基因组数据隐私保护理论与方法综述

刘 海[1),2),3)] 彭长根[1),2),3)] 吴振强[4)] 田有亮[2),3)] 田 丰[4)]

[1)](贵州大学贵州省大数据产业发展应用研究院 贵阳 550025)

[2)](贵州大学计算机科学与技术学院 贵阳 550025)

[3)](贵州大学公共大数据国家重点实验室 贵阳 550025)

[4)](陕西师范大学计算机科学学院 西安 710119)

摘 要 基因组数据已广泛应用于科学研究、医疗服务、法律与取证和直接面向消费者服务.基因组数据不但可以唯一标识个体,而且与遗传、健康、表型和血缘关系密切关联.此外,基因组数据具有不随时间而变化的稳定性.因此,基因组数据管理不当和滥用将会带来人类所担心的隐私泄露问题.针对此问题,除了相关法律法规的监管以外,隐私保护技术也被用于实现基因组数据的隐私保护.为此,本论文对基因组数据的隐私保护理论与方法进行综述研究.首先,本论文根据基因组测序到应用归纳基因组数据的生态系统,并依据基因组数据特点分析其存在的隐私泄露问题.其次,分类总结和对比分析基因组数据存在的隐私威胁,并陈述重识别风险与共享基因组数据的价值之间的均衡模型.再次,分类概述和对比分析量化基因组数据隐私和效用的度量.然后,分析基因组数据生态系统中测序与存储、共享与聚集及应用的隐私泄露威胁.同时,分类介绍和对比分析用于基因组数据的隐私保护方法.针对基因组数据生态系统中存在的隐私泄露问题,根据所使用的隐私保护方法,分类概括和对比分析目前基因组数据隐私保护的研究成果.最后,通过对比分析已有的基因组数据隐私保护方法,对基因组数据生态系统中基因隐私保护的未来研究挑战进行展望.该工作为解决基因组数据的隐私泄露问题提供基础,进而推动基因组数据隐私保护的研究.

关键词 基因隐私；隐私泄露；隐私度量；效用度量；隐私保护

中图法分类号 TP309

A Survey of the Theories and Methods of Privacy Preserving of Genome Data

LIU Hai[1),2),3)] PENG Chang-Gen[1),2),3)] WU Zhen-Qiang[4)] TIAN You-Liang[2),3)] TIAN Feng[4)]

[1)](Guizhou Big Data Academy, Guizhou University, Guiyang 550025)

[2)](College of Computer Science and Technology, Guizhou University, Guiyang 550025)

[3)](State Key Laboratory of Public Big Data, Guizhou University, Guiyang 550025)

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

Abstract Genome data have been widely applied to the scientific research, healthcare, legal and forensic, and direct-to-consumer. Genome data can uniquely identify an individual, and it can closely associate with the inheritance, health, phenotype, and kinship. Furthermore, genome data are stable over time. Thus, the improper management and abuse of genome data will bring about the privacy concerns. To solve this problem, in addition to the supervision of relevant laws and regulations, privacy preserving technologies are also used to achieve the privacy preserving of genome data. To this end, this paper surveys the theories and methods of privacy preserving of genome data. First, this paper induces the ecosystem of genome data from genome sequencing to applications. According to the properties of genome data, this paper also analyzes privacy leakage concerns of

本课题得到国家自然科学基金(U1836205, 62002081, 61662009, 61772008, 61602290)、中国博士后科学基金资助项目(2019M663907XB)、贵州省公共大数据重点实验室开放课题(2018BDKFJJ004)、贵州省科技重大专项计划项目(20183001)、贵州省科技计划项目(黔科合平台人才[2020]5017)资助. **刘海**, 博士, 主要研究领域为隐私保护. E-mail: gzu_liuhai@163.com. **彭长根**(通信作者), 博士, 教授, 中国计算机学会(CCF)会员, 主要研究领域为密码学、信息安全、隐私保护. E-mail: peng_stud@163.com. **吴振强**, 博士, 教授, 中国计算机学会(CCF)会员, 主要研究领域为网络安全、隐私保护、可信计算. E-mail: zqiangwu@snnu.edu.cn. **田有亮**, 博士, 教授, 主要研究领域为博弈论、密码学与安全协议. E-mail: yltian@gzu.edu.cn. **田 丰**, 博士, 副教授, 主要研究领域为云计算、网络安全、隐私保护. E-mail: tianfeng@snnu.edu.cn.

the ecosystem of genome data. Second, this paper sums up the privacy threats to genome data from four aspects of individual identification, linkage attack, genotype inference, and Bayesian inference. This paper makes a comparative analysis of these privacy threats from five aspects of scenario, data type, method, attack efficiency, and threat level. This paper also states the equilibrium model between re-identification risk and the value of sharing genome data. Third, this paper presents the metrics of privacy quantification of genome data from three aspects of inaccuracy, uncertainty and health privacy. This paper also summarizes the metrics of utility quantification of genome data from seven aspects of information loss, chi-square statistics, false positive and false negative, error rate, accuracy rate, expected accuracy rate, and expected interval width. This paper compares and analyzes the privacy and utility metrics of genome data from the aspects of measurement method, measurement formula, protection effect, application scenario, attack difficulty, and adoption rate. Forth, this paper concludes that the ecosystem of genome data consists of sequencing and storage, sharing and aggregation, research and analysis, healthcare, legal and forensic, and direct-to-consumer, and this paper also analyzes the privacy leakage threats of sequencing and storage, sharing and aggregation, and applications of the ecosystem of genome data. At the same time, this paper introduces privacy preserving methods of genome data from four aspects of cryptography, anonymity, differential privacy, and hybrid approach. This paper compares and analyzes the privacy preserving methods of genome data from three aspects of method, property, and protection effect. This paper classifies and covers the existing work of privacy preserving for privacy concerns of the ecosystem of genome data based on the corresponding privacy preserving methods. This paper also makes a comparative analysis of the existing work of privacy preserving of genome data from two aspects of scenario oriented and protection effect of scenario oriented. Finally, this paper compares and analyzes the existing methods of privacy preserving of genome data, and this paper discusses the future challenges to genomic privacy preserving of sequencing and storage, sharing and aggregation, research and analysis, healthcare, legal and forensic, and direct-to-consumer of the ecosystem of genome data. This work serves as a basis of solving the problem of privacy leakage of genome data, and this work promotes the research of privacy preserving of genome data.

Key words genomic privacy; privacy leakage; privacy metric; utility metric; privacy preserving

1 引言

随着高通量测序技术的发展,基因组数据测序成本大幅度降低,进而产生高维大量的基因组数据[1].基因组数据是富含人类重要信息的生物大数据,基因组数据包含完整的 DNA 序列,DNA 序列由腺嘌呤、鸟嘌呤、胸腺嘧啶和胞嘧啶 4 种核苷酸组成.人类基因组大约含有 30 亿由核苷酸组成的碱基对,分布在 23 对染色体上.人类有 99.9%共同的 DNA 序列,大约有 5000 万单核苷酸多态性(SNP).因此,不需要关注整个基因组,而需要关注最常见的 DNA 变异 SNP.由于 SNP 唯一标识个体,进而关联个体的遗传、疾病、表型和血缘关系等信息,并因其具有稳定性常作为基因分析中的遗传标记.正因基因组数据具有唯一性和稳定性的特点,并与人类的遗传、健康、表型和血缘关系密切关联,使得基因组数据显得神秘并具有重要的价值,而且这种神秘感和价值随着时间的推移而越发重要[2].因此,基因组数据已经被广泛应用于科学研究（Scientific Research）、医疗服务（Healthcare）、法律与取证（Legal and Forensic）和直接面向消费者服务（Direct-to-Consumer）.正因基因组数据固有的特点和性质,基因组数据广泛应用于全基因组关联研究（Genome-Wide Association Study,GWAS）、基因组注释、个性化医学、基因组诊断、药物基因组学、药物敏感性预测、相似患者查询、基因检测和刑事取证等,其具体应用如表 1 所示.基因检测可以识别染色体、基因或蛋白质的变化,检测结果可以确认或排除可疑的基因状况,或有助于确定个体发展或遗传疾病的机会.在本论文中,基因检测包括疾病易感性检测、身份检测、血缘关系检测、祖先检测、配偶兼容性检测和亲子鉴定.

表 1 基因组数据应用

应用	定义	实例解释

GWAS	利用基因组中数以百万计的 SNP（图 1）作为分子遗传标记,进行全基因组比较分析或相关分析,识别与疾病相关的遗传变异.	例如,表 2 描述案例-对照组 SNP 基因型值计数,带疾病的作为案例组,无疾病的作为对照组,案例组数和对照组数分别为 $N/2$,案例-对照组中 SNP 值为 0 和 1 的总数分别为 m 和 n .通过计算 χ^2 检验统计和相应的 p 值,以小概率事件发生来评估 SNP 与疾病的关联.
基因组注释	结合生物信息学方法、蛋白质组学和转录组学,对 DNA 序列进行分析,其基因及功能是在基因组序列上挖掘、鉴定、标记和注释的.	例如,根据已知功能基因的注释信息,利用序列相似性原则注释新的基因,将需要注释的序列翻译为氨基酸序列,找到与需注释序列相似度高的蛋白质序列号及其对应的注释号,并获得注释信息作为新的基因注释.
个性化医学	利用基因组数据来进行医学治疗,建立基于个体遗传标记的定制药物疗法.	例如,临床会诊期间医生需要使用一种药物来治疗患者的疾病,需要根据患者的基因组数据和非基因组数据（年龄、身高、体重等）计算个性化剂量.
基因组诊断	利用 DNA 重组技术在分子水平上对人类遗传病的基因缺陷进行检测以诊断遗传病.	例如,单基因遗传病,常染色体上隐性致病基因导致镰刀型细胞贫血症.
药物基因组学	研究基因如何影响个体对药物的反应.	例如,Warfarin 是一种抗凝剂,可以稀释血液以防止血栓形成,Warfarin 敏感性是指个体对 Warfarin 的耐受性较低.
药物敏感性预测	从基因组信息预测最佳治疗方案.	例如,应用线性回归模型预测给定基因表达数据的药物敏感性.
相似患者查询	医生将患者的数据作为输入,并返回医院数据库中与输入患者类似的一组患者.	例如,基因型值向量 x 和 y ,根据 Pearson 相关系数、欧氏距离和余弦相似性函数等比较两名患者之间的疾病相似性.
疾病易感性检测	基于统计学方法检测 SNP 与疾病的关联强度.	例如,根据 SNP 的基因型值（0,1 或 2）对某种疾病易感性的权重,通过加权平均可以计算患者对该疾病的易感性,还可以使用线性回归、逻辑回归模型等方法计算疾病易感性.
身份检测	分析两个 DNA 样本,检测是否存在某些匹配,以表明样本来自同一个体.	例如,一个序列来自犯罪现场,一个序列来自法医数据库,目标是确定两个序列是否都是来自同一个体.
血缘关系检测	检测 DNA 包含关于个体的血亲的信息.	例如,兄弟、姐妹、叔侄、爷孙等的血缘关系检测,有血缘关系的个体比没有血缘关系的个体共享相同等位基因的频率更高,血缘关系密切的个体比血缘关系远的个体更容易共享相同等位基因.
祖先检测	检测两个个体的基因组是否具有超越父母或子女关系的共同生物祖先.	例如,有两个可利用的 Y 染色体 STR（Short Tandem Repeat）序列,一个存储于在线系谱数据库中,另一个由希望确定其祖先的个体拥有,由于 Y 染色体在生殖过程中的稳定性,如果这两个序列共享相同的 Y 染色体 STR 序列,则被认为具有共同祖先.
配偶兼容性检测	比较携带强遗传成分的疾病危险因素的夫妇的基因组,以评估其遗传兼容性.	例如,通过检测夫妇的疾病风险,以此来评估其将来孩子的疾病风险.
亲子鉴定	分析两个样本,观察基因组的特定部分是否有足够的相似性,以表明一个样本来自父母,而另一个样本来自其生物学上的孩子.	例如,孩子 STR 序列的每个基因位点有一个等位基因来自父亲,另一个等位基因来自母亲,则确定其为孩子父母亲.
刑事取证	使用 DNA 序列来鉴定个体的合法身份,可以识别犯罪或灾难受害者,排除或牵连犯罪嫌疑人,或建立人与人之间的生物关系,如亲子关系.	例如,在犯罪现场或受害者身上发现的 DNA 可被执法部门用作追踪犯罪嫌疑人的证据.

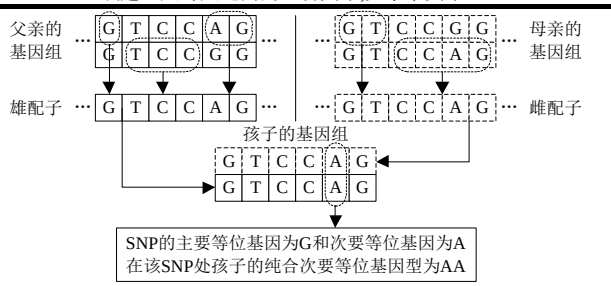

图 1　重组与 SNP

表 2　案例-对照组 SNP 基因型值计数

SNP 基因型值	0	1	2
案例组	α	β	$N/2-\alpha-\beta$
对照组	$m-\alpha$	$n-\beta$	$N/2-m+\alpha-n+\beta$
χ^2 统计	$\frac{(2\alpha-m)^2}{m}+\frac{(2\beta-n)^2}{n}+\frac{(2\alpha-m+2\beta-n)^2}{N-m-n}$		

从基因组数据的测序、存储和共享到广泛应用,基因组数据的生态系统如图 2 所示.患者发送生物样本（例如,血液、唾液等）到测序中心.这里的患者是指基因组被测序的个体,不一定是患病的个体.测序中心可以是医院、测序机构或像 23andMe 这样的服务提供商等.测序中心对生物样本进行测序后,将原始基因组数据发送给数据存储处理中心.数据存储处理中心多指医院或者第三方的数据存储和处理服务器,也可以是患者自己的服务器.数据存储处理中心对基因组数据进行标准生物信息处理,并格式化基因组数据.然后,研究机构可以利用基因组数据进行 GWAS 和基因组注释.医学中心使用基因组数据进行个性化医学、疾病易感性检测和相似患者查询,以便提供更好的医疗服务.在直接面向消费者服务方面,服务提供商 23andMe、Ancestry、MyHeritageDNA 等利用基因组数据进行身份、血缘关系、祖先、配偶兼容性和疾病易感性检测.基因组数据还可以用于实现法律权威机构进行亲子鉴定和刑事取证.此外,由于云计算具有可扩展性,并且提供低成本的计算资源,大规模基因组数据的管理和使用可以通过云计算来实现.

如图 2 所示,由于基因组数据固有的特点和性质,基因组数据具有广泛的应用.然而,因基因组数据固有的敏感特性,使基因组数据的管理和使用会带来人类所担心的隐私泄露问题,例如,载脂蛋白 E（ApoE）基因上的两个特殊 SNP[C/T]（rs7412 和 rs429358）表明易感老年痴呆症（Alzheimer）的风险.结合文献[2-4],根据基因组数据的特点和性质,基因组数据所存在的具体隐私泄露问题如下：

（1）基因组数据唯一标识人类个体,可以进行

个体身份识别.

