Position Paper

A Taxonomy of Genomic Privacy and Beyond

Kay Hamacher
Computational Biology & Simulation
Depts. of Biology, Computer Science, Physics
Technical University Darmstadt
`http://www.kay-hamacher.de`

1. Introduction

'Genomic Privacy' has gained traction in present research, both in computer science - as is evident by this very workshop (GenoPri) and a previous Dagstuhl seminar [1] -, and in life sciences themselves [2,3]. While at present most research addresses single issues, we still lack a coherent and cohesive framework to classify emerging problems and potential (technical) solutions. This position paper tries to give a first overview and taxonomy/classification. The classes are not mutually exclusive; sometimes they have a large overlap, but still a noticeable number of variation. The author proposes that – in the light of previous work on *general* privacy classification schemes [4,5] that focused on the threats – the uniqueness of genomic (non-mutable, very specific) and similar data that changes (epi-genetic, physilogical, tissue data), as well as the concrete questions and research protocols in the life sciences demand a domain-specific taxonomy that is 'orthogonal' to these previous approaches and should be seen as supplemental – most of all, as a threat-based classification will not be applicable due to unknown, potential threats to privacy due to the ever improving bioinformatics tools that might give new and unforeseeable insight and interpretation to genomic data.

2. Taxonomy

Here, we take an orthogonal approach to the, e.g., taxonomy of Solove [4] for “classical data” which is purely process orientated. Due to the uniqueness of genomic (and epigenetic) data, we argue that a taxonomy has to be solely “data type orientated” as we cannot – at present – foresee future processes and workflows in bioinformatics and biostatistics.

2.1. Annotation Data

Clearly, genomic data poses a privacy problem. But the sole genomic sequence is not the biggest problem: its annotation by, e.g., disease markers, the correlation with other sequences (which can be seen as a special case of annotation), or the predictive power of biostatistical models based on genomic data renders the privacy problem an all encompassing challenge. Therefore, one has to see the genomic sequence or portions thereof in its context, that is its annotation by name, DoB, SSN, medical history, expression levels in micro-arrays, blood or tissue samples etc.

2.2. Type of Data

The annotation issue is related to the type of data stored: is it the full genome (discrete alphabet)? Portions of a whole genome, probably just a few SNPs (or any other enumerable sample)? Is this data (at present) stored along and therefore linked to physiological data, such as mRNA expression levels (continuous values)? While the annotation of data correlates sequences to *interpretation* and *meaning*, the data type itself sets *boundaries to technical solutions* and therefore needs to be taken into account, too.

2.3. Underlying Application or Research Question

The genomic sequence itself and its annotations are – in turn – largely determined by the underlying usage. Here, a fundamental difference is noticeable: application in diagnosis, treatment, drug development etc. versus basic research as done in, e.g., biobanks. While the diagnosis *applies* models, epidemiology and biostatistics *develop* models themselves. The data flow is accordingly very different. Data sharing among commercial entities is quite different due to the abuse potential than in public research, which comes with other privacy issues (e.g. the tension between privacy and good scientific practice to publish raw data).

2.4. Time Horizon: Duration of Storage and Retrieval

The application determines another dimension in which privacy issues are involved: how long is the data accessible and/or stored. For simple and repeated tests on single biomarkers, such as phenotype related SNPs, the storage period might be just some weeks, while whole genome data can be expected to be stored indefinitely.

2.5. Beneficiary of Analysis, Modeling, or Research

The usage and the implied storage mode and period is influenced - in turn - by the entity that has interest in the data and its analysis. This could be the general public or an epidemiologist who needs to correlate several genomes, even hundreds to learn new mechanisms or emerging health issues; this might be a pharmaceutical company which needs to develop models for genome specific therapies in personalized medicine, or a couple of parents with a desire to learn more about heritable diseases. The motivation for genomic analysis and implied data handling and analysis is obviously relevant in any assessment of privacy threats due to potential trade-offs and opportunity costs.

