Future Generation Computer Systems 36 (2014) 441–452

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Privacy/performance trade-off in private search on bio-medical data

H. Perl [a,*], Y. Mohammed [b], M. Brenner [a], M. Smith [a]

[a] Distributed Computing and Security Group, Leibniz University Hannover, Schloßwender Straße 5, 30159 Hannover, Germany

[b] Biomolecular Mass Spectrometry Unit, Leiden University Medical Center, Einthovenweg 20, 2333 ZC Leiden, The Netherlands

HIGHLIGHTS

- We improve and elaborate on a new fast and secure exact term search algorithm with private queries.
- The algorithm offers a trade-off between privacy and performance.
- The results can then be further aggregated using Homomorphic Cryptography.
- We expose these functionalities as a ready-to-use Web service.

ARTICLE INFO

Article history:
Received 1 February 2013
Received in revised form
4 November 2013
Accepted 3 December 2013
Available online 16 December 2013

Keywords:
Privacy
Homomorphic Cryptography
Bloom filters
Secure search

ABSTRACT

Outsourcing of biomedical data, especially human patient data, for processing is heavily constrained by legal issues. For instance searching for a biological sequence of amino acids or DNA nucleotides in a library or database of sequences of interest to identify similarities is not something which can easily be outsourced due to the data protection and privacy laws. However, DNA sequencing is becoming a main stream technology, thus it would be desirable to be able to offer computational services without endangering the patient privacy. While data in transit can easily be protected by transport layer security, the data must be stored in the clear during processing. Most algorithms and schemes are either optimized for speed with no consideration for data protection and thus cannot be used to offer services. On the other hand the theoretical Private Information Retrieval (PIR) schemes that protect the privacy of patient data are so slow that they are not feasible for the real world use. Since the search spaces represented for instance by the genome or proteome of complex organisms are immense, fast privacy preserving search algorithms are needed. In the previous work we introduced the foundation for such a privacy preserving genome search engine. In this work, we improve and elaborate on this and present an extensive evaluation and comparison showing that this scheme is both secure and practical. Our approach is based on Bloom filters with a configurable security property that performs more than 2000 times faster than PIR equivalents for large datasets, making it suitable for applications in bioinformatics. The results can then be further aggregated using Homomorphic Cryptography to allow an exact-match searching. In performance tests a search of a 50-nucleotides-long sequence against human chromosomes can be securely executed in less than 0.1 s on a 2.8 GHz Intel Core i7. We offer the entire system as an open source service for the community and offer ready-to-use REST as well as SOAP Web services.

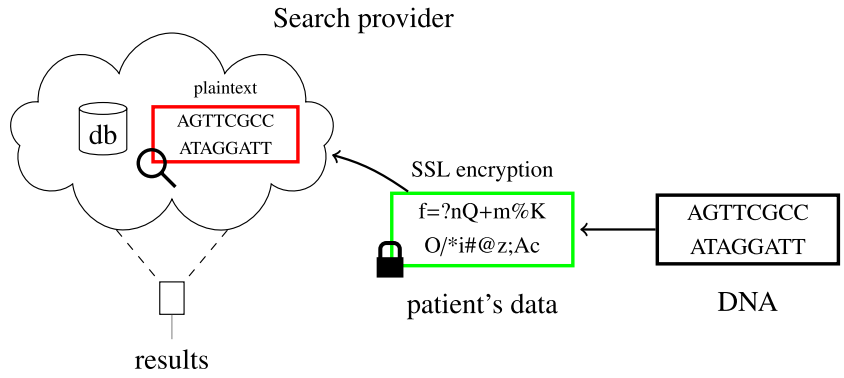
1. Introduction

Trust remains a crucial challenge concerning the outsourcing of computational tasks to the Cloud in the biomedical domain. Data can be transferred via encrypted channels, but as soon as processing begins, the data is disclosed to the Cloud provider that subsequently has reading access to the data and applies the processing algorithm to it in the clear (c.f. Fig. 1). This is a significant hindrance for many data intensive applications that could make good use of Cloud computing but have privacy and data protection requirements that forbid the disclosure of sensitive data to a third party. The search for DNA or a protein sequence in a database is an example of a biomedical application that uses patient related data that may allow re-identification of the subject. Gymrek et al. [1] describe how the re-identification of subjects is possible using their publicly available DNA data that had been thought of as anonymized.