（2）基因组数据具有稳定性,不随时间而变化,撤销或替换基因组数据是不可能的.

（3）基因组数据含有祖先、兄弟姐妹和后代的血缘关系信息.

（4）基因组数据与遗传、表型和易感疾病等密切关联.

（5）基因组数据含有未被提取或未获得的敏感信息内容.

（6）基因组数据用于执法和医疗服务,引起许多伦理道德问题.

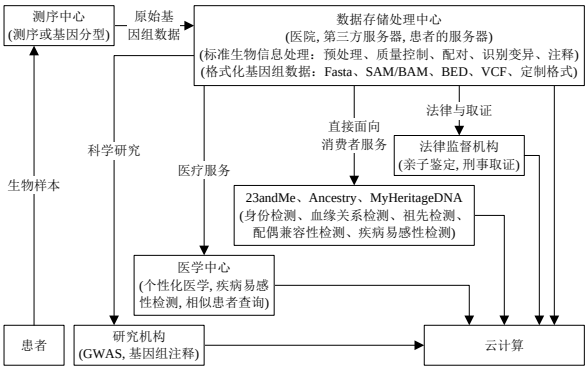

图 2 基因组数据生态系统

即使基于云计算可以实现大规模基因组数据的管理和使用,可是由于云服务器存在不可信的情况而更容易导致基因组数据的隐私泄露.基因组数据是人类身份的最终来源,越来越多的基因检测面临严重的隐私挑战[5].个体的基因组数据与家人的基因组数据相关联容易放大对基因组数据的隐私威胁,从而导致相互关联的隐私泄露风险[6].基因组数据的隐私威胁会导致遗传歧视,遗传歧视可能导致被拒绝医疗保险服务而产生真正和潜在的破坏性结果.遗传歧视还有其他不良影响,例如,由于担心滥用遗传信息,个体可能不愿共享遗传信息来参与科学研究,也拒绝与他们的医疗服务提供者或家庭成员共享其遗传信息[7].基因组数据的隐私泄露也会导致在教育、工作、抵押和婚姻中的基因歧视[8].因此,基因组数据的隐私泄露对和谐稳定的社会发展产生不良影响,直接导致人类的经济损失.甚至基因组数据的隐私泄露将会为基因战争提供机会,威胁国家安全.

基于上述基因组数据的敏感性,使得基因组数据的管理和使用不同于其他类型数据,并且具有挑战性,急需解决基因组数据的安全和隐私泄露问题.通过社区网站 GenomePrivacy.org 共享有关基因组数据安全和隐私保护研究成果,以及该领域的新闻和事件,可以找到活跃的研究团队和公司信息,例如从事该领域研究的土耳其比尔肯大学（Bilkent University）Ayday 实验室,还可以搜索该领域重要的学术论文,为该领域研究人员提供了所需的背景知识和工具.在基因组数据安全和隐私保护研究中,常用的公开可利用的数据库包括国家基因库、HapMap 项目的 1000 Genomes、基因型和表型数据库 dbGaP、NCBI 单核苷酸多态性数据库 dbSNP 等.

目前,已经通过法律法规来解决基因组数据的隐私泄露问题.美国总统布什于 2008 年 5 月 21 日签署《遗传信息非歧视法》（Genetic Information Nondiscrimination Act）禁止在健康保险和就业等方面的遗传信息歧视.欧盟于 2018 年 5 月 25 日开始实施《通用数据保护条例》（General Data Protection Regulation）,该条例旨在防止滥用个人敏感数据,并且明确规定基因组数据和生物数据为敏感类型数据.在对基因组数据进行隐私保护时,除了相关法律法规的监管,也需要结合隐私保护方法,从而使得基因组数据的管理和使用不易遭受隐私威胁.目前,主要使用密码学[9]、匿名[10]和差分隐私[11],以及混合方法实现基因组数据的隐私保护.基于上述隐私保护方法,目前已对基因组数据的隐私保护进行相关的研究,并且对已有工作进行大量的综述研究.文献[4]讨论与人类基因组数据相关的重要隐私泄露问题,并从密码学视角讨论解决基因组数据隐私保护的方法.文献[5]针对基因组数据查询处理的隐私保护进行综述研究.文献[2]针对基因组数据应用中的隐私保护进行综述研究.文献[12]对基于安全多方计算（Secure Multiparty Computation）的基因组数据隐私保护进行综述研究.文献[13]从访问控制、差分隐私和密码学三个方面对基因组数据隐私保护进行综述研究.文献[14]对外包基因组数据到云服务器并进行计数查询的已有安全解决方案进行综述研究.文献[15]讨论基因组数据的差分隐私保护和安全相关的问题.文献[16]从隐私增强技术的视角系统地综述基因组数据隐私保护的研究.文献[17]在半诚实和恶意模型下,讨论和总结基因组数据隐私泄露问题,以及相关的隐私攻击方法,并对最新的基因组数据隐私保护方案进行分类,以便于减轻存在的攻击,同时讨论基因组数据隐私保护研究的挑战和未来的研究方向.遗传数据是独特的数据类型,引发不可避免的关联隐私问题.面对识别个体、预测健康相关问题、甚至了解家族史的隐私风险,使处理和共享遗传数据成为安全和医学专家面临的挑战.文献[18]研究

发现制度信任、对家庭和朋友隐私的关注以及共享遗传数据的可能性与很多因素密切相关,例如携带遗传标记和数据请求者的机构类型,以及人口因素年龄和种族,该工作有助于为共享遗传数据开发全面的解释模型.有别于以上研究综述,根据所陈述的基因组数据生态系统和基因组数据隐私泄露问题,如图 3 所示,本论文根据基因组数据测序和存储、共享和聚集到广泛应用的整个生态系统,综合地从基因数据隐私威胁、基因组数据隐私和效用度量、基因组数据隐私保护模型的关键问题进行分类归纳总结和对比分析,并通过对比分析已有的基因组数据隐私保护方法,对未来基因组数据隐私保护研究挑战进行展望.本论文所用的部分符号及其解释说明如表 3 所示.

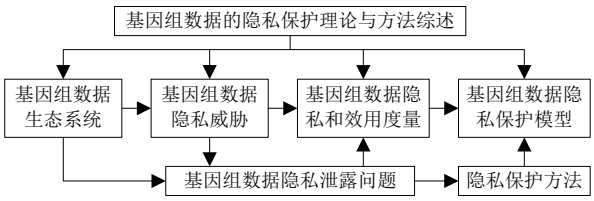

图 3　本论文组织结构

表 3　本论文部分符号及其解释说明

符号	解释说明
x	真实基因型数据集
y	估计基因型数据集
x_i	第 i 个 SNP 的真实基因型 $x_i \in \{0,1,2\}$,且 $x_i \in x$
y_i	第 i 个 SNP 的估计基因型 $y_i \in \{0,1,2\}$,且 $y_i \in y$
$p(x_i = y_i)$	正确地猜测第 i 个 SNP 基因型的概率
W_i	基因隐私（Genomic Privacy）度量中第 i 个 SNP 的权重

2　基因组数据隐私威胁

本节将对基因组数据的隐私威胁进行对比分析,根据已有的工作,主要包括个体识别（Individual Identification）、链接攻击（Linkage Attack）、基因型推断（Genotype Inference）和贝叶斯推断（Bayesian Inference）.这些隐私威胁容易造成严重的后果,例如,遗传信息歧视,从而遭受到敲诈,进而导致经济损失.最后,陈述重新识别风险博弈模型,该模型使基因组数据管理者能够平衡重新识别风险与共享基因组数据的价值.

2.1　个体识别

对于基因组数据的个体识别,攻击者已访问个体基因组数据,并通过统计推断的方法,同时与公共基因组数据集进行匹配,以此实现个体的成功匹配和识别.接下来,本节将归纳和总结基于基因组数据进行个体重识别威胁的主要方法.

在目前的基因组数据个体识别研究中,主要基于统计学方法实现个体的匹配与识别.根据贝叶斯定理和遗传平衡定律,Lin 等[19]表明攻击者访问个体基因组数据,并通过统计推断与公共 SNP 数据集进行匹配,在先验模型中假设研究对象是从群体中均匀抽样的,通过贝叶斯定理计算两个个体之间随机匹配的后验概率,那么小的 SNP 集合可能导致个体的成功匹配和识别,进而公共记录中与该个体相关的其他基因型、表型和其他信息也将泄露.结合个体的等位基因频率与参考群体的等位基因频率之间的距离,Homer 等[20]对高密度 SNP 基因分型微阵列（Genotyping Microarray）使用 t 检验可以准确地确定个体是否存在于复杂的基因组 DNA 混合物中.检验统计量的理论推导如下,DNA 混合物的等位基因频率估计为 M ,且 $M_j = A_j/(A_j + k_j B_j)$ 为参考群体的平均等位基因频率,其中 A_j 和 B_j 分别是第 j 个 SNP 的等位基因 A 和 B 的探针强度,k_j 是 SNP 特有的校正因子用于解释实验偏差,并且易于根据个体基因分型数据计算.由于大多数个体包含两个常染色体 SNP 的基因组拷贝,在基因型 BB、Bb 或 bb 处等位基因 B 的频率值可能分别为 0、0.5 或 1.从 SNP 基因分型阵列出发,让 r_{ij} 是个体 i 和 SNP j 的等位基因频率估计,其中 $r_{ij} \in \{0,0.5,1\}$.差值 $|r_{ij} - M_j|$ 度量混合 M_j 在 SNP j 处的等位基因频率与个体 r_{ij} 在 SNP j 处的等位基因频率的差异.对于每个 SNP j ,差值 $|r_{ij} - Pop_j|$ 度量参考群体的等位基因频率 Pop_j 与个体的等位基因频率 r_{ij} 的差异. Pop_j 的值可以从等摩尔的混合样本阵列或包含不同群体的基因型数据的数据库中确定.利用这两个差值的不同,可以得到用于个体 r_i 的距离测度是

$$D(r_{ij}) = |r_{ij} - Pop_j| - |r_{ij} - M_j|$$

在原假设下个体不在混合物中,因为混合物和参考群体由于具有相似的祖先成分而被假定具有相似的等位基因频率,则 $D(r_{ij})$ 接近于零.在备择假设下, r_i 在祖先上与混合物比与参考种群更相似,因此不太可能在参考种群中,则 $D(r_{ij}) > 0$,在 $D(r_{ij}) < 0$ 的情况下, r_i 在祖先上与参考种群比与混合物更相似,因此不太可能在混合物中.根据中心极限定理,通过对大量 SNP 的抽样,通常期望 $D(r_{ij})$ 服从正态分布.考虑个体 r_i 的样本 t 检验,在所有 SNP 上取样,从而

$$T(r_i) = \frac{E(D(r_i) - \mu_0)}{SD(D(r_i))/\sqrt{s}}$$

其中假设 μ_0 是 $D(r_i)$ 对未在混合物中的个体 r_i 的平均值, $SD(r_i)$ 是所有 SNP j 和个体 r_i 的 $D(r_{ij})$ 的标准偏差,并且 s 是 SNP 的数目.因为随机个体 r_i 应该与

混合物和混合物的参考群体等距离,假设 $\mu_0 = 0$,则

$$T(r_i) = \frac{E(D(r_i))}{SD(D(r_i))/\sqrt{s}}$$

在原假设下 $T(r_i) = 0$,在备择假设下 $T(r_i) > 0$.因此,在原假设下个体 r_i 在混合物和参考群体中的可能性相同,然而在备择假设下个体 r_i 很可能在混合物中.

在 Homer 等工作的基础上,结合基因型频率比的对数和 *t* 检验,Jacobs 等[21]使用基因频率和个体的基因型来推断个体或近亲是否参与 GWAS.通过分别使用 *t* 检验和整数规划两种攻击方法,Wang 等[22]表明个体实际上可以从相对较小的统计数据集中被识别,以此扩展 Homer 等的攻击.聚集回归结果可以揭示敏感信息,例如遗传数据被用来重构已发表的回归估计中使用的原始疾病状态,建议 GWAS 通常发布不超过 500 个回归分析结果[23].Homer 等及其后续的研究表明从大规模的基因组数据中可以识别出某些个体的存在,甚至可以完全恢复其 DNA 序列.Zhou 等[24]提出风险等级系统和方法以确定何时发布以及何时不发布聚集的基因组数据集.根据基因型与疾病、基因型与基因型之间的关联性,并通过使用 χ^2 检验,Cai 等[25]表明能够成功地从 WTCCC（Wellcome Trust Case Control Consortium）数据集的有限发布的关联中识别特定的个体.Sankararaman 等[26]使用似然比检验（Likelihood-Ratio Test）检测群体中的个体.Visscher 和 Hill[27]量化识别的限制,提出似然和回归分析方法实现个体识别.GA4GH（Global Alliance for Genomics and Health）创建 Beacon 项目,Beacon 是用 Yes 或 No 回答等位基因是否存在的查询网络服务器,例如,个体的基因组是否在特定的基因组位置有特定的核苷酸?Shringarpure 和 Bustamante[28]基于似然比检验推断个体是否存在于给定的遗传 Beacon 中.针对此重识别攻击,Raisaro 等[29]使用 Beacon 变更策略、随机翻转策略和每个策略的查询预算来减少 Beacon 重新识别的风险.将 Homer 等的思想扩展到实值 miRNA（microRNA）表达谱,Backes 等[30]使用 L_1 距离和似然比检验两种方法对 miRNA 表达谱进行成员推断,发现还威胁到个体的隐私.在数量性状 GWAS 回归系数或 *p* 值的汇总结果中包含隐私信息,因此 Im 等[31]基于线性回归模型表明对于大量的 SNP,回归系数能够准确地揭示个体参与 GWAS 的程度及其表型.Backes 等[32]通过主成分分析和图匹配,尽管基因表达的变异性,可能在不同的时间点跟踪一个或多个 miRNA 表达谱.此外,使用非统计学方法也可以实现基因组数据的个体识别.利用 128 万人的基因组数据,Erlich 等[33]通过远亲鉴定大约 60%的欧洲血统的个体搜索将导致第三个表亲或更接近的匹配.基于相似性函数度量表型和基因组之间的相似性,Lippert 等[34]应用全基因组测序、详细的表型分型和统计建模预测 1061 个不同祖先的参与者的生物特征.即使系统地隐藏 SNP,von Thenen 等[35]基于连锁不平衡和高阶马尔可夫链的推断算法以非常高的可信度推断感兴趣位置的等位基因以及 Beacon 查询结果,并表明诸如隐藏基因组的某些部分或为用户设置查询预算等对策将无法保护参与者的隐私.

由于基因组数据固有的稳定性和唯一性特点,使得任何两个个体的 DNA 都可以很容易地相互区分,进而易遭到个体重识别攻击.因此,在表 4 中对基因组数据的个体识别方法进行比较,个体的重识别易导致关联的敏感基因型和表型泄露,而且进一步增加亲属的敏感信息泄露风险.因此,急需隐私保护方法应对个体的重识别风险.

