On Epigenomic Privacy: Tracking Personal MicroRNA Expression Profiles over Time (Extended Abstract)

Michael Backes*†, Pascal Berrang*, Anne Hecksteden§, Mathias Humbert*, Andreas Keller‡ and Tim Meyer§
*CISPA, Saarland University
†MPI-SWS
‡Clinical Bioinformatics, Saarland University
§Sports Medicine, Saarland University

***Abstract*—Following the genomic revolution and the consequent deluge of DNA data, a lot of research has been carried out to better understand and protect genomic privacy. However, genomics is only the tip of the iceberg of a broader epigenomic breakthrough currently going on. In order to shed light on privacy issues stemming from epigenomic data, we study how personal microRNA expression profiles can be tracked over time. By relying on principal component analysis and graph matching, we show that, despite the variability of gene expression, it is possible to track one or multiple expression profile(s) at different points in time. Specifically, we show that blood miRNA profiles of healthy athletes collected at a one-week interval can be matched together with a success rate of 90%. We also find out that blood expression profiles are much easier to link over time than plasma profiles that yield a success rate around twice smaller. Our results for plasma microRNA expression profiles are confirmed by another dataset of patients with lung cancer collected over a time period of more than 18 months. This second dataset also shows that a greater time shift between two miRNA expression's databases slightly decreases the attack's success.**

I. Introduction

Since the first sequencing of the full human genome in 2001, tens of thousands of genomes and over a million of 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 or cardiovascular diseases. However, the genome is by far not the only element influencing our phenotype (i.e., traits, diseases, ...). Environmental factors (e.g., lifestyle) also play a crucial role in the development of most common diseases. Epigenomics, which is the study of the key functional elements that regulate gene expression in a cell [11], aims at bridging the gap between the genome and our phenotypic characteristics. Gene expression profiling is a logical complementary step to genome sequencing: the DNA sequence tells us what the cell could possibly do, while the expression profile tells us 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].

Privacy is one of the most important issues stemming from the genomic revolution [1], [7]. Indeed, genomic data contains very sensitive information about individuals' predisposition to certain severe diseases, about kinship, and about ethnicity, which can lead to various sorts of discrimination. Furthermore, genomic data is very stable in time and correlated between family members [5]. A long line of research has already been published about privacy concerns and protection mechanisms related to the genome (surveyed in [3], [8]). With the better understanding for epigenomics, it becomes clear that epigenomic data also contains a vast amount of very sensitive information, which has been largely overlooked. For example, major severe diseases (such as cancers, diabetes, or Alzheimer's [4], [6], [10], [12]) are already identified to be affected by epigenetic changes and a recent study stated that epigenomic alterations could even affect sexual orientation [9]. Furthermore, the epigenome can 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]

At first sight, it could appear that the high variability of the epigenome over time (especially expression levels) is enough to make an individual's epigenomic profile unlinkable over time, thus naturally enhancing epigenomic privacy. This work, however, shows the contrary: individuals are still identifiable and linkable over time periods of several months through their microRNA expression levels. MicroRNA (abbreviated miRNA) plays a crucial role in regulating the transcriptional activity. Initially discovered in the early 2000s, these small RNA molecules of only about 20 nucleotides are biochemically stable and regulate the majority of human genes. Moreover, miRNA has been shown to influence Alzheimer's and Parkinson's disease [10], but almost all cancers as well. A summary of the relation between miRNA and human pathologies is provided in the Human miRNA Disease Database.[2]

UEOP '16, 21 February 2016, San Diego, CA, USA

http://dx.doi.org/10.14722/ueop.2016.23005

[1]The only exception to this rule are Mendelian disorders, such as cystic fibrosis, which are largely determined by our genes.
[2]http://www.cuilab.cn/hmdd

II. Threat Model and 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 epigenomic 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 study two different tracking attacks. The *identification attack* aims to pinpoint a specific miRNA expression profile among n miRNA expression profiles, by observing the targeted profile at another point in time. The *matching attack* refers to the case where the attacker has access to two databases (in most cases of similar size and greater than one) of miRNA profiles at different time instances and wants to match their elements together.

We rely on principal component analysis (PCA) with whitening to pre-process the more than 1000 miRNA real-valued expression levels. We then make use of the Euclidean distance between the miRNA expression vectors projected on the first c principal components. In the identification attack, the adversary simply selects the profile i^* with minimum distance to the targeted profile. In the matching attack, the adversary should find the best assignment between the two databases of expression profiles, which is the one that minimizes the sum of the distance between every matched pair.

This problem boils down to finding an optimal matching on a weighted bipartite graph where each vertex represents a miRNA profile, and where the weight on each edge represents the Euclidean distance between any pair of miRNA profiles. In order to find the optimal assignement efficiently, we make use of the blossom algorithm that finds the minimum/maximum weight assignment in $O(n^3)$.

III. Experimental Results

We evaluate the success of our tracking attacks by using three datasets: (i) the blood miRNA expression levels of 29 athletes at two time points separated by one week, (ii) the plasma miRNA expression levels of the same 29 athletes separated by one week, and (iii) the plasma miRNA expression levels of 26 lung-cancer patients over more than 18 months and eight time points.

a) Identification Attack: First, we compare the success rate in identifying the correct profile over all possible PCA dimensions with the athletes' dataset: We reach a success rate of 76% for the blood miRNAs with 22 and 23 PCA dimensions, and 28% for the plasma samples with 17, 18, 19, and 31 PCA dimensions. In order to validate our findings, we also evaluate the success rate for the plasma-based miRNAs of lung-cancer patients, and get similar results. Over all possible time shifts, we achieve a maximum success rate of 42% with 25 and 39 PCA dimensions, and an average success rate of 22% with 22 dimensions. Finally, we also analyze the effect of time shifts on the attack's success. We notice a slight decrease in the best success rate for increasing time shifts, with highest success rates achieved almost always between consecutive time points.

b) Matching Attack: As for the identification attack, we first compare the success rate in matching profiles over all PCA dimensions for the athletes' dataset: We obtain a succes rate of 90% for the blood with 39 and 40 dimensions, and 48% with 34 dimensions for the plasma samples. We notice that the success rate is higher for the matching than for the identification attack. This is explained by the fact that, by forcing each profile at the first time point to be matched to one and only one profile of the second time point, the (perfect) matching attack rules out the cases where multiple samples of the first time point are matched to the same sample of the second time point. We validate our findings with the third dataset, of lung-cancer patients. Regardless of the time shift, we reach a maximum success rate of 55% with 39 PCA dimensions, and an average success of 30% with 34 dimensions. We finally explore the effect of time distance between expression levels on the success rate. We do not notice a significant trend until 12-month time shift. A slight decrease in success for 15 and 18 months can nevertheless be observed.

IV. Conclusion

In this work, we have presented and studied two new tracking attacks against miRNA expression profiles, considering time shifts from one week to 18 months. We have observed a slight decrease in success when time distance increases, especially for shifts greater than one year. We have also found that blood miRNAs are much more linkable than plasma miRNAs. Finally, we have shown that matching attacks are more successful than identification attacks. This work shows the extent of the threat against miRNA expression data, and paves the way for further research on epigenomic privacy.

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] 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.

[9] 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.

[10] 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.

[11] C. E. Romanoski, C. K. Glass, H. G. Stunnenberg, L. Wilson, and G. Almouzni, "Epigenomics: Roadmap for regulation," *Nature*, vol. 518, no. 7539, pp. 314–316, 2015.

[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, 2007.