软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2019,30(4):1094−1105 [doi: 10.13328/j.cnki.jos.005367]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

SNP 连锁不平衡下的基因隐私保护模型*

刘 海[1], 吴振强[1], 彭长根[2], 雷秀娟[1]

[1](陕西师范大学 计算机科学学院,陕西 西安 710119)

[2](贵州省公共大数据重点实验室(贵州大学),贵州 贵阳 550025)

通讯作者: 吴振强, E-mail: zqiangwu@snnu.edu.cn

摘 要: 人类基因测序技术的快速发展,测序成本大幅降低,使基因数据得到广泛的应用,在全基因组的单核苷酸多态性与疾病关联研究中,单核苷酸多态性与患者的身份、表型和血缘关系等敏感信息相关联,单核苷酸多态性连锁不平衡容易导致患者的隐私信息泄露.为此,基于单核苷酸多态性连锁不平衡相关系数,提出矩阵差分隐私保护模型以实现基因数据和单核苷酸多态性连锁不平衡的隐私保护,同时确保基因数据具有一定的效用.该模型可以实现单核苷酸多态性连锁不平衡下全基因组关联研究中基因数据隐私与效用的权衡,并对单核苷酸多态性连锁不平衡下的基因隐私保护具有促进作用.

关键词: 单核苷酸多态性连锁不平衡;差分隐私;基因隐私;基因数据效用

中图法分类号: TP309

中文引用格式: 刘海,吴振强,彭长根,雷秀娟.SNP 连锁不平衡下的基因隐私保护模型.软件学报,2019,30(4):1094−1105. http://www.jos.org.cn/1000-9825/5367.htm

英文引用格式: Liu H, Wu ZQ, Peng CG, Lei XJ. Genomic privacy preserving framework for SNP linkage disequilibrium. Ruan Jian Xue Bao/Journal of Software, 2019,30(4):1094−1105 (in Chinese). http://www.jos.org.cn/1000-9825/5367.htm

Genomic Privacy Preserving Framework for SNP Linkage Disequilibrium

LIU Hai[1], WU Zhen-Qiang[1], PENG Chang-Gen[2], LEI Xiu-Juan[1]

[1](School of Computer Science, Shaanxi Normal University, Xi'an 710119, China)

[2](Guizhou Provincial Key Laboratory of Public Big Data (Guizhou University), Guiyang 550025, China)

Abstract: The cost of sequencing is substantially decreasing with the rapid development of human genome sequencing technologies. The generated genome data are supporting various applications. The genome-wide associated analysis study between the single nucleotide polymorphisms and diseases may lead to more privacy breaches for considering single nucleotide polymorphisms linkage disequilibrium, because of sensitive information related to single nucleotide polymorphisms including individual identity, phenotype, and kinship. To this end, the matrix differential privacy preserving framework is proposed based on the correlated coefficient of single nucleotide polymorphisms linkage disequilibrium. Therefore, this framework can preserve privacy of genome data and single nucleotide polymorphisms linkage disequilibrium, while ensures a certain genome data utility. And it achieves the trade-off between genome data privacy and utility for single nucleotide polymorphisms linkage disequilibrium in genome-wide association studies. Furthermore, the proposed framework plays an important role for promoting genomic privacy preserving under single nucleotide polymorphisms linkage disequilibrium.

* 基金项目: 国家自然科学基金(61173190, 61602290, 61672334, 61662009); 中央高校基本科研业务费专项资金(2016CBY004, GK201704016, GK201501008); 陕西省重点科技创新团队(2014KTC-18)

Foundation item: National Natural Science Foundation of China (61173190, 61602290, 61672334, 61662009); Fundamental Research Funds for the Central Universities (2016CBY004, GK201704016, GK201501008); Key Science and Technology Innovation Team in Shaanxi Province (2014KTC-18)

本文由"面向隐私保护的新型技术与密码算法专题"特约编辑禹勇教授推荐.

收稿时间: 2017-06-01; 修改时间: 2017-07-13; 采用时间: 2017-08-22

Key words: single nucleotide polymorphisms linkage disequilibrium; differential privacy; genomic privacy; genome data utility

基因数据是富含人类重要信息的生物大数据[1],并且是人类脱氧核糖核酸(deoxyribonucleic acid,简称 DNA)序列的总称.DNA 是生物遗传信息的携带者,与生物的繁殖、遗传及变异密切相关.DNA 序列包含 30 亿由 4 种核苷酸(腺嘌呤 A、鸟嘌呤 G、胸腺嘧啶 T、胞嘧啶 C)组成的碱基对,人类有 99.9%共同的 DNA 序列,其中,大约有 5 000 万单核苷酸多态性(single nucleotide polymorphism,简称 SNP).SNP 是最常见的 DNA 变异,SNP 是指个体 DNA 序列同一位置单个核苷酸变异所引起的多态性.SNP 变异由单个碱基的转换(C↔T,在其互补链上则为 C↔A)或颠换(C↔A,G↔T,C↔G,A↔T)所引起,一般所说的 SNP 变异由碱基转换所致.通常对于每个 SNP 位点具有两个不同核苷酸(称为等位基因),一个是高频率的主要等位基因,一个是低频率的次要等位基因.等位基因是同源染色体上的相同位点控制同一性状的不同形式的基因.位点是染色体上一个基因或标记的位置.SNP 的连锁不平衡(linkage disequilibrium,简称 LD)是一种普遍存在的生物现象,指的是基因序列中任意两个邻近 SNP 之间的等位基因在多代遗传中的非随机组合现象.

随着高通量基因测序技术的发展,测序成本大幅度降低,产生了海量高维的基因数据.基因数据广泛用于科学研究、面向消费者服务和法律与司法鉴定等[2].例如,在全基因组关联研究(genome-wide association studies,简称 GWAS)中可以识别与 SNP 相关的疾病[3].但是,SNP 携带个体健康的隐私敏感信息,并且可以唯一标识人类个体,基因数据使用不当会导致敏感信息泄露[4],例如,载脂蛋白 E(apolipoprotein E)基因的两个 SNP(rs7412 和 rs429358)会增加患老年痴呆症(Alzheimer's disease)的风险.并且在连锁不平衡下,可以从 SNP 相关的敏感信息推断出其他 SNP 相关的敏感信息.因此,本文基于 SNP 连锁不平衡相关系数,提出基因隐私保护模型:矩阵差分隐私(matrix differential privacy,简称 MDP).该模型既可以保护基因数据和 SNP 连锁不平衡的隐私,同时确保基因数据具有一定的效用.

由于 SNP 可以唯一标识人类个体,并且关联表型和血缘关系等隐私敏感信息.如果没有适当地对基因数据进行隐私保护,将会阻碍科学研究的进步和发展,并给人类社会带来巨大影响.例如,在基因组序列中只需 30~80 个独立的 SNP 位点就可以唯一重识别个体[5],进而导致其关联的隐私敏感信息泄露.从 GWAS 中揭示个体的疾病状态[6]可能会导致工作和保险中的基因歧视[7].考虑到具有血缘关系个体之间的基因数据非常相似,可以从 GWAS 中推断个体的亲戚及其相关的表型敏感信息[4].因此需要联合法律法规和隐私保护技术来实现基因数据的隐私保护.目前国内尚未有专门的基因隐私保护法律法规,美国于 1996 年颁布了 HIPAA(health insurance portability and accountability act)禁止基因歧视.