表 4 基因组数据的个体识别方法比较

针对场景	数据类型	方法	攻击效率	威胁程度	关键问题
个体遗传数据与公共 SNP 数据匹配[19]	SNP	贝叶斯定理	评估匹配的后验概率	随着 SNP 数量增加个体重识别风险增加	构建重识别对象在群体中的先验模型
确定个体参与 GWAS 或存在于 Beacon 中[20-22, 25-28, 30]	SNP DNA	假设检验	检验个体是否参与 GWAS 或者存在于 Beacon 中	个体重识别风险与独立单核苷酸多态性的数量成正比	构建基因型-疾病、基因型-基因型检验统计量
数量性状 GWAS 的回归系数或 *p* 值包含隐私信息[31]	SNP	线性回归模型	SNP 的回归系数表明个体的参与程度和表型	在多个表型下回归系数与等位基因频率包含同样数量的个体信息	计算个体贡献的回归系数
攻击者将获得的 miRNA 表达水平在某个时间点与其他 miRNA 表达水平相匹配[32]	miRNA	主成分分析	可随时间跟踪个体 miRNA 表达谱	在基因表达变异下随时间跟踪一个或多个表达谱	基于 miRNA 表达向量之间的欧氏距离构造第 1 个,…,第 *c* 个主成分
基因组数据的身份推断[33]	DNA	远亲匹配	使用人口统计标识符识别远亲家庭搜索的第三个表亲或更接近匹配个体的身份	可以识别公共测序项目的研究参与者	执行远亲家庭搜索
基于表型的基因组识别[34]	DNA	相似性函数	利用全基因组测序数据进行性状预测识别个体	应用全基因组测序、详细的表型和统计模型来预测生物性状	度量未识别基因组表型与已识别表型之间的相似性
确定个体是否在 Beacon 中[35]	SNP	高阶马尔可夫链	基于高阶马尔可夫链对不同基因位点的非独立等位基因进行推断	在个体隐藏 SNP 的情况下以非常高的可信度推断感兴趣位置的等位基因和 Beacon 查询结果	通过使用高阶马尔可夫链模型化 SNP 相关性

2.2 链接攻击

对于基因组数据的链接攻击,攻击者根据已访问且已标识的个体表型特征,使用表型特征和基因组数据之间的已知相关性在基因组数据库中识别特定的基因型,并使用公开可利用的数据集推断其他相关的敏感信息.在表 5 中,本节比较和分析基因组数据的链接攻击.

表 5 基因组数据的链接攻击

文献	隐私威胁	链接数据
Malin 和 Seeney[36]	个体识别 基因型推断	医疗服务数据 特定疾病知识
Malin 和 Seeney[37]	个体识别 基因型推断	位置访问模式
Malin[38]	个体识别 家庭关系识别	系谱知识
Malin 和 Airodi[39]	个体识别 基因型推断	位置访问模式
Nyholt 等[40]	基因型推断	基因组数据
Humbert 等[41]	个体识别 基因型推断	OpenSNP
Alser 等[42]	个体识别 基因型推断	公共基因组数据库 公共非基因组数据库
Li 等[43]	基因型推断	临床蛋白质组数据
Harmanci 和 Gersterin[44]	基因型推断 表型推断	基因型-表型关联性 表达量性状位点

在已有的工作中,可以通过基因组数据链接到公共知识或公开数据集进行个体识别和基因型推断.Malin 和 Sweeney[36]提出 CleanGene 依赖于公开可获得的医疗服务数据和关于特定疾病的知识,以帮助将所标识的个体与 DNA 记录相关联,CleanGene 及其相关实验对于任何试图提供匿名遗传材料用于研究的机构来说都是有用的工具.Malin 和 Sweeney[37]还通过利用患者位置访问模式中的独特特征将基因组数据链接到公开可利用记录中的个体,以此实现轨迹中数据的重新识别,并且可应用于隐私保护的系统测试.Malin[38]利用公开的在线资源链接到个体的家庭关系,从公开可利用的记录中提取系谱知识,并确定特定家庭关系的重识别风险,用来评估家族成员在公开之前的匿名性,并以此设计形式化的隐私保护技术.在现实世界中个体访问不同位置聚集相似信息,当研究多个位置的数据库时,DNA 和已识别的数据库可以建立链接,个体的位置访问模式可以在共享数据库中进行匹配导致轨迹重新识别,因此评估系统中的重新识别风险以开发减轻风险的技术是很重要的,于是 Malin 和 Airodi[39]研究位置访问模式影响匿名基因组数据的重新识别,表明人地分布的不均匀性是影响轨迹重新识别的主要因素之一.Nyholt 等[40]研究在公共领域可获得用于预测缺失数据的基因组数据,表明在完全公开的基因序列中隐藏遗传信息并不简单.

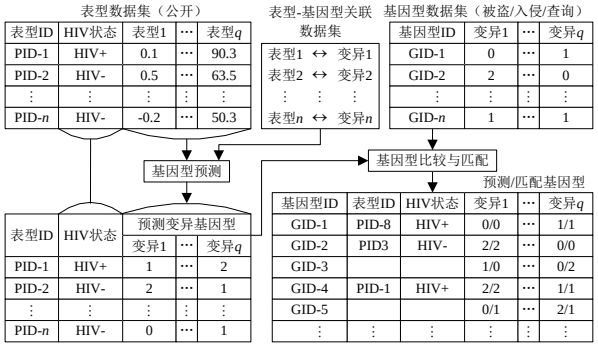

图 4 链接攻击模型

基于各种表型特征,Humbert 等[41]通过在 23andMe 测序的基因组数据库 OpenSNP 上实施去匿名化攻击,结果表明在 50 名参与者的数据库中,采用有监督的方法进行匹配的正确率达到 23%,还分析已知表型性状的数量对攻击成功率的影响,特别是基因型-表型关联的研究使得隐私威胁将变得更加严重.更具体地说,存在两种去匿名化攻击,一是识别攻击,其中攻击者希望识别对应于给定表型的基因型,二是完全匹配攻击,其中攻击者希望匹配多个表型及其相应的基因型.在第一种攻击中,如果攻击者能够访问已识别个体的表型特征,攻击者可以使用表型特征和基因组数据之间的已知相关性在基因组数据库中识别该个体的基因型并推断其他敏感信息.第二种攻击是在加权二部图上找到完全匹配,其中一侧的 n 个顶点表示 n 个不同的基因型,而另一侧的 n 个顶点表示 n 个表型,在完全匹配攻击中,基因型集合为 $G=\{g_1,\ldots,g_n\}$, $g_i=\{g_{i1},\ldots,g_{is}\}$ 表示个体 i 的基因型向量,其中 $g_{ij}\in\{0,1,2\}$,表型集合为 $P=\{p_1,\ldots,p_n\}$, $p_i=(p_{i1},\ldots,p_{it})$ 是个体 i 的表型性状的向量, w_{ij} 是给定基因型 g_i 的表型 p_j 的对数似然比.因为攻击者可以访问包含相同 n 个个体的基因组和表型数据,根据所构建的完全加权二部图,攻击者可以将基因型与其相应的表型相匹配.Alser 等[42]使用公共基因组数据库和其他公共非基因组数据库研究人类基因组数据的广泛跨层隐私攻击策略,包括通过元数据和侧信道攻击进行身份跟踪、系谱三角法进行身份追踪、表型预测的身份追踪、DNA 的属性泄露攻击和完全攻击.Li 等[43]研究在 nsSNP（Non-Synonymous SNP）位点上携带少量次要等位基因的肽可以从单个患者的血液或血清样品中获取的典型临床蛋白质组数据中被识别,进而可以识别患者.如图 4 所示,Harmanci 和 Gersterin[44]研究可以利用公开的基因型-表型关联性,如表达量性状位点（Expression Quantitative Trait Loci,eQTL）来推断基因型,当使用高维的许多表达水平时,这种链接是准确的,并且由此产生的链接可以揭示敏感信息.在链接攻击模型中,攻击者旨在窃取基因型数据集中关于一组个体的敏感信息.在 QTL 数据集

中,eQTL 数据集的丰度使得它们最适合于链接攻击.在 eQTL 数据集中,每个条目包含表达水平和基因型之间的基因、变异和相关系数.该模型可用于在发布之前估计大规模基因组数据集的隐私泄漏.

DNA 包含个体健康和表型的信息,还包含个体血亲的信息.如今,越来越多的基因型、表型数据可以公开获取和使用.因此,在表 5 中使用各种公开可利用数据进行链接攻击,当攻击者访问已标识个体的表型特征,使用表型特征和基因组数据之间的已知相关性在基因组数据库中识别特定个体的基因型,并使用公开可利用的数据集推断其他相关的敏感信息.如果链接的公开基因型、表型数据越多,越容易遭到隐私威胁,而且泄露的敏感信息越多.在链接攻击中主要解决的关键问题是建立基因型数据与表型数据的关联模型,以此进行个体识别和基因型推断攻击.因此,如果要打破数据孤岛,促使数据的融合分析,以此进行基因组数据研究和分析,那么解决基因组数据的链接攻击是关键的挑战.

2.3 基因型推断

基因型推断是指对未观察到的基因型进行重构或预测,本节归纳和总结基因组数据的基因型推断方法.

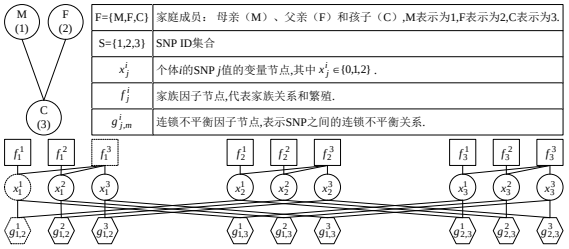

图 5 母亲、父亲、孩子使用 3 个 SNP 的因子图

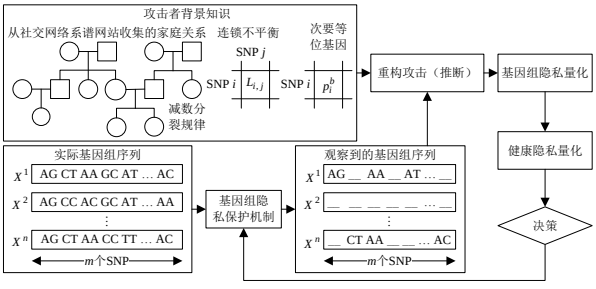

图 6 量化亲属基因组隐私框架

在基因型推断研究中,目前的工作使用不同的方法实现基因型的重构或预测.Malin 和 Sweeney[45]构造简单的基于知识的模型,不需要领域专家,并且在数据很少或没有训练数据的情况下非常有用,从临床表型推断患者的基因型以提高治疗效率.Cassa 等[46]基于条件概率度量患者的 SNP 基因型泄露时其兄弟姐妹的基因型存在的泄露风险,并相当准确的推断出兄弟姐妹的 SNP 基因型,在常见变异的 SNP 处使用非常低的匹配数足以确认兄弟姐妹,从已发布的序列数据可以导出兄弟姐妹的身份.Gitschier[47]根据 Y 染色体的单体型分析,确定贡献 HapMap 项目（CEU 组）的犹他州人（Utahns）是否与其中的个体有关,虽然 CEU 的贡献者似乎都不是男性亲属,可通过类似的过程来预测 CEU 参与者的姓氏.Gymrek 等[48]通过剖析 Y 染色体上的 STR 和查询娱乐性遗传系谱数据库,可以从个体基因组中恢复姓氏,并显示姓氏与其他类型的元数据的组合,可以使用三角测量个体的身份,该技术的关键特性是它完全依赖于免费的、可公开访问的互联网资源.Samani 等[49]基于 k 阶马尔可夫链模拟高阶 SNP 连锁不平衡,并对具有隐藏 SNP 的个体基因型数据进行推断攻击,并通过 SNP 相关性推断攻击来量化基因组隐私的框架.基于贝叶斯定理,Bakes 等[50]表明由基因组变异影响的甲基化区域的子集足以推断出个体的部分基因组,并进一步将该 DNA 甲基化谱映射到相应的基因组,具体包括学习攻击依据甲基化水平与基因型之间的关系推断基因型,匹配攻击归结为在加权二部图上找到最佳的基因型与甲基化谱的顶点匹配.此外,基于机器学习方法也可以实现基因型的重构或预测,Humbert 等[51]通过基于图模型和信念传播（Belief Propagation）的高效重构攻击,攻击者可以推断其基因组被观察到的个体亲属的基因组,主要依靠孟德尔遗传定律和核苷酸之间的统计关系,为了量化所提出的推断攻击导致的基因组隐私水平,讨论基因组隐私度量的可能定义,特别介绍健康隐私（Health Privacy）量化疾病的易感程度.在图 5 中构建父亲、母亲和孩子的因子图（Factor Graph）,并在图 6 中结合信念传播算法量化亲属基因组隐私.利用基因组数据的特殊性、从网络上获得的背景知识以及个体之间的家庭关系,可以推断出共享和非共享基因组的隐藏部分,存在的工作考虑基因组中的简单相关性,以及孟德尔定律和个体及其家庭成员的部分基因组,Deznabi 等[52]对现有的基于基因组隐私的推断攻击进行改进,主要通过观察到的马尔可夫模型和单倍型之间的重组模型来考虑基因组中的复杂相关性,还利用个体的表型信息,提出有效的信念传播算法来考虑所有上述背景信息进行推断,在信息量显著减少的情况下改进了推断攻击.

DNA 中包含重要、未知的信息,而且不随时间的推移而降低,反而随时间的推移而增加,并且基因

组数据具有稳定性和唯一性,基因组数据与血亲、健康表型关联.因此,在表 6 中通过各种方法可以推断基因型.基因型推断可以重构或预测未观察到的基因型,进而可以关联更多敏感信息,例如个体及亲属的遗传疾病和退行性疾病,导致更严重的隐私泄露.因此,面对基因型推断导致的隐私泄露问题,更急需要灵活实现基因组数据的隐私保护.

