Conference Paper

Context trees for privacy-preserving modeling of genetic data

Author(s):
Kusters, Lieneke; Ignatenko, Tanya

Publication Date:
2016

Permanent Link:
https://doi.org/10.3929/ethz-a-010646761

Rights / License:
In Copyright - Non-Commercial Use Permitted
Context trees for privacy-preserving modeling of genetic data

Lieneke Kusters and Tanya Ignatenko
Eindhoven University of Technology
Department of Electrical Engineering
Eindhoven, The Netherlands
Email: c.j.kusters@tue.nl

Abstract—In this work, we use context trees for privacy-preserving modeling of genetic sequences. The resulting estimated models are applied for functional comparison of genetic sequences in a privacy-preserving way. Here we define privacy as uncertainty about the genetic source sequence given its model and use equivocation to quantify it. We evaluate the performance of our approach on publicly available human genomic data. The simulation results confirm that the context trees can be effectively used to detect similar genetic sequences while guaranteeing high privacy levels. However, a trade-off between privacy and utility has to be taken into account in practical applications.

I. INTRODUCTION

With recent advances in genome sequencing technologies, genetic data is increasingly being used in everyday applications. Genes are studied to better understand inherited diseases, such as cancer; personalized medicine based on genetic analysis is used to determine the most effective medicine for patients; genetic material is used to identify individuals in forensic investigations. Therefore, more and more genetic sequences are being collected for research and analysis purposes. Several projects [1], [2], [3] have been initiated to identify and catalog similarities and differences between genetic sequences and relate these findings to medical conditions. The human reference genome has been reconstructed as a general reference and is a representative example of the human genome.

Disclosing one’s genetic data is often beneficial for medical purposes, however, this information is also considered to be privacy sensitive. Genetic sequences contain health-related information as well as information about one’s ancestry. Furthermore, genetic sequences are unique identifiers of human beings. Therefore privacy concerns are raised as this information can be misused by insurance companies, employers and forensic institutions. Remarkably, one’s genome is unique inherited information, of which disclosure is irreversible, thus affecting not only its owner but also the owner’s family members for many generations. As such, in modern society protection of genetic data becomes a crucial problem.

A. Related work

Traditionally, privacy-sensitive information (e.g. in medical profiles) is protected by data anonymization techniques through removing or aggregating the information that can be used to identify the corresponding individual (e.g. a name, birth date or address). Uniqueness of genetic data makes traditional anonymization techniques insufficient, as the data itself uniquely identifies its owner. Therefore, recently, privacy-preserving approaches for genome analysis became the focus of a research community. These approaches aim at analysis of genetic data while protecting the privacy of individuals that are involved.

Erlich et al. provided an overview of possible privacy breaches based on the information that can be derived from genetic databases, and reviewed a number of techniques for their prevention [9]. These techniques range from access control and data anonymization techniques to cryptographic solutions. There is also a large branch of research focusing on privacy-preserving sequence comparison. Works in this direction focus on privacy-preserving edit-distance computation between genetic sequences, and typically deploy secure multiparty computation and homomorphic encryption techniques [5], [11]. This approach has however a number of problems. First, due to computational complexity, these algorithms do not scale well to large sequences, while genetic sequences can reach three billion base pairs in length. Therefore, to implement these techniques, outsourcing of the expensive computations to the cloud is used, examples of the corresponding secure protocols can be found in [6], [7]. Moreover, since these approaches are based on cryptographic techniques, their security relies on hardness of the underlying problems and attacker’s computational power limitations. However, advances in computer power do not guarantee that such cryptographic techniques remain secure in the future. Therefore, given that the data is unique and inherited, information theoretic security is a desirable property for genomic data protection. Finally, edit-distance does not provide sufficient information to draw conclusions about functional similarity of genetic data to be used in genetic analysis.

B. Our contribution

In this paper we study generative statistical modeling of genetic sequences applied to sequence comparison and the privacy-preserving properties of these models. Note that genetic sequences can be seen as a natural code that sequentially encodes amino acids and proteins. Therefore models that take into account source memory, such as e.g. hidden Markov models, have been successfully used in genetic sequence
analysis [20]. Here we focus on the context-tree models, proposed and studied by Willems et al. [19]. These models are a special case of Markov chains and are closely related to the concept of k-mers\(^1\) used in Bioinformatics for statistical analysis of genetic data [4].

