Computational Biologist

Interview by Kristi Birch

Aravinda Chakravarti, PhD

Professor of Medicine, Pediatrics, Molecular Biology & Genetics
Johns Hopkins University School of Medicine

Professor of Biostatistics
McKusick-Nathans Institute of Genetic Medicine

ISTOCK

Aravinda Chakravarti was getting his undergraduate degree in math and statistics when he asked a biology professor to teach him molecular biology on the side. Twenty-five years later, he found himself on the advisory council of the National Human Genome Research Institute, which carried out the NIH's role in the International Human Genome Project, a massive undertaking that sequenced the three billion base pairs in human DNA. Today, he uses his quantitative skills to perform computational analyses on vast biological datasets to discover how genes interact with each other and the environment to cause diseases ranging from heart disorders and diabetes to mental illness and autism. Here, he explains his work and what he thinks is next in his rapidly changing field.

What is computational biology? How is it different from bioinformatics?

Bioinformatics is the science of the storage, retrieval, management, and statistical analysis of large datasets in the biological sciences. Computational biology is computation on those datasets to draw biological inferences. Physicists compute, financial people compute, sociologists compute, as do we biologists. The important difference is the domain of knowledge, the principles of biology, not the hardware. To be a computational biologist, you have to be a biologist in your heart and in your soul.

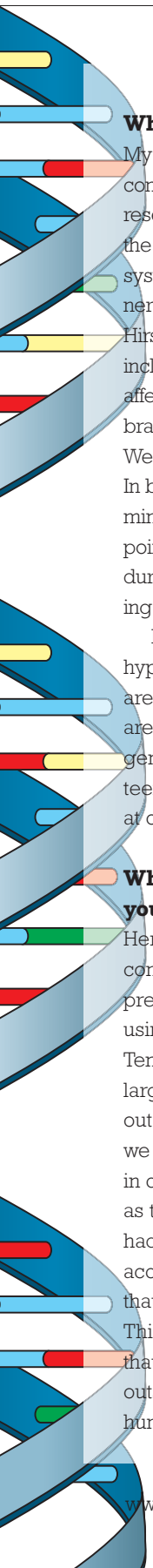
The importance of both of these areas has increased exponentially in the past few years with our ability to examine the genome in a variety of ways—by sequencing the genome or assessing the genetic activity of each gene at the RNA level or understanding gene regulation by methylation—in many individuals, both with and without disease.

In your lab, you look for genes implicated in certain diseases. How do you use computational biology to do that?

We examine the genomes of hundreds to thousands of people to identify which genes are enriched or deficient in patient groups compared with controls. We study not only the structural differences (the presence or absence of a specific gene and the number of copies of each) among genomes, but also the impact these structural differences have on gene function. We use both bioinformatics tools and computational biology methods to search for those differences between individuals at the DNA and RNA levels. Bioinformatics looks at huge volumes of data to distill its features in some essential ways. Computational biology requires us to search these data against a genetic model of disease so that relevant inferences can be made.

We used to do these studies one gene at a time. Now we do them an entire genome at a time. Pretty soon we will be doing these experiments at the level of individual DNA nucleotides. This increase in resolution has not only led to the boon in biological discoveries but changed the way we do science. In the past, most of our time was spent on doing data collection and little on interpretation. Now, it's the reverse. We spend days or weeks just gathering the data and then weeks or months doing intense analyses of the large data set. Often, we have to invent new analytical methods and tools, since the path is still not well trodden. New technologies (both genomic and computational) that allow us to sequence entire genomes at a much lower cost, or look at functional changes in gene activity, have rapidly altered what we can do in human disease discovery.

At the beginning of the Human Genome Project, some scientists said we would find nothing, and others said it would be boring or meaningless. It has been anything but. We find new treasures of biology every day.



What is your laboratory working on now?
My laboratory works on two classes of disorders: congenital and later-onset. In the first, we are researching two developmental disorders of the nervous system, one of the central nervous system (autism) and the other of the enteric nervous system (a rare intestinal disorder called Hirschsprung disease). Each of these diseases includes a variety of genetic aberrations that affects the development of neurons either in the brain (autism) or the gut (Hirschsprung disease). We have identified new genes involved in both. In both situations, we have identified only a minority of the genes, but already they have pointed out critical features of what goes wrong during development. Hopefully, this understanding will guide our thinking about future therapies.

In the second class of disorders, we study hypertension and sudden cardiac death. We are wondering why these cardiac problems are manifested later in life. Why aren't the genetic changes manifested earlier, like in the teen years? Why does it occur in some people at one age, and other people at another age?

Which of your findings are you most proud of?
Here are three: About 20 years ago, I contributed to the analysis that led to the precise mapping of the cystic fibrosis gene using methods that are commonplace today. Ten years ago, my group was the first to do large-scale studies on human genes to figure out how variable human gene sequences were: we showed 1 in 1,000 nucleotide differences in comparing any two human genomes, such as the two we receive from our parents. People had done it before, but not so extensively or accurately. We used microarray technologies that are, once again, commonplace today. Third, we were one of the first groups to show that we could identify functional changes outside the protein-coding part of the gene in human diseases: we identified an exceedingly common variant that predisposes its bearers to Hirschsprung disease.

The sequencing of the human genome is the basis for much of your work. What's the next big discovery that needs to happen to further the field?
I think the next ten years will be critical for us to understand the human genome's regulatory code, which determines gene expression, and how the genome specifies an organism. Humans have at least 200 cell types. How are they different, and how do these differences make them function in different ways? How do cells organize to make tissues—a whole heart, for example? Right now, genetics and cell biology are disconnected. I believe that genetics will play a crucial role in understanding the biology of cells so that we understand the functional connectivity all the way from DNA to RNA to proteins to organelles to cells to tissues. You can go to Google Maps and zoom in from miles above the earth to your own backyard. Biology is unable to do that today. We need an "app" for that!

What advice do you have for students interested in the field?
Many people believe that if they haven't taken enough rigorous courses in biology, they are out of the game. But biology is changing rapidly and what you learn in high school is a good start. By the time you get out of college, you'll need to learn so much all over again because our understanding is being altered not cumulatively but radically.

Genetics will remain challenging for a long time. If you need a bigger challenge, consider neuroscience *and* genetics. Figuring out the molecular and functional basis of neural growth, development, and action will be fun for a long, long time. Most importantly, find something you enjoy, and then you don't have to go to work a single day of your life. That's what I tell my children. i

What computational biologists do
Computational biologists use computer science, applied math, and statistics to address biological problems. With the completion of the Human Genome Project, they now can use computing methods to analyze billions of base pairs in human DNA. They also use computing to address other biological problems, such as how plants respond to changing environmental conditions.

Where they work
Computational biologists work in academia and in private and government research institutes.

Education required
Until recently, there were few graduate degrees in computational biology, which is why most people in the field have graduate degrees in computer science and/or biology. Colleges and universities now offer programs in computational biology and bioinformatics, and future practitioners will need graduate degrees in them to be competitive.

Job outlook
According to the Bureau of Labor Statistics, employment of biological scientists should increase much faster than the average for all occupations. And with the sequencing of the human genome, the quantity of DNA sequence data now available greatly exceeds the tools available to process them, making computational biologists especially in demand.

Salary range
A 2009 U.S. Bureau of Labor Statistics survey shows that the mean annual salary for biologists is $60,000. Salaries for computational biologists, however, can be much higher. An August 2010 search on SalaryList.com yielded results ranging from $45,000 to $199,000.

For more information
International Society for Computational Biology
www.iscb.org

American Institute of Biological Sciences
www.aibs.org

National Human Genome Research Institute
http://genome.gov