Privacy preserving genome analysis using context trees

Citation for published version (APA):

Document status and date:
Published: 01/01/2016

Document Version:
Publisher’s PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:
• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher’s website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain.
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:
www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:
openaccess@tue.nl
providing details and we will investigate your claim.
Privacy preserving genome analysis using context trees

Lieneke Kusters  Tanya Ignatenko
Eindhoven University of Technology

Introduction  Genome analysis has many applications of which well known examples are identification and personalized medicine. However, the genetic data should be treated with care as it can reveal information that is considered privacy sensitive, such as kinship, ethnicity, and predisposition to certain diseases. Such information can be misused for genetic discrimination, for example by employers and insurance companies.

Recently, more and more genetic data is being collected and analyzed, and protection of this sensitive information becomes a high priority. The protection of the genetic data has many specific challenges [EN14]. Most importantly, the genetic data is unique and reveals information that can uniquely identify the corresponding individual. Therefore, traditional anonymization techniques are not applicable. Proposed solutions vary from cryptographic techniques, to techniques that guarantee information theoretic privacy.

We propose to use compression techniques which we apply for genetic sequence comparison, while at the same time information theoretic privacy is guaranteed.

Methods  In this work we focus on sequences that correspond to genes, and thus encode certain functionalities. We assume that the codes are sequential and use context trees [WST95] to model the sequences. A context tree is a statistical model which stores the probabilities of symbols given their context. The context of a symbol is in this case defined by its preceding symbols in the sequence. We can vary the model complexity by increasing or decreasing the depth \( D \) of the tree, where \( D \) corresponds to the length of the context.

We evaluate both utility and privacy performance of the context tree models. We evaluate the utility performance of our models on distinguishing sequences corresponding to different genes. We construct the model corresponding to each sequence and then estimate the sequence similarity based on KL-Divergence [CT06] of the respective tree models. Finally, we use a threshold to distinguish whether a sequence corresponds to the same or to a different gene. The privacy performance results from the generality of the models. That is, each tree model actually represents a set of sequences that correspond to the same class. An adversary cannot distinguish the actual source sequence from any other sequence in the same class, and thus uncertainty remains about the original sequence. We measure the resulting privacy performance as equivocation [SRP13], defined as \( E(x) = H(x) = \log_2 |T| \), with \( |T| \) the number of sequences that correspond to the same model.

Results and conclusion  We perform experiments on annotated genes in the human genome. We construct context tree models of various complexities corresponding to each sequence and evaluate the performance on distinguishing between similar and non-similar sequences. Furthermore, we calculate the equivocation corresponding to each model. The results can be seen in the Figures above. Clearly, increased model complexity results in improved classification performance, while at the same time privacy performance decreases. Therefore, a trade off must be considered between privacy and utility performance, and an appropriate model complexity must be selected depending on the application.

References