除了专门的基因隐私保护法律法规外,还需要隐私保护技术来实现基因数据的隐私保护.由于基因与人类的敏感信息密切相关,基因-疾病关联分析中目前主要有 3 类基因隐私保护方法,包括密码学[8−11]、安全计算[12,13]和差分隐私[14−18].

为了从分布式基因数据中分析罕见疾病,Chen 等人[8]提出隐私保护分布式协作框架 PRINCESS,并使用 AES-GCM(advanced encryption standard in galois counter mode)加密所有基因数据,PRINCESS 为了保护健康信息的隐私对加密数据执行安全的分布式计算.在使用 AES-GCM 加密基因数据时,由于密钥分发通信代价高而使加解密受限,并且不可信用户解密后通过分析基因数据导致患者隐私泄露.因此,为了防止不可信用户解密后分析基因数据导致的隐私泄露,使用同态加密直接对密文进行计算.Ayday[9]使用 Paillier 密码系统和 Honey 加密方法保护基因数据的隐私.为了发现罕见变异与疾病易感性的关系,基于惩罚似然的确切逻辑回归(exact logistic regression)减少偏差的方法,Wang 等人[10]在同态加密的确切逻辑回归的基础上提出 HEALER 框架,便于在 GWAS 中安全地实现小抽样的罕见疾病变异分析.为了实现查询和结果的隐私保护,Shimizu 等人[11]基于加法同态加密的不经意传输(oblivious transfer)隐藏序列查询和感兴趣的基因区域.由于同态加密基于有限域数学理论,计算效率非常低,并且在不可信用户解密后同样面临隐私泄露的问题.

在人类基因序列之间,安全计算编辑距离(edit distance)在医学的个人基因数据和公共健康领域呈现出许多有趣的应用.Wang 等人[12]结合基因编辑距离近似算法和隐私集合差大小协议设计隐私编辑距离协议,并基

于此,设计全基因组安全相似患者查询系统 GenSets.最近的工作表明,个体的微生物 DNA 序列与人类个体标识相符合,并且可以关联基因数据集中敏感属性的实际身份.目前,DNA 隐私保护分析工具不满足微生物测序研究的要求.为了解决微生物测序的隐私问题,Wagner 等人[13]使用安全计算实现宏基因组分析.基因数据的安全计算中计算效率低,而且通信代价高.

从基因数据选择到 GWAS 统计值的隐私保护,差分隐私[14]已经广泛应用于基因数据.例如,在 DNA 数据选择过程中,Zhao 等人[15]利用连锁不平衡对高维单体型降维到单体型块,并通过对单体型块的次要等位基因计数加噪音产生差分隐私实验数据集,不但保护了患者的隐私,而且保证了 DNA 数据的效用.在隐私保护数据选择中仅仅通过对次要等位基因计数加噪音来实现差分隐私.由于隐私攻击对参与 GWAS 患者的隐私具有潜在的威胁,Cai 等人[16]提出了差分隐私技术是一个有希望的研究方向,差分隐私通过注入随机噪音到基因型频率、基因型-疾病关联性和基因型-基因型关联性统计值.但并未考虑 SNP 的连锁不平衡性质.假设 GWAS 的基因数据是不相关的,Tramèr 等人[17]考虑更多合理的背景知识作为先验分布,提出有界先验差分隐私用于 GWAS 中每个 SNP 列联表的 χ^2-统计值达到效用与隐私的平衡.不过,同样没有考虑基因数据中 SNP 的连锁不平衡性质.然而,在挖掘最重要的 SNP 的所有差分隐私方法中都具有准确度或计算效率的缺点,为此,Simmons 和 Berger[18]使用等位基因检测统计值的输入扰动和自适应边界的方法来克服准确性问题.总的来说,在 GWAS 中的差分隐私保护研究仅仅考虑添加噪音到统计值,而没有考虑 SNP 的连锁不平衡性质,并且没有对原始基因数据进行隐私保护.

但是,基于 GWAS 中的统计值和 SNP 的连锁不平衡,可以推断出患者的隐私信息.因为 SNP 连锁不平衡是同一染色体上相互邻近的等位基因可能同时遗传到后代.那么从一个 SNP 位点的敏感信息可以推断出其他 SNP 位点相关的敏感信息.例如,在 SNP 连锁不平衡下,观察到的 SNP 越多,基因隐私保护强度越低[4].现有的工作主要有两方面的局限性:(1) 没有从基因数据而仅仅是从 GWAS 中的统计值上实现患者的差分隐私;(2) 没有考虑 SNP 连锁不平衡下的基因数据隐私保护.另外,由于基因型数据只包含数值 0、1 和 2,如果对基因数据直接使用差分隐私机制将导致基因数据效用灾难,详见第 4.4 节基因数据的效用分析.

为了解决此问题,本文提出基因数据和 SNP 连锁不平衡的矩阵差分隐私保护模型.首先将单核苷酸多态性二倍体基因数据进行矩阵存储,然后在连锁不平衡下基于严格的差分隐私定义实现二倍体基因数据以及 SNP 连锁不平衡的不可区分性,最后运用模余运算进行二倍体基因数据的置换.矩阵差分隐私保护模型不仅满足差分隐私,而且确保一定的基因数据效用.同时,矩阵差分隐私保护模型可以扩展到基因数据的其他应用领域.本文主要贡献如下.

(1) 结合 SNP 二倍体基因数据的矩阵存储、SNP 连锁不平衡下严格的差分隐私定义和模余运算,提出矩阵差分隐私保护模型作为基因隐私保护的新方法.

(2) 基于拉普拉斯机制和高斯机制,在 SNP 连锁不平衡相关系数下,设计矩阵差分隐私保护模型的算法,实现基因数据与 SNP 连锁不平衡的隐私保护.

(3) 矩阵差分隐私保护模型确保基因数据效用在区间$[R_0,1]$中,其中,R_0 表示当隐私预算最小时矩阵差分隐私下噪音矩阵中模 3 余 0 元素数量的百分比值.

本文第 1 节介绍基因背景知识以及矩阵计算、模余运算和差分隐私的预备知识.第 2 节提出矩阵差分隐私保护模型.第 3 节对矩阵差分隐私进行理论分析.第 4 节分别对矩阵差分隐私的隐私保护和基因数据效用进行实验分析.第 5 节对全文进行总结.

1 预备知识

首先介绍基因的背景知识.然后介绍矩阵计算、模余运算和差分隐私的预备知识.

1.1 基因组

尽管人类的 DNA 大部分是相同的,但是产生的变异大约有 5 000 万,其中 SNP 是人类最常见的 DNA 变异.由于每个 SNP 位点的两个核苷酸分别从父亲和母亲的基因中遗传而来,因此可能是高频率的主要等位基因,也

可能是低频率的次要等位基因.每个 SNPg_i具有次要等位基因的频率为 p_{maf}^i,对于一个个体的基因型 SNP 的次要等位基因频率表示为 m 维向量 $(p_{maf}^1, p_{maf}^2, ..., p_{maf}^m)$.用 B 表示主要等位基因,b 表示次要等位基因,$B,b\in\{A,C,G,T\}$,并且编码 BB 为 0,Bb 为 1,bb 为 2.考虑 SNP 序列作为个体的基因数据,称为二倍体基因型,其中,每个基因型取值属于集合{0,1,2}.因此,单倍体基因型对应一条染色体,而二倍体基因型对应一组染色体.

