De: Gloria Lligadas <gloria.lligadas@crg.es>
Asunto: **Thank you again**
Fecha: 27 de noviembre de 2007 09:21:01 GMT+01:00

23 archivos adjuntos, 145 KB Guardar Pase de diapositivas

Roderic Guigó on the Informatics Challenges Of ENCODE and a Steady Rise of Genome Data

[November 23, 2007]

Roderic Guigó

Biologist
Coordinator of the Centre de Regulació Genomica, Professor at Universitat Pompeu Fabra, Barcelona

ATLANTA — Roderic Guigó has overseen the Bioinformatics and Genomics program at the Centre de Regulació Genomica in Barcelona, Spain since 2005, and in tandem, has been a professor at neighboring Universitat Pompeu Fabra since 1999.

He has also worked at the Los Alamos National Lab as a postdoctoral fellow in its' theoretical biology and biophysics group with James Fickett, where he worked on genome analysis-related problems such as estimating a genome's protein-coding density.

Prior to that, as a post-doctoral fellow, he worked with Temple Smith at Boston University's biomolecular engineering research center, doing research on gene identification, protein sequence pattern analysis and molecular evolution.

In This Week's Issue bio1NFORM

For the past several years, Guigó has concentrated on the ENCyclopedia of DNA Elements, or ENCODE, project. A participant in the pilot phase of the project since the first grants were awarded in 2003 [*BioInform* 10-20-03], he has recently geared up to explore the entire human genome as part of the full-scale ENCODE project, which kicked off last month.

Guigó spoke at the 6th Georgia Tech-Oak Ridge National Laboratory International Conference on Bioinformatics, held here last week. His talk focused on the transcriptional complexity of the human genome and its relationship to ENCODE.

BioInform sat down with Guigó during the conference. Following is an abbreviated version of that conversation.

You studied with Temple Smith. How was that?

Great.

I spoke to him at ISMB and he doesn't like the term 'bioinformatics.'

I know, but that's the name that has finally established itself, right? [Although] 'informatics' … in Europe … is more [often equated with] 'computational science.' For example, in Barcelona, the school of computer science is called 'the school of informatics.'

Years ago, you and Temple Smith went on to publish a number of papers. What can you tell *BioInform* about, say, 'Inferring correlation between database queries: Analysis of protein sequence patterns.' [IEEE Transactions on Pattern Analysis and Machine Intelligence, 15:1030-1041 (1993)]?

That paper has almost never been cited, but it is one of the most challenging papers I've written. It was more about interrogating the databases in general. I think we had a nice application to biology. It was intellectually very challenging. I had to make an effort to understand the solution. Very few people have noticed this paper; for me it was a great period of time. I was challenged, and I think I was able to be up to the task, but as it happens this paper [has not been cited often].

What informatics challenges have you faced and will you face on the extended version of the ENCODE project? What have you learned from the pilot project that will tackle those challenges?

In terms of informatics, the challenge [is that] the amount of data is going to be much larger. There will be varied and continuous data, so just to keep track of all the data … the interaction between the different groups that are participating in the project so that there is understanding and communication. … I think that from the informatics standpoint, this is one of the challenges.

The ENCODE pilot has helped us anticipate the difficulties, learn from the errors, and identify the most cost-effective strategies before attacking the entire human genome. Cost of errors in 1 percent of the human genome is less than the cost of errors in the entire human genome.

Plus, I guess in the pilot — a lot of things happened that we hadn't expected.

How so?

These four years, I think, [have] helped us to [move ahead to the full-scale project]. … If we were to start just with the entire genome, we'd have made many more mistakes … than if we'd just started with 1 percent. [That] was a sort of warm-up for the entire genome, so many lessons have been learned.

For instance, one very basic thing is that different groups need to work with the same biological material, so data can be correlated — for instance, the same cell lines and same tissue on the experiments. So data tracking is going to be a large problem.

How long have you been working on the full-scale version of the ENCODE project?

