eIUS: Bioinformatics Experience Report

One interviewee; text = quotes;

Background/ Research area/Research question(s)

- I'm a bioinformatician and I've been working in bioinformatics for about eight or nine years now, and the thing that interests me the most, really, is how you can use computer science technologies, for answering biological questions; how can you get further in biology by inputting it into your computer science technology, so I've been working in a variety of different fields, addressing those questions.[...] I spend a lot of time doing training in the community, I spend a lot of time doing consultations, showing people how to use our tools, helping them when they get stuck with things, but at the same time as that I have my own projects, using our tools so that I become a super user as well, but also in things that I'm interested in anyway.

- The Human Genome Project was a big media event a few years ago, but that was part of a much bigger revolution in biology which really changed the way that biologists did their work. Before that people would spend ten or twenty or thirty years looking at one gene in one organism or one cell, and then they'd build their whole careers on that one thing, and they become very focussed and very detailed knowledge about small amounts of things. But when we started looking at the whole Genome of an organism, whether it was human or something else, then the way that people though about science changed somewhat, and when we started publicising all of this data - it's all publicly available, it's all out there in the world somewhere and you don't have to have passwords for it, you don't have to do anything, you can just go to a website and download all Genomes, all Proteomes, some wonderful stuff - it really did change everything. So what I see my job as being, is about interpreting that data, finding patterns in that data, looking at very interesting things in that data, because at the moment a lot of it is just strings of letters and numbers, and by itself it's not interesting, you have to find within that, the good bits.

- a lot of bioinformatics is about comparing what you know in one species, one organism, to everything else, so you can look at gene sequences and see if they're the same or different. And you can understand a little bit more about the differences between the organisms by the differences in their generic content, you can understand a lot about diseases if you look at the sequences of genes and then mutated sequences of genes in diseased individuals, and so it's all about aligning data and then finding the differences in that data.

- I've been working with a mouse geneticist recently, and she is trying to find mouse models for human diseases - because obviously you can't do lots of genetic tests on humans, it's just not possible, but maybe you can with a mouse - and so she's looking for similarities between abnormalities in human disease and abnormalities that you see in a mouse - which is quite tricky because there are definitely some differences between a human and a mouse - but what she's doing is she's looking at it from a genomic perspective, she's aligning regions of similarity up between human and mouse. But she's a lab biologist and she doesn't really understand how to do this in Insilico world, she doesn't understand how to manipulate the data, she doesn't understand what tools are available, so I've been helping her with that, because that's what I do, so we've been coming together and working together on that.

- the question she was asking was, how do I model Crohn's disease in mice, how do I find a mouse that's producing the same kind of symptoms as a human Crohn's disease patient. And so they have

these regions of genome that they've identified in Crohn's disease patients as being implicated in the disease - and it's a complex disease, so it's not just one gene, we're looking at several different genome locations with lots of genes in there - and she's trying to find those regions of similarity in the mouse. And then I'm doing a lot statistical processing of the levels of similarity between those genes - whether the orders of the genes are the same, whether the expression levels of the genes are the same, all sorts of different things - to try and determine whether the mouse that looks like it might be suffering from the same kind of symptoms is actually a good enough model for trying to make the disease better, like find things to treat the disease, that kind of thing.

- myGrid is the name of the project, and within myGrid there are lots of different components and they're all designed to support the Insilico life cycle process form beginning to end. Taverna, think of that as the main user interface to all of the other myGrid components, because they tend to fit into the back somewhere. So a lot people don't even know about the myGrid project, even if they're using Taverna, it's just a tool with some other tools attached, but myGrid is the name of the umbrella project over the top. And myExperiment is a spin-off of the myGrid project, because once we had lots of users, and they were all producing workflows, they started sharing them by email or on our mailing list and stuff. And it became very obvious that actually people were reusing other people's workflows instead of starting from scratch all the time, reinventing the wheel, they could just take somebody else's and use it again, so we needed somewhere that was central that people could put their stuff.

- The myGrid project started back in 2001, it was one of the eScience UK, eScience pilot projects, and everything was going to be open source form the beginning, it was just a policy decision right then and there, the Taverna workbench came out in the end of 2003, the first version, so we've been around for a while now, and it took a while to get to that point.

Research Lifecycle

Literature review

- in biology, generally, you have to follow the scientific method, which is, you have to start from what other people know, formulate some kind of hypothesis about the thing that you're interested in, design your experiments; proper methodological 'this is what I'm going to do, this is how I'm going to do it' and then prove it, so that's always the way it has to be really.[...] First of all you've got to establish that nobody else has answered the question before, and then you've got to establish who is working in similar areas that can give you some insight [...] Most biology journals are available online now so I haven't actually been to the library for a long time I have to say.[...] At the university we have access to every journal I can think of; it's very rare that I get an abstract for a paper and then find that I can't get in without buying it, it's usually free and available.