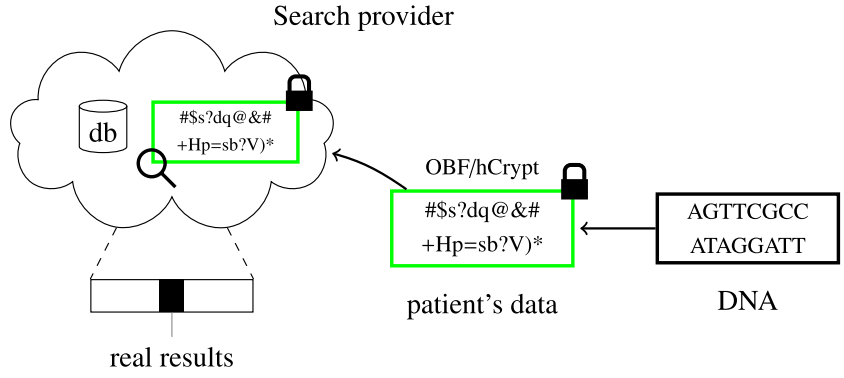
Homomorphic Cryptography (hCrypt, c.f. [2] for Gentry's breakthrough work) is a theoretically promising technology that provides a solution to this problem by allowing a remote party to execute arbitrary algorithms on encrypted data. This would allow the outsourcing of privacy constrained computational tasks to

* Corresponding author. Tel.: +49 511 762 79 58 69.
E-mail addresses: perl@dcsec.uni-hannover.de (H. Perl), y.mohammed@lumc.nl (Y. Mohammed), brenner@dcsec.uni-hannover.de (M. Brenner), smith@dcsec.uni-hannover.de (M. Smith).

0167-739X/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.future.2013.12.006



(a) Encryption of transfer channels only.



(b) This article: obfuscation of the search query.

Fig. 1. Achieving privacy of patients' data.

Cloud resources without the need to establish trust, as the Cloud provider could not read the encrypted data. However, the performance of plain hCrypt for large data sets makes it unfeasible for most real world problems if used as a stand-alone solution, since recrypts are necessary after every AND gate and recrypts are a very expensive operation. Even in a recent scheme by Coron et al. [3], a recrypt operation takes 51 s for $\lambda = 62$ bits of security. Thus, the overhead from hCrypt housekeeping quickly dominates the real work.

The family of Private Information Retrieval (PIR) schemes establishes a cryptographic protocol that allows a user to search the database without revealing which item was queried. While some progress has been made in terms of runtime and communication complexity by Boneh et al. [4], Gentry and Ramzan [5], and Kushilevitz and Ostrovsky [6], searching the database is still at best bound linearly to the size of the database. For applications in biology and bioinformatics with large datasets, PIR schemes are still not feasible, as shown by our performance comparison.

1.1. Our contributions

Since hCrypt and PIR both solve the problem of private search in theory only, we have developed a Bloom filter based approach to privacy preserving search on databases [7] that allows for the feasible outsourcing of searching a sequence in a database by introducing a privacy/performance trade-off: instead of a hard security guarantee, the search result is hidden in noise. The ratio between real search and the noise can be configured so that it conforms to data protection regulations. The huge benefit of this somewhat weaker security guarantee is a performance increase of a factor of 2650 for the time it takes to execute a query compared to the recent PIR schemes (see Section 10.1). In detail, the runtime complexity of our scheme is in $\mathcal{O}(\log |A| + |s| + |R|)$ for a database A with the search term s and the results set R. This qualifies the algorithm to search large data sets.

The search algorithm operates on an Obfuscated Bloom filter (OBF) of the search term. In this paper, we prove that it is cryptographically hard to deduce the search term from the OBF or the results set. Furthermore, the additive property of Bloom filters is used to combine a set of queries into one that matches any search term of the set. This makes searching in a stream as efficient as searching in a set of discrete strings. We present a real world application of this principle from the biomedical domain and show that protected and secured search of encrypted DNA sequence queries in the complete human chromosomes is feasible. The output of this Bloom filter based search algorithm is then large enough to conform to the data protection regulations, as well as small enough to allow further processing using hCrypt.

In this article, we also present an analysis of how the Bloom filter search performs against Private Information Retrieval schemes on real-world datasets. In performance tests a search of a 50 nucleotides long sequence against human chromosomes can be securely executed in less than 0.1 s on a 2.8 GHz Intel Core i7. Finally, we show how the Bloom filter search can be integrated into the existing e-Science ecosystems as a Web service. We implement the Bloom filter search as a ready-to-use REST as well as a SOAP Web service.

1.2. Outline

In Section 2, we consider alternative methods to solve the problem of privacy preserving search possibilities and evaluate those against the Bloom filter search. In Section 3, a common notation for homomorphic encryption schemes and Bloom filters is established. Obfuscated Bloom filters are introduced in Section 4 as a crucial component to the search. In Section 5, we discuss the Bloom