Aus dem Institut für Neuro- und Bioinformatik der Universität zu Lübeck
Direktor: Prof. Dr. rer. nat. Thomas Martinetz

Privately computing the intersection of two SNP Sets

Untersuchung der genetischen Ursachen des Divry-Van Bogaert Syndroms

Bachelorarbeit
im Rahmen des Studienganges Medizinische Informatik
der Universität zu Lübeck

vorgelegt von
Niklas Jobst

ausgegeben und betreut von
PD Dr. rer. nat. Amir Madany Mamlouk

mit Unterstützung von
Prof. Dr. rer. nat. Jeanette Erdmann,
Dr. Ingrid Braenne, Benedikt Reiz

Das Praktikum wurde im Institut für theoretische Informatik der Universität zu Lübeck absolviert.

Lübeck, den November 12, 2017

Contents

1 Einleitung **1**
1.1 Abstract . . . 1
1.2 Genetische Marker . . . 1
1.2.1 SNPs . . . 1
1.2.2 INDELs . . . 1
1.3 Personalisierte Medizin . . . 1
1.4 Anwendung . . . 2

2 Methoden **3**
2.1 Bloom Filter . . . 3
2.2 Kryptosysteme . . . 3
2.2.1 Homomorphie . . . 3
2.2.2 Elgamal . . . 3
2.2.3 Pailier . . . 4
2.2.4 Goldwasser-micali . . . 4
2.3 Implementierte Algorithmen . . . 4
2.3.1 Algorithmus 1 - Elgamal . . . 4
2.3.2 Algorithmus 2 - Paillier . . . 4
2.3.3 Algorithmus 3 - Goldwasser-Micali . . . 4

1 Einleitung

1.1 Abstract

Ziel dieses Praktikums war es die Frage zu erörtern, wie zwei Partein die Ähnlichkeit ihrer DNA berechnen können, ohne, dass dabei eine der Parteien Informationen über den genetischen Code der jeweils anderen erlangt.

Die Grundlagen für diese Berechnungen basieren auf bereits existierenden Methoden, mit welchen der Schnitt zweier Mengen unter Sicherung der Privatsphäre berechnet werden kann.

Im Zuge dieses Praktikums werden wir drei dieser Methoden mit Bezug zum gegebenen Anwendungsfall implementieren und deren Effizienz miteinander vergleichen:

- R.Egert et al. : Privately Computing Set-Union and Set-Intersectio Cardinality via Bloom Filters, LNCS volume 9144, 2015
- A.Davidson et al. : An Efficient Toolkit for Computing Private Set Operations, LNCS volume 10343, 2017
- S. K.Debnath et al. : Secure and Efficient Private Set Intersection Cardinality Using Bloom Filter, LNCS volume 9290, 2015

1.2 Genetische Marker

Bestimmte klar definierte Sequenzen und Positionen im genetischen Code können dazu genutzt werden Personen zu identifizieren. Der genetische Code ist bei allen Menschen zu ca.99% gleich. ogenannte

1.2.1 SNPs

1.2.2 INDELs

1.3 Personalisierte Medizin

In der personalisierte Medizin werden individuelle Eigenschaften von Personen berücksichtigt die

1.4 Anwendung

In der Personalisierten Medizin sind Therapien bestimmte genetische Profile gekoppelt. Um festzustellen, ob eine Therapie für einen Patienten zulässig ist, muss daher zunächst sein genetischer Code mit dem für diese Therapie notwendigem verglichen werden. Derzeit werden diese Vergleiche ohne die entsprechenden Datensicherheits-Vorkehrungen vorgenommen. Ziel dieses Praktikums war es durch Anwendung der genannten Methoden die Sicherung der Privatsphäre bei der Durchführung eines solchen Vergleichs zu erhöhen.

2 Methoden

2.1 Bloom Filter

Alle diese Methoden basieren auf sogenannten Bloomfiltern. Hierbei handelt es sich um eine Technik um festzustellen, ob bestimmte Daten in einem Datensatz vorhanden sind oder nicht. Sie bestehen aus einem mit Nullen vorinitialisiertem m Bit langen Array und k Hashfunktionen, welche auf die Positionen des Arrays abbilden.

Zur Initialisierung werden auf jedes Element des Datensatzes alle k Hashfunktionen angewendet. Die zur Ausgabe der Hashfunktionen korrespondierenden Bits im Array werden darauf hin auf Eins gesetzt.

Soll für ein Datenelement geprüft werden, ob dieses Teil des Datensatzes ist, werden alle Hashfunktionen auf dieses angewendet.
Nur wenn alle Positionen im Array an den korrespondierenden Punkten der Ausgabe dem Wert Eins entsprechen wird angenommen das sich das Element im Datensatz befindet.

Diese Überprüfung ist nicht resistent gegenüber

2.2 Kryptosysteme

2.2.1 Homomorphie

Homomorphie bezeichnet eine Eigenschaft von Kryptosystemen. Ein Kryptosystem ist genau dann homomorph gegenüber einer mathematischen Operation, wenn Berechnungen im Ciphertext mit dieser Operation denen im KLartext entsprechen.

2.2.2 Elgamal

Bei Elgamal handelt es sich um ein im Jahr 1985 vom Kryptologen Taher Elgamal entwickeltes Public-Key-Verschlüsselungsverfahren. Elgamal ist eine Erweiterung des Diffie-Hellmann Schlüsselaustausches.

Elgamal ist homomorph gegenüber der Multiplikation

$$E(m_1 * m_2) = (E(m_1) * E(m_2))$$

2.2.3 Pailier

2.2.4 Goldwasser-micali

2.3 Implementierte Algorithmen

2.3.1 Algorithmus 1 - Elgamal

2.3.2 Algorithmus 2 - Paillier

2.3.3 Algorithmus 3 - Goldwasser-Micali