- in biology we have a service called PUBMed, which is a catalogue of many different journals, and when we start with PUBMed I don't really mind which journal or particular paper it is, it's the content of the paper that I'm interested in, so obviously if it's in "Nature" then you think oh, it's a bit more prestigious.[...] simple keyword searches once you start finding authors, you'd have a particular interest in an area you might search for more papers by particular authors and use that kind of pivotal browsing.

Data collection

- it's all experimental data that's been published, so we're looking at human genome sequences, mass genome sequences and features of those sequences.[...] they do get updated very frequently, yeah, every twenty four hours there's an update to the major [genome] database.

- The software we produce [Taverna] is a workflow management system [which allows me to] design my experiment as workflow: I need to get this data from here, this data from here, compare it in some way, usually in various ways, and then I need to take these comparisons and do some statistical analysis on them, or I need to take these comparisons and display them in a way that I can then show the biologist the results. So I automate the whole process, then if I want to look every day to see whether there's new data, I can just run on the website again.[...] [Taverna] allows you to connect local and remote resources, whether they're web services or database cores, or local Java scripts, it allows you to connect things to other things - because in bioinformatics the data is freely available but it's spread out all over the world, it's in different formats, you're accessing it in different ways - so Taverna allows you to pull that all together in an automated fashion. In the olden days, people would go to web pages and cut and paste pieces of data into web forms and wait for the result, and then take them and put them in another web form and wait for the result, which is fine where you're doing one gene at a time, but we're talking about whole regions of genome now, so that would take weeks to do one region and then weeks to do another region, whereas with the workflow you can do that in a few hours.

- [for example] I find myself a human gene and I want to know the equivalent gene in a mouse, the homologous gene in a mouse. So I have to use some kind of analysis tool that will search my human gene sequence against the mouse gene database. Once I've got that gene, I want to then find out if the gene next to it on the chromosome in the human and the mouse is, first, the same, and then, also, homologous to one another. And then I do that all the way along until I run out of homology between the two. So that means I'm doing the same analysis to each piece of data, but I don't want to go back to the same web page and do that over and over again, I want to just give it a whole list of human genes and say, find me a list of mouse genes, and then once I have those genes I want to find out things like, percentage identity, whereabouts on each chromosome of the mouse this is, what's known in that region already, so you can search against the literature, you can search against disease databases, and they can tell you if something has been discovered at a particular position on the genome. But you don't want to do that for every single piece of data again, you want to do it automatically. You can do it by hand, and I did do once upon a time when I started bioinformatics, that was kind of the way to do it.

- [a workflow] is a very modular thing. Small parts of it will be useful in other experiments, you might start at the same point in lots of different experiments, and then go off at different tangents below, so the workflow itself is a very re-useable document, you can give it to other people and they can run their workflow, or you can run it with different data; all sorts of things.

- all the data is publicly available so anyone can collect it whenever they like

Data analysis

- there are lots of different bioinformatics tools, which do statistical analysis, or just manual comparison between the letters in the sequences, to tell you how many are the same and how many are different, all that kind of thing, and again, lots of different steps involved to get to your final answer.

- [if Taverna didn't exist people could still do comparisons between genome databases manually] or write more programmes to connect to the web pages, but then you have to be a programmer as well. So the reason Taverna is becoming popular in bioinformatics is, some bioinformaticians are very good programmers, some are very good biologists who know about biological data, and there's not really much of an overlap. I'm a bit more in the middle because I work in the computer science department and I do a bit of programming when I can't get away with it.

- [to do the kind of things Taverna does you only need Internet access] - It's the power of the web services paradigm. Things are just there and you're connecting them, you don't even need a username

or password for many of these services, they're just there. [With the opposite, of course, that if something is wrong with the Internet, you're stuck really] Yeah, but with many of these processes you would be stuck anyway. You can download all of the data and then do things locally as well.

- I don't [make use of high power computing services, although] we collaborate with some people who do. It generally depends how big your data is, and then how long it takes to do the processing that you need to do on your data. We have some astronomers who use Taverna, and their data, you know, we say we've got huge data and they laugh at us, because, you know, [comparatively] it's not huge at all. But their data is much more uniform. We have very heterogeneous data, we have lots of different types of data, and integrating it and making it [available] in formats that different services can understand is a big challenge, whereas in astronomy the size of the data is the problem, because everything is standardised.

