CSIRO

Australian e-Health Research Centre

Genomator: creating synthetic data based on logic-solving

We've developed an SAT-solver based algorithm to generate synthetic data that is hallucination-resistant and provable private

The challenge

Synthetic data is used in clinical applications to increase representation of rare events for machine learning, make data shareable without privacy implications, or create unique test cases for quality control.

However, current computational approaches for creating synthetic representations from real data have limitations that hamper their applicability. Specifically, statistical methods like Markov Chains rely on haplotype-shuffling, which risks of inadvertently exposing genomic segments that hold sensitive or personally identifiable information of the participants.

Similarly, Artificial intelligence (AI) approaches such as RBM or GAN generate novel sequences that can contain variant combinations that are potentially not viable in humans, i.e. 'hallucination'.

Our response

We developed Genomator, a SAT-solving based approach, whose logic-solving engine prevents 'fake' data generation (hallucination-free) with the capability to generate provably private data.

SAT solvers are a class of algorithms used to solve problems by (SAT)isfying constraints between Boolean variables. This ensures the efficient deductive construction of synthetic genomes from input data that ensures no unrealistic variant combinations are created and no rare variant combinations are leaked, as this can compromise privacy of participants.

As SAT-solving is a reversible operation, we created Reverse Genomator, a tool for determining if an individual's information was used to create synthetic data, which, in-turn, can be used to quantify privacy absolutely. Together with the observation that there is a trade-off between accuracy and privacy, Genomator has a slider that can produce data sets that cover the full spectrum, from creating highly accurate cohorts for rare disease research to creating provable private representations of marginalized populations.

Genomator creates de-novo genomic sequences that are 70% more accurate and are scalable to the three billion letters of the human genome.

Benefits

Genomator is the first method that can tailor the accuracy-privacy trade-off while being scalable to whole genome sequencing data.

Genomator requires no training time, and has been shown to replicate the structure of multiple genomic populations with accuracy.

In the clinical setting, Genomator was shown to create synthetic data that preserve population-specific pharmacogenomic markers.

Updates at: bioinformatics.csiro.au/Genomator

Janet Fox, AEHRC Business Development Manager
+61 7 3253 3646 | +61 466 779 797 | janet.fox@csiro.au | aehrc.csiro.au

B&M | 24-00284