2.6. Classification by Externalities: Technical vs. Non-Technical

Genomic sequences, models based on them, and analysis thereof do not exist in an isolated universe; on the contrary, any analysis must also be understood with regard to the environment of actors, beneficiaries, and other parties involved. The concept of externalities necessarily has to be broad and somewhat imbalanced as it “collects” various stakeholders and their interests.

2.6.1. Technical

Here, we need to distinguish between a) privacy ensuring or enabling technologies in genomic privacy, such as secure computations [6] and other techniques [7,8], and b) the technologies required and applied for the bioinformatics analysis purpose. The latter imply requirements on privacy-by-design-mechanism that are frequently discussed. E.g., while simple SNP-analysis boils down to string matching and comparison, more involved analysis of, e.g., cancer based on systems biological models [9] involve very sophisticated models of machine learning and simulation.

2.6.2. Organizational

Typically, in basic research such as biobanks and beyond several data handling entities are involved and need to exchange data. Is the cooperation based on NDAs? How are mechanisms codified, if at all? These questions also largely influence potentials problems to (individual) privacy.

2.6.3. Legal

Closely connected, is the legal setting in which the above mentioned entities and the patients act. While harmonization is under way for business data and even data with sensitive content (such as social network data), still some country-specific regulations exists for genomic data that are in conflict to the ones in other countries.

2.6.4. Ethical & Political

While the legal framework is an expression of ethical and political concerns, in today's fast changing scientifc world, they lag behind the current development. Therefore, one should – when discussing genomic privacy – also take into account the overall ethical and political attitude to anticipate what might be codified in the law one day.

2.6.5. Economical

Major players in the field of genomic analysis are a) companies, b) basic research institute, c) entities such as governmental agencies which could license/sell data, and d) finally individuals who decide to follow economic incentives. Therefore, every potential solution needs also to consider the economics of solutions.

3. Outlook

This position papers wants to encourage discussion about classification of privacy problems and solutions. From my point of view, it would be most desirable to “ground” the current discussion and provide for a platform on which we can asses (technological) solutions in regard to their efficiency, applicability, and open questions that should be addressed as soon as possible before potential lock-in effects take hold in genomic privacy as in other areas.

4. References

[1] Hamacher, Hubaux, Tsudik. *Genomic Privacy* (Dagstuhl Seminar 13412)}}, Dagstuhl Reports (ISSN 2192-5283) vol.3, pp. 25-35, 2014. http://drops.dagstuhl.de/opus/volltexte/2014/4426/

[2] Li, Zou, Liu, Peng, Chen. *New threats to health data privacy*, BMC Bioinformatics 12(2011)S7.

[3] Lauss, Schröder, Dabrock, Eder, Hamacher, Kuhn, Gottweis. *Towards Biobank Privacy Regimes in Responsible Innovation Societies*, Biopreservation and Biobanking, 11:319-323, 2013.

[4] Solove. *Taxonomy of Privacy*. University of Pennsylvania Law Review, Vol. 154, No. 3, p. 477, January 2006; GWU Law School Public Law Research Paper No. 129

[5] Paintsil, Fritsch. *A Taxonomy of Privacy and Security Risks Contributing Factors* .in Privacy and Identity Management for Life , Fischer-Hübner, Duquenoy, Hansen, Leenes, Zhang (eds.), Springer Berlin Heidelberg, pp. 52-63, 2011

[6] Franz, Hamacher, Jha, Katzenbeisser, Schröder. *Secure Computations on Non-Integer Values with Applications to Privacy-Preserving Sequence Analysis*, Information Security Technical Report 17(3):117-128, 2013.

[7] Wang, Wang, Li, Tang, Reiter, Dong. *Privacy-Preserving Genomic Computation Through Program Specialization.* Proceedings of the 16th ACM Conference on Computer and Communications Security (CCS), 2009.

[8] Ayday, Raisaro, Laren, Jack, Fellay et al. *Privacy-Preserving Computation of Disease Risk by Using Genomic, Clinical, and Environmental Data*. USENIX Security Workshop on Health Information Technologies (HealthTech '13), Washington, USA, 2013.

[9] Hornberg, Bruggeman, Westerhoff, Lankelma. *Cancer: A Systems Biology disease*, Biosystems 83(2006):81-90.