NATURE | NEWS

Spiking genomic databases with misinformation could protect patient privacy

Technique that adds noise to genetic data would enable much faster access to large data sets.

Anna Nowogrodzki

15 August 2016

Large genomic databases are indispensable for scientists looking for genetic variations associated with diseases. But they come with privacy risks for people who contribute their DNA. A 2013 study[1] showed that hackers could use publicly available information on the Internet to identify people from their anonymized genomic data.

To address those concerns, a system developed by Bonnie Berger and Sean Simmons, computer scientists at the Massachusetts Institute of Technology (MIT) in Cambridge, uses an approach called differential privacy. It masks the donor's identity by adding a small amount of noise, or random variation, to the results it returns on a user's query. The researchers published their results in the latest issue of *Cell Systems*[2].

The system calculates the statistic that researchers want — such as the chance that one genetic variation is associated with a particular disease, or the top five genetic variations associated with an illness. Then it adds random variation to the result, essentially returning slightly incorrect information. For example, in a query for the top five genetic variations associated with a disease, the system might yield the top four genetic variations and the sixth or seventh variation.

The user would not know which of the results to their query is more correct than another, but they could still use the information. It would just be much harder for someone to work out the patient information behind the data.

“When you induce a little noise in the system, in many ways it’s not that different from noise in the data to begin with,” says Bradley Malin, a computer scientist at Vanderbilt University in Nashville, Tennessee. “It is reliable to a certain degree.” The US Census Bureau and US Department of Labor have been adding noise to their data in this way for decades, he says.

Faster access

The privacy of individuals in a data set employing this technique remains intact as long as the database is big enough — containing information from a few thousand individuals or more — and if researchers stay within their 'privacy budget', which limits the number of questions they can ask. Users would not be able to ask about hundreds or thousands of locations in a genome.

A database protected by this technique could be instantly searchable. Currently, getting permission to access databases administered by agencies including the US National Institutes of Health can take months.

Simmons and Berger say that even with the noise, the system's answers will be close enough to be useful for asking a few targeted questions. “It's meant to be used to get access to data sets that you might not have access to otherwise,” says Simmons.

For example, if researchers analysing a small data set found a genetic variation associated with a disease, this system could allow them to verify that association using a much larger data set that they otherwise couldn’t access. It could also let researchers preview a data set to determine its usefulness before going through a time-consuming application process for full access.

That queasy feeling

“I think it's a really excellent mathematical work,” says Yaniv Erlich, a computational biologist at Columbia University in New York City. “It’s nice on paper. But from a practical perspective I'm not sure that it can be used.”

One of his concerns is the system’s question limitation. What researchers want these days is to examine the top 10 or top 100 genetic variations associated with a disease, Erlich says, not just 5.

Also, “people don’t like to put noise in their data” because a lot of hard work goes into generating the information, Erlich says. The noise issue could also have troubling implications for clinical decisions based on such information.

Malin adds that there is a very small probability that the system would introduce a large amount of noise in answer to a query. “That’s what makes people a little queasy.”

But Simmons is trying to improve the system, attempting to add less noise while achieving the same privacy. And Berger is working with the Broad Institute of MIT and Harvard in Cambridge to determine ways of decreasing privacy risks, possibly by using differential privacy techniques. This would be useful if the institute decided to release aggregate genomic data from its databases more widely.

“In the end that’s what we really care about,” Simmons says, “making this data as widely accessible as possible.”

Nature | doi:10.1038/nature.2016.20407

References

1. Gymrek, M., McGuire, A. L., Golan, D., Halperin, E. & Erlich, Y. *Science* **339**, 321–324 (2013).
2. Simmons, S., Sahinalp, C. & Berger, B. *Cell Systems* **3**, 54–61 (2016).