Malin B. Why methods for genomic data privacy fail and how we can fix it. For circulation at the AAAS Annual Meeting. Seattle, WA, Feb. 2004.

Why Methods for Genomic Data Privacy Fail and How We Can Fix It

Bradley Malin
Data Privacy Laboratory, School of Computer Science
Carnegie Mellon University, Pittsburgh, PA 15213-3890
malin@cs.cmu.edu

The increasing integration of patient-specific genomic data into clinical practice and research raises serious privacy concerns. Various privacy enhancing techniques that "de-identify" the data through the removal of explicit identifiers, such as name or Social Security number, have been proposed and deployed. While advocates of these systems have the best of intentions, they are fundamentally flawed, due to a lack of formal modeling and proofs of privacy. These systems fail to account for information that can be inferred from genomic data, as well as, the environment into which data is shared. This research addresses the extent to which these systems are susceptible to computational "re-identification" attacks. The attacks we employ exploit residual information in genomic data and relates it to explicit identity. Though susceptibility varies, each of the protection methods studied is deficient in their protection against re-identification. Our findings stress the need for genomic data privacy protection methods that allow for provable guarantees of privacy.

Affiliated with the sharing or disclosure of person-specific health information is a serious concern for an individual's privacy. In the past, such data was stripped of explicitly identifying information (e.g. name, residence, Social Security number, etc.). However, such *ad hoc* methods of "de-identification" were shown to be insufficient for protecting the anonymity of the data subjects. This is because, though explicit identity is obscured, there remain many features, such as demographics, that can be uniquely "re-identified" to named publicly available records. To combat unintended disclosures, technological and computational methods for providing privacy have evolved from simple "well, it looks anonymous" de-identification methods to formal computational models amenable to logical proofs of privacy. [11, 13, 14]

Currently, the biomedical world finds itself in the midst of a genomics revolution. The genomic data of an individual is increasingly being collected, stored, and shared in both the research and clinical environment. Genomic data provides opportunities for health care that until recently were severely limited or nonexistent. For example, as a diagnostic tester of certain diseases, early confirmation can initiate lifesaving treatment, raise the standard of living, and help facilitate in family planning. Beyond gross diagnostics, there is gathering evidence that suggests variation in our genome influences our body's susceptibility to disease and ability to metabolize drugs.

At the same time, genomic data poses complex privacy problems. Many people fear that information gleaned from their genome will be misused, abused, or result in social stigma. [12] For individuals afflicted with particular diseases, diagnostic confirmation provides little hope or comfort because no cure or proven treatments exist. Moreover, an individual's genomic data, unlike much standard clinical information, retains specific information about family relations. As a result, there are many social and legal pressures to protect the privacy of an individual's genomic status. Without proper guarantees of anonymity, not only will patients be less willing to provide data, but many data collectors will be unable to share genomic data for worthwhile endeavors. In recognition of this situation, the genomic privacy is considered one of the major challenges for the biomedical community. [1]

Privacy protection methods for genomic data must address the question, "How can an individual's DNA be separated from explicit identity, such that the relationship can not be established without permission?" Despite the fact that the human genome consists of over 3 billion base pairs, the variation of which can uniquely characterize an individual, there exists no central registrar that maps genomes to named identities. So, genomic uniqueness is not sufficient criteria for revealing identity. Over the past several years, many proposed, and deployed, protection techniques have been based on this premise. However, these systems neglect the aforementioned research in computational privacy in favor of more *ad hoc* methods of simple de-identification and strong security protocols. It is assumed an adversary must “crack" the encryption or steal the encryption keys in order to learn the identity of a DNA sequence. If this is impossible, then the protocol must protect privacy.

The previous is a naïve and dangerous view of privacy protection. Privacy can not be protected by security procedures alone. Such claims of anonymity are fundamentally flawed for several reasons.

First, discussions about the protection capabilities of a particular technique fail to model social or environmental factors about which data is shared. For example, a schema that accounts for the protection of identities from a single institution can fail when multiple institutions are releasing data. We have demonstrated exactly how this can occur with our model of “trail re-identification”. [8] In short, when an individual visits multiple locations, de-identified genomic data and identifiable clinical data can be linked through patterns in the sets of locations visited.

Second, these methods fail to account for residual information that can be inferred from genomic data. The ability to infer identifying features from genomic data is exemplified by our prior research into genotype-clinical phenotype relations. At a first order level, there exist an increasing number of clinical features that are directly dependent on DNA sequence variation. [7] From a more fine-grained perspective, we developed a general model with the capability to learning patient-specific genomic data from publicly available longitudinal medical information. [9] The model relates a disease’s symptoms to particular clinical states of the disease. Appropriate weighting of the symptoms is learned from observed diagnoses to subsequently identify the state of the disease presented in hospital visits. Currently, this approach is applicable to any simple genetic disorder with defined clinical phenotypes. The efficacy of our model was demonstrated by inferring specific DNA mutations of clinically positive Huntington’s disease patients.

In order to better understand the state of genomic data privacy protection systems, we recently performed a general system susceptibility analysis. Specifically, five published privacy protection systems for biological/genomic data were studied. Several of the more sophisticated techniques (deCODE, Gent) advocate the use of pseudonyms to protect privacy. [3, 4] In general terms, pseudonymization converts the explicitly identifying features of an individual into an encrypted or random value. Other systems (Quebec) utilize denominalization, which separates genomic data from nominal, named or familial, information. [5] In addition, others (De-ID) use simple methods of de-identification of identifiers, sometimes accompanied with the use of a random ID number, as an identifier. [2, 15]

Each of the system designs was tested against known re-identification attacks. Details of the system analyses can be found in [10]. The first test (Family), determines whether or not family structures (i.e. pedigree information) can be inferred from the data. Such information can be linked to identified genealogical information that can be constructed from public records. The second test (Trail), assesses if the data is potentially traceable over multiple locations, and thus susceptible to trail analysis. The third test (Gen-Phen) concentrates on inferences that can be made from the genomic data itself, and ascertains if it is relatable to physiological traits. Finally, the fourth test (Dictionary) is a cryptanalysis trick based on the dictionary attack. When pseudonyms are based on demographic or known features of an individual, then the encryption itself may be susceptible to further attacks.

