www.bioscience.heacademy.ac.uk