journal homepage: www.intl.elsevierhealth.com/journals/cmpb

Improvements on a privacy-protection algorithm for DNA sequences with generalization lattices

*Guang Li**, *Yadong Wang*, *Xiaohong Su*

School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, People's Republic of China

ARTICLE INFO

Article history:
Received 29 April 2010
Received in revised form
9 February 2011
Accepted 21 February 2011

Keywords:
Personal DNA database
Privacy protection
k-Anonymity

ABSTRACT

When developing personal DNA databases, there must be an appropriate guarantee of anonymity, which means that the data cannot be related back to individuals. DNA lattice anonymization (DNALA) is a successful method for making personal DNA sequences anonymous. However, it uses time-consuming multiple sequence alignment and a low-accuracy greedy clustering algorithm. Furthermore, DNALA is not an online algorithm, and so it cannot quickly return results when the database is updated. This study improves the DNALA method. Specifically, we replaced the multiple sequence alignment in DNALA with global pairwise sequence alignment to save time, and we designed a hybrid clustering algorithm comprised of a maximum weight matching (MWM)-based algorithm and an online algorithm. The MWM-based algorithm is more accurate than the greedy algorithm in DNALA and has the same time complexity. The online algorithm can process data quickly when the database is updated.

1. Introduction

With the development of genotyping technology, DNA sequences are increasingly becoming a part of the patient medical record [1], and an increasing amount of personal DNA sequences are being collected. Databases of personal DNA sequences are also being developed. The collection of DNA occurs at many different kinds of institutions: at research sites for clinical trials and basic research, at hospitals for diagnostic testing, and at commercial companies, where gene discovery is of high commercial value.

At the same time, genomic data pose complex privacy problems. The genetic information of an individual is as personally revealing as a fingerprint, if not more revealing [2]. Many people fear that information gleaned from their genomic data, such as a health situation or a family member relationship, will be misused or abused to influence their employment and insurance status or simply to cause social stigma [3,4]. In addition to social pressures, there are legal mechanisms for protecting genomic data privacy, such as the Privacy Rule of the Health Insurance Portability and Accountability Act in the United States and the Data Protection Act of 1998 in the European Union. Thus, without an appropriate guarantee of anonymity, not only will patients be less willing to provide data but also many data collectors will be unable to share genomic data for worthwhile endeavors. Given this situation, genomic privacy is considered one of the major challenges for the biomedical community [5,6].

Unfortunately, contrary to popular belief, the protection of a patient's anonymity in genomic data is not as simple as removing or encrypting explicit identifying attributes, such as name or social security number [7–10]. To resolve this problem, a DNA sequence anonymity method called DNA lattice anonymization (DNALA) has been presented [11]. It is based on the *k*-anonymity principle.

* *Corresponding author at*: Room 434, the A17 Student Dormitory, Science Park, Harbin Institute of Technology, No. 2 Yikuang Street, Nangang District, Harbin, Heilongjiang 150001, People's Republic of China. Tel.: +86 15846591735.
E-mail address: hit6006@126.com (G. Li).
0169-2607/$ – see front matter © 2011 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.cmpb.2011.02.013

Although DNALA can protect the privacy of personal DNA data, it has some shortcomings. Specifically, it uses the multiple sequence alignment, which is time-consuming, together with a low-accuracy greedy clustering algorithm. Furthermore, DNALA is not an online algorithm and, thus, cannot quickly return results when the database is updated.

Li et al. [12] improved the original DNALA method by replacing the multiple sequence alignment with global pairwise sequence alignment to save time; they also replaced the greedy clustering algorithm with stochastic hill-climbing method to improve the precision of the clustering. However, the stochastic hill-climbing method is not an online algorithm; moreover, there is still potential for improvements in precision.

Our study in this paper focused on overcoming the shortcomings in DNALA. When calculating distance matrix, this study used the method in Li et al. [12] by replacing the multiple sequence alignment in DNALA with a global pairwise sequence alignment. For clustering, it replaced the greedy algorithm in DNALA with a hybrid algorithm consisting of a maximum weight matching (MWM)-based algorithm and an online algorithm. Compared with DNALA, our new method is faster, especially when the database is updated, and it can achieve more accurate results.