(laughs) Just for a few days. … I don't know when it started — maybe Oct. 1, maybe Nov. 1. We have had only one [conference] call [regarding the full-scale project]. And the first meeting is going to [take place Nov. 28-29 in Bethesda, Md.].

What are your thoughts on new experimental platforms: high-density chips, next-gen sequencers, and so on? Will these types of technologies, some of which have been developed since the launch of the initial, pilot version of ENCODE, propel the project forward?

Well that's the thing. Some of these technologies exploded during the pilot project — like tiling arrays. So, we have a good feeling for the data that's produced by tiling arrays. Still, we've got challenges to really understand this data. We have not experienced the developed project [with the use of] … these new technologies, next-gen sequencing technologies. This is something that I guess is going to surely be weighed substantially on the ENCODE project…

The new genomic technologies provide us with an unprecedented ability to survey the transcriptional activity of the human genome, and to some extent, that's the goal of the ENCODE project — to see which are the actors in the human genome sequence, which regions of the human genome have activity. And these sequencing platforms will, I think, really help to get a much higher-resolution picture of what those actors are of the genomic sequence.

And of course, there are many challenges now in bioinformatics. I think, in a sense with these new sequencing technologies and platforms it's like what happened … [for example] with microarrays 10 years ago. There was an explosion of people moving to work with microarray data. Now, there is an explosion of people working on algorithms to deal with high-throughput sequencing data because this data is going to be used for many

different applications — from *de novo* sequencing, but also to transcriptome activity, from interactions between DNA and proteins for structure with the DNA, for sequencing of metagenomic communities, like epigenomics project[s].

Do you think we have enough people working on those algorithms? It seemed to me at the conference I was just at, Genome Informatics [at Cold Spring Harbor], every single research institute was hiring. Is there a shortage of top talent?

There is a shortage — I think because of technical challenges to deal with the data, but also the amount of data we produce. I mean, there will be a shortage of people. I think so.

Why do you think that is? Did it take [the field] by surprise, the amount of data that would have to be dealt with?

Well, I don't know if it took everybody by surprise. But I think that even if you knew this was going to happen — when it happened, it was always, to some extent [still surprising]. … I don't think we are completely ready to deal with all of the data we are able to produce now.

That's what I am hearing from people in [the field]. I wonder what the net effect of that will be. Does it mean the people who are handling the data are over-worked and it will lead to attrition in the field?

The thing is, many data is not analyzed. Many data will be produced and only superficially analyzed. There is so much data now out there that could be mined. … Maybe there [are] some very important facts about the biology that [are] already there, that [are] available somewhere, but the problem is there are not enough human beings. It's not just a computational problem where you need more powerful computers, but in particular, you need more human beings trained to understand the data and technically trained to be able to refer biologically significant information from the sequence data.

And also, it needs some adjustment from the more biologically oriented researchers, because they can now do things at a scale that could not be imagined only two or three years ago.

Can you give me an overview of the software tools — be they gene analysis, visualization, or gene regulation — that you've developed in your lab, and how much of the development is performed by you versus your students?

(laughs) This has changed over the years. I used to do everything; now I do nothing, right? In terms of coding, … there has been a progressive transition from one situation to another. So the tools we have been mostly producing have been tools for gene prediction — for analyzing genome sequences and inferring the structure of the protein-coding genes, the genetic sequences …

We have … visualization tools which have been used to produce large printouts of the genome maps like the human genome, or the fly genome, or … some other genomes.

We have been working with some algorithms to compare and align the promoter regions.

How many people [are] in the group in Barcelona? I saw 12 in the photo on your site. Are there about 12?

Yes.

Is there a freedom for your research outside the US that doesn't exist here? Are there more regulations here?

Such as for stem cells, there are other countries that are more open to research on this. But in terms of genome sequencing, this is something I really don't know about. [In terms of] sequencing of pathogens which cause disease, while ….I do not know to what extent, I guess there should be some restrictions around the distribution of this data because in principal you could use [the research knowledge] to engineer more dangerous possibilities. You could use the sequence, in theory, to create strands that are more resistant to antibiotics and things like that.