表 6　基因组数据的基因型推断方法比较

针对场景	数据类型	方法	攻击效率	威胁程度	关键问题
推断个体基因型[45]	SNP	知识算法	从临床表型推断基因型	推断任何简单遗传性疾病临床表型的基因型	将诊断与疾病的症状相关联构建基于知识的初始模型
推断同胞 SNP 基因型[46]	SNP	条件概率	度量患者的 SNP 基因型泄露兄弟姐妹身份的风险且可以准确推断兄弟姐妹的 SNP 基因型	在常见变异 SNP 处以非常低的匹配数确认兄弟姐妹且从已发布的序列数据中获得兄弟姐妹身份	在两个同胞之间的家庭关系和选定的 SNP 下确定第二个同胞携带特定基因型的先验概率变化
确定参与 HapMap 计划（CEU 组）的犹他州人是否与任何个体有关[47]	SNP	Y 染色体单体型分析	确定犹他州人的身份和预测 CEU 参与者的姓氏	通过 Y 染色体单体型分析预测 CEU 参与者的姓氏	对 Y 染色体的分析
从个体基因组中恢复姓氏[48]			将姓氏与其他类型的元数据年龄和状态结合进行三角测量目标的身份	对公共测序项目中多个参与者的身份进行高概率追踪	
推断个体基因型[49]	SNP	k 阶马尔可夫链	基于不同 SNP 的相关性对具有隐藏 SNP 的个体基因型数据进行推断攻击	推断能力随着 SNP 相关性阶数的增加而逐渐提高	构建 SNP 相关性的马尔可夫链模型
识别个体 DNA 甲基化谱[50]	DNA 甲基化	贝叶斯定理	小部分受基因组变异影响的甲基化区域足以推断出基因组的某些部分	小部分受基因组变异影响的甲基化区域足以推断出某人基因组的某些部分,进一步将 DNA 甲基化图谱映射到相应的基因组.	确定先验基因型概率和条件概率
推断个体亲属的基因组[51]	SNP	信念传播算法	根据孟德尔定律和核苷酸之间的统计关系推断个体隐藏的基因型或个体基因组被观察到的其亲属的基因组	从观察到的个体亲属基因组中推断出个体的基因组数据	构建父亲、母亲、孩子家庭关系的因子图
推断共享基因组的隐藏部分[52]			通过观察到的马尔可夫模型和单倍型之间的重组模型来考虑基因组中的复杂相关性可以推断出共享和非共享基因组的隐藏部分	从个体及其家庭成员公开可利用的基因组数据中推断出个体的基因组数据,能够高效、准确地推断出个体的隐私敏感变异.	

2.4　贝叶斯推断

贝叶斯推断是指利用贝叶斯理论推断基因型或识别个体.基因型和疾病之间的 GWAS 表明许多基因位点有助于观察到亲属之间的相似性,GWAS 主要集中在发现具有致病性多态性的基因或调控区.全基因组标记数据也可用于预测遗传值,从而预测表型,因此 Lee 等[53]基于贝叶斯方法利用标记数据预测表型,从全基因组标记数据中预测未观察到的表型方面是成功的,并且明显优于基于近亲表型的预测.RNA 谱分析可用于捕获与 eQTL 相关的许多基因的表达模式,Schadt 等[54]基于 RNA 表达数据提出贝叶斯方法预测 SNP 基因型,并且预测的基因型可以准确和唯一地识别群体中的个体.GWAS 中的数据隐私是关键但尚未被充分探索的研究领域,挖掘 GWAS 统计数据威胁到更广泛人群的隐私,应采用隐私保护机制.因此,Wang 等[55]通过挖掘公开的 GWAS 统计信息构建贝叶斯网络,并对特征推断攻击和身份推断攻击进行评估,表明它们不仅针对 GWAS 参与者而且针对普通个体也可以进行隐私攻击.此外,GWAS 越来越受到人们的关注,以了解遗传变异如何影响不同的人类性状,Zhang 等[56]提出构建三层贝叶斯网络的方法,明确地揭示 SNP 与来自公共 GWAS 目录的性状之间的条件依赖关系.利用 GWAS 统计量建立贝叶斯网络的关键问题是确定具有多个父变量的变量的条件概率表.采用因果影响的独立性模型,假设每个父变量的因果机制是相互独立的,然后根据贝叶斯网络中捕获的依赖关系,提出四个推断问题,即已知基因型下的性状推断、已知性状下的基因型推断、已知性状下的性状推断、已知基因型和性状下的身份推断,并给出有效的公式和算法.这些推断问题的目标是任何个体,而不限于 GWAS 参与者.这意味着可以从 GWAS 统计模型中推断出有意义的信息,并且需要适当的隐私保护机制来保护 GWAS 参与者和个体的遗传隐私.通过从 GWAS 中提取风险等位基因的摘要统计量（Summary Statistics）构建三层贝叶斯网络.所构建的贝叶斯网络能够明确地捕捉 SNP 与其相关性状之间的条件依赖关系,将其作为推断的背景知识.如图 7 所示,性状集合 T 包含 m 个性状,SNP 集合 S 包含 n 个 SNP,每个具体的性状 $T_k \in T$ 关联 SNP 子集 S_k,每个关联 SNP $S_{kj} \in \mathrm{S}_k$.对于每个 SNP S_j,两个节点 S_j^g 和 S_j^a 分别表示其基因型和等位基因.每个性状 $T_k \in \{0,1\}$,其中 1 表示参与者的性状存在,0 表示参

与者的性状不存在.$S_j^a \in \{0,1\}$,其中 1 表示 SNP 具有风险等位基因,否则为 0.$S_j^g \in \{0,1,2\}$,其中 0 表示纯合非风险等位基因,2 表示纯合风险等位基因,1 表示杂合子.为了构建三层贝叶斯网络,首先估计等位基因先验概率

$$P(S_j^a = s_j) = P(S_j^a = s_j \mid T=0)P(T=0) + P(S_j^a = s_j \mid T=1)P(T=1)$$

基于遗传平衡定律,每个 SNP 基因型 S_j^g 的先验概率

$$P(S_j^g = s_j) = \begin{cases} P(S_j^g=0)^2, & s_j = 0 \\ P(S_j^g=1)^2, & s_j = 2 \\ P(S_j^g=0)P(S_j^g=1), & s_j = 1 \end{cases}$$

其次,对于每个 SNP 需要明确条件概率

$$P(S_j^a = s_1 \mid S_j^a = s_2) = \begin{cases} 1, & 2s_1 = s_2 \\ 0.5, & s_2 = 1 \\ 0, & \text{其他} \end{cases}$$

最后,需要确定关联 SNP S_k 每个性状 T_k 的条件概率表,其中

$$P(T_k = 0 \mid \mathrm{S}_k^a = \mathrm{s}^a) = \frac{P(T_k=0)\prod_{S_{kj}\in S_k} P(S_{kj}^a = s_j^a \mid T=0)}{\prod_{S_{kj}\in S_k}\sum_{S_{kj}^g} P(s_{kj}^g)P(s_{kj}^a \mid s_{kj}^g)}$$

对于 $\mathrm{S}\subseteq S$ 的任何赋值 S^g,$\mathrm{T}\subseteq T$,联合概率为

$$P(\mathrm{s}^g, \mathrm{t}) = \prod_{S_j\in \mathrm{S}_1} P(S_j^g)\times \sum_{\mathrm{S}_2^g,\mathrm{S}_3^g,\mathrm{S}_3^g} \Big(\prod_{S_j\in \mathrm{S}_2\cup \mathrm{S}_3} P(s_j^g)P(s_j^a \mid s_j^g)\prod_{T_k\in \mathrm{T}} P(\mathrm{t}_k \mid Par(T_k))\Big)$$

其中 S_1 表示 S 中不关联 T 的 SNP,S_2 表示 S 中关联 T 的 SNP,S_3 表示关联 T 但不在 S 中的 SNP,$Par(T_k)$ 表示三层贝叶斯网络中变量 T_k 的父节点集合.给定观察变量集 S_y^g 和 T_y 的赋值 s_y^g 和 t_y,变量集 S_x^g 和 T_x 的任何期望赋值 s_x^g 和 t_x 的条件概率为

$$P(\mathrm{s}_x^g, \mathrm{t}_x \mid \mathrm{s}_y^g, \mathrm{t}_y) = \frac{P(\mathrm{s}_x^g, \mathrm{t}_x, \mathrm{s}_y^g, \mathrm{t}_y)}{P(\mathrm{s}_y^g, \mathrm{t}_y)}$$

通过公式 $P(\mathrm{s}^g, \mathrm{t})$ 可以计算联合概率 $P(\mathrm{s}_x^g, \mathrm{t}_x, \mathrm{s}_y^g, \mathrm{t}_y)$ 和 $P(\mathrm{s}_y^g, \mathrm{t}_y)$.基于所构建的三层贝叶斯网络,四种推断方法如下.

(1) 给定基因型的性状推断.目标 v 的基因型表示为向量 $\mathrm{s}_v^g = (s_{v1}^g, s_{v2}^g, \ldots, s_{vn}^g)$,其中 s_{vj}^g 表示 SNP j 的基因型.假设攻击者已经得到目标的基因型分布,目的是利用构建的贝叶斯网络推断目标具有特定性状的概率.特定性状的先验概率可从文献或网络中检索.然后,通过从目标基因型推断出目标具有此性状的后验概率为

$$P(\mathrm{t} \mid \mathrm{s}_v^g) = \sum_{\mathrm{Q}^a} \Big(\prod_{S_j\in \mathrm{Q}} P(s_j^a \mid s_{vj}^g) P(\mathrm{t} \mid \mathrm{q}^a)\Big)$$

其中 Q 表示关联性状 T 的 SNP.

(2) 给定性状的基因型推断.使用 $\mathrm{s}_v^g = (s_{v1}^g, s_{v2}^g, \ldots, s_{vn}^g)$ 表示任意基因型分布,给定目标性状 T_v 的一个子集及其赋值 t_v.通过所构建的贝叶斯网络,在给定个体相关性状信息的情况下,获得个体具有特定基因型的后验概率

$$P(\mathrm{s}_i^g \mid \mathrm{t}_v) = \frac{\prod_{S_j\in \mathrm{Q}} P(s_{ij}^g)\sum_{\mathrm{Q}^a}\Big(\prod_{S_j\in \mathrm{Q}} P(s_{ij}^g)P(s_j^a \mid s_{ij}^g)\Big)\prod_{T_k\in \mathrm{T}_v} P(t_k \mid Par(T_k))}{\sum_{\mathrm{Q}^g,\mathrm{Q}^a}\Big(\prod_{S_j\in \mathrm{Q}} P(s_j^g)P(s_j^a \mid s_j^g)\Big)\prod_{T_k\in \mathrm{T}_v} P(t_k \mid Par(T_k))}$$

其中 Q 表示与 T_v 性状相关的 SNP.

(3) 给定性状的性状推断.给定目标的性状 t_v,目标有新性状的概率

$$P(t_{new} \mid \mathrm{t}_v) = \sum_{\mathrm{Q}^g} P(t_{new} \mid \mathrm{q}^g) P(\mathrm{q}^g \mid \mathrm{t}_v)$$

其中 Q 是与 t_{new} 和 t_v 相关联的 SNP 集合.

(4) 给定基因型和性状的身份推断.当目标的某些性状可用时,推断匿名基因型数据库中的记录属于目标的概率.攻击者可以访问包含目标基因型记录 s_v^g 的匿名基因型数据集 R,攻击者还知道目标拥有的性状 t_v 的子集,然后攻击者可以获得数据库中的每个基因型记录 s_i^g 对应于目标的概率

$$P(\mathrm{s}_i^g = \mathrm{s}_v^g \mid \mathrm{t}_v) = \frac{P(\mathrm{s}_v^g \mid \mathrm{t}_v)}{\sum_{i=1}^{|R|} P(\mathrm{s}_i^g \mid \mathrm{t}_v)}$$

以此攻击者能够从匿名数据集中识别目标的记录.

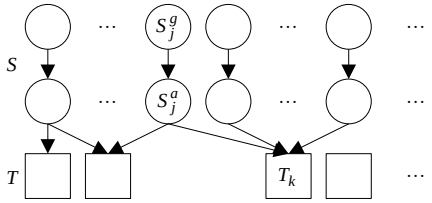

图 7 性状及其关联 SNP 的三层贝叶斯网络

针对给定基因型推断性状、给定性状推断基因型、给定性状推断性状、给定基因型和性状推断身份[53-56],在大规模基因表达研究和全基因组关联研究中,根据 SNP 基因型和表型数据的依赖关系构建贝叶斯网络推断表型和基因型,进而能够准确地识别个体.在贝叶斯推断中目标可以是任何个体,而不限于 GWAS 参与者.贝叶斯推断方法的关键问题是利用背景知识的条件概率建立贝叶斯网络,需要大规模的数据进行训练,以此构建稳定、良好的贝叶斯网络,进而更加准确地实现基因型推断或个体识别.

2.5 重识别风险博弈

博弈论是指参与者同时或先后选择可能的策略,通过策略实施后每个参与者都将获得相应的收

益.重识别风险博弈模型使管理者能够平衡重识别风险与共享数据的价值,然后数据管理者决定是否发布基因组数据.

Mastermind 是在两个参与者之间的博弈,即编码器和译码器之间的博弈.Goodrichp[57]研究 Mastermind 对基因组数据的攻击,基因组查询者 Bob 使用基因组字符串 Q 执行 Mastermind 博弈,Q 的所有者 Alice 以隐私保护的方式与 Bob 所提供的 Q 进行比较,可使得 Bob 发现 Q 的完全身份.攻击场景是 Alice 重复地参与 Q 的隐私保护比较,以迭代地将 Q 与 Bob 所提供的字符串进行比较.因此,取决于对 Q 的结构了解,Bob 利用 Mastermind 攻击有效地确定 Q.

利用自然假设,接收方只有在潜在收益大于成本的情况下才尝试重新识别.Wan 等[58]提出博弈模型使发布者能够平衡重新识别风险和共享数据的价值.形式化发布者和攻击者的收益函数如表 7 所示,对于未被攻击的固定数据共享策略 g,发布者和攻击者的收益函数分别为 $v(g)$ 和 0.在攻击的情况下,针对固定数据共享策略 g,对于发布者和攻击者的收益函数分别为 $v(g)-L\pi(g)$ 和 $L\pi(g)-c$,其中 L 是发布者被成功地重新识别一条记录的损失,$\pi(g)$ 表示攻击者重新识别一条记录的概率,c 是攻击者对一条记录发起重新识别攻击的成本.该博弈模型决定是否有关去标识符数据的共享策略.

表 7　固定数据共享策略 g 的发布者和攻击者的收益函数

参与者	收益函数	无攻击	攻击
发布者	$U_p(g)$	$v(g)$	$v(g)-L\pi(g)$
攻击者	$U_a(g)$	0	$L\pi(g)-c$

针对基因组数据共享和隐私泄露风险之间的矛盾问题[57, 58],在共享之前通过构建隐私泄露风险与共享数据的价值之间的均衡博弈模型,有助于选择合适的隐私保护方法解决数据共享中的隐私泄露问题,进而解决因隐私泄露而不能挽回损失的问题.如果攻击者的收益大于所期望的收益函数值,那么共享的基因组数据易遭到隐私攻击.进一步,如果攻击者的收益函数远大于所期望的收益函数,那么攻击者具有完全控制共享基因组数据的能力.在基因组数据的重识别风险博弈中,使隐私泄露风险与共享数据的价值之间达到均衡是需要解决的关键问题.不过,考虑基因组数据具有高维、大规模,且连锁不平衡的特点,构建完美解决隐私泄露风险与共享数据的价值之间的均衡博弈模型是具有挑战的.