Application of context-tree models for genetic sequence comparison was proposed in [10]. Furthermore, it was shown in [12] that context trees can be applied to model coding and non-coding parts of genetic sequences. Following the Minimum Description Length (MDL) principle [15], in both [10] and [12] compression rate of genetic sequences given an estimated genetic sequence model was used to assess similarity of the compressed and modeled sequences.

In our setting, we assume that a genetic database is composed of genetic sequence models and some associated relevant metadata. This database can be used to search genetic sequences that are functionally similar to a query sequence. Here genetic models are the only genetic information that an attacker can deduce from the database, and therefore, we analyze privacy as the privacy of an underlying genetic source sequence given its model. Such models in general do not correspond to a single sequence but to a class of sequences. Clearly, the larger the class is, the better the privacy guaranties are. The privacy-preserving properties of the models are characterized by the uncertainty about the underlying source sequence and thus correspond to the entropy of the source model. Hence our approach provides information-theoretic security guarantees.

II. BACKGROUND AND NOTATION

Genetic sequences are described in terms of four symbols from the alphabet \(\mathcal{A} = \{A, C, G, T\}\) that correspond to the DNA building blocks, called nucleobases, i.e. Adenine, Cytosine, Guanine and Thymine. Although genomes of different individuals are similar, each individual genome is unique. Differences between genetic sequences result from genetic variations that include substitution of one nucleotide with another one and insertion or deletion of a subsequence of nucleotides.

In the following, we denote by \(x_N^n = x_1x_2x_3\ldots x_N \in \mathcal{A}^N\) a genetic sequence of length \(N\). Furthermore, a sequence of arbitrary length is denoted by \(x\), and a source sequence for which we construct the statistical model is denoted by \(\tilde{x}\).

III. SEQUENCE MODELING

In this section we first introduce the concept of context trees, see e.g. [19], that we apply to model genetic sequences. Then, we describe our approach to estimate the model corresponding to a given sequence.

A. Context tree model for genetic sequences

A context tree is a tree structure whose nodes represent the memory of the source. Given an observed source sequence, each node in the tree is associated with a specific context that represents the past of the current symbol in the observed sequence. The leafs of the tree represent the contexts of maximum memory and have associated conditional probabilities, characterizing probability of generating a symbol from the source alphabet given its observed context.

More precisely, a genetic tree source is described in terms of a context-tree model \(\langle S_T, P_T \rangle\) represented by a quaternary-tree structure (resulting from the quaternary alphabet for genetic sequences), see e.g. Figure 1. The tree model \((S_T, P_T)\) defines the leafs or contexts \(s \in S_T\) and the conditional probability distribution \(P_T = \{p_T(x|s) : x \in \mathcal{A}, s \in S_T\}\), associated with occurrence of a symbol given its context. For an observed sequence \(x\), the context of a symbol \(x_i\) is defined by its at most \(D\) preceding symbols in reversed order, i.e. \(x_{i-D}x_{i-1}\ldots x_{i-1}\). Therefore, given a tree model, we can define a mapping \(\omega_{S_T}(x_i^{i-D}) \rightarrow s \in S_T\) that maps \(D\) preceding symbols of \(x_i\) to a leaf in the tree model and thus retrieve \(p_T(x_i|s)\), the probability of occurrence of symbol \(x_i\) after context \(s\) was observed. Then, given a tree source model, we can determine the probability of sequence \(x\) being generated by this source as follows:

\[
\Pr(x|(S_T, P_T)) = \prod_i p_T(x_i|\omega_{S_T}(x_i^{i-D})).
\]

In general, for a given sequence, we do not know the actual source model that has generated this sequence. Therefore, we have to find a good estimate for it, i.e. given this sequence we have to estimate the corresponding contexts in the tree and the associated conditional distribution. In the following, we present the algorithm to estimate such models.

