Hessisches Kompetenzzentrum
für Hochleistungsrechnen

# SecBic - A Platform for Genome Privacy Based on Homomorphically Biclustering Analysis

Researchers
Shokofeh VahidianSadegh

Principal Investigator
Prof. Dr. Lena Wiese

Project Term
2022 - 2022

Clusters
Lichtenberg II Cluster Darmstadt

Additional Software
Pyfhel, Biclustlib

Institute
Database Technologies and Data Analytics

University
Goethe Universität Frankfurt am Main

## Introduction

Massive amounts of newly generated gene expression data have been used to further enhance personalised health predictions. High Performance Computer (HPC) enable us testing partially some intermediate computations in our research beside a single sever in our research group.

## Methods

We proposed SeCCA (Secured Cheng and Church Algorithm) in which a large volume of gene expression data is encrypted by using one of highly efficient cryptographic approach – Fully Homomorphic Encryption (FHE).

## Results

For the first time, we achieved homomorphically biclustering analysis of gene expression data starting with Cheng and Church algorithm and a given number of biclusters. Homomorphic encryption operations are capable of calculating the maximum mean squared residue accepted, and mean squared residues score in both phases (node deletion and addition) for an input data matrix which are important criteria in our algorithm with predefined parameters. Our experiment reveals a meaningful analysis of the sample yeast gene expression data and synthetic data based on a constant bicluster model with 5 biclusters and

subsets of rows and columns (genes and conditions, respectively). While this demonstration shows the feasibility of computation directly over encrypted gene expression data in Cheng and Church algorithm, however, on the downside, we observe significantly increased execution time for a couple of hours. The complexity of the algorithm arises by encrypted multiplication increasing the size of the output polynomial, which slows down operations. The execution is still efficient in this study for 5 biclusters and given data set consisting of 2884 genes and 17 conditions. However, for large sizes of biclusters (e.g., 100 biclusters), there has to be optimisation in time performance to obtain the result within acceptable time frames. Further, no appropriate homomorphic operator can currently be used to implement the condition to add/remove nodes from biclusters; changing the expressions to make them run on homomorphic encryption is necessary. We proposed to execute the necessary evaluation of conditional branching on the client side: hence, with this way of implementation, our proposed SeCCA secured Cheng and Church algorithm only works partially on the cloud side. In a similar way, as Pyfhel does not support operations to find the maximum and minimum value of the data set with homomorphic operations, we continue with encrypting their plaintext values. The quality of the bicluster needs to be assessed by measures or evaluation functions. A good quality measure for bicluster should be taken into account in practice as criterion to compare the resulting biclusters from both encrypted/ nonencrypted approaches. In literature, a number of evaluation measures have been proposed: external validation measures as computations dependent on prior knowledge of ground truth data (e.g., recovery and relevance score) and internal ones which are useful for real data sets for which ground truth is unknown (e.g., average Spearman’s Rho, transposed Virtual Error). Suitable external evaluation measures assess generated biclusters in terms of similarity and accuracy of the results. Actually, no single quality measure has been proposed that can be useful for finding all types of biclusters from the gene expression data and is utilized to extract specific types of biclusters. However, measures including Clustering Error (CE) have been used in the empirical analysis when the true biclustering structure is known. Hence, evaluation of correctness of our results were executed by using CE.

## Discussion

We achieved homomorphically biclustering analysis of gene expression data. As a future work, we plan to improve the overall accuracy of SeCCA and enhance the scalability of our approach (in particular regarding the number of biclusters) so that usage of HPC is highly beneficial for our future prospects.

## Publications

VahidianSadegh, S.; Wiese, L.; Brenner, M.: SeCCA: Homomorphic Encryption Based Privacy Preservation Scheme for Biclustering Algorithm, 34th Crypto-Day 2022, Weiden i.d.Opf., Germany, June 9-10, 2022.

VahidianSadegh, S.; Wiese, L.; Brenner, M.: Homomorphically Biclustering Analysis for Gene Expression Data, CrossFyre Workshop 2022, Passau, Germany, October 6-7, 2022.

## *Last Update:* 2023-09-11 16:49