3　基因组数据隐私与效用度量

针对基因组数据存在的隐私泄露威胁,需要量化基因组数据的隐私泄露风险.同时,通过使用隐私保护技术对基因组数据进行隐私保护后,也需量化基因组数据隐私和效用.为此,可以使用基因组数据的隐私度量和效用度量来分别量化隐私泄露风险和数据效用.由于基因组数据隐私度量和效用度量方法不唯一,因此本节归纳和总结常用的隐私和效用度量方法,其具体的隐私度量和效用度量如下.需要说明的是,本节中的估计基因型指的是攻击者推断的基因型或者是使用隐私保护方法后的基因型.

3.1　基因组数据隐私度量

本论文将基因组数据隐私度量分为不正确性、不确定性和健康隐私三类.不正确性通过计算每个 SNP 的估计值与真实值之间的距离来量化攻击者的攻击效果.攻击者的不确定性表示推断 SNP 基因型值所包含的信息量.健康隐私主要关注导致特定疾病的 SNP 的贡献.下面分别具体介绍各种基因组数据隐私度量方法.

（1）不正确性

保护基因组数据隐私的关键步骤是量化保护后的基因组数据引起的隐私损失.不正确性指的是基因组数据保护后与保护前的差值、误差或距离.在表 8 中,不正确性可以用作基因组数据的隐私度量,不正确性度量包括汉明距离、编辑距离、集合差、期望估计误差、泛化强度、平均误差和标准化均方误差.在不正确性的隐私度量中,随着基因组数据保护后与保护前的差值、误差或距离等增加,隐私保护效果越好.

（2）不确定性

不确定性也可以作为基因组数据隐私度量方法,其中信息熵量化随机变量中包含的信息量,作为隐私度量表示攻击者的不确定性.在表 9 中,主要介绍信息熵、标准化熵、互信息和非对称熵,以及基因隐私作为基因组数据的隐私度量方法.对于其他不确定性的隐私度量方法,主要是基于信息熵、标准化熵、互信息和非对称熵的变式,本论文省略这些变式的介绍,详细的形式化和定义详见文献[65].

（3）健康隐私

由于基因型和疾病相关联,个体的健康状态是敏感信息.因此,可以使用健康隐私来度量个体的疾病状态.针对不同的基因型-疾病关联性,Humbert 等

[51]提出健康隐私量化个体的疾病状态.S_d 表示与疾病 d 相关联的 SNP 的 ID 的集合.关于疾病 d 的个体 i 的健康隐私度量可定义为

$$D_d^i = \sum_{k\in S_d} c_k G_i^k \Big/ \sum_{k\in S_d} c_k$$

其中 G_i^k 是 SNP k 对个体 i 的基因隐私度量,使用前面提到的基因隐私度量进行计算,而 c_k 是 SNP k 对疾病 d 的贡献.当然,根据基因与疾病之间的关联性,可以基于基因型的非线性组合或等位基因组合定义其他健康隐私度量.面向疾病易感性的隐私度量,为了量化个体的健康隐私以表明对不同疾病的易感性趋势.健康隐私度量值越大越容易关联某种疾病,但因基因组研究可以改变关于 SNP 对疾病贡献的知识,健康隐私度量在给定的时间是有效的,不能用于评估未来的健康隐私.因此,健康隐私常用作疾病易感性检测中的隐私度量.

由于隐私度量方法多种多样,而且在具体使用隐私度量时,目前没有统一的标准.因此,在基因组数据的具体应用中,研究人员可能根据自己所熟悉和掌握的专业知识,使用不同的隐私度量方法.文献[65]较全面地总结用于基因组数据的现有隐私度量方法,感兴趣的读者可以深入阅读和了解相关的基因隐私度量方法.

表 8 基因组数据的不正确性隐私度量及比较

度量方法	度量公式	保护效果	应用场景	攻击难度	采用率
汉明距离	$\lvert\{x_i \in x, y_i \in y : x_i \neq y_i\}\rvert$	x 和 y 之间不同基因型的位点数目	攻击者推断个体的 SNP	汉明距离越大则将估计基因型变换成真实基因型所需替换的基因型个数越大	在便携式和普及的智能环境中基因检测的隐私度量[59]
编辑距离	$\lvert\{z : x \cup z = y \text{或} x \setminus z = y\}\rvert$	将 x 转换为 y 所需的插入和删除的基因型数目	攻击者推断个体的 SNP	估计基因型转换为真实基因型所需增加和删除的基因型数越大则它们之间的编辑距离很大	序列配对和相似序列查询的隐私度量[60-63]
集合差	$\lvert(x \setminus y) \cup (y \setminus x)\rvert$	度量两个基因型集 x 和 y 的对称差的大小	攻击者推断个体的 SNP	估计基因型与真实基因型之间的共同基因型数越小则集合差越大	相似序列查询的隐私度量[62]
期望估计误差	$\sum_{x_i\in\{0,1,2\}} p(y_i)\lVert y_i - x_i\rVert, x_i \in x, y_i \in y$.	量化每个 SNP 的估计值与真实值之间的期望距离	攻击者推断个体的 SNP	计算真实基因型和估计基因型之间的期望估计误差越大则推断基因型的正确性越低	量化亲属基因隐私[51]
泛化强度	$\sum_a s_a \Big/ (\sum_a h_a - m)$, m 表示碱基的数量, h_a 表示碱基 a 在泛化层次结构中的层数, s_a 是发布者泛化 a 的策略.	属性等价类的大小最少为 h_a	攻击者推断个体的 SNP	估计基因型的等价类大小比真实基因型的大则泛化程度强	基因组数据共享[58]
平均误差	$\sum_{x_i\in\{0,1,2\}} \lvert y_i - x_i\rvert \Big/ \lvert x\rvert, x_i \in x, y_i \in y$.	真实基因型值与估计基因型值差的绝对值的平均值	攻击者推断个体的 SNP	估计基因型值与真实基因型值之间的绝对误差越大则平均误差越大	量化基因隐私[49]
标准化均方误差	$\sum_{x_i\in\{0,1,2\}} \lVert y_i - x_i\rVert^2 \Big/ \lVert x\rVert^2$	以输入基因型 x 的平方标准化输出基因型 y 与输入基因型 x 之间的平方差	攻击者推断个体的 SNP	估计基因型值与真实基因型值之间的差异程度越大则标准化均方误差越大	微生物 DNA 分析的隐私度量[64]

表 9 基因组数据的不确定性隐私度量及比较

度量方法	度量公式	保护效果	应用场景	攻击难度	采用率
熵	$H(x) = -\sum p(x_i)\log_2 p(x_i)$	推断基因型值的不确定性	推断个体的 SNP	随着熵增大不确定性越大	量化亲属基因隐私[51]
标准化熵	$H_0(x) = \log_2 \lvert x\rvert$	推断基因型值的平均不确定性	推断个体的 SNP	随着标准化熵增大平均不确定性越大	量化亲属基因隐私[51]
互信息	$I(y;x) = H(y) - H(y\mid x)$	真实基因型 x 与估计基因型 y 之间共享多少信息	推断个体的 SNP	互信息越小真实基因型与估计基因型之间共享信息越少	量化亲属基因隐私[51]
非对称熵	$\sum_{i=1}^{\lvert x\rvert} \frac{p(x_i = y_i)(1 - p(x_i = y_i))}{(-2w_i + 1)p(x_i = y_i) + w_i^2}$,如果 $y_i = 0$,则 $w_i = (1 - r_i)^2$;如果 $y_i = 1$,则 $w_i = 2r_i(1 - r_i)$;如果 $y_i = 2$,则 $w_i = r_i^2$;其中 r_i 表示次要等位基因的频率.	推断基因型值的不确定性	推断个体的 SNP	估计基因型是基于群体范围内的次要等位基因频率获得每个 SNP 的不同最大熵值	基因检测的隐私度量[66]
基因隐私	$-\sum \log_2(p(x_i = 1) + p(x_i = 2))W_i$,其中 W_i 表示第 i 个 SNP 的权重.	衡量每个 SNP 对疾病的贡献	推断个体的 SNP	对 SNP 以变异形式出现具有 1 个或 2 个次要等位基因时的情况进行加权估计	疾病易感性检测的隐私度量[67]

3.2 基因组数据效用度量

本节概述已有工作中常用的基因组数据效用度量方法,具体的效用度量方法和度量公式如下.

（1）信息损失

文献[58]给定发布策略 g,该策略将属性 f 泛化到区间 $[q_l(f,g),q_h(f,g)]$.假设属性 f 的原始值 A_f 是该区间中的均匀分布随机变量,则 $A_f=q$ 的概率

$$p(q,f,g)=\frac{1}{q_h(f,g)-q_l(f,g)}$$

则信息损失为每个属性的熵的综合,其计算公式为

$$IL(g)=-\sum_f\sum_q p(q,f,g)\log_2 p(q,f,g)$$

（2）卡方统计

在GWAS研究中,Mohammed等[68]使用 χ^2 检验来评估差分隐私数据的效用,χ^2 检验的计算公式为

$$\chi^2=\sum_i^r\sum_j^c\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

其中 r 是行数,c 是列数,O_{ij} 是观察频率,而 E_{ij} 是期望频率. χ^2 检验统计量提供观察频率与期望频率的接近程度的度量.

（3）假阳性和假阴性

假阳性是当不存在基因型时返回错误查询结果,而当存在基因型时返回错误查询结果则会发生假阴性.Wang等[62]使用假阳性与假阴性的比率和错误率来度量准确性.错误率度量隐私集合差大小协议的准确性,它被定义为 $|u-v|/u$,其中 u 表示实际大小,并且 v 是安全协议输出的大小.

（4）准确率

设 x 是 SNP 基因型数据集,而 y 是使用某些差分隐私算法随机扰动 x 获得的 SNP 基因型数据集.因此,Simmons和Berger[69]利用准确率 $|x\cap y|/|x|$ 作为效用度量,其越接近 1,基因组数据的效用越好.进一步,定义期望准确率（Expected Accuracy）作为基因组数据的效用度量方法,其计算公式为

$$EA=\sum_{\{x_i:x_i\in x,y_i\in y,x_i=y_i\}}p(x_i=y_i)x_i\Big/|x|$$

（5）期望区间宽度

只要区间包括对应于真实输入的输出,如果区间的宽度较小,则隐私保护机制 M 的输出更有用.因此,Kusano 等[70]定义期望区间宽度来度量基因组数据的效用,其计算公式为

$$U(M)=-E_{x\in X}[|M(x)|]=-\sum_{x\in X}|M(x)|\,p(x)$$

其中 $|\cdot|$ 表示区间的宽度.

在表 10 中对上述基因组数据的效用度量方法进行比较,表明对于使用隐私保护方法后保持个体SNP 的度量效果,从攻击的视角分析各种度量方法的攻击难度.效用度量也表明隐私保护后接近真实基因型数据的程度,以此说明隐私保护方法的数据效用.此外,根据上述度量方法在目前研究成果中的应用,分析各种度量方法在不同应用中的采用率.

表 10 基因组数据的效用度量及比较

度量方法	度量效果	应用场景	攻击难度	采用率
信息损失	泛化真实基因型到共享基因型的程度越小则信息损失越小	保持个体的SNP	真实基因型的泛化程度越低则容易推断个体的SNP	基因组数据共享[58]
卡方统计	度量估计基因型频率与真实基因型频率的接近程度	保持个体的SNP	估计基因型频率与真实基因型频率以小概率发生显著接近则容易推断个体的SNP	基因组数据共享[68]
假阳性与假阴性比率	假阳性是对不真实基因型的错误估计 假阴性是对真实基因型的错误估计	保持个体的SNP	假阳性错误和假阴性错误之间具有互斥性,则假阳性与假阴性之间的比率过大或过小都容易推断个体的SNP.	相似患者查询[62]
错误率	衡量正确估计基因型与真实基因型之间的准确性	保持个体的SNP	错误估计基因型的数量占真实基因型数量的比例越小则容易推断个体的SNP	相似患者查询[62]
准确率	正确估计的基因型数量占真实基因型数量的比例	保持个体的SNP	估计基因型与真实基因型共享越多的基因型则容易推断个体的SNP	GWAS[69]
期望准确率	正确估计基因型的加权占真实基因型数量的比例	保持个体的SNP	估计基因型与真实基因型共享越多的基因型则容易推断个体的SNP	基因组数据共享、GWAS、相似患者查询等
期望区间宽度	发布包含真实输出值的区间而不是发布真实输出值	保持个体的SNP	发布包含真实输出值的区间越小则容易推断个体的SNP	发布疾病易感性值[70]

4 基因组数据隐私保护

本节分析基因组数据测序和存储、共享和聚集,以及应用中的隐私泄露问题,并从基于密码学、匿名和差分隐私方法,以及混合方法介绍和分析基因组数据的隐私保护.

4.1 基因组数据隐私泄露威胁

结合图2的基因组数据生态系统,根据第2节中存在的基因组数据隐私威胁,下面从基因组数据测序到应用过程中,分析基因组数据存在的隐私泄露问题.

（1）基因组数据测序与存储

在相关法律法规的监管下,基因数据测序中心

是可信的.但是,在利益驱使下,考虑测序人员具备基因组学背景知识,人为因素可能导致基因组数据测序中的敏感数据泄露.同样在隐私保护法律法规监管下,数据中心是可信的.因此,数据中心从测序中心获得原始基因组数据后进行存储和管理.但是,考虑到基因组数据具有稳定性的特点,不随时间变化而变换,而随时间的推移其价值越来越重要,因为设备故障、恶意攻击等不可预测的行为会导致基因组数据的隐私泄露.此外,基因组数据具有唯一性,而且关联遗传、健康、表型和血缘关系.为此,在基因组数据测序中进行脱敏处理,以及在存储中实现基因组数据的长期安全是非常有必要的.因此,已有工作使用密码学、匿名方法,以及混合方法实现基因组数据测序与存储的隐私保护.

（2）基因组数据共享与聚集

基因组研究、基因组分析和基因组计算需要聚集基因组数据,以此通过聚集大量的基因组数据来鉴定基因组数据各种应用中的复杂性状.而聚集大量的基因组数据之前,需要个体和机构参与基因组数据的共享.但是由于基因组数据固有的唯一性,并且关联遗传、健康、表型和血缘关系的敏感性特点,基因组数据的共享必然会导致隐私泄露问题.因此,目前已有相关工作对此展开研究,并取得一系列的研究成果.主要使用密码学、匿名、差分隐私和混合方法实现基因组数据共享与聚集中的隐私保护.

（3）基因组数据研究与分析

在基因型与疾病的关联研究与分析中,GWAS是研究基因型-表型关联的最常见类型之一.随着新的人类遗传学研究继续向许多实验室和医疗机构迅速扩展,对参与其中的个体的隐私保护的关注日益增加.即使是最小的实验室也必须面对如何在他们的研究与分析中保护参与者基因组数据的唯一性,以及关联遗传、健康、表型和血缘关系等敏感信息的问题.已有工作通过密码学和差分隐私方法实现基因组数据研究与分析的隐私保护.

