Poster: Tracking Personal MicroRNA Expression Profiles over Time

Michael Backes*†, Pascal Berrang*, Anne Hecksteden§, Mathias Humbert*, Andreas Keller‡ and Tim Meyer¶
*CISPA, Saarland University, lastname@cs.uni-saarland.de, †MPI-SWS
‡Clinical Bioinformatics, Saarland University, andreas.keller@ccs.uni-saarland.de
§Sports Medicine, Saarland University, a.hecksteden@mx.uni-saarland.de
¶Sports Medicine, Saarland University, sportmed@mx.uni-saarland.de

***Abstract*—The decreasing cost of molecular profiling tests, such as DNA sequencing, and the consequent increasing availability of biological data are revolutionizing medicine, but at the same time create novel privacy risks. The research community has already proposed a plethora of methods for protecting *genomic* data against these risks. However, the privacy risks stemming from *epigenetics*, which bridges the gap between the genome and our health characteristics, have been largely overlooked so far, even though epigenetic data such as microRNAs (miRNAs) is no less privacy sensitive. This lack of investigation is attributed to the common belief that the inherent temporal variability of miRNAs shields them from being tracked and linked over time.**

In this work, we show that, contrary to this belief, miRNA expression profiles can be successfully tracked over time, despite their variability. Specifically, we show that two blood-based miRNA expression profiles taken with a time difference of one week from the same person can be matched with a success rate of 90%. We furthermore observe that this success rate stays almost constant when the time difference is increased from one week to one year. In order to mitigate these linkability threats, we propose and thoroughly evaluate two countermeasures: (i) hiding a subset of disease-irrelevant miRNA expressions, and (ii) probabilistically sanitizing the miRNA expression profiles. Our experiments show that the second mechanism provides a better trade-off between privacy and disease-prediction accuracy.

I. Introduction

Since the first sequencing of the human genome in 2001, tens of thousands of genomes and over a million genotypes have been sequenced. The knowledge of our genetic background enables to better predict, and thus anticipate, the risk of developing several diseases, including cancers and cardiovascular and neurodegenerative diseases. Moreover, the genomic research progress enables the development of personalized treatment through pharmacogenomics, studying the effect of the genome on drug response. One of the most important negative counterparts of this genomic revolution is the threat towards genomic privacy [1], [7]. Indeed, genomic data contains very sensitive information about individuals' predisposition to certain severe diseases, about kinship, and about ethnicity, all of which can lead to various sorts of discrimination. Furthermore, genomic data is very stable in time and correlated between family members [5]. Due to these issues, a lot of research has already been carried out to improve the genomic-privacy situation (surveyed in [3], [9]).

However, our genome is by far not the only element influencing our health. Environmental factors (e.g., pollution, diet, lifestyle,...) often play a crucial role in the development of most common diseases. Epigenetics (or epigenomics), transcriptomics, and proteomics aim to bridge the gap between the genome and our health characteristics. Multi-omics research is a logical complementary step to genome sequencing: the DNA sequence tells us what the cell could possibly do, while the epigenome and transcriptome tell what it is actually doing at a given point in time. Using a computer analogy, if the genome is the hardware, then the epigenome is the software [2].

Despite the growing importance of epigenetics in the biomedical community, privacy concerns stemming from epigenetic data have received little to no attention so far. With the increasing understanding of epigenetics, it becomes clear that epigenetic data contains a vast amount of additional sensitive information, and can thus yield potential privacy risks. For example, major severe diseases (such as cancers, diabetes, or Alzheimer's [4], [6], [11], [12]) are already identified to be affected by epigenetic changes and a recent study stated that epigenetic alterations could even affect sexual orientation [10]. Furthermore, epigenetic data can potentially tell us more about whether someone is carrying a disease at a given point in time, compared to the genome that only informs about the *risk* of getting certain diseases.[1] In this work, we focus on microRNAs, an important element of the epigenome discovered in the early 1990s. MiRNAs are small RNA molecules that regulate the majority of human genes. Studies of miRNA expression profiles have shown that dysregulation of miRNA is linked to neurodegenerative diseases, heart diseases, diabetes and the majority of cancers [4], [6], [8], [11], [12]. Therefore, miRNA expression profiling is a very promising technique that could enable more accurate, earlier and minimally invasive diagnosis of major severe diseases. As a consequence, it will certainly be increasingly used in medical practice.

It is widely believed in the biomedical community that the miRNA expression levels are varying sufficiently to invalidate any linkability attempts over time. This work, however, shows the contrary: despite their temporal variability, microRNA expression profiles are still identifiable and linkable after time periods of several months.

[1]The only exception to this rule are Mendelian disorders, such as cystic fibrosis, which are largely determined by our genes.

II. Attacks

We study here the temporal linkability of personal miRNA expression profiles by presenting and evaluating two different attacks. We consider a passive adversary who can get access to miRNA expression levels of one or multiple individuals and wants to match them with other miRNA expression levels at some point in time. This epigenetic information could be collected online (publicly shared by the research community, like in the Gene Expression Omnibus), or be leaked through a major security breach, e.g., of a hospital server. We first study an *identification attack*, which pinpoints a specific miRNA expression profile in a database of multiple expression profiles by knowing the targeted profile at another point in time. Second, we study a *matching attack*, which tracks a set of miRNA expression profiles over time.

