Michael Hoffman: Noisy Genomic Data, Made Clearer

December 2010/January 2011
By Tracy Vence

Type size:
SHARE
Email
Printer-friendly version
RSS Feed

Title: Senior fellow, University of Washington
Education: PhD, University of Cambridge, 2008
Recommended by: Francis Collins, National Institutes of Health

Michael Hoffman was a plant biochemist doing immunohistochemistry in Karen Browning's lab at the University of Texas at Austin when the *Arabidopsis* genome was published. Now a senior fellow in Bill Noble's group at the University of Washington, Hoffman is applying machine-learning techniques to genomic data sets generated as part of the National Human Genome Research Institute's ENCODE project.

As part of his work with ENCODE, Hoffman is refining a computational tool called Segway, "which does a simultaneous segmentation and clustering of functional genomics data" from ChIP-seq and other experiments, he says, "and tries to find patterns" within multiple data tracks. "It's really an attempt to see what you can find if you throw everything at a computer," he adds.

Hoffman's transition from biochemistry to computational biology, though not exactly simple, was facilitated in large part by the support of Ewan Birney, his PhD supervisor at the University of Cambridge. Birney "really encourages a lot of creativity in how you look at genomic problems," Hoffman says. That emphasis has paid off, especially as he's now applying electrical engineering and computer science techniques to analyze noisy biological data, he adds.

According to Hoffman, the current genome-wide screening technologies generate a fair amount of uncertainty. "You [can] sequence a few hundred base pairs … but it only tells you so much because you're excising that sequence from its neighborhood," he says. The more researchers understand about how the genome is regulated, the larger the role its neighborhood appears to play, he says. "Potentially, interactions with the other arm of the chromosome, or some other chromosome," could be functionally important.

Exacerbating the issue, Hoffman adds, is the fact that the majority of machine-learning techniques he uses were originally developed for natural language processing. "Computer scientists had a big advantage when developing these — they knew what the right answer was," he says. "A problem in genomics is that you never really know what the right answer is. The best you can hope to do is compare [your results] against what you already know."

Looking ahead

For Hoffman, finding the right answers is a race against time. He expects that over the next five to 10 years, the inundation of genomic information will only get worse. Wet lab researchers ought to learn how to perform their own computational analyses, "otherwise there will be a tremendous backlog of data. ... Right now I'm not sure there are enough bioinformatics geeks to go around," Hoffman says.

Publications of note

Hoffman says his best work to date appeared in *Genome Research* in 2010. In a paper he co-authored with Birney, Hoffman describes Sunflower, a package that models transcription factor binding and provides an "interesting ... look at the selective pressure that may have caused" binding competitions in the human genome in the past.

And the Nobel goes to ..

1. **GenomeWeb Daily News**
UPDATE: Danaher Buying Beckman Coulter for $6.8B
2. **GenomeWeb Daily News**
XDx, LabCorp Expand AlloMap Deal
3. **GenomeWeb Daily News**
AB Sciex, ARUP to Develop Hormone Analysis Methods
4. **GenomeWeb Daily News**
1000 Genomes Team Catalogs Structural Variants in Human Genomes
5. **GenomeWeb Daily News**
In Brief This Week: Beckman Coulter; Agilent; Avantra Biosciences, TGen; Eagle Genomics; Exiqon; and More

In this issue of Genome Technology

Conferences, Meetings & Deadlines

Advances in Genome Biology and Technology 2011 Annual Meeting
Feb 2-5 / Marco Island, Fla.
AGBT

1st IEEE International Conference on Computational Advances in Bio and Medical Sciences
Feb 3-5 / Orlando, Fla.
IEEE Computer Society

MSACL 2011
Feb 5-9 / San Diego
MSACL

Gene Expre
Technical C
Tech Guide

If the Nobel Prize committee chose to honor Hoffman, he hopes it'd be for teaching "a computer to understand genomic regulation with the same degree of accuracy as we can understand speech. If a computer could predict how genes were going to be regulated, or how a developmental program is organized, with [a high] level of accuracy, that would be really fantastic."

Related Stories

DNAnexus Adds Variant Analysis Tool, Joins PacBio Partner Program amid Mounting Interest in Cloud
February 4, 2011 / BioInform

Aurora Health Care Turns to Oracle to Mine Biorepository Data for Clinical Trials
February 4, 2011 / BioInform

Biogemma Invests in Compute Infrastructure to Address Analysis Challenges Posed by Plant Genomics
February 4, 2011 / BioInform

Illumina to Market GenoLogics' Geneus LIMS
February 4, 2011 / BioInform

Ion Torrent to Distribute DNAStar, Partek Software; Taps BioTeam for Support
February 4, 2011 / BioInform

9th Annual ASM Biodefense and Emerging Diseases Research Meeting
Feb 6-9 / Washington, D.C.
ASM

ASM Conference on Viral Genome Replication
Feb 6-9 / Banff, Canada
ASM

Abstract & Registration Deadlines

more

Do you think large labs tha translational r over smaller don't?

- Yes. All this enhancing tra work has me v the fate of my
- Maybe. I gu propose collal large labs just
- Maybe, but will likely work the interests c labs.
- No. NIH kno translational r powered by ba discoveries at small labs.

Vo
View R

Science	Business	Funding	Genome Technology Magazine
By sifting through genome data for nearly 200 individuals, researchers from the 1000 Genomes Consortium turned up tens of thousands of structural variants in the human genome, including nearly 2,000 changes affecting full genes or exons. The findings are helping pinpoint parts of the genome particularly prone to such variation.	Amid a restructuring program that will result in 4,800 job cuts, mostly in its pharma business, Roche reported that its molecular diagnostic sales increased 1 percent to CHF 1.19 billion ($1.27 billion) for 2010. Sales for its applied science business, which includes sequencing and microarray tools, were roughly flat year over year at CHF 868 million.	Demand by academic institutions in the US for life science tools has not weakened, with spending for such instruments increasing 13 percent year-over-year during the last three months of 2010, according to Leerink Swann. About $1 billion is being consistently disbursed each quarter, suggesting stimulus funding won't be exhausted until June 2012.	University of Michigan researchers uncovered familial relationships within HapMap3 populations. Using RELPAIR and allele-sharing analyses, the researchers confirmed the known relatives in the population and found 25 unreported parent-offspring pairs, 33 unexpected full sibling pairs, 118 unreported second-degree relative pairs, and five parent-parent-offspring trios.