（4）基因组数据医疗服务

近年来,随着基因测序成本的大幅度降低,进而产生大量的基因组数据,基因组数据广泛应用于推进医学研究、改进临床程序和医疗服务,使得GWAS、诊断检测、个性化医学和药物发现领域发生革命性变化.可是,由于人类基因组数据的固有的唯一性和关联遗传、健康、表型和血缘关系的敏感性特点,这就带来基因组数据医疗服务中的隐私威胁挑战.在目前的工作中,基于密码学、差分隐私方法实现基因组数据医疗服务的隐私保护.

（5）基因组数据法律与取证

在亲子鉴定和刑事取证中,都可以基于 DNA 进行搜索与匹配,以此实现亲属鉴定和犯罪嫌疑人识别.但是,因为 DNA 具有唯一性,而且关联遗传、健康、表型和血缘关系的特点,在 DNA 搜索与匹配中会导致个体的隐私泄露,甚至导致家庭的基因组数据隐私泄露问题.针对此问题,目前主要使用密码学实现基因数据的法律与取证中的隐私保护.

（6）基因组数据直接面向消费者服务

基因组数据直接面向消费者服务包括个性化医学、疾病易感性检测、身份检测、血缘关系检测、祖先检测、配偶兼容性检测、亲子鉴定等.直接面向消费者服务使得个体能够直接参与基因组数据的收集和处理,甚至分析.然而,因为基因组数据固有的唯一性,且关联遗传、健康、表型和血缘关系的特点,直接面向消费者服务不仅泄露个体本身的敏感信息,也会泄露其亲属的敏感信息.由于亲属的基因组数据高度相关联,某些家庭成员可能反对透露该家族的任何基因组数据.因此,急切需要对直接面向消费者服务的基因组数据进行隐私保护.在目前的工作中,主要使用密码学和匿名方法实现基因组数据直接面向消费者服务中的隐私保护.

具体地,在表 11 中针对基因组数据从测序产生到应用过程中的隐私泄露问题,通过使用相应的方法实现基因组数据的隐私保护.

表 11　基因组数据安全与隐私需求及保护方法

基因组数据生态系统	安全与隐私保护需求	主要方法
基因组数据测序与存储	长期安全 个体敏感信息泄露	密码学 匿名 混合方法
基因组数据共享与聚集	个体敏感信息泄露	密码学 匿名 差分隐私 混合方法
基因组数据研究与分析	个体敏感信息泄露	密码学 差分隐私
基因组数据医疗服务	个体敏感信息泄露	密码学 差分隐私
基因组数据法律与取证	个体敏感信息泄露 亲属敏感信息泄露	密码学
基因组数据直接面向消费者服务	个体敏感信息泄露 亲属敏感信息泄露	密码学 匿名

4.2　基于密码学的基因组数据隐私保护

针对基因组数据测序和存储、共享和聚集,以及应用中的安全和隐私保护研究,主要需要解决的关键问题是基因组数据的机密性、完整性、可用性、可认证性和不可否认性.机密性要求基因组数据不被非授权攻击者获取,使攻击者实施非法窃听的被

动攻击行为无法得逞,实现机密性的最根本方法是采用高强度加密算法对基因组数据进行加密保护.完整性要求基因组数据不被非授权攻击者篡改,即信息在存储或传输的过程中不被修改、插入或删除,因现实中很难防止攻击者对基因组数据的篡改行为,主要采用检测与恢复措施.可用性要求基因组数据和相关资源可以持续有效并且合法用户可以访问和使用,可用性用来应对基因组数据传输的中断攻击.可认证性要求基因组数据的来源或本身能够被正确地标识且该标识不能被伪造,可认证性用来应对基因组数据的假冒、伪造、重放攻击,为了获得认证的持续保证,通常需要将可认证性和完整性结合起来使用.不可否认性要求基因组数据处理的参与方事后不能否认的行为,基因组数据发出后,接收方能够证实该基因组数据的确来源于声称的发送方,发送方不可否认,基因组数据被接收后,发送方也能够证实该基因组数据的确发送到指定的接收方,接收方也不可否认,而且接收方对接收到的基因组数据的处理行为也不可否认,不可否认性用来应对基因组数据处理行为的否认攻击.这里着重介绍常用于实现基因组数据安全和隐私保护的密码学方法,及其目前用于基因组数据隐私保护的研究工作.除非另有说明,在本节介绍目前的工作中主要考虑半诚实攻击模型下的基因组数据安全和隐私保护.

4.2.1 基于对称和非对称加密的基因隐私保护

（1）对称加密和非对称加密

在密码系统中,Enc是加密算法,Dec是解密算法,m表示消息,因此$Dec(Enc(m,k_e),k_d)=m$.对称加密的加密密钥k_e和解密密钥k_d是相同的,即$k_e=k_d=k$.DES、AES是经典的对称密码系统.例如通过AES-GCM（AES in Galois Counter Mode）[71]加密基因组数据和计算结果的隐私保护国际合作框架,用于不可信计算服务提供商分析分布在不同大陆的罕见疾病遗传数据,并以安全的方式将结果返回给数据所有者,确保其机密性和完整性.非对称加密使用公私钥对,其中公钥k_e是公开的,而私钥k_d是保密的.RSA、ElGamal、Paillier和Rabin是常用的非对称密码系统.例如,椭圆曲线提升ElGamal加密（Elliptic-Curve Lifted-ElGamal Encryption）[72]用于研究人员发送安全的查询到服务器,以便于实现等位基因频率的隐私保护数据挖掘.基于非对称加密可以实现数字签名,发送者使用私钥签名,即$Sig(m,k_d)$,接收者使用公钥验证签名,即$Ver(Sig,k_e)$,使得$Ver(Sig(m,k_d),k_e)=m$.RSA数字签名、ElGamal数字签名是经典方案.数字签名可用于身份验证和保证消息的完整性.例如,通过椭圆曲线数字签名算法（Elliptic Curve Digital Signature Algorithm）认证数据所有者的身份,可以防止不可信计算服务提供商从未经授权的数据所有者接收伪造或恶意数据[71].

（2）保序加密

保序加密（Order Preserving Encryption,OPE）是保留明文的数字顺序的确定性加密方案[73].对于$A,B\subseteq \mathrm{N}$,且$|A|\leqslant|B|$,如果$i,j\in A, f(i)>f(j)$当且仅当$i>j$,那么函数$f:A\to B$是保序的,其中N是所有非负整数集合.对于所有密钥空间$k\in K$,如果$Enc_{\mathrm{OPE}}(k,m)$是从明文空间$m\in M$到密文空间$c\in C$的保序函数,那么关于M到C的确定性加密方案是保序的.例如,保序加密方案可以用来加密重要核苷酸的位置.基于保序加密,可以实现加法同态性质的保序加密,即加法保序加密（Additive OPE）[74].例如,序列比较需要进行数值加法和数值比较,使用加法保序加密保证数个密文相加后仍能保序.

（3）基因组数据测序与存储

无论是存储、分析,还是在传输中,敏感基因组数据都应该保持安全.然而,基因组数据的安全存储、传输和使用由于文件的大小和工作流程的多样性而变得复杂,文献[75]在存档前使用计数器模式的AES单独加密基因组数据,实现基因组数据在存储和传输过程中的端到端安全,并提供高级随机访问模式.因极为敏感的性质,需要保证基因组信息的机密性、完整性和真实性,Hosseini等[76]结合AES和Shuffling机制快速加密基因组数据,以维护在传输和存储中基因组数据的安全性,从而增强抵御复杂性攻击的安全性,并保证基因组数据的机密性、完整性和真实性.生物医学数据可以通过密码协议进行共享、管理和分析,而不必透露任何特定记录的内容,实际的密码协议需要包含多个第三方,这在信任或带宽限制的情况下可能并不总是可行的,为解决该问题,如图8所示,针对不可信平台,Canim等[77]基于AES通过使用加密硬件,敏感基因组数据仅在加密硬件中解密以防止泄露,从而不同组织以安全的方式共享和存储基因组数据到集中的站点,并进行安全的查询.为了解决配对的原始基因组数据的隐私处理问题,Ayday等[78]在SAM（Sequence Alignment/Map）文件中通过排列短读（Short Read）的位置来修改基因组的结构,然后在短读的位置使用保序加密,以此从Biobank检索加密的短读,而不向Biobank透露请求的范围,并对短读的内容使用AES加密,用于隐藏Biobank中超出医疗单位请求范围的加密短读的特定部分,然后再将其提供给医疗单位,以此将患者的加密SAM文件存储在Biobank

中,并在保护患者基因组隐私的同时,向医疗单位提供所需的核苷酸范围.文献[78]提出的隐私保护系统没有提供有效的压缩方法,针对没有标准的基因组数据存储解决方案能够实现压缩、加密和选择性检索,Huang 等[79]使用保序加密来加密短读的位置信息,使用对称加密来加密各个位置的压缩敏感内容,用于安全存储压缩配对的基因组数据,因此在检索过程中只输出查询范围内的序列和元数据,不存在泄露任何未授权位置信息的风险.基因组数据需要长期保护完整性、真实性和机密性,因此 Braun 等[80]用信息论隐藏承诺来保护机密性、完整性和真实性,使用先应秘密共享（Proactive Secret Sharing）来存储机密数据,在先验秘密共享协议中通过量子密钥分发和一次一密的信息论隐私信道实现安全存储系统,由于使用分布式存储系统,因此适合于保护云中非常敏感的数据.虽然文献[80]的方案是为存储大文件的数据库而设计的,但它不适合于保护具有许多小条目的基因组数据库,而这些条目又需要单独访问,因此 Buchmann 等[81]使用无条件隐藏承诺、Merkle 哈希树和数字签名来保护数据的完整性,同时保持机密性,该方案允许查询和完整性证明基因组中的特定位置,而其余的数据仍不公开,且无法推断有关相邻位置的信息,患者基因组的随机部分由授权的医生和临床医生定期访问,该方案支持更新完整性保护,以防所使用的密码方案在不久的将来变得不安全.

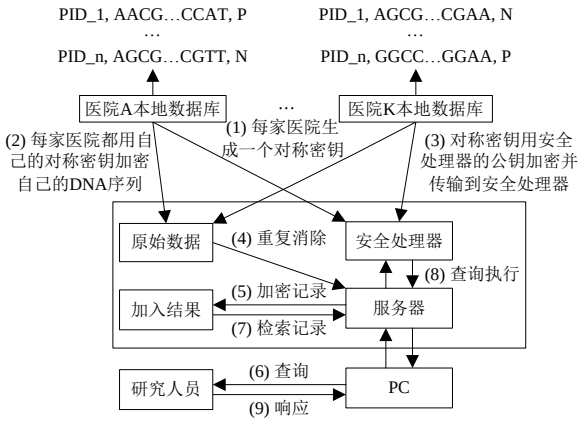

图 8 基因组数据管理框架

（4）基因组数据共享与聚集

DNA 样本通常用患者标识符或假名进行标记,如果在传输过程中被截获,则允许潜在的链接到身份和隐私临床信息,为此在图 9 中 Cassa 等[82]提出安全传输外部生成序列数据的加密方案,该方案不需要任何患者标识符、公钥基础设施或密码传输,使用从遗传指纹中提取出来的共享对称密钥,用于在测序实验室（Sequencing Laboratory）加密完整的基因组序列,也用于生物存储中心（Biorepository Center）解密完整的基因组序列,它确保序列在整个传输过程中加密,如果出现遗传指纹泄漏的安全故障,而加密过程的唯一性确保每个样本是独立的,并且不会导致一组样本的加密失败.GA4GH 开发并运营 MME（Matchmaker Exchange）平台,该平台允许研究人员通过多个联邦数据库查询罕见遗传病的发现,查询包括与罕见疾病相关的基因变异,但是研究人员可能不愿意使用该平台,因为他们所做的查询会透露给其他研究人员,这就产生对隐私和竞争优势的担忧.为解决此问题,Oprisanu 和 Cristofaro[83]提出支持 MME 平台内匿名查询的框架 AnoniMME,该框架基于隐私信息检索协议允许研究人员匿名地查询联邦平台,具体地研究人员使用非对称密码加密查询,其他研究人员可通过提供加密的联系人详细信息来响应查询.

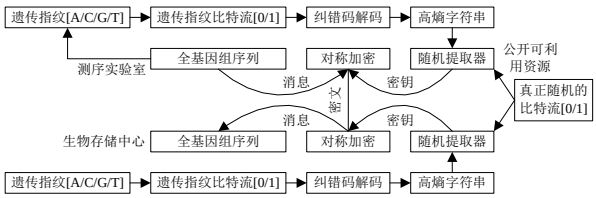

图 9 共享序列数据的密码体系结构

（5）基因组数据研究与分析

在 GWAS 的隐私保护研究中,需要保护参与者的遗传和表型敏感信息,Gulcher 等[84]使用 AES 开发与 Icelandic 数据保护委员会（Data Protection Commission）直接合作的第三方加密系统,并且已经将加密系统合并到样本收集和存储软件中,这样就减少不便,而且提高安全性.人类遗传学最近从 GWAS 转向基于下一代测序（Next Generation Sequencing,NGS）数据的研究,对于 GWAS 通过跨组织交换摘要统计信息进行元分析,NGS 研究每个基因的多个潜在原因等位基因之间的组间关系,在数据交换过程存在遗传信息的隐私泄露风险,许多 NGS 关联的评分方案依赖于每个变异的频率,因此需要交换序列变异的身份,由于这种变异通常很少见,可能会暴露其携带者的身份,并危及隐私.因此,Singh 等[85]提出 MetaSeq 协议用于多个合作方对全基因组测序数据进行元分析,并对所有合作方的每个基因的罕见变异进行关联评分,以此解决罕见的测序等位基因频率计数的问题,在不泄露等位基因身份和计数的情况下对测序数据进行元分析,从而保护样本身份.各方使用对称加密方法加密基因

和变异的身份,当在案例和对照组中传输有关频率计数的信息时,交换的数据不会传递变异的标识,因此不会暴露载体标识,该方法适用于来自多个研究的公开的外显子组测序数据,模拟表型信息进行强大的元分析.如图 10 所示,利用云服务商进行人类基因组分析是高风险的,因为人类基因组数据包含人类个体及其疾病易感性的可识别信息,于是 Zhao 等[86]基于 AES 提出用于加密个体的基因组序列的站点式加密方法,其可以在公共云上进行基因组特征的安全搜索.此外,跨全基因组关联研究的元分析是发现遗传关联的常用方法,然而由于隐私和机构审查委员会的考虑,个体不能广泛共享敏感数据,使得研究人员不能确认每项研究代表独特的人群,这可能会导致夸大的测试数据和假阳性结果,为此 Turchin 和 Hirschhorn[87]基于单向加密哈希实现个体的隐私保护,允许在不共享个体数据的情况下识别重叠的参与者,实现个体级别数据所需的安全性,同时保护识别相同个体所需的特异性和敏感性.在大多数人类 DNA 分析之前读取映射（Read Mapping）,将数百万个短序列与参考基因组配对,这涉及评估数百万亿序列对的编辑距离,因此需要外包给低成本的商业云,考虑对人类基因组数据的主要威胁是对 DNA 来源的个体识别,Chen 等[60]提出在混合云上实现安全和可扩展的读取映射的新方法,混合云包括公共商业云和组织内的私有云,公共云在短读子串和参考序列的带密钥的哈希值之间寻找精确匹配,以大致定位基因组上的读取,私有云从这些位置的种子以找到正确的配对来扩展种子,该方法有效地对抗已知的推断攻击,并且也容易扩展到数百万的读取.