B. Model estimation

In order to estimate a tree model corresponding to sequence \(x\), we assume the model depth (or memory) to be \(D\). Now to estimate the sequence statistics, we process the sequence \(x\) in a sequential way. For each symbol in the sequence we find its context, defined by its \(D\) preceding symbols, and count the total number of occurrences \(n^x_s(x)\) of a symbol \(a \in \mathcal{A}\) with each observed context in the sequence. Here each distinct observed context becomes a leaf in the tree. Moreover, for the first symbol we assume some initial context \(s_1\). Now, given the contexts and the counts, the conditional probability distribution \(p_T(x|s)\) is estimated as follows:

\[
p_T(x|s) = \frac{n^x_s(x) + 1/2}{\sum_{a \in \mathcal{A}} n^x_s(a) + |\mathcal{A}|/2}, \quad x \in \mathcal{A}, s \in S_T.
\]

Fig. 1. An example of the context-tree model \((S_T, P_T)\) of depth 2, with root \(\lambda\) and set of leafs \(S_T = \{T, G, C, AT, AG, AC, AA\}\) with corresponding conditional probabilities, given by \(P_T = \{p_T(|s) : \forall s \in S_T\}\).
The source model of the sequence $x$ is now defined by $(S_T, P_T)_x$, where $S_T$ corresponds to the distinct contexts observed in the processed sequence $x$ and represents the leaves of the tree model, and $P_T = \{p_T(x|s), x \in A, s \in S_T\}$ corresponds to the conditional probability distribution associated with each of these leaves.

IV. Sequence comparison

In order to apply our context-tree models for sequence comparison, we need to define a measure to evaluate similarity of an arbitrary genetic sequence $x$ to a source sequence $\hat{x}$, given only its model $(S_T, P_T)_x$. The MDL principle [15] states that the model that describes the data in the shortest possible way is the model that most probably generated this data. Observe that compression rate characterizes the amount of bits required to describe data in a concise way. Furthermore, the techniques presented in the previous section originally aimed at finding good source coding distributions, and are characterized by asymptotically optimal performance. Thus based on the MDL principle and the fact that a model that closely relates to an observed sequence also results in a low compression rate, we use compression rate to characterize sequence similarity.

The compression rate of a sequence $x_1^M$, given a tree model $(S_T, P_T)_x$ is estimated as the logarithm of the probability of the sequence being generated by this model:

$$R(x_1^M|(S_T, P_T)_x) = -\frac{1}{M} \sum_{i=1}^{M} \log_2(p_T(x_i|\omega_{S_T}(x_{i-1}^i))). \quad (3)$$

For our problem, we use parameter $p_T$ equals 1/4, when there is no leaf in $S_T$ associated with a subsequence in $x_1^M$.

Clearly, compression rate of a sequence given its model depends on both accuracy and complexity of this model. Therefore, to take into account model complexity we will use compression rate of a sequence $\hat{x}$ given a source model of $\hat{x}$ relative to the compression rate of this sequence given its own model. Thus in this work we use the normalized information gain as a similarity measure, defined as

$$\Delta R(\hat{x}) = \frac{R(\hat{x}|(S_T, P_T)_{\hat{x}}) - R_{\hat{x}}}{R_{\hat{x}}}, \quad (4)$$

where $R_{\hat{x}} = R(\hat{x}|(S_T, P_T)_{\hat{x}})$ is the compression rate of the sequence $\hat{x}$ given its own model.

Note that while estimating the compression rate for the query sequence, the best (lowest) compression rate is achieved with the model that corresponds to the query sequence itself. Furthermore, the more similar the query sequence is to the source sequence, the more similar their models are, and thus the better the compression that is achieved using the model corresponding to the source sequence. Therefore, for statistically similar sequences the resulting information gain will be small. Also the opposite is true, and for sequences with very distinct models, the information gain achieved with the correct model will be high.

We will use information gain estimates based on the context-tree models to determine whether the genetic sequences are (functionally) similar or not. The utility of our approach is tested on the publicly available genetic data to which mutations are introduced, in Section VI.

V. Privacy analysis

Another aspect of genetic sequence modeling that we study in this paper is privacy. Here we regard privacy as uncertainty about the source sequence $\hat{x}$, given its corresponding model $(S_T, P_T)_x$. For the attack model we assume that an attacker only has access to the sequence models. Note that the sequence models actually correspond to a class of sequences, therefore the attacker’s best strategy is to reconstruct the most probable sequence that can be generated by the given sequence model. However, the context-tree models correspond to a class of sequences that are all generated by this model with equal probability. Thus the attacker cannot distinguish the original sequence $\hat{x}$ from the other sequences that can be generated by the model.

