American Journal of
Biomedical Science & Research

www.biomedgrid.com
ISSN: 2642-1747

Minireview

Analysis of Privacy Protection Methods for DNA Motif Finding

Yongting Zhang[1,2], Minyu Shi[1,2], Huanhuan Wang[1,2] and Xiang Wu[1,2]*

1School of Medical Informatics, Xuzhou Medical University, China

2School of Computer Science and Technology, Dongguan University of Technology, China

***Corresponding author:** Xiang Wu, School of Medical Informatics, Xuzhou Medical University and Dongguan University of Technology, 523808, Dongguan, China.

To Cite This Article: *Xiang Wu.Analysis of Privacy Protection Methods for DNA Motif Finding. Am J Biomed Sci & Res. 2019 - 6(3). AJBSR. MS.ID.001023. DOI:* 10.34297/AJBSR.2019.06.001023.

Received: October 30, 2019 ; **Published:** November 20, 2019

Abstract

DNA motif finding is a repetitive expressive sequence fragment that found in given DNA sequence sets and its precise location is of significance to fully comprehend the regulation mechanism of genetic expression. Motif finding is the key to grasp the mechanism of genetic transcriptional regulation, however, the security and privacy issues of motif finding are so overwhelming that we must to pay more attention to it. In our paper, we simply overviewed the methods for protecting motif finding privacy from three broad perspectives: Controlled access, anonymity method, and ε -differential privacy.

Keywords: DNA Motif Finding; Privacy Protection; Privacy Disclosure

Introduction

The implementation of the genome project makes DNA sequence analysis become a top priority for Bioinformatics. Sequence alignment and motif finding are the two main directions of biological sequence analysis. DNA sequence motif is the short and recurred patterns in DNA sequences that are assumed to have the biological function [1-3]. In 1975, Professor Pribnow used the early multi-sequence comparison methods to analyze the promoter region of yeast, found a TATA box that refers to a highly conserved and consistent 10 bp patterns, which was the first time that people found motifs [4]. Recently, people used stochastic evolution methods for motif finding. After more than 40 years of development, the research on the method of motif finding has grown exponentially. National scientific research projects such as motif finding in genetic researches are growing at an annual rate of 30 ; by the 2020 publication year, the number of papers about genetic researches in the direction of motif finding published as many as 64,116 ; at the same time, thousands of DNA motif finding algorithms [5-9] and platforms [10-12] have also been developed.

In fact, DNA sequences analysis gives the access to make sense of amounts of information about a person's characteristics, function, illnesses, and personality disorders and his or her genetic relatives [13-14] which are very privately. These private information's are easily leaked in the mining of DNA sequences.

In this paper, we firstly simply investigated the privacy leakage types of DNA motif finding as well as some methods for its discovery process. Then, we overviewed the above content from three broad perspectives: controlled access, anonymity method, and differential privacy. This paper is organized as follows. Section 2 summarized the current privacy protection methods for DNA motif finding. Section 3 briefly reviewed the main content of this paper.

Methods

With the explosive growth of DNA data, making full use of data is the only way to increase the value of DNA data. However, the privacy protection of DNA data has clearly become a bottleneck in the development of DNA sequence analysis especially in motif finding. How to solve the privacy leakage when finding results sharing is an urgent problem to be solved. Therefore, through reviewing the current motif finding privacy protection technologies, we think that there are three main ways to address privacy leakage: controlled

access [15-17], anonymity method [18-20], and -differential privacy [21-24].

Controlled Access

The access control method is the same as the dbGaP file download in the NCBI database, which allows the user to obtain and manipulate the specified data after having the approval and within the granted operation authority. Also, Controlled access to protect DNA motif finding results privacy relies on central control. In most DNA motif finding web platforms, users obtain experimental results by contacting the manager. For instance, [25] developed a MEME algorithm-based motif finding Web platform, and there is a "mailbox sending" option in the key steps of obtaining motif finding results. This one-to-one controlled access method guarantees the privacy of the publication of the DNA motif finding results. However, a lot of communication work, longer verification and audit times become potential bottlenecks for this method.

k-Anonymity Method

The main idea of this method is that by generalizing/concealing the target data. Each record in the published DNA data set has records that are indistinguishable from each other on the Quasi-Identifier. The probability that the attacker discriminates the individual's private information from the published data set is less than, thus effectively protecting the personal privacy of the data owner. For example, [26] used the k-anonymity-based method before performing DNA motif finding, and successfully protect the privacy of DNA data sharers. However, due to the particularity of DNA sequence data, it is easy to overgeneration data by applying k-anonymity method in this field, which makes DNA data analysis lose its value.

ε -Differential Privacy

Although it is widely believed that improper use of DNA data can reveal personal privacy, it is still uncertain what types of privacy leakage is caused by what information or background knowledge [27] an attacker might launch an attack. These can be solved by -differential privacy, which is a powerful method for current applications in the field of DNA motif finding privacy protection. -differential privacy requires that the results of any analysis cannot be relied on any single data record, and similarly in the process of performing DNA motif finding, referring to any single DNA sequence. For instance, in [24], author proposed a high-utility motif finding algorithm based on -differential privacy. Their solution was that make use of the closed frequent pattern set to reduce redundant motifs of result sets and obtain accurate motifs results, then use -differential privacy to protect motif finding results. Therefore, when motif finding results are shared, -differential privacy can ensure that the privacy information of it is not disclosed even if when the attacker mastered the background information of all the data except a certain DNA sequence. However, the use of -differential privacy in DNA motif finding has problems such as large redundancy of results.