We rely on principal component analysis to pre-process the miRNA expression levels, and on a maximum weight assignment algorithm for the matching attack. We thoroughly evaluate our linkability attacks by using three different longitudinal datasets: (i) the blood-based miRNA expression levels of 29 athletes at two time points separated by one week, (ii) the plasma-based miRNA expression levels of the same athletes at two time points separated by one week, and (iii) the plasma-based miRNA expression levels of 26 patients with lung cancer over more than 18 months and eight time points. Our experimental results notably show that blood miRNA expression profiles are about twice as easy to track over time than plasma miRNA profiles, and that the matching attack is more successful than the identification attack: We reach a success rate of 90% with blood and a success rate of 48% with plasma miRNAs in the matching attack whereas, in the identification attack, we reach a success rate of 76% with blood and 28% with plasma miRNAs. Moreover, we demonstrate that only 10% of the miRNAs are sufficient to achieve similar success rates as with all miRNAs. With the third dataset containing plasma-based miRNA expression profiles, we observe that the attack achieves a similar success rate from one-week time shifts to 12-month time shifts.

III. Defenses

We present two defense mechanisms to counter the linkability of personal miRNA expression profiles: (i) hiding a subset of the miRNA expressions, e.g., those that are not relevant for medical practice, and (ii) sanitizing miRNA expression profiles by adding noise in a differentially private and distributed manner. While the first countermeasure would especially be useful in a clinical setting, in which the disease-relevant miRNAs are already known, the second countermeasure is intended to be better suited for the biomedical research community. In this context, as one of the objective is to discover associations between miRNAs and diseases, it is impossible to restrict the released data to only a few miRNAs.

We evaluate our protection mechanisms with the first aforementioned blood-based miRNA profiles of athletes and a fourth, also blood-based, miRNA dataset of more than 1,000 participants that includes information about 19 diseases (at a single point in time). The former is used to measure how temporal linkability is reduced with our countermeasures, whereas the latter helps us evaluate the evolution of accuracy (i.e., utility) in predicting patients' diseases from their miRNA expressions with a support vector machine (SVM) algorithm. The experiments show that it is possible to decrease linkability by at least 50% for almost no loss of accuracy ($< 1\%$) for the majority of diseases with the noise mechanism. Moreover, our results demonstrate that the noise mechanism provides better privacy-utility trade-offs than the hiding method in 17 out of 19 of diseases, while allowing more flexibility in the data usage for biomedical researchers. This finding is reinforced by the fact that an adversary can use correlations between miRNA expressions to infer more miRNAs than those actually shared with the first countermeasure.

IV. Conclusion

This work demonstrates that personal miRNA expression profiles can be successfully tracked over time, especially when these expressions are measured from blood samples. This study sheds light on a widely overlooked problem, privacy risks stemming from epigenetic data, bringing it to the attention of both the biomedical and computer security research communities. Note also that our linkability attacks can also be used for preventing the mixing of miRNA samples in the clinical setting. Our work also shows that probabilistically sanitizing the expression profiles is a promising technique that could be applied on other types of longitudinal biomedical data to enhance the privacy of their owners.

References

[1] E. Ayday, E. De Cristofaro, J.-P. Hubaux, and G. Tsudik, "Whole genome sequencing: Revolutionary medicine or privacy nightmare?" *Computer*, pp. 58–66, 2015.

[2] J. Cloud, "Why your DNA isn't your destiny," *Time*, January 2010.

[3] Y. Erlich and A. Narayanan, "Routes for breaching and protecting genetic privacy," *Nature Reviews Genetics*, vol. 15, pp. 409–421, 2014.

[4] A. P. Feinberg and M. D. Fallin, "Epigenetics at the crossroads of genes and the environment," *JAMA*, vol. 314, pp. 1129–1130, 2015.

[5] M. Humbert, E. Ayday, J.-P. Hubaux, and A. Telenti, "Addressing the concerns of the Lacks family: quantification of kin genomic privacy," in *Proceedings of the 2013 ACM SIGSAC CCS*, 2013, pp. 1141–1152.

[6] P. A. Jones and S. B. Baylin, "The epigenomics of cancer," *Cell*, vol. 128, pp. 683–692, 2007.

[7] Z. Lin, A. B. Owen, and R. B. Altman, "Genomic research and human subject privacy," *SCIENCE-NEW YORK THEN WASHINGTON-*., pp. 183–183, 2004.

[8] J. Lu, G. Getz, E. A. Miska, E. Alvarez-Saavedra, J. Lamb, D. Peck, A. Sweet-Cordero, B. L. Ebert, R. H. Mak, A. A. Ferrando *et al.*, "Microrna expression profiles classify human cancers," *nature*, vol. 435, no. 7043, pp. 834–838, 2005.

[9] M. Naveed, E. Ayday, E. W. Clayton, J. Fellay, C. A. Gunter, J.-P. Hubaux, B. A. Malin, and X. Wang, "Privacy in the genomic era," *ACM Computing Surveys (CSUR)*, vol. 48, p. 6, 2015.

[10] T. Ngun et al., "Abstract: A novel predictive model of sexual orientation using epigenetic markers," in *American Society of Human Genetics 2015 Annual Meeting*, 2015.

[11] I. A. Qureshi and M. F. Mehler, "Advances in epigenetics and epigenomics for neurodegenerative diseases," *Current neurology and neuroscience reports*, vol. 11, pp. 464–473, 2011.

[12] L. D. Wood, D. W. Parsons, S. Jones, J. Lin, T. Sjöblom, R. J. Leary, D. Shen, S. M. Boca, T. Barber, J. Ptak *et al.*, "The genomic landscapes of human breast and colorectal cancers," *Science*, vol. 318, pp. 1108–1113, 2007.