In order to measure the privacy level of our models, we define privacy as a function of the number of sequences corresponding to the model of the underlying source sequence. This measure of privacy is known as equivocation [16] and is given by

$$E(\hat{x}) = H(\hat{x}|(S_T, P_T)_{\hat{x}}) = \log_2(K), \quad (5)$$

where $K$ is the number of sequences that are generated by the model $(S_T, P_T)_x$. The next step is to estimate the equivocation of our context-tree models.

A. Type-class cardinality

The type-class $T_x$, see [8] and [14], of a sequence $\hat{x}$ with respect to a tree model is defined as the set of sequences that have the same symbol counts $n_a^\hat{x}$ in the leaves of their tree model as the source sequence $\hat{x}$, i.e.: $T_{S_T}(\hat{x}) = \{\hat{x} \in A^N : n_a^\hat{x}(\hat{x}) = n_a^\hat{x}(\hat{x}), \forall s \in S_T, a \in A\}. \quad (6)$

Note that our context-tree model only contains conditional probabilities and no symbol counts in its leaves. Therefore, to construct the type-class, one first needs to estimate the counts from the available conditional probabilities. This requires knowledge of the length of the original sequence used to construct the model. While this information is in general not available, we can assume that one (e.g. the attacker) can approximate or make an assumption about the sequence length $N$, approximate the counts, and then construct the type-class. Clearly, since all the sequences from the type-class are equiprobable from an attacker point of view, the cardinality of the type-class provides us with the number of sequences $K$ that are generated by the corresponding model, and thus can be used to get the equivocation estimate.

We use Whittle’s formula [18] to calculate the cardinality of a type-class, given the transition frequency matrix $F_S$ (see the next subsection for the details) and the initial $s_1$ and final $s_L$ contexts corresponding to the source sequence:

$$K = |\mathcal{T}_S(\hat{x})| = CS \prod_{s \in S} F_S(s, s)! \prod_{t, v \in S} F_S(t, v)!, \quad (7)$$
where $F_S(s,*) = \sum_{v \in S} F_S(s,v)$ is the number of symbols emitted with context $s$, and $C_S$ is the cofactor of entry $(s_1, s_1)$ of $I - F_S$, with $F_S$ being the transition frequency matrix with normalized rows.

B. Transition frequency matrix

Given the context-tree model and the symbol counts in its leaves, an element of the transition frequency matrix $F_S(t,v)$ gives the number of times context $t$ is being followed by context $v$ in the source sequence, i.e.

$$F_S(t,v) = |\{i: 1 \leq i \leq N, s_i = t, s_{i+1} = v\}|,\quad t,v \in S_T,$$

where the context sequence $s_1^L = \{s_1, \ldots, s_L\}$ is the concatenated set of contexts that occur successively in the source sequence.

For the context-tree models described in Section III, the next context $s_{i+1}$ follows from the current context $s_i$ and current ($i^{th}$) symbol $x_i$ in the sequence: $s_{i+1} = x_i s_i^{D-1}$, where $D$ is the depth of the tree. Therefore, the transition frequency matrix can be constructed directly from the model counts $\hat{n}_{s,v}$, $s \in S_T$, as

$$F_S(t,v) = \begin{cases} \hat{n}_{s,v}, & v = a t_1^{D-1}, \ a \in A, \ t,v \in S_T; \\ 0, & \text{otherwise}. \end{cases}$$

In order to apply Whittle’s formula and evaluate the type-class size, besides the transition frequency matrix $F_S$, we also need to know the initial and final contexts that occurred in the source sequence. We assume that the initial context $s_1$ is predefined and that it is known. The final context $s_L$ can then be derived from the transition frequency matrix and the initial context using the flow conservation equations:

$$F_S(*,s) + \delta_{s,s_1} = F_S(s,*) + \delta_{s,s_L}, \quad s \in S,$$

$$\delta_{s,t} = \begin{cases} 1 & \text{if } s = t \\ 0 & \text{otherwise} \end{cases},$$

where $F_S(*,s) = \sum_{t \in S} F_S(t,s)$ is the number of transitions into context $s$. Now, given the transition frequency matrix and initial and final contexts, the privacy-preserving properties of our context-tree models are given by equivocation (5) estimated with the help of Whittle’s formula (7).