Collaboration: Discuss/compare results

- the first step is to share the data back with the lab biologist that you're working with, and we've done a little bit of that already, and then hopefully the next stage is that she takes the interesting bits back to her laboratory, she does some more research and then comes back to me with more questions and then we do this backwards and forwards a few times […] Talking to each other, physically. I prefer to have a proper meeting, [but] emails would work, [also] Skype chatting, we do a lot of Skype chatting.[...] In our project we have people who don't work on site, and we have Skype meetings with them almost every day [...] and we have a big meeting every week.

- There's another component of myGrid which plugs into Taverna, which you can store your data in, but that again is local until you say otherwise. [...] You can choose between having your own store on your computer, which is what I tend to do; they're quite small pieces of data, you can email them to your collaborators, or you can put them on a secure web page that they can download from, but that's all until publication.

- In biology *when* you share your data is very important. If you share your data before publication stage, then maybe somebody else can do something with it, publish it, and then you won't be able to publish, because there are lots of people competing, if you like. So although [bioinformatics] is very open, biology tends to be a bit more closed off, because the experiments themselves in the lab might take months, and so if you spent months in the lab doing something, you don't want somebody else to come along and get the credit for it. So we have to be careful as bioinformaticians not to kind of destroy that, so we don't tend to publish anything until we publish the paper that describes the results.

Dissemination

- the idea is that you publish enough data that somebody else can come along and repeat your experiment, and verify your experiment, and that's the ideal situation. But actually if you take a paper at random out of the database you'll find that that's not always the case. And especially now as data is so much larger, the underlying databases are updated so frequently that actually even if you do publish everything in it's entirety, if you try and run the same analysis again, some of the searches that you were running, you'll be searching over different data anyway, so you have to be aware of that when you come to look at the results.

- It's becoming the practice that you have your general article which describes your experiment, it describes your results and it shows you the key results in maybe tables and graphs and things like that, but then you point to supplementary material, which tends to be in the raw data files or some

proportion of the raw data files after you've tidied up [...] A lot of the journals have a supplementary material page now, and when you upload your paper you have to also upload your supplementary material.

- the unique thing about Taverna is that you could also publish your workflow, because that's a protocol, it's the description of your experiment, and for that we have this site called myExperiment, which you might've come across, which is a workflows repository, and so we're trying to encourage people to upload their workflows there so that other people can reuse them in an easier way.

- we have two different types of users, we have the scientist users who just share workflows, and then we have people who do something in the workbench, find something they can't do, build a bit of code to allow them to do that, and then of course it gets shared back to the rest of the user group because it's open source, they submit it back to us, so yeah, we get a lot of contributions.

- the social computing aspect of myExperiment means that people comment on each other's workloads. The reputation of a person who put the workflows up is important, and that's the same as in Wikipedia I suppose, if you see something written that's not right, you change it, then you're putting your own reputation on the line really, you explain: I know more about this thing than you do, so I'm going to correct you.[...] You can comment or you can ask questions of the workflow author, and you can rate it, we have a star rating system.[...] And you don't have to share [your workflow] with the world, you can share it with just small groups of people or just your friends, depending on what you want to do as well, so people like that as well.

- [Important journals/publications:] "Nature", "Nucleic Acids Research", "Bioinformatics", "Cell Biology"

Other important elements about/in the research

- researcher working on a different workflow system, which uses grid computing to analyse medical imaging data, so the data is much bigger, the problems are slightly different and they're much more like physicist astronomer problems because the data is all the same, it's all about lining up the images so they match and how you tell when the images are aligned, but there's an anomaly in one image, that kind of problem, medical image data.

- we have two different groups of astronomers [on Taverna], there's the AstroGrid project, and they're all based in various places in the UK; Nick Walton is our main contact, and he's based somewhere, I think it's in Cambridge but I don't know the institute name. [The others] are based largely in Scandinavia, so that's probably harder to get to [...] that group's called Sampo, so they've got very similar aims to AstroGrid, but they're going about it in slightly different ways, but they're both using Taverna in different ways.[...] Yeah, that group's quite interesting because we didn't know anything about them until they already had produced workflows, they didn't come and ask us how to do anything, they just went away and did it. Because when you go into physics and chemistry, the level of computational skill tends to be much higher than in biology, so you just take the tools and do stuff with them, and they don't really need any extra help at all.

- on myExperiment there are a couple of social science workflows actually.[...] somebody called [name], I don't know if you're familiar with her [...] She's been building workflows in social science, and they're on myExperiment, if you go there