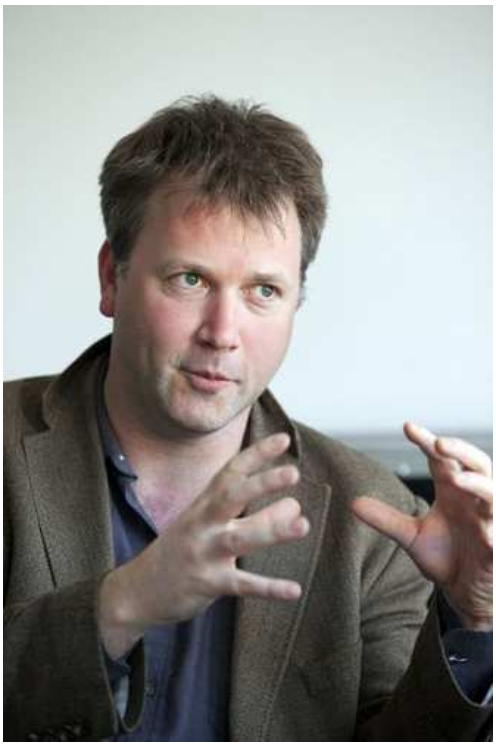
Data in the DNA: transforming biology and data storage

FEATURE | OCTOBER 9, 2013 | BY ANDREW PURCELL

Ewan Birney, associate director of the European Bioinformatics Institute (EBI), will be giving a keynote talk at the upcoming EUDAT 2nd Conference, due to take place from 28 to 30 October in Rome. He speaks to iSGTW *ahead of the event…*



Biology has become a 'big-data science', says Birney. Image courtesy Duncan Hull, Flickr (CC BY-SA 2.0).

What's going to be the theme for your talk?

I'm going to talk about the transformation that's happened in biology: it's become a 'big-data science'. It's always been a science about information, but over the last decade the pace of data generation has been remarkable. Technologies, such as DNA sequencing for instance, have been halving in price every 6 months — it's amazing. The cost of sequencing a human genome back in 2005 was a couple of million US dollars, whereas today it's just $5,000-$10,000. Basically, the data generation part of biology has become a lot cheaper and has moved the bottlenecks on to what I describe as 'blue-' and 'white-collar problems'.

'Blue-collar problems' are getting the data, making sure the data is 'straight', shipping the data around the world, aggregating the data, *etc.* These are all conceptually quite easy to do, but now that we're at the petabyte-scale — we have 5 PB of scientific data at EBI and we 'spin' about 30 PB of data — they can actually be quite tricky. If you go back about a decade ago, molecular biology really didn't have too many of these blue collar challenges, but now they're a big issue. You can't do the research credibly on your laptop; you must have server farms and other systems to help you handle the data.

'White-collar problems', on the other hand, are things like the actual analysis of the data. This ends up being very exciting and is the very reason for which we generate the data: we don't generate petabytes of data just to stick it on disks!

The session in which you'll be giving your talk at the EUDAT conference has been given the title '*Life and Earth sciences at a cross-road: community-driven flagship initiatives in the EU and USA*'. In what way do you feel that life and earth sciences can be said to be 'at a cross-road'?

Personally, I'm not so sure about 'a cross-road', but we've certainly passed a watershed in terms of really becoming a data-intensive science. A decade ago, only a very small proportion of people working in the field were doing what you could call 'data-intensive research', but now the vast majority of life scientists do have to think about how to deal with large data sets.

And what are you looking forward to most at the event?

I'm very interested in finding out more about data-dispersion technologies, as well as about data transfer and data I/O on a large scale. A core part of the 'blue-collar problems' I mentioned earlier involves receiving data, storing data, aggregating data, and distributing it. Any technology which can help with any part of this is going to be very, very useful. I'm particularly interested in meeting people who are pushing fundamental storage technologies.

EUDAT is a very useful initiative, particularly in terms of helping to share expertise across different areas of science. I'm really interested in picking up on some of this expertise at the event and understanding the trends in how data is handled, by learning about what other scientific groups have done and how they have solved particular problems. Of course, at the same time, we'll be contributing information on how we've solved particular problems at the conference, as well.

You've also recently made headlines around the world with research you've conducted into the viability of using DNA as a data storage medium. Perhaps you could tell the *iSGTW* readers a little more about this…

I was in a pub in Hamburg, drinking a Hefeweizen beer with a colleague of mine, Nick Goldman, and we were thinking to ourselves: how are we possibly going to solve the problem of storing the ever-growing amounts data generated globally? What we need, we thought to ourselves, is an electricity-free, dense storage system, that's also digital. So, that's how we came up with the idea of DNA. People have recovered DNA that's over 700,000 years old; you've just got to keep it cold, dry and in the dark.

What sort of data have you been able to store using DNA molecules so far?

We've now stored on DNA: Shakespeare's sonnets, a snippet of Martin Luther King's 'I have a dream' speech, a PDF of Watson and Crick's paper identifying the double helix structure of DNA, a picture of the EBI, and a self-referential piece of code explaining how we'd done the encoding. Out of these five files, we were able to recover four of them from the DNA with no problem at all. With one, we did have a problem, but we were able to work out how to fix this.

So, how useful do you think DNA storage has the potential to become? Is it really practical?

DNA is remarkable: just one gram of DNA can store about a petabyte's worth of data, and that's with the redundancy required to ensure that it's fully error tolerant. It's estimated that you could put the whole internet into the size of a van! You can also copy trivially. The only problem at the moment is cost: it's prohibitively expensive to write DNA. Nevertheless, this technology is expected to come down in price dramatically over the coming years. The only question is: how quickly will it come down in price?

DNA could be really useful for storing data sets on a really long timescale. The earliest writings ever discovered are under 6,000 years old, so data stored on DNA would outlast everything we know about today. DNA data storage could be really useful for storing things like films, governmental archives, key scientific data, and so on. For example, it could be really useful for climate change research to store an archive of Earth-observing satellite images from previous decades.

If DNA synthesis costs come down, it's definitely going to become a credible technology. Of course, it's worth pointing out that I'm actually part of a team now thinking about commercializing this technology, so we're not exactly impartial observers!

Average:

Your rating: None Average: 4 (248 votes)

About the Author »

Andrew Purcell

Editor

Andrew Purcell is the editor of iSGTW and is based at CERN, near Geneva.

RELATED TERMS: big data data DNA EBI EMBL EUDAT Europe European Bioinformatics Institute Ewan Birney handling storage biology data management systems high-performance computing health and medicine

Comments

ADD NEW COMMENT

Post new comment

Subject:

Comment: *

Input format

By submitting this form, you accept the Mollom privacy policy.

SAVE PREVIEW