VI. EXPERIMENTAL RESULTS

In order to evaluate the performance of context trees for privacy-preserving sequence modeling, we apply our method to subsequences of the human reference genome [17]. We use a set of distinct genomic subsequences, each corresponding to a different gene. Moreover, for each gene we have generated a set of similar sequences by applying a predefined number of mutations. It is well-known that on average the genomes of human individuals are 99.5% similar [13]. These 0.5% of genetic variations come from different types of mutations, such as single nucleotide variants, indels (insertions or deletions of a block of nucleotides of length < 100), and large-scale structural variants. Furthermore, most genetic variation occurs in a limited region of the human genome. In this work, we use a simplified mutation model, where we apply single nucleotide mutations of 1 per 1000, 1 per 100, and 1 per 10 base-pairs, as well as indels of length 10, occurring once per 1000 base-pairs, in order to simulate a set of similar sequences.

A. Utility performance

First, we evaluate the utility performance of the models for genetic sequence comparison as a function of tree-depth (characterizing model complexity) and sequence length. We start with estimating the compression rate of the source sequence, see Figure 2. Note that as the compression rate is measured in bits per base-pair (bp), a compression rate smaller than 2 corresponds to actual compression of the sequence. On average the compression performance of the context-tree models improves when larger tree-depth is used. Since compression rate indicates how well the model fits the data, we may conclude that including larger memory in the models and thus increasing its complexity helps to better describe the gene data. This can be supported by data interpretation, since genes correspond to coding regions of the genomes. However, compression rate of individual sequences does not provide information about the specificity of the model when distinguishing between different sequences.

Next, we evaluate the performance of the context-tree models in distinguishing whether the sequences are similar or not. We have evaluated the normalized information gain for sequences with various rates of mutation in comparison to the source sequence, see Figure 3. Furthermore, we have also applied the context-tree models to estimate the normalized information gain corresponding to the other genes, see
Figure 3. Based on our results, we conclude that the context-tree models can be applied to efficiently distinguish between similar (sequences with mutation rate smaller than 1/100) and non-similar sequences by defining an appropriate threshold. Note that models with higher complexity perform better in distinguishing between similar and distinct sequences. This is not surprising given the compression results discussed above.

Finally, we evaluate the performance of the context-tree models for distinguishing sequences with different functionalities. In this experiment we use sequences corresponding to different genes (or functionalities). Furthermore, we have simulated a set of sequences with similar functionalities by introducing the following mutations: 1 mutation per 1000, 1 mutation per 100, and 10 indels per 1000 base-pairs in comparison to the source sequences. Figure 4 shows the resulting ROC curves demonstrating the performance of models with different complexity. We conclude that a perfect distinction can be achieved with the models of depth 3 or larger, but also the models of depth 2 achieve an acceptable performance.

B. Privacy performance

Finally, we evaluate the privacy-preserving properties of our models in terms of equivocation. We plot equivocation as a function of sequence length in Figure 5. We see that with our models we can achieve very high privacy levels. In contrast to the information gain, improved equivocation is achieved for smaller tree-depth. Therefore there exists a trade-off between privacy and utility that should be taken into account when selecting models for privacy-preserving genetic sequence comparison.

VII. CONCLUSIONS

In this paper we have studied the use of context-tree models for privacy-preserving modeling of genetic sequences in application to sequence comparison. We have focused on functional sequence similarity that can be expressed as statistical similarity of compared sequences, and used normalized information gain as a similarity measure. Furthermore, privacy of the context-tree models is given in terms of equivocation, that characterizes uncertainty about the source sequence given its model. Based on the experimental results, we can conclude that context-tree models can be successfully applied for privacy-preserving sequence comparison resulting in both good discriminating and privacy performance. However, since

ACKNOWLEDGMENT

This work has been partially funded by the EC via grant agreement no. 611659 for the AU2EU project. The authors would also like to thank dr. J.F.J. Laros for the fruitful discussions on genomic data and its analysis.

REFERENCES