In Table 1, we report a general overview of the susceptibility of the privacy protection systems. None of the systems studied are impervious to re-identification. However, it is interesting to note that the overall susceptibility is different for each system. Based on the analyses above, it can be concluded that pseudonymization and naïve de-identification strategies are not sufficient mechanisms for the protection of identities.

	Protection System			
Attack	**deCODE**	**Gent**	**Quebec**	**De-ID**
Family	Yes	No	Yes	Yes
Trail	No	Yes	No	Yes
Gen-Phen	No	Yes	Yes	Yes
Dictionary	Yes	Yes	No	No

Table 1. General susceptibility of privacy protection models to re-identification attacks.

Yet, this finding does not imply that pseudonyms and third party solutions are worthless in the pursuit of genomic data privacy protection. Rather, to an extent, these systems do provide certain privacy protections. For instance, pseudonyms serve as a first-order protector and deterrent. It is conceivable that an adversary, who approaches re-identification in a non-computational manner, will be deterred by the simple obscuring of explicitly identifiable information. In addition, datasets devoid of linkage capabilities severely limit the types of research that can be performed. It is often the case where researchers may need to request additional information about a subject. From another point of view, a subject may wish to remove their data from a research study or find out information about how their data is being handled. In this respect, pseudonyms provide an extremely valuable service by accounting for future research, applications, and auditing capabilities that would be virtually impossible to handle without a linkage mechanism.

And yet, something must be done to protect the identities of the data subjects. This research is a call to arms for the biomedical community. The next generation of methods must account for multiple environments of data sharing, as well as various types of inferences that can be made from the shared data. Furthermore, these methods must be developed in a more scientific and logical manner, such that formal proofs about the protection capabilities and limitations afforded by the specific method can be constructed. Though proofs may be difficult to derive in the face of uncertainties about the sharing environment, especially when the data itself holds latent knowledge to be learned at a later point in time, researchers can validate their approaches experimentally against known re-identification attacks, such as those discussed above.

On the flipside though, researchers should not remain content with their proofs and experiments. New re-identification attacks will be developed by those in the academic community, as well as by the adversaries that reside outside of the public realm. As such, researchers must continue to innovate and develop new methods re-identification for testing their protection techniques. These methods may be new types and more robust forms inferential or location-based techniques or completely new models that have yet to be discovered. Regardless, without the development of new protection and re-identification methods, researchers will continue to rely upon untested and possibly dangerous methods of privacy protection.

The development of new identity protection strategies is paramount for continued data sharing and innovative research studies.

References

[1] Altman RB and Klein TE. Challenges for bio-medical informatics and pharmaco-genomics. *Annu Rev Pharmacol Toxicol*. 2002; 42: 113-133.

[2] Burnett L, et. al. The “GeneTrustee”: a universal identification system that ensures privacy and confidentiality for human genetic databases. *Journal of Law and Medicine*. 2003; 10(4): 506-513.

[3] de Moor GJ, Claerhout B, and de Meyer F. Privacy enhancing technologies: the key to secure communication and management of clinical and genomic data. *Meth Info Med*. 2003; 42: 148-153.

[4] Gaudet D, et. al. Procedure to protect confid-entiality of familial data in community genetics and genomics research. *Clin Genet*. 1999; 55: 259-264.

[5] Gulcher JR, et. al. Protection of privacy by third-party encryption in genetic research. *Eur J Hum Genetics*. 2000; 8: 739-742.

[7] Malin B and Sweeney L. Determining the identifiability of DNA database entries. In *Proc AMIA Symp*. 2000; 547-551.

[8] Malin B and Sweeney L. Re-identification of DNA database entries through an automated

linkage process. In *Proc AMIA Symp*. 2001; 423-427.

[9] Malin B and Sweeney L. Inferring genotype from clinical phenotype through a knowledge based algorithm. In *Pac Symp Biocomp*. 2002; 41-52.

[10] Malin B. Why pseudonyms don't anonymize: a computational re-identification analysis of genomic data privacy protection systems. *Data Privacy Lab Working Paper LIDAP-WP-19*. Carnegie Mellon University, Pittsburgh, PA. Nov. 2003.

[11] Øhrn A and Ohno-Machado L. Using Boolean reasoning to anonymize databases. *Artif Intell Med*. 1999; 15(3): 235-254.

[12] Rothstein MA, ed. Genetic secrets: protecting privacy and confidentiality in the genetic era. New Haven: Yale University Press. 1997.

[13] Sweeney L. Replacing personally-identifying information in medical records, the Scrub system. In *Proc AMIA Symp*. 1996: 333-337.

[14] Sweeney L. Three computational systems for disclosing medical data in the year 1999. *Medinfo*. 1998; 9(pt 2): 1124-1129.

[15] Wylie JE and Mineau GP. Biomedical databases: protecting privacy and promoting research. *Trends in Biotech*. 2003; 21(3): 113-116.

Bradley Malin received a BS in biological sciences (2000), MS in knowledge discovery and data mining (2002), MPhil in public policy and management (2003), and is currently a doctoral candidate in the School of Computer Science at Carnegie Mellon University. His research interests span the fields of data privacy, bioinformatics, medical informatics, bio-metrics, data mining, and digital rights management. His publications on the topic of DNA privacy have received several awards from the American Medical Informatics Association.