Privacy and Sharing of Genomic Data

Mario Südholt
IMT Atlantique

Séminaire Cybersécurité
IMT, 10 nov. 2017
Objective: analyze human genome for diseases

Genomic Wide-Association Studies (GWAS)
Associations between genetic variations and specific traits
- Ex.: BRCA genes and risk of breast cancer

Sharing of aggregate data
- Simple client-server architectures
- Initially: no privacy problems known

© Pasieka Science Photo Library
Attacks for identification of individuals in genetic DBs with aggregate data

[Homer et al., 2008]: identification in large aggregate data sets
[Sankaraman et al., 2009]: upper bound on detection power
[Wang et al., 2009]: identification in small data sets

Result: severe restrictions to public data sharing

Ex.: gwascentral.org
  - frequency info not available
  - Large data sets only available on request
    (“Data Sharing Statement” of GWAS Central)

Generally: restrictions on
  - Sharing system architecture
  - Queries on genomic DBes
Need for more advanced sharing

Geneticians interested in more advanced sharing possibilities
Allow sharing of larger data sets
Enable collaborative work on rare variants / uncommon diseases

Advanced sharing architectures
Sharing of raw data via e.g. trusted party
Quicker access to data in the cloud

Use advanced sharing techniques
Support confidentiality and integrity efficiently
Support for ownership and traceability properties
How to support such sharing scenarios?
Methods for the construction of architectures/processes/queries
• Means for design and programming
• Ensure basic privacy guarantees
• Optimize architectures/applications
• Formal verification of advanced privacy properties

How to ensure privacy properties?
Multiple privacy enhancing techniques:
• Encryption: a/symmetric, homomorphic, attribute-based, …
• Client-side computing: compute associations within local perimeter
• Data fragmentation
• Watermarking for ownership and traceability properties
COMPOSITIONAL CONSTRUCTION OF GENOMIC APPS

Declarative scenario definitions with privacy types

Definition of watermarking operator

Verific./Optim. of genetic applications for privacy/efficiency

Algebraic theory: watermarking laws

\[
\begin{align*}
\text{decrypt}_{(s,a)} \circ \text{crypt}_{(s,a)} \circ \text{detect}_{(s,a)} \circ \text{watermark} & \equiv \\
\text{detect}_{(s,a)} \circ \text{decrypt}_{(s,a)} \circ \text{crypt}_{(s,a)} \circ \text{watermark}
\end{align*}
\]

if \( \text{dom}(p) \cap a = \emptyset \)

\[
\text{detect}_{(s,a)} \circ \sigma_p = \sigma_p \circ \text{detect}_{(s,a)}
\]
3. CHALLENGES

- How to harness/integrate other PETs (e.g., differential privacy)
- Which kind of genetic data and analyses can be safely outsourced?
- What about new analyses?