Conclusions

DNA sequence analysis will deepen our understanding of human health or disease and plays a major role in discovering the cause of disease and achieving prevention, diagnosis and personalized medical treatment. But the rich information contained in the DNA sequence is easily leaked out during the motif finding. In this article, we describe techniques for protecting human genetic privacy in three ways: Controlled access, anonymity method, and differential privacy. Of course, these are not perfect methods for privacy protection methods for DNA motif finding, because the problem is always imposed.

In this context, the future direction is clearly, and struggle will focus more on effective solutions to the problem of genetic privacy and security. We foresee that it is very necessary to seriously investigate and adopt varieties of methods for DNA motif finding privacy protection.

References

1. D' Haeseleer P (2006) What are DNA sequence motifs? Nature Biotechnology 24(4): 423-425.
2. Tran N T L, Huang CH (2014) A survey of motif finding Web tools for detecting binding site motifs in ChIP-Seq data. Biology Direct 9(1): 4.
3. Zaslavsky E, Singh M (2006) A combinatorial optimization approach for diverse motif finding applications. Algorithms for Molecular Biology 1(1): 13.
4. David Pribnow (1975) Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter. Proc Nat Acad Sci USA 72(3): 784-788.
5. Hertz G Z, Hartzell G W, Stormo G D (1990) Identification of Consensus Patterns in Unaligned DNA Sequences Known to be Functionally Related. Comput Appl Biosci 6(2): 81-92.
6. Saurabh S, Martin T (2003) YMF: a program for discovery of novel transcription factor binding sites by statistical overrepresentation. Nucleic Acids Research 31(13): 3586–3588.
7. Pesole G, Prunella N, Liuni S, Attimonelli M, Saccone C (1992) WORDUP: an efficient algorithm for discovering statistically significant patterns in DNA sequences. Nucleic Acids Research 20(11): 2871-2875.
8. Lawrence C E, Reilly A A (1990) An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. Proteins Structure Function and Genetics 7(1): 41-51. Bailey TL, Elkan C (1995) Unsupervised Learning of Multiple Motifs in Biopolymers Using Expectation Maximization[C]// Machine Learning.
9. William T, Rouchka EC, Lawrence CE (2003) Gibbs Recursive Sampler: finding transcription factor binding sites. Nucleic Acids Research 31(13): 3580-3585.
10. Kurtz S, Choudhuri J V, Ohlebusch E, Schleiermacher C, Stoye J, et al. (2001) RE Puter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Research 29(22): 4633-4642.
11. Li N, Tompa M (2006) Analysis of computational approaches for motif discovery. Algorithms for Molecular Biology 1(1): 8.
12. Phan V, Furlotte NA (2008) Motif Tool Manager: a web-based framework for motif discovery. Bioinformatics 24(24): 2930-2931.

13. Roche PA, Annas GJ (2006) DNA testing, banking, and genetic privacy. N Engl J Med 355: 545-546.

14. Takai-Igarashi T, Kinoshita K, Nagasaki M, Tanaka H, Yamamoto M, et al. (2017) Security controls in an integrated Biobank to protect privacy in data sharing: rationale and study design. BMC Medical Informatics and Decision Making 17(1): 100.

15. Kim J W, Kim D H, Jang B (2018) Application of local differential privacy to collection of indoor positioning data. IEEE Access 6: 4276-4286.

16. Johnson A, Shmatikov V (2013) Privacy-preserving data exploration in genome-wide association studies. KDD pp. 1079-1087.

17. Simmons S, Berger B (2016) Realizing privacy preserving genome-wide association studies. Bioinformatics 32(9): 1293-1300.

18. Simmons S, Sahinalp C, Berger B (2016) Enabling privacy preserving GWASs in heterogeneous human populations. Cell syst 3(1): 54-61.

19. Huang Z, Hubaux J P, Ayday E (2015) Differential Privacy with Bounded Priors: Reconciling Utility and Privacy in Genome-Wide Association Studies. ACM Sigsac Conference on Computer and Communications Security. ACM 1286-1297.

20. Uhlerop C, Slavković A, Fienberg SE (2013) Privacy-Preserving Data Sharing for Genome-Wide Association Studies. Journal of Privacy & Confidentiality 5(1): 137-166.

21. Yu F, Ji Z (2014) Scalable privacy-preserving data sharing methodology for genome-wide association studies: an application to i DASH healthcare privacy protection challenge. Bmc Medical Informatics & Decision Making 14(1): S3-S3.

22. Yu F, Rybar M, Uhler C, Stephen E Fienberg (2014) Differentially private logistic regression for detecting multiple-SNP association in GWAS databases. International Conference on Privacy in Statistical Databases. Springer Cham pp. 170-184.

23. Wang X, Lin P, Ho J (2018) Discovery of cell-type specific DNA motif grammar in cis-regulatory elements using random Forest. Bmc Genomics 19(1): 929.

24. Wu X, Wei Y, Mao Y (20181) A differential privacy DNA motif finding method based on closed frequent patterns. Cluster Computing (21): 1-13.

25. Bailey T L, Johnson J, Grant C E, et al. (2015) The MEME Suite. Nucleic Acids Research 43(W1): W39-W49.

26. Malin B A (2005) Protecting genomic sequence anonymity with generalization lattices. Methods of Information in Medicine 44(5): 687-692.

27. Malin BA (2005) An evaluation of the current state of genomic data privacy protection technology and a road map for the future. J Am Med Inform Assoc 12(1): 28-34.