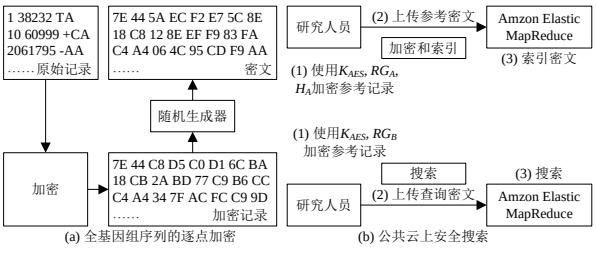

图 10　基因组特征搜索密码协议

（6）基因组数据法律与取证

DNA 测序技术和人类遗传学的进步导致廉价基因检测的出现,尤其是针对某些疾病的个体易感性检测,虽然这类信息通常很有价值,但它的可用性引起人们对基因信息隐私的严重担忧,当遗传信息被收集到数据库中这些担忧会进一步加剧.为此,Bohannon 等[88]使用对称加密和访问控制实现法医 DNA 数据库搜索与匹配的隐私保护,用于将未知的犯罪者与潜在的嫌疑犯进行匹配,该方法表明如何执行法医 DNA 数据库,以便只有合法的查询是可行的,对于合法的法医查询,属于目标个体的敏感信息已经以犯罪现场的血液或组织样本的形式提供给查询代理.无限访问数据库将无法提取任何个体的信息,除非已经知道该个体的必要遗传信息.通过开发通用的解决方案框架,并展示如何实现数据库处理某些情况下丢失或不正确的 DNA 检测,该框架适用于基于部分已知或部分正确密钥加密信息的一般问题,其安全性基于标准的密码假设.

（7）基因组数据直接面向消费者服务

在基因组数据直接面向消费者服务中,出于隐私和可用性的要求,在图 11 中 Naveed 等[89]提出受控函数加密（Controlled Functional Encryption）,使用从权威机构获得的密钥,仅获得加密数据的函数值,并且客户端每次计算密文的函数时都向权威机构发送新的密钥请求,以便于解密不同的密文,受控函数加密可用于实现个性化医学、患者相似性、亲子鉴定和血缘关系检测的隐私保护.

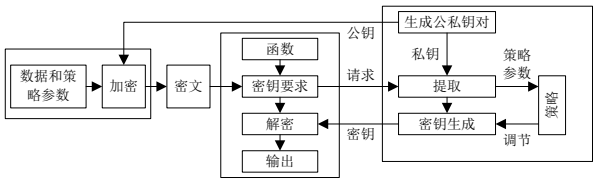

图 11　受控函数加密

在表 12 中,从特点和保护效果两方面对比分析对称密码、非对称密码和其他密码学方法,并从面向场景和面向场景的保护效果两方面比较分析基于对称加密和非对称加密的基因组数据安全和隐私保护.相比于非对称加密,对称加密的加解密效率高.但是在非安全信道中,不能保证对称加密的密钥交换的安全性.所以在基因组数据的实际应用中,将两者混合使用,非对称加密用于共同密钥协商,而对称加密用于加密基因型数据.此外,将对称加密或非对称加密与其他密码技术相结合可以实现基因组数据的安全存储、隐私查询等功能.结合认证、秘密共享、密钥协商、完善保密、数字签名和 Merkle 哈希树等密码技术可以实现基因组数据长期存储的永久安全.带密钥的哈希可用于保证基因组数据的完整性和进行身份验证,因此可以实现跨样本检测重叠个体和公共云上的人类基因组序列映射的安全和隐私保护.受控函数加密保证基因组数据的隐私和可用性,可用于实现基因检测的安全和隐私保护.

表 12 基于对称加密和非对称加密的基因组数据安全和隐私保护

方法	特点	保护效果	面向场景	面向场景的保护效果
对称加密	发送方和接收方共同协商密钥来加密和解密基因型数据,如果一方的密钥泄露则加密基因型数据不安全,并且在每对用户之间共同协商唯一密钥,使收发双方的密钥数量巨大,密钥管理成为两方的负担,而且陌生用户之间不便进行保密通信,还无法实现抗抵赖的数字签名,不过对称加密计算量小通常比非对称加密快.	当分组长度足够大、密钥量足够大和密码变换足够复杂时,使用对称加密可以实现基因型数据的安全和隐私保护.	基因组数据存储	文献[75]为基因组数据在存储和传输过程中提供端到端的安全,同时提供随机访问模式.
				文献[77]基于安全加密硬件安全地存储、共享和查询临床基因组学数据,消除对多个第三方的需求.
			基因组数据共享	文献[82]利用每个个体的基因型数据来加密全基因组序列,确保序列在整个传输过程中加密.
			Iceland 基因研究	文献[84]通过直接与 Iceland 数据保护委员会合作开发第三方加密系统,将其合并到样本收集和存储软件中,最大限度地减少不便,增强安全性.
			基于序列关联研究的元分析	文献[85]解决联合成员之间稀有测序等位基因频率计数的难题,通过各方加密基因和变异的身份,在不公开等位基因特性和计数的情况下,对测序数据进行元分析从而保护样本的特性.
			公共云上的基因组数据计算	文献[86]提出站点加密方法来加密整个人类基因组序列,可以在公共云上进行基因组特征的安全搜索.
			法医 DNA 数据库	文献[88]提出基于对称密码技术的法医 DNA 数据库信息隐私保护方法,允许恢复个人的身份和犯罪记录,但只有当基因型出现在一定数量的位点上的嫌疑人的 DNA 是确定的.
对称加密	同上	同上	基因组数据存储	文献[76]提出快速安全的基因组数据加密工具 Cryfa,以维护传输和存储基因组数据的安全性.
Shuffling 机制	Shuffling 机制是 ESA 模型的核心思想,Encoder随机编码基因型数据,可信 Shuffler 以随机顺序的方式输出随机编码的基因型数据,Analyzer 分析 Shuffler 输出的基因型数据.	实现 Shuffling 的 Mixnet 方法是可扩展和稳健的隐私保护协议,可用来隐藏基因型数据的来源.		
对称加密	同上	同上	基因组数据存储	文献[78]提出用于存储、检索和处理配对的原始基因组数据的隐私保护系统.
保序加密	确定性对称加密方案,保持基因序列的数字顺序.	用来加密核苷酸的位置,除了顺序外没有泄露基因型值的附加信息.		文献[79]提出节省空间、保护隐私和有效的基因组数据存储方案能够实现压缩、安全和选择性检索.
非对称加密	允许不同患者或医疗单位之间的认证以及基因组数据的安全共享,不过针对大规模基因型数据远比对称密码计算量大.	密钥不能太短也不能太长,使用基于数学困难假设的非对称密码加密基因型数据,可以实现基因型数据的安全和隐私保护.	基因组数据共享	文献[83]基于隐私信息检索提出 AnoniMME 框架支持匿名查询,同时支持 MME 的功能.
隐私信息检索	用户向服务器提交查询请求,在查询过程中服务器不知道用户具体查询信息及检索出的数据项.	研究人员向服务器查询与疾病相关的基因变异,利用隐私信息检索可以避免查询信息及查询结果的隐私泄露.		
承诺方案	承诺方案有两个基本性质,隐藏性（Hiding）指承诺值不会泄露任何关于基因型数据 x 的信息,而绑定性（Binding）是指任何恶意的承诺方都不能将承诺打开为不是基因型数据 x 的消息且验证通过,即接收方可以确信基因型数据 x 是和该承诺对应的消息.	承诺阶段,承诺方选择基因型数据 x ,以密文的形式发送给接收方,意味着自己不会更改 x .打开阶段,承诺方公开消息 x 与相当于密钥的盲化因子,接收方以此来验证其与承诺阶段所接收的消息是否一致.	基因组数据长期存储	文献[80]提出基因组数据长期安全存储解决方案,可以在不确定的时间内同时保护基因组数据的完整性、真实性和机密性.
先应秘密共享	根据秘密共享（同表 13）的方法将基因组数据分解成秘密份额,然后定期对份额更新.	只要保证给定时间段内获得的份额数不超过预定门限数,就可在更新阶段恢复被攻破的份额,重建系统安全,新旧份额不存在任何相关性,攻击者无法根据旧份额推断出现有份额的任何信息.		
量子密钥分发	通信的双方能够产生并共享随机的、安全的密钥来加密和解密消息,利用量子力学特性来保证通信安全性.	如果有第三方试图窃听密码,必须用某种方式测量它,任何对量子系统的测量都会对系统产生干扰,而这些测量就会带来可察觉的异常,通过量子叠加态或量子纠缠态来传输信息,通信系统便可以检测是否存在窃听.		
一次一密	正因每次加密时密钥都要变化,使得密钥的传递和分发变得困难.	如果密钥流完全随机产生且长度至少和基因序列长度相同,则可实现绝对安全的一次一密.		
数字签名	保证基因型数据的安全,实现	验证基因型数据的发送者是真实	基因组数	文献[81]提出基因组数据的长期保护方案,该方案

	发送者和接收者的身份认证并保持匿名性,确保基因型数据的完整性,满足签发数字签名的不可否认性,以及数字签名的可验证性.	的,而不是冒充的,同时验证基因型数据的完整性,保证基因型数据在传送和存储过程中没被篡改、重放和延迟等.	据长期存储	使用无条件隐藏承诺、Merkle 哈希树和数字签名来保护数据的完整性,同时保持机密性.
承诺方案	同上	同上		
Merkle 哈希树	相比于对多个基因型数据做一次数字 Hash 签名,Merkle 哈希树结构具有一次签名大量认证的优点.	基于 Merkle 哈希树的数字签名方案在安全性上仅仅依赖于哈希函数的安全性,且不需要太多的理论假设,这使得基于 Merkle 哈希树的数字签名更加安全、实用.		
带密钥的哈希	使用哈希函数,同时结合加密密钥,通过对不同输入长度的基因型数据进行哈希计算,得到固定长度的输出,而且是单向的.	保证基因型数据的完整性,同时用于发送基因型数据的身份验证.	跨样本检测重叠个体	文献[87]提出 Gencrypt 用于保护和比较个体水平的数据,以识别不同基因型数据集中重叠的个体.
			公共云上的人类基因组序列映射	文献[60]提出在混合云上实现安全且可扩展的读取映射的方法.
受控函数加密	解决隐私和可用性的矛盾,且只依赖于成熟和高效的加密工具,受控函数加密对于任意函数是有效的.	半诚实中央机构不了解基因型数据的任何信息,数据所有者明确指定的策略参数除外,数据所有者也不了解关于如何使用其基因型数据的任何信息,除非他们的策略将被强制执行.	个性化医学、患者相似性检测、亲子鉴定和血缘关系检测	文献[89]将受控函数加密用于个性化医学、患者相似性检测、亲子鉴定和血缘关系检测应用中基因组数据的隐私保护.

4.2.2 基于安全多方计算的基因隐私保护

（1）安全多方计算

安全多方计算允许两方或更多参与方在各自的秘密输入上进行联合计算函数,而不泄露各自的敏感信息[64].例如,在安全两方计算中,一方持有输入 x,另一方持有输入 y,两方都希望计算公开的函数 $z=f(x,y)$,而不向双方透露输出 z 之外的任何信息.此外,安全多方计算框架还涉及混淆电路（Garbled Circuit）、同态加密（Homomorphic Encryption）、秘密共享[90]、不经意传输协议（Oblivious Transfer Protocol）等多种密码技术.秘密共享是在一组参与方中分配秘密的方法,每个参与方被分配秘密的份额,只有当足够数量、不同类型的份额组合时,才能重构秘密,而且单个份额本身没有任何用处,(t,n)-秘密共享方案要求在 n 份份额中至少需要 t 份份额组合才能重构出秘密.例如,使用秘密共享对获取的基因组数据进行安全存储,判定案例和对照组,并进行安全统计检测,实现安全 GWAS.

（2）不经意传输协议

不经意传输协议是发送方将潜在多条消息中一条消息发送给接收方,但是发送方不知道已经发送哪一条消息,而且接收方除了接收到对应的那一条消息外不了解发送方的其他消息[61].例如,如图 12 所示的 $(1,n)$ 不经意传输协议,发送方输入消息 $(m_1,m_2,\ldots,m_n)$ 到可信第三方（Trusted Third Party,TTP）.接收方输入索引 $i(1\le i\le n)$ 到 TTP.接收方从 TTP 获得消息 m_i,但不了解发送方的任何其他信息,发送方也不知道接收方接收到哪一条消息.

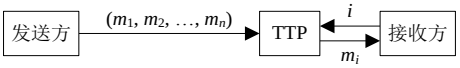

图 12 $(1,n)$不经意传输协议

（3）基因组数据研究与分析

虽然 DNA 序列可以帮助诊断遗传性疾病,并用于个性化治疗,但会对患者的隐私造成重大威胁.在基因组数据分析中,为了对加密的基因组数据进行计算,同时保持基因组所有者的完全控制,Deuber 等[91]基于支持任何多项式计算函数的混淆电路提出存储在云中的加密数据计算系统 METIS,该系统将可接受计算的决定权交给数据持有者,并且该系统的数据持有者不受计算过载的影响,其通信复杂度与输入数据的大小无关,仅与电路输出的大小成线性关系.联邦基因组数据分析为有效的 GWAS 提供跨机构协作的希望,但因数据是跨机构交换的,引起患者对隐私和医疗信息保密性的担心,Constable 等[92]基于混淆电路提出联合基因组数据集的 GWAS 框架,该框架允许分布式系统中的两方相互执行安全的 GWAS 计算,但不公开各自的隐私数据,并将其用于等位基因频率计数和 χ^2 统计计算,该框架在实现高效、安全的跨机构 GWAS 计算方面具有一定的应用前景.用于基因组分析的隐私保护数据挖掘近年来得到广泛的研究,Fisher 精确检验是 GWAS 中统计假设检验的重要方法,Hamada 等[93]基于安全多方计算构造用于 GWAS 的 Fisher 精确检验的高效隐私保护算法,所提出的安全 Fisher 检验算法的通信复杂度为 $O(N^{1.7})$,其中 N 是样本大小.基因与复杂疾病关联研究需要对大量基因组数据进行有针对性的研究,而大规模的基因组数据聚集引起许多隐私问题,为此 Kamm 等[90]提出使得敏感数据在多个独立实体之间秘密共享的数据采集系统,如图 13 的安全 GWAS 工作流程,在不侵犯单个捐赠者的隐私和不向第三方泄露数据的情况下进行 GWAS 分析.安全 GWAS 包括三个主要阶段,数据采集、案例对照组的形成和统计检验.数据以安全编码的方式收

集和存储,安全编码案例和对照组信息并进行安全存储,以便于进行统计分析.图 13（a）描述用于基因型和表型数据的安全存储的两种可供选择的方案,场景 1 描述由实验室将基因型数据进行安全存储,而供体本身输入表型数据的情况.场景 2 描述不同的基因库发送所选择的基因型和表型数据进行存储,从而可以对更多的数据进行联合分析.图 13（b）描述如何形成案例和对照组,在最简单的情况下,研究可以不受限制地访问表型数据,从而可以形成案例组和对照组.在更复杂的环境中,研究人员没有权利访问表型数据,并且主机必须使用安全多方计算来构建案例组和对照组.最后,执行统计检验所需的时间取决于计算等位基因频率和评估检验统计.

