George Church discusses the quest for quality bioengineering

Interviewed by J.C. Louis, jclouis@att.net

George Church, Professor of Genetics
Harvard Medical School

George Church is a Professor of Genetics at Harvard Medical School (http://www.hms.harvard.edu) and Director of the Lipper Center for Computational Genetics (http://arep.med.harvard.edu). His research focuses on the integration of biosystems modeling with high-throughput data for haplotypes, RNA arrays, proteomics and metabolites. He began research at Duke University (http://www.duke.edu) after his BA in Chemistry and Zoology, and completed his PhD at Harvard in Biochemistry and Molecular Biology with Walter Gilbert. He was briefly CEO of BioGen (http://www.biogen.com), developing the first direct genomic sequencing method in 1984 – work that helped lay the foundation for the Human Genome Project. He later helped found the Stanford MIT (http://www.wi.mit.edu/Genome/natureg.html) and Waltham Genome Centers. He invented the broadly applied concepts of molecular multiplexing and tags, homologous recombination methods and the DNA array synthesizer. His automated sequencing and annotation software resulted in the first commercially sold genome sequence (the human pathogen *Helicobacter pylori*).

Please say a few words about how your first scientific interests influenced your later journey into computational genomics.
Ever since I was a kid, I was always looking for connections between computing, biology and chemistry. That first great connection was in crystallography, during the early 1970s. Crystallography had key components that are now very much a part of biology and computing. The essentials common to both fields lay in automated data collection; a solid biophysical theory by which to model the data, as well as a computational way to share models. That is what we want in the life sciences today - automated data collection, a biophysical theory that works and computer-ready sharing of data and models.

There's a strong undercurrent of engineering that informs your practice of functional genomics. Can you suggest how the applied emphasis of engineering influences the basic genomic research you pursue?
It's often the way in engineering that it's not so much discovering one big thing but discovering lots of little things. For example, the Human Genome Project was not so much about discovering, but rather about developing resources for quantitative assays of most RNAs and proteins. Data collection, generating a hypothesis and perturbing a system are not new. Rather, the task of systems biology should be more about data mining...we want to collect enough data up-front to do many rounds of hypothesis-driven science.

'The goal is to assimilate, optimize or modify a system to serve a larger end within the field.'

How do you view the distinction between hypothesis- and discovery-driven science? And can you briefly describe the interplay between data, models or simulations in hypothesis formation and subsequent experimentation?
We need to collect enough data systematically to do hypothesis-driven science. I don't really see the choice [as being] so much between hypothesis and discovery science but rather between data mining and experimental practice. We have more solid biophysical theories combined with enough data to fuel them. Crystallography provided more than enough data to solve the structure in that field. Functional genomics is still trying to solve the system without quite enough data, but that's okay because there is enough data coming. The key challenge ahead remains how to get enough cost-effective data to solve the structures in this field.

What is the range of biophysical models available?
Population genetic models, evolutionary models and neurobiological models are each historically limited by the data. Because of the computing revolution, we can get to a point where we have a lot of data. The data is of course from different fields. Neural data might come from functional magnetic resonance imaging, and ecological data might be in form of microbial data in populations that impact the carbon-cycle and toxic clean-up. Biochemical network modeling might measure all the proteins, RNA and metabolites in various environments. We do not have to invent new fields of mathematics or chemical kinetics. [We] just need slightly more clever algorithms, enough computing power and better apparatus for data collection. If the quantity of data over-determines the system, you do not need as sophisticated data.

'Together, we are pushing the technology while phrasing interesting questions about biological systems.'

Can you say briefly describe the work of Lipper Center in the larger scheme of computational genomics?
The Lipper Center tries hard to integrate facts rather than merely connect them. That is, we try to fit sequences and mutation patterns with molecular concentration, localization and structural data in the context of physical models where the pieces fit together. The goal is to assimilate, optimize or modify a system to serve a larger end within the field.

How do you see yourself positioned between the academic world and the private sector?
I see myself at the interface between academia and the commercial world, rather than exclusively in one or the other. I try to work with the best academic minds – students and post-docs. Together, we are pushing the technology while phrasing interesting questions about biological systems. If we need to scale-up, we often establish the technology in the academic sector for transfer to the commercial sector so that the whole field can advance.

1478-5282/03/$ – see front matter PII: S1478-5282(02)02266-3