在人类基因组序列中,每个序列可以表示为有序的 SNP 序列 $g_1,g_2,\ldots,g_m$ 序列,其中,每个 $g_i\in\{0,1,2\}$.假设 g_i 与 g_j 相互连锁不平衡,(B,b)和(D,d)分别是 g_i 和 g_j 的等位基因.假设$(p_1,1-p_1)$和$(p_2,1-p_2)$分别是(B,b)和(D,d)的等位基因概率.这里,等位基因频率即是等位基因的概率.如果 g_i 和 g_j 相互独立,那么个体在 g_i 和 g_j 的主要等位基因是 B 和 D 的概率为 p_1p_2.然而,由于 g_i 和 g_j 的关联性,因此连锁不平衡系数为 $LD=P(BD)-P(B)P(D)$,其中,在连锁不平衡下,$P(BD)$等于在 SNP 位点 i 和 j 的等位基因 B 和 D 共同出现在群体中的频率,并使用 $r_{ij}=LD/\sqrt{p_1(1-p_1)p_2(1-p_2)}$ 作为 SNP 连锁不平衡的相关系数,当 $r_{ij}=1$ 时,表示最强的 SNP 连锁不平衡[4].

1.2 矩阵计算

对于两个 $n\times m$ 矩阵 $S=(s_{ij})_{n\times m}$ 和 $T=(t_{ij})_{n\times m}$,其中,$1\leqslant i\leqslant n$,$1\leqslant j\leqslant m$.S 和 T 之间的加运算定义为$(c_{ij})_{n\times m}=(s_{ij})_{n\times m}+(t_{ij})_{n\times m}$,其中,$c_{ij}=s_{ij}+t_{ij}$.另外,$round(S)$表示运用四舍五入规则将矩阵 S 中的元素取整的近似运算.

1.3 模余运算

给定整数 s、t、q 和 r,余数 $r=s-qt$ 表示为 $r\equiv s \bmod t(0<r<t)$,该运算称为模余运算.如果任意整数 $s_i(1\leqslant i\leqslant k)$除以 t 的余数都是 r,那么集合 $R=(s_1,s_2,\ldots,s_k)$构成一个等价类.因此,从集合 R 中选择一个整数 s_i 满足等式 $r=s_i-q_it$ 的概率是 $1/k$.

1.4 差分隐私

根据两个相同的概率分布是不可区分的,对于个体数据的集合,差分隐私[14]确保一个攻击者的能力是相同的,独立于任何个体是否在数据集中.因此,在同样大小的数据集之间,邻近数据集仅只有一个不同.也就是说,两个邻近数据集 X_1 和 X_2 的汉明距离(Hamming distance)为 $d(X_1,X_2)=1$.其中,差分隐私定义如下.

定义 1(差分隐私). 给定$\varepsilon\geqslant 0$,如果有任意两个邻近数据集X_1和X_2,对于拥有全背景知识的攻击者,随机机制 M 的任意输出 $S\subseteq Range(M)$使得 $\Pr[M(X_1)\in S]\leqslant e^{\varepsilon}\Pr[M(X_2)\in S]+\delta$, 那么 M 是(ε,δ)-差分隐私.

其中,$1-\delta\in[0,1]$是 M 满足(ε,δ)-差分隐私的概率,并且,如果$\delta=0$,那么 M 是ε-差分隐私.

为了实现差分隐私机制,需要计算查询函数 f 的敏感度,查询函数 $f:X\to R^k$ 的敏感度是

$$\Delta f=\max_{d(X_1,X_2)=1}\|f(X_1)-f(X_2)\|_1 \tag{1}$$

另外,差分隐私具有后处理(post-processing)和并行组合(parallel composition)[19]的性质.

性质 1(后处理). 随机机制 $M:X\to R$ 关于数据集 X 是(ε,δ)-差分隐私,$f:R\to R'$是一个随机映射,那么 $f\circ M:X\to R'$是(ε,δ)-差分隐私.

性质 2(并行组合). 随机机制 M_i 满足(ε_i,δ)-差分隐私,数据集 X_i 是 X 的子集,且 $X_i\cap X_j=0(i\neq j)$,那么 M_i 的并行组合满足$(\max\{\varepsilon_i\},\delta)$-差分隐私.

2 基因隐私保护模型

首先引入 SNP 连锁不平衡下对基因数据的攻击模型,接下来提出基因隐私保护模型:矩阵差分隐私.

2.1 攻击模型

因为通过 SNP 可以识别个体及其相关的敏感信息.假设攻击者已经观察到隐藏的 SNP,并且攻击者是 honest-but-curious.攻击者可以通过成对的 SNP 连锁不平衡获得敏感信息,例如相邻两个位点 i 和 j 的 SNPg_i和 g_j,它们之间存在 SNP 连锁不平衡,如果 g_i 与某种疾病易感性相关,那么 g_j 也与该疾病相关.

2.2 矩阵差分隐私保护模型

在 SNP 连锁不平衡下,由于基因数据的隐私保护需求,我们首先给出基因隐私保护模型——矩阵差分隐私,如图 1 所示,该模型主要包括 3 部分.第 1 部分为编码 SNP 二倍体基因数据并用矩阵存储.第 2 部分为对已编码的 SNP 二倍体基因数据进行随机扰动,同时满足基于 SNP 连锁不平衡下的差分隐私.第 3 部分为使用模余运算置换随机扰动的 SNP 基因数据.其中,各个部分的主要思想如下.

在第 1 部分,B 表示主要等位基因,b 表示次要等位基因,根据等位基因的频率,将主要等位基因 B 编码为 0,次要等位基因 b 编码为 1,并且 $B,b\in\{A,C,G,T\}$,编码基因型 BB 为 0,Bb 为 1,bb 为 2.那么对于 n 个个体,每个个体有 m 个 SNP,用矩阵表示为 $X=(x_{ij})_{n\times m}(1\leqslant i\leqslant n,1\leqslant j\leqslant m)$,且 $x_{ij}\in\{0,1,2\}$表示第 i 个个体第 j 个位点的 SNP 基因型.

第 2 部分是对 SNP 二倍体基因数据进行随机扰动,并且满足 SNP 连锁不平衡下的差分隐私.图 2 所示为 SNP 二倍体基因型数据随机扰动的主要思想,根据 SNP 连锁不平衡下的差分隐私扰动机制,将 SNP 二倍体基因型矩阵元素 $x_{ij}\in\{0,1,2\}$分别以概率 p_1、p_2 和 p_3 进行随机扰动.这里,p_1、p_2 和 p_3 是 SNP 连锁不平衡下差分隐私随机噪音对应的概率.

如图 2 所示,第 3 部分对随机扰动的二倍体基因型数据进行模余运算,使其具有 SNP 二倍体基因型数据的语义,并根据等位基因频率和基因型编码置换为相应的基因型.

