Analysis and Protection of Sensitive Information in Gene Expression Datasets

Arif Harmanci, Jieming Chen, Dov Greenbaum, Mark Gerstein

ABSTRACT

With the unprecedented increase in the size of genomic datasets, the quantification and protection of privacy-sensitive information is a vital issue to be addressed for protection of anonymity of the participants of the scientific studies.

In this paper, we present a framework for analysis and protection of private information in the gene expression datasets. We present a general scenario where the gene expression datasets can be exploited to predict eQTL genotypes to link independently distributed anonymized datasets by an adversary to re-identify individuals. We first analyze the amount of leakage of genetic information for each eQTL SNP when predicted using the gene expression datasets. We propose a simple method to predict eQTL genotypes from gene expression datasets. We then utilize the prediction method for low frequency multiple SNP genotype prediction, which can be used to de-identify individuals. Using publicly available gene expression dataset we illustrate that a significant fraction of the samples are vulnerable to de-identification. As a remedy for the privacy loss, we focus on anonymization of the gene expression datasets and present a method for anonymizing the gene expression dataset. We illustrate that the datasets can be anonymized with very small amount of loss in the biological information.

1 BACKGROUND

[[Introduction goes here]]

[[Define sensitive information: Anything that the individuals do not want leaked]]

[[Definition of privacy that we will use is twofolds: Identifiability of SNP genotypes and identifiability of individuals.]]

[[Our novelty is in first presenting a quantification methodology that can be extended for future studies and also for showing the linking attack in the context of expression-genotype association with a simple attack]]

2 RESULTS

2.1 Overview of the Privacy Breaching Scenario

Figure 1 illustrates the privacy breaching scenario that is considered. The breach occurs by linking two datasets such that one of the datasets contains the individual identities and corresponding genotypes and the second dataset contains the gene expression levels and sensitive information (e.g. disease status) about each individual. The second dataset is assumed to be anonymized by removal of the individual identities to protect the individuals. The adversary gains access to both datasets and links the datasets to associate the sensitive information to individuals. While performing the linking "attack" the adversary utilizes publicly available databases. In the considered scenario, the eQTL databases are utilized which enable linking the expression levels to the genotypes.

2.2 Identifiability of eQTL SNP Genotypes

We first analyze in a general setting the amount of identifiable genetic information using the linking of gene expressions with genotypes. For this, we utilize the mutual information based metric that has been applied for quantifying privacy loss. Figure 2a and 2b show the distribution of privacy loss for all the eQTLs given the gene expression levels. It can be seen that, given the expression levels, there is significant loss of genetic information compared to random associations, which is as high as 20% for some of the eQTLs.

[[Introduce the entropy based measure and MI based leakage measure]]

2.2.1 Extremity Attack

This analysis is useful for getting quantification of leaked genetic information from gene expression datasets. To predict the eQTL genotypes from gene expression levels, we propose using a method that we name "extremity attack". In this attack, given one gene whose expression level correlates with a variant. The prediction utilizes a statistic we termed $extremity$ of gene expression level which quantifies how extreme an individual's gene expression level is away from the mean of the distribution. Given the gene expression level, e, for a $extremity$ is defined as following:

$$extremity(e) = \frac{rank\ of\ e}{\#\ of\ individuals} - 0.5.$$

Extremity is bounded between -0.5 and 0.5. Figure 3a illustrates the extremity attack. The adversary utilizes the extremity and the gradient of association between the gene expression level to assign a genotype to the associated variant.

[[Figure 3bc shows the accuracy of extremity attack with different extremity and correlation thresholds.]]

2.3 Identifiability of Individuals using Low Frequency Multiple SNP Genotypes

Identifiability of individuals requires generating features that are unique to certain individuals in the sample. The eQTLs are not suitable for identifying individuals since they are common variants. Although each eQTL genotype is common, the frequencies for genotypes of SNP combinations can be small and these combinations, which can be well predicted, can be utilized for identifying individuals. By using a similar approach in Section 2.1, we first quantify the amount of individual identifying information that is leaked in the gene expression datasets. This quantification enables us to evaluate the bounds on the amount of information that can be extracted from the expression levels about the genotype data, which can then be utilized for anonymization of expression dataset so as to guarantee a quantifiable privacy level in the released dataset.

