QnAs with Donald Geman

Farooq Ahmed, *Science Writer*

With the proliferation of "omics" technologies, personalized medicine—which tailors treatment to an individual's genomic profile—promised a revolution in care. That revolution, says applied mathematician Donald Geman, has been slow to arrive. Geman has spent nearly four decades devising statistical methods for a variety of applications. He recently teamed up with an interdisciplinary group of scientists at The Johns Hopkins University, where he is a professor of applied mathematics and holds appointments at the university's Institute for Computational Medicine and Center for Imaging Science. Geman helped engineer an algorithm that reduces data complexity and may assist in differentiating between certain forms of cancer. This work builds on his earlier research in computer vision, leveraging his experience with pattern-recognition problems. PNAS recently spoke to Geman, who was elected to the National Academy of Sciences in 2015, about his current research.

Donald Geman. Image courtesy of Donald Geman.

PNAS: We have been hearing about personalized medicine for at least a decade. What are some of the challenges researchers face?

Geman: Personalized medicine hasn't yet reached a point where it is deployable on a mass scale, and there are many reasons for this. Genomic data are incredibly complex due to the interactions among gene products, as well as heterogeneity both within and between patients. All of it contributes to the challenge of moving personalized medicine forward.

Of course, when genome-wide omics data first became available, there were, perhaps, some irrational expectations about the level and timing of the impact on healthcare. And, there may also be many omics data tests that are currently working their way through the long clinical trial pipeline. On the plus side, the discovery of genetic variants has had a clinical impact, for example, in the treatment of single-gene disorders, such as in Huntington's disease.

PNAS: The statistical analysis of omics data are one of the keys to correctly interpreting the information generated by these new disciplines. How does the method described in your Inaugural Article (1) improve upon existing ones?

Geman: The goal of this paper was to radically simplify the data and shift the analysis from the population level to individual profiles. The reason for this shift is that we see an amazing amount of person-to-person heterogeneity in cancer genomics data,

This is a QnAs with a member of the National Academy of Sciences to accompany the member's Inaugural Article on page 4545.

Published online April 16, 2018.

www.pnas.org/cgi/doi/10.1073/pnas.1804880115

whether that's variability in mutations, transcription levels, or epigenetic states. Two patients, for example, may have the same cancer phenotypically but different types and levels of underlying cellular dysregulation.

To address these issues, we came up with the idea of generalizing divergence from baseline. It's the same concept as when you get a blood test: there's a range of normal for individual factors, and anything outside that range is considered abnormal. In cancer genomics, we can do the same by assigning a baseline range for gene expression or other variables. We define what's "normal" for any part of an omics profile and then binarize the data so that anything outside of the baseline parameter is declared "divergent" or "dysregulated." This helps us identify the pattern of dysregulation for each individual with a particular cancer phenotype.

PNAS: How does simplifying the data help with diagnosis?

Geman: Binarization allows us to more efficiently find the particular molecular variables and interactions that distinguish one cancer phenotype from another. In effect, we have massively reduced the space of potential biomarkers and prediction rules, which is very liberating from a statistical perspective. Moreover, the reduction of complexity might make it feasible to find predictors based on small subsets of mixed omics variables that are related mechanistically: for example, variables that cooperate in some aspect of gene regulation.

PNAS: What attracted you to cancer genetics?

Geman: I came to this work after many years in image analysis and computer vision, which I still think about. But I was not personally motivated by the applications for computer vision, like autonomous vehicles, automated manufacturing, and surveillance. On the other hand, I was very drawn to the challenge of applying statistics and machine learning to medicine. When I started, I had no idea how much I'd have to recalibrate and learn from conversations with people outside of my domain, people at the medical school at [The Johns Hopkins University], for example. But, there are links between computational vision and genomic medicine because both are pattern-recognition problems.

Maybe the biggest difference is the sample size. Compared with disciplines in which machine learning has been most successful, such as computer vision or speech recognition, the number of labeled samples in cancer genomics is extremely small. A small sample size combined with a huge number of features makes for a formidable technical challenge. In fact, we believe that today's tabula rasa learning, a method of learning from scratch, is not well-suited to understanding disease.

In computer vision, we use sequential adaptive testing, a method similar to decision trees or the once popular 20 questions game played by children. I have always been amazed by the power of 20 questions, by how fast we can nail something down if we ask the right questions in the right order. It works very well in practice, and the decision making is transparent, which is another plus in cancer genomics or any computational biology field. We give an example of a decision tree for discriminating between two cancer subtypes in the Inaugural Article (1).

PNAS: As a mathematician who has explored a variety of problems in your career, what have you experienced working across disciplines?

Geman: It is hard! In general, data scientists do not adapt their strategies to the specific learning scenarios they encounter in other areas, including molecular medicine. They come with a set of tools and often have only a secondary interest in the applications; relatively few know or care about the underlying mechanisms. Conversely, biologists and physicians are handicapped because they're unfamiliar or even mystified by the mathematical methodology: the data science and statistics.

I don't think it's possible to move forward without working in a group where each discipline keeps the other's feet on the ground. Our group brings together people from across disciplines: oncologists, molecular biologists, computer scientists, statisticians, and mathematicians. I find that it's the only way to make any progress and to make sure that we're working on real problems.

[1] Dinalankara W, et al. (2018) Digitizing omics profiles by divergence from a baseline. *Proc Natl Acad Sci USA* 115:4545–4552.