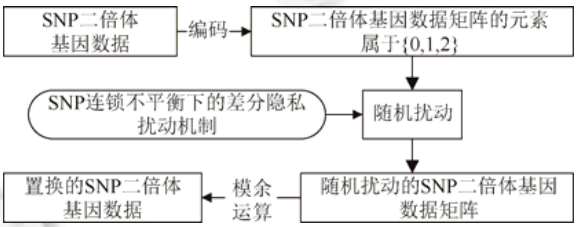

Fig.1 The genomic privacy preserving framework for SNP linkage disequilibrium

图 1 SNP 连锁不平衡下的基因隐私保护模型

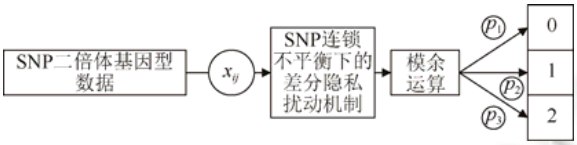

Fig.2 The differential privacy perturbation mechanism for SNP linkage disequilibrium

图 2 SNP 连锁不平衡下的差分隐私扰动机制

2.3 矩阵差分隐私

矩阵 $X=(x_{ij})_{n\times m}$ 表示 n 个个体的 SNP,每个个体的 DNA 序列有 m 个 SNP,其中,$x_{ij}\in\{0,1,2\}$表示个体 i 的 SNPg_j 的随机变量.特别地,$X_i=(x_{i1},x_{i2},\ldots,x_{im})$表示个体 i 的 SNP 序列取值.

$(x_{ij})^1_{n\times m}$ 的邻近矩阵 $(x_{ij})^2_{n\times m}$ 与 $(x_{ij})^1_{n\times m}$ 只有一个元素不同.对于查询第 i 个个体的第 j 个位点的 SNP 的函数 f,考虑 SNP 连锁不平衡,查询函数 f 的敏感度为

$$\Delta f=\max_{d((x_{ij})^1_{n\times m},(x_{ij})^2_{n\times m})=1}\left\|f((x_{ij})^1_{n\times m})-f((x_{ij})^2_{n\times m})\right\|_1 \tag{2}$$

因为 $x_{ij}\in\{0,1,2\}$,所以查询函数 f 的敏感度为$\Delta f=2$.

本文用 $r_i^{\max}(1\leqslant i\leqslant n)$ 表示个体 i 的所有 SNPg_i 和 g_j 的连锁不平衡的相关系数 $r_{ij}(i\neq j,1\leqslant i\leqslant m,1\leqslant j\leqslant m)$的最大值,$\max\{r_i^{\max}\}^n$ 是所有 n 个个体连锁不平衡的相关系数 $r_i^{\max}$ 的最大值.为了构建矩阵差分隐私,需要在 SNP 连锁不平衡下产生满足差分隐私的随机噪音矩阵 $Y=(y_{ij})_{n\times m}$,且噪音 y_{ij} 服从尺度参数为 $\Delta fc/\max\{r_i^{\max}\}^n\varepsilon$ 的概率

分布 $\pi(\Delta fc/\max\{r_i^{\max}\}^n\varepsilon)$.

下面结合矩阵加运算、SNP 连锁不平衡下的差分隐私定义和模余运算,给出矩阵差分隐私的定义.

定义 2(矩阵差分隐私). 给定$\varepsilon\geqslant 0$,任意两个邻近矩阵 $(x_{ij})^1_{n\times m}$ 与 $(x_{ij})^2_{n\times m}$, 对于具有全背景知识的攻击者,M 的任意输出 $S=(s_{ij})_{n\times m}\subseteq Range(M)$,使得 $\Pr[M((x_{ij})^1_{n\times m})\in S]\leqslant e^{\max\{r_i^{\max}\}^n\varepsilon}\Pr[M((x_{ij})^2_{n\times m})\in S]+\delta$. 那么,随机机制

$$M=[(x_{ij})_{n\times m}+round((y_{ij})_{n\times m})]\ \bmod\ 3 \tag{3}$$

是 $(\max\{r_i^{\max}\}^n\varepsilon,\delta)$-差分隐私.

另外,由于个体 i 的 SNP 序列值表示为向量 $X_i=(x_{i1},x_{i2},\ldots,x_{im})$.类似地,下面我们来定义向量差分隐私.

定义 3(向量差分隐私). 给定$\varepsilon\geqslant 0$,任意两个邻近向量 $(x_{ij})^1_{1\times m}$ 与 $(x_{ij})^2_{1\times m}$, 对于具有全背景知识的攻击者,M 的任意输出 $S_i=(s_{ij})_{1\times m}\subseteq Range(M)$,使得 $\Pr[M((x_{ij})^1_{1\times m})\in S_i]\leqslant e^{r_i^{\max}\varepsilon}\Pr[M((x_{ij})^2_{1\times m})\in S_i]+\delta$. 那么,随机机制

$$M=[(x_{ij})_{1\times m}+round((y_{ij})_{1\times m})]\ \bmod\ 3 \tag{4}$$

是 $(r_i^{\max}\varepsilon,\delta)$-差分隐私.

因此,向量差分隐私是矩阵差分隐私的特例.下面给出矩阵差分隐私的通用算法 1.其中,概率分布$\pi(\Delta fc/\max\{r_i^{\max}\}^n\varepsilon)$ 可以是拉普拉斯分布和高斯分布,即噪音矩阵$(y_{ij})_{n\times m}$ 是由拉普拉斯机制(Laplace mechanism,简称 LM)和高斯机制(Gaussian mechanism,简称 GM)[20]产生的.相应的常数 c 分别为 1 和 $\sqrt{2\ln(1.25/\delta)}$.由于 SNP 二倍体基因型矩阵存储$(x_{ij})_{n\times m}$中元素 $x_{ij}\in\{0,1,2\}$,这里暂且将 x_{ij} 看作字符型,简单地定义基因型 x_{ij} 的效用函数为 $u:x_{ij}\to x_{ij}$,也就是说,$u(x_{ij}=0)=0,u(x_{ij}=1)=1$ 和 $u(x_{ij}=2)=2$,那么效用函数的敏感度为$\Delta u=2$,因此在指数机制下选取基因型值 0、1 和 2 的概率分别正比于 1、$e^{\varepsilon/4}$ 和 $e^{\varepsilon/2}$.因为 SNP 基因型矩阵及其对应的效用矩阵的元素都是 0、1 和 2,所以通过指数机制选择基因型值 0、1 和 2 的随机性较差,那么在 SNP 基因型数据的这种效用函数定义方式下,使用指数机制将导致基因型数据及其相关的敏感信息泄露,因此本文没有考虑指数机制(exponential mechanism,简称 EM)[20].

算法 1. 在 SNP 连锁不平衡下的矩阵差分隐私.

输入:SNP 二倍体基因型矩阵$(x_{ij})_{n\times m}$,且 $x_{ij}\in\{0,1,2\}$.初始化ε,δ和Δf;

输出:随机扰动和置换的 SNP 二倍体基因型矩阵$(s_{ij})_{n\times m}$.

1: 计算 SNP 连锁不平衡的相关系数 r_{ij}

2: 生成噪音矩阵$(y_{ij})_{n\times m}$,且 $y_{ij}\sim\pi(\Delta fc/\max\{r_i^{\max}\}^n\varepsilon)$

3: $(s_{ij})_{n\times m}=[(x_{ij})_{n\times m}+round((y_{ij})_{n\times m})]\ \bmod\ 3$

3 矩阵差分隐私的分析

下面从理论上分析矩阵差分隐私的性质.

定理 1. 矩阵差分隐私是 $(\max\{r_i^{\max}\}^n\varepsilon,\delta)$-差分隐私.