The rest of this paper is organized in the following manner. Section 2 introduces some related work, while Section 3 provides details on the original DNALA method. Section 4 explains the principal ideas behind our method, and Section 5 presents the improvements for calculating the distance matrix. Section 6 discusses the MWM-based clustering algorithm, and Section 7 discusses the online clustering algorithm. Finally, Section 8 presents the results of the experiments, and Section 9 presents the conclusions.

2. Related work

2.1. *k*-Anonymity

The k-anonymity principle is a privacy protection principle proposed by Samarati and Sweeney [13–15]. It has been extensively studied in recent years [16–19].

The key idea behind k-anonymity is to make individuals indistinguishable in a released table. A data set complies with k-anonymity protection if each individual's record stored in the released data set cannot be distinguished from at least $k-1$ other individuals, whose data also appear in the data set. This protection guarantees that the probability of identifying an individual based on the released data in the data set does not exceed $1/k$. The larger the value of k is, the better the protection.

Generalization and suppression are the two main methods for achieving k-anonymity. Generalization means replacing a value by a less specific, more general value that is faithful to the original one, for example, by using "Europe" to replace "Netherlands" or using "[20–30]" to replace "27". Suppression means removing data from a table so that they are not released. Suppression can be done for tuples [13] or attributes [15,19]. Generalization and suppression can be used alone or in combination.

Recently, some improvements over k-anonymity have been proposed [20], including p-sensitive k-anonymity [21], (α, k)-anonymity [22], l-diversity [23] and t-closeness [24]. They all can be achieved by similar methods for achieving k-anonymity.

2.2. Genomic data privacy protection

For the research needs of and potential benefits to health care [25], personal genomic data should be shared. But as we showed above, because of social concerns and public policy, person-specific genomic records must be shared in a manner that preserves the anonymity of data subjects.

There are three main ways for protecting privacy for personal genomic data. They are data deidentification, data augmentation and methods based on cryptology.

The deidentification method protects privacy by removing or encrypting person-specific identifiers, such as name and social security number, which are initially associated with genomic records. Recent studies have shown that data deidentification is not enough to protect privacy [7–10].

Data augmentation is often achieved by generalization. It protects privacy by making each record be indistinguishable from some other shared records [11,26,27]. DNALA is a data augmentation method that performs generalization on DNA sequences [11]. The study by Zhen et al. [26] mainly focused on single nucleotide polymorphisms but not DNA sequences. Loukides et al. [27] performed generalization on diagnoses codes but not on DNA sequences.

Cryptology-based methods do not access the original data. They maintain data utility by using privacy-preserving data querying methods that can be applied to genomic sequences [28,29].

2.3. Sequence alignment

In bioinformatics, sequence alignment is a way of arranging the sequences of DNA, RNA or protein to identify regions of similarity that may be a consequence of functional, structural or evolutionary relationships between the sequences. This paper only considers global sequence alignment.

From the perspective of computer science, global sequence alignment is the process of adding gaps to sequences and making them as similar as possible, not least by giving them the same length. The added gaps should be as few as possible.

We assume that sequence alignment is done for n sequences. If $n=2$, it is called pairwise sequence alignment, and if $n>2$, it is called multiple sequence alignment.

For example, if using "#" to represent a gap, then the alignment for "AAACGTTT" and "AAACGCTTT" is as follows.

A	A	A	C	G	#	T	T	T
A	A	A	C	G	C	T	T	T

In addition, the alignment for "AAACGGCTTT", "AAACGTTT" and "AAACGCTTT" is the following.

A	A	A	C	C	G	C	T	T	T
A	A	A	#	C	G	#	T	T	T
A	A	A	#	C	G	C	T	T	T

For pairwise sequence alignment, the ordinary dynamic programming algorithm has time complexity $O(nm)$, where