Individual identification is basically generating a feature that can distinguish, or discriminate, an individual from all others in the dataset. Since we are aiming to do this via prediction of genotypes from expression levels, we need a measure that captures discriminative power of the genotypes and also enable predictability measure. To quantify the individual identifying information, we use surprisal, measured in terms of self-information of the genotypes:

$$III(g) = I(G = g) = -\log(p(G = g))$$

where G is an eQTL variant and g ($g\epsilon\{0,1,2\}$) is a specific genotype and $p(G = g)$ is the frequency of the genotype in the sample set and III denote the individual identifying information. Assessing this relation, the genotypes that have low frequencies have high identifying information, as expected. Given multiple eQTL genotypes, assuming that they are independent, the total individual identifying information is simply summation of those:

$$III(\{G_1 = g_1, G_2 = g_2, \dots, G_N = g_N\}) = -\sum_{i=1}^{N} \log\big(p(G_i = g_i)\big).$$

The individual identifying information after the gene expression levels are revealed is basically the conditional III given the gene expression levels:

$$III_{remaining}(\{G_1 = g_1, G_2 = g_2, \dots\} \mid \{E_1 = e_1, E_2 = e_2, \dots\}) = -\sum_{i=1}^{N} \log(p(G_i = g_i \mid E_i = e_i))$$

where E_i represents the gene expression level for the i^th gene, which is associated with the genotype of G_i. The leakage in III is the remaining III after expression levels are revealed:

$$III_{leaked} = III - III_{remaining}.$$

[[Figure 4a shows the predictability versus discrimination power of the top eQTLs]]

[[Figure 4bc individual identifying information and leakage quantification]]

2.3.1 Identifiability of Individuals by Extremity Attack in k-Anonymity Framework

[[Fig. 5a; Distribution of the maximum of absolute extremity over all the samples. How well does expression extremity identify individuals? It is mostly uniform except for some samples.]]

Although we showed that there is a significant leakage of individual identifying information in the gene expression levels, we do not have a way to extract the information. This can be performed using prediction models, similar to the ones built previously~\cite{Schadt et al}.

We will utilize previously presented extremity attack for identification of individuals.

To formalize the analysis using the low frequency multi-SNP genotypes, we utilize the k-anonymization framework. K-anonymization formalizes a way to identify the number of vulnerable individuals and also to ensure the anonymization, which is presented in Section 2.5. Briefly, in order to identify the individuals that are vulnerable to the linking attack, we identify the individuals that have the low frequency multiple SNP genotypes such that all the SNP genotypes are highly predictable using the expression dataset.

[[External information: 1 bits of gender information can be easily predicted from ; how does this change vulnerability; this justifies the fact that we need “buffering” in anonymization to protect against unaccounted external information that may cause increased vulnerability.]]

2.4 Anonymization

[[Do anonymization for all possible parametrizations to decrease the privacy loss to minimum]]

[[k-anonymization formality for guaranteeing anonymity]]

3 METHODS

3.1 Quantification of Genotype Information Content and Loss of Privacy

[[MI and entropy based definition of IC and Loss of Privacy]]

[[Must justify with MI computation method]]

3.2 Extremity Attack

[[Define the extremity attack: Correlation and extremity parameters]]

3.3 K-Anonymization

[[Define k-anonymization]]

[[Present in detail the anonymization procedure that we propose]]

4 CONCLUSION AND DISCUSSION

In this paper we present a simple framework for quantification of the sensitive information leakage in the linking attack scenarios. The premise of sharing genomic information is that there is always an amount of leakage in the sensitive information. We believe that this quantification methodology can be utilized for more extensive analysis of the leakage in sensitive information for high level correlations in the genomic datasets. The quantification can be further developed for guaranteeing bounds on anonymized datasets.

[[How does this framework compare to other formalities? For example differential privacy? It is similar but differential privacy does not enable quantification of the leakage.]]

We also presented a simple attack that is based on using extremity statistic to predict genotypes that can implicate the sensitive information. Compared to previous approaches, this statistic is very easy to compute.