证明:让 $(x_{ij})^1_{n\times m}$ 与 $(x_{ij})^2_{n\times m}$ 是邻近矩阵,因此有 $d((x_{ij})^1_{n\times m},(x_{ij})^2_{n\times m})=1$. 噪音矩阵 $round((y_{ij})_{n\times m})$ 中元素 y_{ij} 服从尺度参数为 $\Delta fc/\max\{r_i^{\max}\}^n\varepsilon$ 的概率分布 $\pi(\Delta fc/\max\{r_i^{\max}\}^n\varepsilon)$. 在矩阵差分隐私中,噪音矩阵是由拉普拉斯机制和高斯机制产生的.所以,根据性质 2,对于个体 i,使用噪音向量$(y_{ij})_{1\times m}$ 扰动基因数据$(x_{ij})_{1\times m}$ 是满足 $(r_i^{\max}\varepsilon,\ \delta)$-差分隐私的.这里,对应于邻近矩阵 $(x_{ij})^1_{n\times m}$ 与 $(x_{ij})^2_{n\times m}$, 添加到 $(x_{ij})^1_{n\times m}$ 与 $(x_{ij})^2_{n\times m}$ 的噪音矩阵$(y_{ij})_{n\times m}$ 服从期望为 0 的概率分布 $\pi(\Delta fc/\max\{r_i^{\max}\}^n\varepsilon)$, 则有

$$\Pr[(x_{ij})^1_{1\times m}+(y_{ij})_{1\times m}]\leqslant e^{r_i^{\max}\varepsilon}\Pr[(x_{ij})^2_{1\times m}+(y_{ij})_{1\times m}]+\delta \tag{5}$$

由性质 2 可知,不等式 $\Pr[(x_{ij})^1_{n\times m}+(y_{ij})_{n\times m}]\leqslant e^{\max\{r_i^{\max}\}^n\varepsilon}\Pr[(x_{ij})^2_{n\times m}+(y_{ij})_{n\times m}]+\delta$ 成立.由性质 1,下面两个不等式成立:

$$\Pr[(x_{ij})^1_{n\times m}+round((y_{ij})_{n\times m})]\leqslant e^{\max\{r_i^{\max}\}^n\varepsilon}\Pr[(x_{ij})^2_{n\times m}+round((y_{ij})_{n\times m})]+\delta \tag{6}$$

$$\Pr[((x_{ij})^1_{n\times m}+round((y_{ij})_{n\times m}))\bmod 3]\leqslant e^{\max\{r_i^{\max}\}^n\varepsilon}\Pr[((x_{ij})^2_{n\times m}+round((y_{ij})_{n\times m}))\bmod 3]+\delta \tag{7}$$

所以不等式 $\Pr[M((x_{ij})^1_{n\times m})\in S]\leqslant e^{\max\{r_i^{\max}\}^n\varepsilon}\Pr[M((x_{ij})^2_{n\times m})\in S]+\delta$ 成立.因此,矩阵差分隐私机制 M 满足 $(\max\{r_i^{\max}\}^n\varepsilon,\delta)$-差分隐私. □

为了分析矩阵差分隐私的效用,因为 $S=(s_{ij})_{n\times m}\subseteq Range(M)$, 本文使用 $U=\left|(x_{ij})_{n\times m}\cap(s_{ij})_{n\times m}\right|\Big/\left|(x_{ij})_{n\times m}\right|$ 度量矩阵差分隐私机制的效用[18].

定理 2. 矩阵差分隐私的效用在[R_0,1]区间,R_0 表示隐私预算ε最小时矩阵差分隐私下噪音矩阵中模 3 余 0 元素数量的百分比值.

证明:首先考虑 3 种极端的情况.

(1) 当噪音矩阵 $Y=(y_{ij})_{n\times m}$ 的所有元素满足 $round(y_{ij})\bmod 3=0$ 时,$round((y_{ij})_{n\times m})$的所有元素都模 3 同余 0.因此,在$(x_{ij})_{n\times m}$ 与$(s_{ij})_{n\times m}\subseteq Range(M)$之间的所有 SNP 二倍体基因型数据相同.因此,矩阵差分隐私机制的最大效用为 1.

(2) 当噪音矩阵 $Y=(y_{ij})_{n\times m}$ 的所有元素满足 $round(y_{ij})\bmod 3\equiv 1$ 时,$(0+1)\bmod 3\equiv 1$,$(1+1)\bmod 3\equiv 2$ 和$(2+1)\bmod 3\equiv 0$.因此,$(x_{ij})_{n\times m}$ 与$(s_{ij})_{n\times m}\subseteq Range(M)$之间的所有 SNP 二倍体基因型取值都不相同,此时矩阵差分隐私机制的效用是 0.

(3) 当噪音矩阵 $Y=(y_{ij})_{n\times m}$ 的所有元素满足 $round(y_{ij})\bmod 3\equiv 2$ 时,$(0+2)\bmod 3\equiv 2$,$(1+2)\bmod 3\equiv 0$ 和$(2+2)\bmod 3\equiv 1$.因此,$(x_{ij})_{n\times m}$ 与$(s_{ij})_{n\times m}\subseteq Range(M)$之间的所有 SNP 二倍体基因型取值也都不相同,此时矩阵差分隐私机制的效用是 0.

上述证明中考虑(2)和(3)两种极端情况,使矩阵差分隐私下基因数据的效用为 0.然而,由于噪音的随机性,矩阵差分隐私下基因数据的最小效用是大于 0 的,详见第 4.4 节基因数据的效用分析.下面考虑第 4 种情况.

(4) 在矩阵差分隐私中,由于隐私预算ε越小,邻近基因数据矩阵 $(x_{ij})^1_{n\times m}$ 与 $(x_{ij})^2_{n\times m}$ 的不可区分性越好,进而矩阵差分隐私保护越强,那么基因数据的效用达到最低.在矩阵差分隐私中基因数据的效用与模 3 余 0 的噪音数量的百分比值是一致的.也就是说,如果隐私预算ε最小,矩阵差分隐私产生模 3 余 0 的噪音数量百分比值为 $R_0(1>R_0>0)$,那么基因数据的最小效用为 R_0.反之,隐私预算ε越大,基因数据效用可达到百分比值 1.

综上,由于噪音的随机性,矩阵差分隐私机制的效用属于区间[R_0,1]. □

定理 3. 考虑连锁不平衡、矩阵加运算和模余运算的计算复杂度分别为 $O(n\times m^2)$、$O(n\times m)$和 $O(n\times m)$.矩阵差分隐私的计算复杂度如下:(1) 当 $n=m$ 时,矩阵差分隐私的计算复杂度为 $O(n^3)$;(2) 当 $n>m$ 时,矩阵差分隐私的计算复杂度为 $O(nm^2)$;(3) 当 $n<m$ 时,矩阵差分隐私的计算复杂度为 $O(nm^2)$.

证明:在矩阵差分隐私中,产生随机噪音是有效的,忽略其计算复杂度,而计算连锁不平衡、矩阵加运算和模余运算分别需要 $8n\times(m^2-m)$、$n\times m$ 和 $n\times m$ 次运算,考虑 3 种情况.

(1) 当 $n=m$ 时,矩阵差分隐私的计算复杂度为 $O(n^3)$.

(2) 当 $n>m$ 时,矩阵差分隐私的计算复杂度为 $O(nm^2)$.

(3) 当 $n<m$ 时,矩阵差分隐私的计算复杂度为 $O(nm^2)$. □

总之,矩阵差分隐私满足差分隐私的定义,同时具有效用属于区间[R_0,1],其中,R_0 是矩阵差分隐私下隐私预算最小时噪音矩阵中模 3 余 0 元素数量的百分比值,并且矩阵差分隐私的计算复杂度是多项式时间的.

4 实验分析

本文在矩阵差分隐私下选择拉普拉斯分布和高斯分布来进行实验分析.首先进行噪音分析,然后与拉普拉斯机制和高斯机制比较分析矩阵差分隐私保护模型的隐私和效用.在所有的实验分析中,考虑 SNP 二倍体基因型数据的特点,初始化 SNP 连锁不平衡的相关系数为 $r_{ij}=1$ 和敏感度$\Delta f=2$.另外,分别初始化隐私预算$\varepsilon=0.001$ 和

概率值δ=0.01.

4.1 数据集

国际人类基因组单体型图计划(Int'l Hapmap Project)的数据是公开可用的[21],本文使用 2010 年 5 月发布的阶段 III 的 165 个 CEU(utah residents with northern and Western European ancestry from the CEPH collection)群体的 22 号染色体的基因型和频率数据集.在实验分析之前,基于频率数据集预处理基因型数据集,将 SNP 二倍体基因型数据编码为 0、1 和 2.在 CEU 基因型数据集中,将丢失的数据'NN'用 0 代替.本文分别选择 500、1 000 和 1 500 个 SNP 位点进行实验分析.

4.2 噪音分析

在矩阵差分隐私中,尺度参数为 $\Delta fc/\max\{r_i^{\max}\}^n\varepsilon$ 的拉普拉斯机制(LM)和高斯机制(GM)产生的噪音矩阵为 $(y_{ij})_{n\times m}$.在两种机制下,图 3 所示分别计算矩阵 $round((y_{ij})_{165\times500})$、$round((y_{ij})_{165\times1000})$和 $round((y_{ij})_{165\times1500})$模 3 余 0 的噪音数量的百分比值 R.可以观察到,模 3 余 0 的噪音数量百分比值随着隐私预算的增加而增加,而不随噪音数量的大小而发生变化.这个结果为解释隐私和基因数据效用的实验结果奠定了基础.随着隐私预算的增加,拉普拉斯机制与高斯机制相比,所有模 3 余 0 的噪音数量的百分比值明显更快地增加.当隐私预算ε=7 时,拉普拉斯机制的 R 值将达到 80%,而高斯机制的 R 值才达到 40%.这是因为,在相同的隐私预算下,拉普拉斯分布与高斯分布相比,基于拉普拉斯机制的矩阵差分隐私产生的噪音矩阵中模 3 余 0 的元素更多.

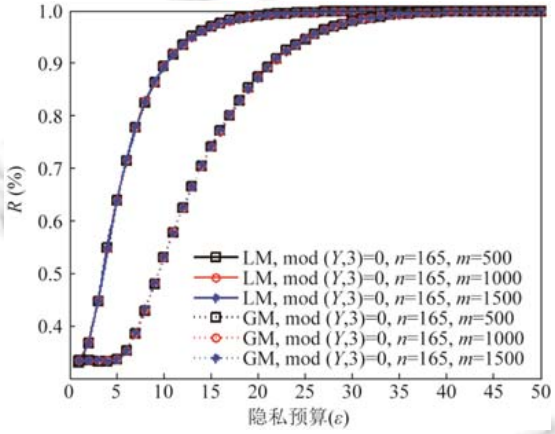

Fig.3 The percentage of noises matrix entries module 3 satisfying the residue to be 0 for matrix differential privacy

图 3 矩阵差分隐私下噪音矩阵模 3 余 0 的元素数量的百分比值

4.3 隐私分析

为了评估基因隐私保护模型的隐私,对于拥有全背景知识的攻击者,本文定义标准化期望估计误差作为隐私度量.因为元素 x_{ij} 在矩阵差分隐私下的随机扰动元素为 s_{ij},因此,定义基因数据的隐私度量为

$$E=\frac{\sum_{1\leqslant i\leqslant n}\sum_{1\leqslant j\leqslant m}P(s_{ij})\left\|s_{ij}-x_{ij}\right\|_1}{mn} \tag{8}$$

通过比较,我们来分析矩阵差分隐私与拉普拉斯机制、高斯机制的标准化期望估计误差.如图 4 和图 5 所示,矩阵差分隐私、拉普拉斯机制和高斯机制的标准化期望估计误差都随隐私预算的增大而减小.主要原因是,隐私预算越大,拉普拉斯分布和高斯分布的方差越小,矩阵差分隐私产生模 3 余 0 的噪音越多.因此拉普拉斯机制和高斯机制直接添加噪音到 SNP 基因型数据会导致效用灾难,而矩阵差分隐私通过噪音模余运算提高了 SNP 基因型数据的效用,见第 4.4 节矩阵差分隐私的效用分析.由此,矩阵差分隐私实现了基因数据的隐私保护,不过,隐私保护强度显然低于拉普拉斯机制和高斯机制.另外,由图 4 和图 5 可知,随着隐私预算的增加,高斯机制

的标准化期望误差较拉普拉斯机制要大,为了更好地权衡隐私和效用,可以选择拉普拉斯机制实现矩阵差分隐私.

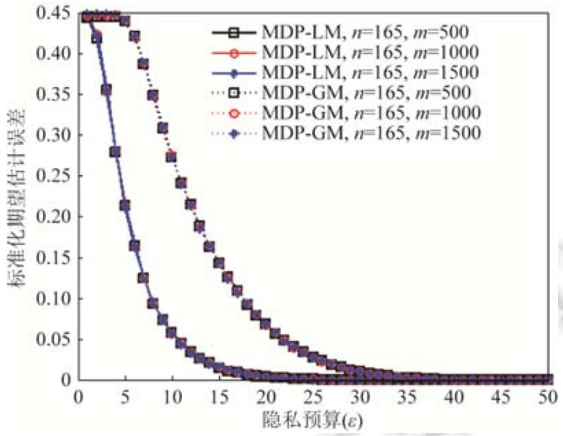

Fig.4 The normalized expected estimation error for matrix differential privacy

图 4 矩阵差分隐私下的标准化期望估计误差

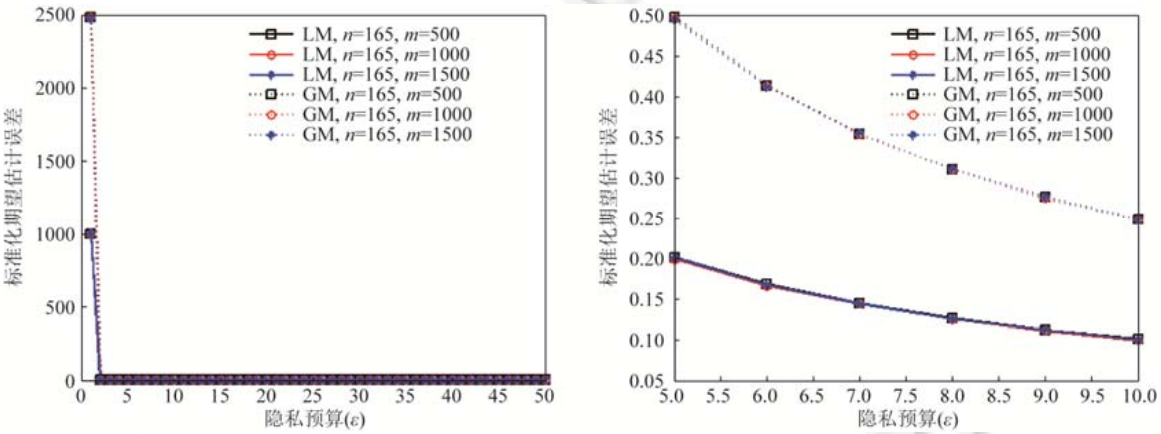

Fig.5 The normalized expected estimation error for Laplace mechanism and Gaussian mechanism

图 5 拉普拉斯机制和高斯机制下的标准化期望估计误差

因此,根据 SNP 连锁不平衡下差分隐私的不可区分性,矩阵差分隐私实现了 SNP 基因型数据和 SNP 连锁不平衡的隐私保护.

4.4 效用分析

尽管矩阵差分隐私可以实现 SNP 基因型数据的隐私保护,考虑到 SNP 基因型数据的分析,因此还需要分析 SNP 基因型数据的效用.在矩阵差分隐私中,对于原始的 SNP 基因型数据$(x_{ij})_{n\times m}$和扰动后的 SNP 基因型数据$(s_{ij})_{n\times m}$,根据$U=\left|(x_{ij})_{n\times m}\cap(s_{ij})_{n\times m}\right|/\left|(x_{ij})_{n\times m}\right|$作为效用度量方法实验分析基因数据的效用.

如图 6 所示,随着隐私预算的增加,矩阵差分隐私保护模型下的基因数据效用递增,并且增长到 100%保持不变.这是因为,随着隐私预算增大,拉普拉斯分布和高斯分布的方差变小,矩阵差分隐私产生模 3 余 0 的噪音就更多.当隐私预算较小时,基于拉普拉斯机制的矩阵差分隐私可以实现更好的基因数据效用,以此保证较好的计算不可区分性,进而实现更好的差分隐私保护.例如,当ε=7 时,基于拉普拉斯机制的基因数据效用可以达到 80%,而基于高斯机制的基因数据效用为 40%,这与图 3 中拉普拉斯机制和高斯机制产生噪音矩阵的四舍五入近似值模

3 余 0 的噪音数量的百分比值是一致的.而图 7 中随着隐私预算的增加,基因组数据的效用保持 0 不变.这是因为,拉普拉斯机制和高斯机制直接添加噪音到基因数据,破坏了基因数据效用,导致基因数据效用灾难.由此可知,矩阵差分隐私比拉普拉斯机制和高斯机制更适合于基因数据的隐私保护.

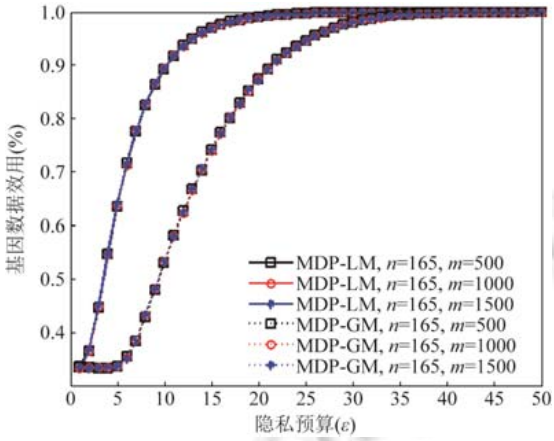

Fig.6 The genome data utility for matrix differential privacy

图 6 矩阵差分隐私下的基因数据效用

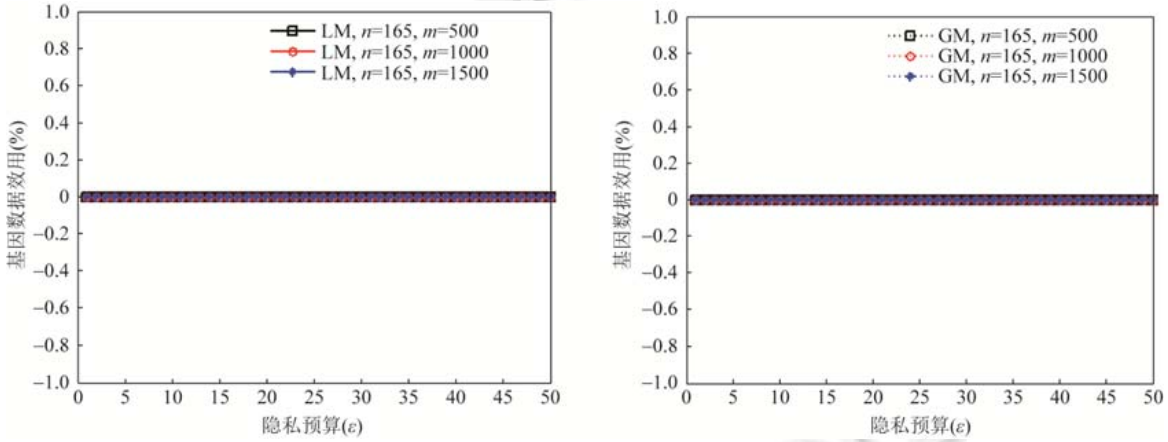

Fig.7 The genome data utility for Laplace mechanism and Gaussian mechanism

图 7 拉普拉斯机制和高斯机制下的基因数据效用

因此,矩阵差分隐私相比于拉普拉斯机制和高斯机制更适合于基因数据的隐私保护,保证了基因数据和 SNP 连锁不平衡的隐私保护与基因数据效用之间的权衡.在表 1 中,通过比较分析,总结矩阵差分隐私与拉普拉斯机制、高斯机制的相关性质.其中,最小效用 R_0 表示矩阵差分隐私在最小隐私预算下所有模 3 余 0 的噪音数量的百分比值.

Table 1 The comparison among matrix differential privacy, Laplace mechansim and Gaussian mechanism

表 1 矩阵差分隐私与拉普拉斯机制、高斯机制的比较

机制	理论基础	扰动过程	置换过程	是否满足差分隐私	基因数据效用
矩阵差分隐私	概率分布不可区分性	添加四舍五入取整噪音	模余运算	是	$[R_0,1]$
拉普拉斯机制	概率分布不可区分性	直接添加噪音	–	是	0
高斯机制	概率分布不可区分性	直接添加噪音	–	是	0

5 结 论

为了保护 SNP 连锁不平衡下基因关联的敏感信息,本文提出了矩阵差分隐私保护模型.该模型满足差分隐私,同时保证基因数据效用在$[R_0,1]$区间,其中,R_0是矩阵差分隐私在隐私预算最小时噪音矩阵中模 3 余 0 的噪音数量的百分比值,并且矩阵差分隐私是多项式时间计算有效的.

对于基因数据,基因隐私保护模型在连锁不平衡下保证隐私是可行的.通过结合矩阵加运算、SNP 连锁不平衡下差分隐私的定义和模余运算,提出了向量差分隐私和矩阵差分隐私,并且向量差分隐私是矩阵差分隐私的特例.根据矩阵差分隐私的性质,为了疾病标记发现,基因隐私保护模型可以用于 DNA 数据集的差分隐私选择[15];在 GWAS 中,矩阵差分隐私也可以对基于隐私编辑距离相似患者查询提供隐私保护[12];矩阵差分隐私阻止从 GWAS 统计值中识别特定的个体[16];并且,矩阵差分隐私可以实现隐私保护罕见疾病变异分析[8];矩阵差分隐私在基因组串搜索中是有效的隐私保护方法[11].更进一步说,在矩阵差分隐私下可以实现宏基因组分析[13].因此,矩阵差分隐私可以推广到基因数据收集、搜索和序列配对等应用的隐私保护中.

