MedicalXpress

Protecting privacy in genomic databases

August 9 2016, by Larry Hardesty

Researchers from MIT's Computer Science and Artificial Intelligence Laboratory and Indiana University at Bloomington describe a new system that permits database queries for genome-wide association studies but reduces the chances of privacy compromises to almost zero. Credit: Christine Daniloff/MIT

Genome-wide association studies, which try to find correlations between particular genetic variations and disease diagnoses, are a staple of

modern medical research.

But because they depend on databases that contain people's medical histories, they carry privacy risks. An attacker armed with genetic information about someone—from, say, a skin sample—could query a database for that person's medical data. Even without the skin sample, an attacker who was permitted to make repeated queries, each informed by the results of the last, could, in principle, extract private data from the database.

In the latest issue of the journal *Cell Systems*, researchers from MIT's Computer Science and Artificial Intelligence Laboratory and Indiana University at Bloomington describe a new system that permits database queries for genome-wide association studies but reduces the chances of privacy compromises to almost zero.

It does that by adding a little bit of misinformation to the query results it returns. That means that researchers using the system could begin looking for drug targets with slightly inaccurate data. But in most cases, the answers returned by the system will be close enough to be useful.

And an instantly searchable online database of genetic data, even one that returned slightly inaccurate information, could make biomedical research much more efficient.

"Right now, what a lot of people do, including the NIH, for a long time, is take all their data—including, often, aggregate data, the statistics we're interested in protecting—and put them into repositories," says Sean Simmons, an MIT postdoc in mathematics and first author on the new paper. "And you have to go through a time-consuming process to get access to them."

That process involves a raft of paperwork, including explanations of how

the research enabled by the repositories will contribute to the public good, which requires careful review. "We've waited months to get access to various repositories," says Bonnie Berger, the Simons Professor of Mathematics at MIT, who was Simmons's thesis advisor and is the corresponding author on the paper. "Months."

Bring the noise

Genome-wide association studies generally rely on genetic variations called single-nucleotide polymorphisms, or SNPs (pronounced "snips"). A SNP is a variation of one nucleotide, or DNA "letter," at a specified location in the genome. Millions of SNPs have been identified in the human population, and certain combinations of SNPs can serve as proxies for larger stretches of DNA that tend to be conserved among individuals.

The new system, which Berger and Simmons developed together with Cenk Sahinalp, a professor of computer science at Indiana University, implements a technique called "differential privacy," which has been a major area of cryptographic research in recent years. Differential-privacy techniques add a little bit of noise, or random variation, to the results of database searches, to confound algorithms that would seek to extract private information from the results of several, tailored, sequential searches.

The amount of noise required depends on the strength of the privacy guarantee—how low you want to set the likelihood of leaking private information—and the type and volume of data. The more people whose data a SNP database contains, the less noise the system needs to add; essentially, it's easier to get lost in a crowd. But the more SNPs the system records, the more flexibility an attacker has in constructing privacy-compromising searches, which increases the noise requirements.

The researchers considered two types of common queries. In one, the user asks for the statistical correlation between a particular SNP and a particular disease. In the other, the user asks for a list of the SNPs in a particular region of the genome that correlate best with a particular disease.

In the first case, the system returns a widely used measure of correlation called a p-value. Here, the p-value would be modified—augmented or reduced by some random factor—in order to ensure privacy.

In the second case, the system has some chance of returning not the top-scoring SNPs in a given region, but several of the top-scoring SNPs and maybe one or two lower-scoring ones. To calculate the probability that a given SNP will make it into the results, the researchers use a measure called the Hamming distance, which indicates how far away a lower-scoring SNP is from the one that it's replacing. This turns out to yield more useful results than relying on the p-value. Finding an efficient algorithm for calculating Hamming distances on the fly is one of the system's chief innovations.

Ironing out differences

The other is that the system corrects for a problem common in population genetics called population stratification. "The standard example is that a particular SNP is closely linked to being lactose intolerant," Simmons explains. "Let's say that people in East Asia are more likely to be lactose intolerant than someone in, say, Northern Europe. But also Northern Europeans tend to be taller than people from East Asia. A naive method would suggest that this particular SNP has an effect on height, but it's really a false correlation."

The researchers' algorithm assumes that the largest variations in a given population are the results of differences between subpopulations, filters

those differences out, and hones in on the ones that remain.

"Since Homer's attack in 2008, the biomedical community has been debating to what extent and to whom genomic and phenotypic databases should be made accessible," says Jean-Pierre Hubaux, a professor of computer science at the École Polytechnique Fédérale de Lausanne, referring to a paper by Nils Homer, then a graduate student at the University of California at Los Angeles, on determining whether a given person's genetic data is present in a database. "In parallel, Cynthia Dwork and other computer scientists have developed the concept of differential privacy, the theory of which is now well-understood. The authors of this paper make a crucial contribution, because they provide concrete examples of how differential privacy can be used to protect the privacy of genome-wide association studies in heterogeneous human populations. Hopefully, this will encourage the biomedical community to test this promising approach at large scale and, if it's successful, define best practices and develop related tools."

More information: Sean Simmons et al. Enabling Privacy-Preserving GWASs in Heterogeneous Human Populations, *Cell Systems* (2016). DOI: 10.1016/j.cels.2016.04.013

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology