Homomorphically Biclustering Analysis for Gene Expression Data

Shokofeh VahidianSadegh[1][0000−0002−6464−6842] and
Lena Wiese[1][0000−0003−3515−9209]

Institute of Computer Science
Goethe University Frankfurt
Frankfurt, Germany
VahidianSadegh@mathematik.uni-frankfurt.de
lwiese@cs.uni-frankfurt.de

Abstract. Ever-increasing genomic data emphasise a pressing need for capturing the underlying patterns in gene expression datasets using efficient and precise methods [1]. Machine learning techniques are employed to gain information from these gene expression data. Clustering aims to group items into subsets with similar profiles by identifying and grouping items (genes or conditions) into clusters based on distance and similarity functions [2]. Although traditional clustering algorithms as one-way clustering methods group observations according to similarities among all variables at the same time [3]. Some studies elaborate on the fact that a biological process may be active only under subsets of genes as well as subsets of samples. Accordingly, a traditional clustering method is not able to answer critical research questions [4]. Hence it is of paramount importance to go beyond the traditional clustering prototype and apply a more adapted technique like the biclustering [5] which simultaneously clusters the rows and columns of a matrix for sorting heterogeneous data into homogeneous blocks [6]. Biclustering proves to achieve tremendous success in revealing potential diagnostic biomarkers and has been commonly applied to genomic datasets [7]. Biclustering algorithm namely Cheng and Church Algorithm (CCA) for the first time introduced biclustering as a new paradigm to simultaneously cluster both genes and conditions. Since then, many other such algorithms have been published but mostly compare their works with this algorithm which has been a foundation and proof-of-concept implementation of biclustering. The concept of bicluster in this algorithm refers to a subset of genes and a subset of conditions with a high similarity score, which measures the coherence of the genes and conditions in the bicluster. It also returns the list of biclusters for the given data set.
Moreover, large biomedical studies require the data to be subsequently shared for joint analysis. Patients' biomedical data are sensitive and despite the importance of data sharing, it demands caution and patient privacy due to the inherently private nature of medical data [8]. Recent efforts have been made to encrypt private and personal genomic data files and analysis on encrypted data; however, there is a noticeable research gap in the area of processing data by biclustering algorithms that are privacy-preserved. Homomorphic encryption can provide privacy for

genomic data by allowing data to remain encrypted even during computation.
In our work, we proposed a comparative privacy-preserving framework consisting of the Cheng and Church Algorithm using homomorphic encryption over yeast cell cycle expression as well as synthetic data collections to process gene expression data while keeping private data encrypted. We evaluated the resulting biclusters from the aforementioned biclustering algorithm with the non-encrypted original counterpart in terms of similarity by external evaluation measures such as Clustering Error and Campello Soft Index [9] then compare the time performance of encryption, decryption, and the total amount for the implemented biclustering algorithm. We applied Pyfhel [11] as a python wrapper for the Microsoft SEAL library and used BFV [10] as a fully homomorphic encryption scheme in our project. Despite proving the applicability of homomorphic encryption over biclustering algorithms but with some limitations including finding minimum and maximum values of the encrypted data matrix and in particular, branching over conditionals in encrypted program flows appears to be less efficient when using Pyfhel then we instead propose a multi-round interactive execution of the conditional checking part by involving the data owner. In our recent work, we decided to benefit from a different fully homomorphic encryption library called concrete-numpy as a python implementation of CONCRETE (Concrete Operates oN Ciphertexts Rapidly by Extending TfhE) [12] which is a Zama's variant of the TFHE scheme [13] and provides an extensive number of operations such as programmable bootstrapping to work with ciphertexts. Of particular importance is that our method introduces for the first time a less non-interactive privacy-preserving biclustering algorithm over gene expression data by TFHE fully homomorphic encryption scheme.

Keywords: Biclustering Algorithm · Gene Expression · Homomorphic Encryption.

References

1. Orzechowski, Patryk and Pańszczyk, Artur and Huang, Xiuzhen and Moore, Jason H runibic: a Bioconductor package for parallel row-based biclustering of gene expression data Bioinformatics **34**(24) 4302–4304 (2018)
2. Rocha, Orlando and Mendes, Rui JBiclustGE: Java API with unified biclustering algorithms for gene expression data analysis Knowledge-Based Systems **155** 83–87 (2018)
3. Tu, Wangshu and Subedi, Sanjeena A family of mixture models for biclustering Statistical Analysis and Data Mining: The ASA Data Science Journal **15**(2) 206–224 (2022)
4. Padilha, Victor Alexandre and de Leon Ferreira, André Carlos Ponce and others Experimental correlation analysis of bicluster coherence measures and gene ontology information Applied Soft Computing **85** 105688 (2019)
5. Maâtouk, Ons and Ayadi, Wassim and Bouziri, Hend and Duval, Béatrice Evolutionary biclustering algorithms: an experimental study on microarray data Soft Computing **23**(17) 7671–7697 (2019)

6. Ngo, Michelle N and Pluta, Dustin S and Ngo, Alexander N and Shahbaba, Babak Conjoined Dirichlet Process arXiv preprint arXiv:2002.03223 (2020)
7. Orzechowski, Patryk and Moore, Jason H EBIC: an open source software for high-dimensional and big data analyses Bioinformatics **35**(17) 3181–3183 (2019)
8. Oestreich, Marie and Chen, Dingfan and Schultze, Joachim L and Fritz, Mario and Becker, Matthias Privacy considerations for sharing genomics data EXCLI journal **20** 1243 (2021)
9. Padilha, Victor A and de Carvalho, André CPLF.: A Comparison of Hierarchical Biclustering Ensemble Methods. In: 2017 Brazilian Conference on Intelligent Systems (BRACIS), pp. 318–323. (2017)
10. Brakerski, Zvika and Gentry, Craig and Vaikuntanathan, Vinod (Leveled) fully homomorphic encryption without bootstrapping ACM Transactions on Computation Theory (TOCT) **6**(3) 1–36 (2014)
11. Ibarrondo, Alberto and Viand, Alexander.: Pyfhel: Python for homomorphic encryption libraries. In:Proceedings of the 9th on Workshop on Encrypted Computing & Applied Homomorphic Cryptography, pp. 11–16. (2021)
12. Chillotti, Ilaria and Joye, Marc and Ligier, Damien and Orfila, Jean-Baptiste and Tap, Samuel.: CONCRETE: Concrete Operates oN Ciphertexts Rapidly by Extending TfhE. WAHC 2020–8th Workshop on Encrypted Computing & Applied Homomorphic Cryptography **15** (2020)
13. Chillotti, Ilaria and Gama, Nicolas and Georgieva, Mariya and Izabachène, Malika.: TFHE: fast fully homomorphic encryption over the torus. Journal of Cryptology **33**(1) pp. 34–91 (2020)