在矩阵差分隐私中,可以通过行划分、列划分或者其他快速矩阵计算方法[22]降低其计算复杂度,进而提高计算效率.另外,考虑高阶的 SNP 连锁不平衡,Samani 等人[23]表明了对隐藏 SNP 的个体基因数据具有更强的推断攻击.Tramèr 等人[17]考虑有界先验知识的差分隐私,并应用于 GWAS.通过孟德尔定律、基因变异之间的统计关系和基因与表型之间的统计关系,在个体的基因组或表型被观察到的情况下,Humbert 等人[4]详述了重构攻击推断该个体的亲戚的基因组.相比较考虑攻击者的背景知识,本文仅考虑了 SNP 连锁不平衡下基因隐私保护.在下一步的工作中,研究 SNP 连锁不平衡下具有先验知识的基因隐私保护模型,除了考虑成对 SNP 连锁不平衡外,还需要考虑高阶的 SNP 连锁不平衡,并考虑攻击者更多的先验知识,包括可利用的基因数据、个体的血缘关系以及重组规则等.

References:

[1] Li Y, Chen L. Big biological data: Challenges and opportunities. Genomics, Proteomics & Bioinformatics, 2014,12(5):187–189. [doi: 10.1016/j.gpb.2014.10.001]

[2] Naveed M, Ayday E, Clayton EW, Fellay J, Gunter CA, Hubaux JP, Malin BA, Wang X. Privacy in the genomic era. ACM Computing Surveys (CSUR), 2015,48(1):6:1–44. [doi: 10.1145/2767007]

[3] Wagner I. Evaluating the strength of genomic privacy metrics. ACM Trans. on Privacy and Security (TOPS), 2017,20(1):2:1–34. [doi: 10.1145/3020003]

[4] Humbert M, Ayday E, Hubaux JP, Telenti A. Quantifying interdependent risks in genomic privacy. ACM Trans. on Privacy and Security (TOPS), 2017,20(1):3:1–31. [doi: 10.1145/3035538]

[5] Lin Z, Owen AB, Altman RB. Genomic research and human subject privacy. Science, 2004,305(5681):183. [doi: 10.1126/science.1095019]

[6] Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, Pearson JV, Stephan DA, Nelson SF, Craig DW. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genetics, 2008,4(8):1–9. [doi: 10.1371/journal.pgen.1000167]

[7] Gottlieb S. US employer agrees to stop genetic testing. British Medical Journal, 2001,322(7284):449. [doi: 10.1136/bmj.322.7284.449/a]

[8] Chen F, Wang S, Jiang X, Ding S, Lu Y, Kim J, Sahinalp SC, Shimizu C, Burns JC, Wright VJ, Png E, Hibberd ML, Lloyd DD, Yang H, Telenti A, Bloss CS, Fox D, Lauter K, Ohno-Machado L. PRINCESS: Privacy-protecting rare disease international network collaboration via encryption through software guard extensions. Bioinformatics, 2017,33(6):871–878. [doi: 10.1093/bioinformatics/btw758]

[9] Ayday E. Cryptographic solutions for genomic privacy. In: Proc. of the Int'l Conf. on Financial Cryptography and Data Security. Berlin, Heidelberg: Springer-Verlag, 2016. 328–341. [doi: 10.1007/978-3-662-53357-422]

[10] Wang S, Zhang Y, Dai W, Lauter K, Kim M, Tang Y, Xiong H, Jiang X. HEALER: Homomorphic computation of exact logistic regression for secure rare disease variants analysis in GWAS. Bioinformatics, 2016,32(2):211–218. [doi: 10.1093/bioinformatics/btv563]

[11] Shimizu K, Nuida K, Rätsch G. Efficient privacy-preserving string search and an application in genomics. Bioinformatics, 2016, 32(11):1652–1661. [doi: 10.1093/bioinformatics/btw050]

[12] Wang XS, Huang Y, Zhao Y, Tang H, Wang X, Bu D. Efficient genome-wide, privacy-preserving similar patient query based on private edit distance. In: Proc. of the 22nd ACM SIGSAC Conf. on Computer and Communications Security. New York: ACM, 2015. 492–503. [doi: 10.1145/2810103.2813725]

[13] Wagner J, Paulson JN, Wang X, Bhattacharjee B, Bravo HC. Privacy-preserving microbiome analysis using secure computation. Bioinformatics, 2016,32(12):1873–1879. [doi: 10.1093/bioinformatics/btw073]

[14] Dwork C, Pottenger R. Toward practicing privacy. Journal of the American Medical Informatics Association, 2013,20(1):102–108. [doi: 10.1136/amiajnl-2012-001047]

[15] Zhao Y, Wang X, Jiang X, Ohno-Machado L, Tang H. Choosing blindly but wisely: Differentially private solicitation of DNA datasets for disease marker discovery. Journal of the American Medical Informatics Association, 2015,22(1):100–108. [doi: 10.1136/amiajnl-2014-003043]

[16] Cai R, Hao Z, Winslett M, Xiao X, Yang Y, Zhang Z, Zhou S. Deterministic identification of specific individuals from GWAS results. Bioinformatics, 2015,31(11):1701–1707. [doi: 10.1093/bioinformatics/btv018]

[17] Tramèr F, Huang Z, Ayday E. Differential privacy with bounded priors: Reconciling utility and privacy in genome-wide association studies. In: Proc. of the 22nd ACM SIGSAC Conf. on Computer and Communications Security. New York: ACM, 2015. 1286–1297. [doi: 10.1145/2810103.2813610]

[18] Simmons S, Berger B. Realizing privacy preserving genome-wide association studies. Bioinformatics, 2016,32(9):1293–1300. [doi: 10.1093/bioinformatics/btw009]

[19] McSherry FD. Privacy integrated queries: An extensible platform for privacy-preserving data analysis. In: Proc. of the 2009 ACM SIGMOD Int'l Conf. on Management of Data. New York: ACM, 2009. 19–30. [doi: 10.1145/1559845.1559850]

[20] Dwork C, Roth A. The algorithmic foundations of differential privacy. Foundations and Trends® in Theoretical Computer Science, 2014,9(3-4):211–407. [doi: 10.1561/0400000042]

[21] NCBI retiring HapMap Resource. https://www.ncbi.nlm.nih.gov/variation/news/NCBI_retiring_HapMap/

[22] Golub GH, Van Loan CF. Matrix Computations. 4th ed., Baltimore: The Johns Hopkins University Press, 2012. 1–104.

[23] Samani SS, Huang Z, Ayday E, Elliot M, Fellay J, Hubaux JP, Kutalik Z. Quantifying genomic privacy via inference attack with high-order SNV correlations. In: Proc. of the 2015 IEEE Security and Privacy Workshops. IEEE, 2015. 32–40. [doi: 10.1109/SPW.2015.21]

刘海(1989－),男,贵州遵义人,博士生,主要研究领域为生物医学大数据隐私保护.

彭长根(1963－),男,博士,教授,博士生导师,CCF 专业会员,主要研究领域为密码学,信息安全,大数据隐私保护.

吴振强(1968－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为网络与信息安全,分布式计算,数据隐私保护.

雷秀娟(1975－),女,博士,教授,博士生导师,CCF 高级会员,主要研究领域为生物信息计算,智能计算.