David Haussler

Human genome pioneer David Haussler talks about the evolving role of annotated data repositories.

When the Human Genome Project was assembling DNA sequences to play catch-up with Celera (Rockville, MD, USA), the public initiative turned to David Haussler's group at the University of California, Santa Cruz. Since then, Haussler's team has built and maintained the UCSC Genome Browser (http://genome.ucsc.edu/) (*Genome Res.* **12**, 996–1006, 2002), a repository for storing genome sequences and annotations, such as genes and transcripts, as well as a tool for data analysis and visualization. Remarkably, Haussler's roots in computational biology—not that it was a discipline at the time—can be traced back to the intellectual hotbed of a weekly meeting for graduate students run by Andrzej Ehrenfeucht at the University of Colorado at Boulder, the same meeting that shaped the careers of Gary Stormo and Gene Myers, two other pillars in the field of bioinformatics.

How did that weekly meeting at UC Boulder influence you?

David Haussler: Gene and I and Gary all went to that meeting. Gene and I were in the computer science department, and Gary was with Larry Gold in molecular biology. At the time, the total amount of DNA that was available was the complete sequence of phi X, and a couple of other short viruses and snippets of *Escherichia coli*. You could save all of it on a little floppy disk. Right there in that seminar we started thinking, "Well, how do we apply computers to analyze DNA sequences?"

I actually went to Boulder to work with Andrzej Ehrenfeucht on logic and theoretical computer science. Andrzej is a polymath. He is off-the-charts brilliant, and knows so much about so many different fields. But I tell you, I had not intended to study DNA, but in that group we covered so many areas that we were completely bombarded by new things, and that's certainly where it started.

How has the creation of annotated data repositories changed biology over the past decade?

DH: Biology is increasingly an information-driven field. But molecular biology has historically been an information-limited field, where the means of communication have been the traditional means: presenting a talk at a conference and publishing a paper in a journal. That is an incredibly limited way to get information.

To transcend this, all of the data need to be unified in one format that's easily accessible—Google for the genome—where you can just make queries and find what you want. We tried to build that with the UCSC Genome Browser, and we think that that has had an enormous impact because it has given people orders of magnitude more information at their fingertips than they had before.

> "All of the data need to be unified in one format that's easily accessible—Google for the genome—where you can just make queries and find what you want."

Can you provide an example?

DH: Let's take comparative genomics. On the [UCSC Genome] Browser you can see the human genome compared with almost 50 other vertebrate genomes. For any gene or non-coding regulatory element, you can look at the corresponding element in these other vertebrates, to the extent that it exists. And so it's very common to look and say, "here's an enhancer for this gene and look, all the mammals have it and the nonmammals don't!" Increasingly, now we want to supplement that evolutionary information with experimental evidence that says, hey, this DNA hypersensitivity track says that this [chromatin] is open and active during neural development. And, wow! This neural developmental transcription factor actually has a binding site in this enhancer.

You can see how it snowballs. It's incredibly exciting to layer on these other functional data and to combine them with the evolutionary data, and you start to get a picture now that says, wow, this was really an innovation in mammals and it has something to do with brain development and it's particularly controlled by these transcription factors, and it turns on these genes at certain times in development. And now biology [not computer science] is happening. We're talking about storing, literally, millions upon millions of these types of data in an accessible format.

Where do you see analytic tools going from here?

DH: Our feeling is that, if you can answer a question interactively on the web, you should be able to do that, because seldom do people ask exactly the right question they wanted to the first time out. If it's a more complicated question, you should be able to easily download the data you need to determine the answer offline.

We also interact with a tool called Galaxy (*Genome Biol.* **11**, R86, 2010), which is a workflow management system. It provides a visual interface that allows you to set up a pipeline of processing activities that you can run, edit, store and exchange with your friends. We have found that for more professional programmers, or people who want to go deeper, the combination of the UCSC Browser with the Galaxy workflow management system is a sweet deal.

What about erroneous annotation in biological databases—how can this be tackled?

DH: What we would like to stimulate is as much community feedback and community correction as possible. One of the great models for this is Wikipedia. Are there errors in Wikipedia? You bet. Do they get fixed? Well, there are a lot of eyes on it. And the more eyes on it, the more they do get fixed. Getting rid of errors is mainly a function of how many eyes you can put on it. There needs to be a mechanism, an easy feedback mechanism, so that when somebody spots an error then something can actually be done about it. Definitely much more can be done in this vein, but I think the more we go toward easy and direct community feedback, the better.