Biggest beast in big data forest? One field's astonishing growth is 'genomical'

July 7 2015

Who's about to become the biggest beast in the Big Data forest? A group of math and computing experts have arrived at what they say is a clear answer. It's not You Tube or Twitter, two social media sites that gobble up awesome quantities of bandwidth and generate hard-to-grasp numbers of electronic bytes every day. And it's not astronomy or particle physics, two of the highest-tech sciences that have long been at the leading edge of data generation and processing.

No, the alpha beast in the Big Data forest, the experts say in the July 7 issue of *PLOS Biology*, turns out to be genomics—a science that didn't exist 15 years ago and is only now just beginning to break out from the field to generate the most electronic bytes per year relative to all other fields. Recognizing that it is about to accelerate away from other formidable data hogs, say the experts, is a necessary first step in a grand-challenge problem - figuring out how to capture, store, process and interpret all that genome-encoded biological information, stripped down to symbolic and, by themselves meaningless, ones and zeros.

"For a very long time, people have used the adjective 'astronomical' to talk about things that are really, truly huge," says Michael Schatz, an associate professor at the Simons Center for Quantitative Biology at Cold Spring Harbor Laboratory (CSHL) and a co-author of the PLOS paper. "But in pointing out the incredible pace of growth of data-generation in the biological sciences, my colleagues and I are suggesting we may need to start calling truly immense things 'genomical' in the years just ahead."

All of the fields the team compared, from social media on the Internet to astronomy, are generating huge quantities of electronic data - on the order of tens to hundreds of petabytes per year. A petabyte is one quadrillion bytes - 10 followed by 15 zeros; it's 1000 times more bytes than a terabyte, the amount of storage you might have on your home computer. And, says the team - which is also composed of a number of data experts from the University of Illinois at Champaign-Urbana - all of the fields are on rapidly upward-sloping growth curves. YouTube actually generates the most data right now, about 100 petabytes a year. But genomics is not far behind and growing much more rapidly. At the current rate, the quantity of genomics data produced daily is doubling every 7 months. By 2025, that figure will range between 2 and 40 exabytes per year, the team estimates, depending on the rate of doubling. One exabyte is the equivalent of 1000 petabytes, about a million times more data than you can store on your home computer.

Schatz and colleagues describe genomics as a "four-headed beast." They refer to the separate problems of data acquisition, storage, distribution and analysis. Like data that flows over the Internet, biological data that is the raw material of genomics is highly distributed. That means it's generated and consumed in many locations. Unlike Internet data, however, which is formatted according to a few standard protocols, genomic data is compiled in many different formats, a fact that threatens its broad intelligibility and utility.

This problem grows in importance as the quantity of data increases. As Schatz explains, much of the torrent of big data from biology will take the form of human genome sequences, as well as related medical information that also depends on sequencing technology. This related information takes the form of both snapshots and the equivalent of movies, and concerns, for instance, levels of gene messages, or transcripts, in specific tissue samples, as well as the identity and levels of protein in samples.

If all the human sequence data so far generated were put in a single place - about 250,000 sequences—it would require about 25 petabytes of storage space. That is a manageable problem, Schatz says. But by 2025, the team expects as many as 1 billion people to have their full genomes sequenced (mostly, people in comparatively wealthy nations). This poses an exabyte-level storage problem.

At some point, sequences in full may not need to be stored. In particle physics, data is read and filtered as it is generated, greatly minimizing storage requirements. But this parsing is not entirely practical for biological information, mainly because the question of which sequences can be safely thrown out is much harder to decide. Conceivably, a billion sets of individual data will need to be preserved if it is to be an aid to future physicians.

Schatz is especially interested in the problem posed by obtaining hundreds of millions, even billions of human full-length genome sequences. The problem is not really speed, which will grow rapidly and predictably, he says, but rather in figuring out how to align and represent different genomes so that they might be compared - and compared in very efficient, smart ways.

"The point of sequencing a billion genomes is not really to make a billion separate lists saying, 'If you have these variants, you have the following risks.' Of course, individuals will want to look at the list of DNA variants they possess. But the real power of having 1 billion human genomes comes from ways of comparing them and combining layers of analysis. Our belief is, by combining all this information, patterns will emerge - in the same way that when Mendel grew tens of thousands of pea plants, at the dawn of genetics 150 years ago, he was able to formulate laws of inheritance by looking a patterns for how specific traits were inherited."

"Genomics is a game-changing science in so many ways," Schatz says.

"My colleagues and I are saying that it's important to think about the future so that we are ready for it."

More information: "Big Data: Astronomical or Genomical?" appears July 7, 2015 in *PLOS Biology*.

Provided by Cold Spring Harbor Laboratory