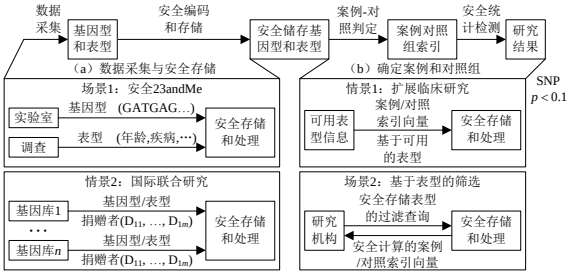

图 13 安全 GWAS

Bogdanov 等[94]通过正确处理群体分层（Population Stratification）和患者群体的固有遗传差异,提高全基因组关联研究的质量,该方法使用主成分分析来降低基因组数据的维数,以便在感兴趣的性状和基因组的某些位置之间获得较少的虚假相关性,虽然这种方法在实际的基因组分析中很常见,但它并没有在隐私保护的环境中使用,因此利用秘密共享来进行主成分分析,以此实现基因组数据的隐私保护.对基因数据隐私的担忧可能会阻止个体将其基因组贡献给科学研究,并可能阻止研究人员与科学界共享数据,因此 Cho 等[95]基于秘密共享提出大规模安全 GWAS 的方案,该方案有助于成千上万的个体中进行质量控制和群体分层校正,同时保持潜在基因型和表型的机密性,该方案有助于将数据提供给科学界,并有可能实现安全的基因组众包,允许个体在不泄露隐私的情况下为研究贡献其基因组数据.共享基因组数据对于支持 GWAS 等科学研究至关重要,然而研究中个体参与者的隐私可能会受到损害,导致严重的担忧和后果,例如对数据的访问受到过度限制,因此 Xie 等[96]基于秘密共享提出安全元分析协议实现大联盟遗传关联研究,在不泄露个体参与者信息和站点级别的关联摘要统计下有效而准确地分析各个子研究站点之间的遗传关联.除了人类DNA外,科学家们现正在研究人体内微生物的 DNA.个体的微生物 DNA 集合与个体的 DNA 一致,并且可以用于将真实世界的身份与研究数据集中的敏感属性关联起来.但是,目前DNA隐私保护分析工具不能满足微生物测序研究的要求.为了解决围绕微生物测序的隐私问题,Wagner 等[64]基于混淆电路实现元基因组分析的隐私保护,允许对组合数据进行比较分析,而不泄露任何单个样本的特征计数.

（4）基因组数据医疗服务

在相似患者查询中,持有患者基因组数据的医生可能会试图找到其他拥有相似基因组数据的患者,并利用这些患者的数据帮助诊断和找到有效的治疗方法.然而,在相似患者查询中隐私泄露的影响是相当大的,因此 Asharov 等[97]使用安全多方计算以隐私保护的方式进行相似患者查询,提出相似患者查询近似计算的高效安全计算方法.编辑距离是相似患者查询中重要和经常使用的度量,然而由于个体基因组数据的隐私泄露问题,许多基因组编辑距离的新用途的范围和规模都受到很大的限制,如图14所示Wang等[62]提出高效率和高精度的隐私编辑距离协议,该协议是基因组编辑距离近似算法和隐私集合差大小协议的组合,基于隐私编辑距离实现安全相似患者查询,并提出全基因组的隐私保护相似性查询系统,能够支持搜索大规模的、分布式的基因组数据库.外包基因组数据帮助数据所有者消除本地存储管理问题,为保障数据的私隐和安全,数据拥有者必须在进行外包前将敏感数据加密,但由于基因组数据量巨大,执行安全、高效的查询是具有挑战性的任务,因此 Mahdi 等[98]提出基于前缀树的索引算法来支持相似的患者查询,通过 AES 和混淆电路来保证数据隐私、查询隐私和输出隐私,整体计算是可扩展的,并且对现实生活中的生物医学应用是足够有效的.计算生物学中的许多基本任务涉及对单个 DNA 和蛋白质序列的操作,即使对这些序列匿名,也容易受到重识别攻击,并且可能泄露高度敏感的个人信息,为此 Jha 等[61]提出相对有效的基因组数据的隐私保护计算,用于两个序列之间的编辑距离和 Smith-Waterman 相似性得分的隐私计算,而且对于分布式数据集上的许多动态编程算法可以实现高效的隐私保护.因为同源基因来自共同祖先 DNA 序列,所以同源基因搜索会导致个体及祖先的隐私泄露,为此Wang等[99]使用安全计算提出同源基因搜索的隐私保护方法,该方法利用人类基因组的关键特性,即绝大部分基因组在人类之间共享,而相对较少的基因组是敏感的,该框架将敏感数据部分分配给数据提供者,将公共数据部分分配给数据使用者,该框架在数据提供者和数据使用者之间分配计算任务,并让数据提供者基于安全多方计算处理与敏感数据相关的小部分任务,当数据使用者没有

敏感数据输入时避免昂贵的安全多方计算,并显著降低计算开销,可以极大地促进大型生物计算问题的安全多方计算.数以千计的单基因疾病已经产生明确的基因诊断和潜在的基因治疗目标,在半诚实模型下,Jahadeesh 等[100]基于安全多方计算识别因果变异,并发现先前未识别的疾病基因和变异,同时保护所有参与者及其最敏感的基因组隐私信息.

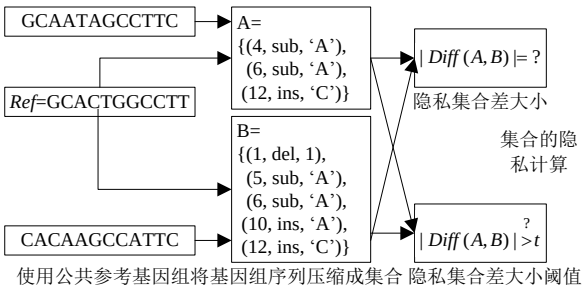

图 14 人类基因组编辑距离的安全协议

（5）基因组数据法律与取证

在隐私 DNA 匹配问题的驱动下,Katz 和 Malk[101]使用安全多方计算实现 DNA 匹配的隐私保护,一方 P_1 持有文本 T,另一方 P_2 持有模式 p 和一些附加信息 y,对于所有位置 j,P_2 想要了解 $\{f(T,j,y)\}$,其中 p 为 T 中的子串,旨在针对恶意的 P_2 提供完全安全的协议,该协议还针对恶意的 P_1 保护隐私.由于基因组数据的处理是高度敏感的,因此需要对基因组数据的隐私增强技术进行研究,如图 15 所示 Karvelas 等[102]提出灵活的机制用于整个基因组序列的隐私处理,该机制可以支持任何查询,其基本思想是将 DNA 存储在几个小的加密块中,使用 ORAM（Oblivious Random Access Machine）机制不经意的访问所需的块,最后运行安全的两方协议以在检索到的加密块上隐私地计算所需的函数,将所有敏感信息隐藏起来,只向合法方透露最终结果.在相似序列查询中,因为基因组数据唯一地识别个体,包含个体患特殊疾病风险的敏感数据,甚至包括其家庭成员的敏感信息,这引起严重的隐私问题,为此 Schneider 和 Tkachenko[63]基于安全多方计算提出用于相似序列查询的高效隐私保护协议,可用于外包基因组数据库中查找基因相似的个体.在隐私保护外包计算中,为了不向计算能力正在被使用的远程代理透露数据或计算结果,Atallah 和 Li[103]针对广泛应用的序列比较问题基于安全计算研究安全外包机制,并为用户提供将序列比较安全外包给远程代理的有效协议,使得代理对用户的两个隐私序列或比较结果一无所知.

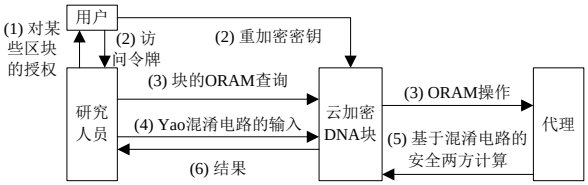

图 15 基因组数据隐私保护计算体系结构

（6）基因组数据直接面向消费者服务

在基因组数据直接面向消费者服务中,受医学和社会应用两方面的激励,其目的在于便携式和智能手机环境下有效地隐私计算基因检测的可行性,于是 Cristofaro 等[59]使用安全计算汉明距离和隐私集合交运算设计基因组工具包 GenoDroid,用于 Android 平台上实现基因检测的隐私保护.随着人类基因组测序技术的发展,生物和医学研究的步伐加快,与健康相关的广泛应用和服务变得越来越普遍和廉价,然而因基因组包含个人极其敏感的信息,数字化的基因组序列带来严重的隐私问题,为此 Shen 等[104]基于不经意传输研究高效、隐私的人类基因组集合交叉协议,该协议用于安全地进行亲子鉴定和祖先鉴定,而不泄露任何额外的个人基因组信息.

在表 13 中,对目前基于安全多方计算的基因组数据安全和隐私保护的研究工作进行比较分析.然而,安全多方计算具有计算和通信开销的瓶颈,在基因组数据研究与分析、医疗服务、法律与取证、直接面向消费者服务的安全和隐私保护中,研究高效的安全多方计算协议和减少通信轮数,并且实现基因组数据的安全和隐私保护是急需解决的问题.

表 13 基于安全多方计算的基因组数据安全和隐私保护

方法	特点	保护效果	面向场景	面向场景的保护效果
安全多方计算	避免各自敏感基因型数据泄露进行联合计算,不过联合计算很容易成为性能瓶颈,特别是针对现实中数百万个核苷酸的基因组计算.	允许各方基于各自的输入基因型安全地计算联合函数,而不向对方透露输入基因型,任何一方只能获得联合计算的结果,但不知道中间值.	GWAS	文献[93]提出有效的隐私保护算法,用于 GWAS 的 Fisher 精确检验.
			相似患者查询	文献[62]提出高效率和高精度的隐私编辑距离协议,以此实现和评估全基因组安全相似患者查询系统.
				文献[97]提出用于查询相似患者基因组数据的隐私保护协议.
			相似序列比较	文献[61]提出用于计算基因组序列之间的编辑距离和 Smith Waterman 相似度得分的隐私保护协议.
			基因组数据计算	文献[99]根据基因组数据的敏感性水平来划分基因组,并使用安全多方计算对敏感数据进行隐私保护计算.
			基因组诊断	文献[100]使用安全多方计算实现基因组诊断,并保护参与者的隐私.
			DNA 匹配	文献[101]设计用于安全文本处理的有效协议,用于隐私 DNA 匹配.

			序列查询	文献[63]提出相似序列查询的高效隐私保护协议,可用于在外包的基因组数据库中查找基因相似的个体.
			序列比较	文献[103]提出序列比较的安全外包计算协议,使用户能够安全地将序列比较外包给两个远程代理,代理不会获得任何信息.
			基因检测	文献[59]在现代便携和普及的智能手机环境中设计隐私敏捷计算基因检测工具 GenoDroid.
安全多方计算	同上	同上		
ORAM	ORAM 使得在查询过程中完全忽略了访问的基因组数据,仅知道访问的数据量而不知道基因组数据的位置.	ORAM 将基因组数据以加密的形式存储在服务器上,实现访问模式的隐私保护.	序列查询	文献[102]将 ORAM 技术与安全的两方计算方案相结合,以提供对外包的、全序列 DNA 的隐私保护计算,从而提供充分的查询灵活性.
混淆电路	允许分别持有输入基因型 x 和 y 的两方对函数 f 的任意布尔电路进行求值,除了输出计算结果 $f(x,y)$ 之外,而不泄露关于其输入基因型的任何信息.	只要对输入基因型进行适当的编码,就可以安全地计算任何函数,而不泄露输入基因型.	基因组数据的第三方计算	文献[91]提出 METIS 密码系统,允许对加密的基因组数据进行安全计算,其重要特点是数据所有者控制计算的函数类型.
			联邦基因组数据集的 GWAS 分析	文献[92]提出基于联邦基因组数据集的隐私保护 GWAS 框架,允许分布式系统中的双方相互执行安全的 GWAS 计算,不会将隐私数据暴露在外部.
			微生物测序研究	文献[64]实现隐私保护微生物分析,允许对组合数据进行比较分析,而不揭露任何单个样本的特征计数.
			相似患者查询	文献[98]提出基于加密数据的相似患者搜索的安全有效方法,数据处理和查询执行阶段都不会揭露任何敏感的基因组数据.
秘密共享	如果预先指定数量的参与方聚集各自的输入基因型,则允许多个参与方重构敏感基因型,否则不能重构敏感基因型.	个体可以自由地将其基因组数据贡献给计算方,而不允许任何参与方访问原始基因型数据.	大规模 GWAS	文献[90]提出安全 GWAS,在不侵犯捐赠者隐私和不向第三方泄露数据的情况下进行 GWAS 研究.
				文献[94]利用秘密共享解决主成分分析问题,通过正确处理种群分层提高隐私保护全基因组关联研究的质量.
				文献[95]提出安全的大规模全基因组分析的方案,该方案有助于进行质量控制和群体分层校正,同时保证潜在基因型和表型的机密性.
				文献[96]提出安全元分析协议,用于支持涉及隐私或机密性的不同数据站点之间的联合研究.
不经意传输	主要使用对称加密和少量公钥加密,允许接收方隐私地从发送方获得指定的基因型,接收方不了解其他基因型数据,发送方不知道接收方获得哪一个基因型.	发送方持有基因型序列数据 $(x_1,x_2,\ldots,x_n)$,接收方有索引 i,在协议执行结束后,发送方不了解接收方的索引 i,接收方接收到基因型 x_i,但不了解其他基因型数据.	亲子鉴定和祖先鉴定	文献[104]主要研究有效的隐私保护人类基因组集合交叉协议,便于亲子鉴定和祖先鉴定安全进行,而不泄露任何额外的个体基因组信息.

4.2.3 基于同态加密的基因隐私保护

（1）同态加密

同态加密可以对密文进行运算,使得先运算后加密和先加密后运算的结果相同.对于加法同态要求满足

$$Enc(m_1)\oplus Enc(m_2)=Enc(m_1+m_2)$$

Paillier 是加法同态密码系统.而对于乘法同态要求满足

$$Enc(m_1)\otimes Enc(m_2)=Enc(m_1*m_2)$$

RSA 和 ElGamal 是乘法同态密码系统.如果加密系统同时满足加法同态和乘法同态,那么该密码系统是全同态加密系统,Gentry 算法是全同态加密[105].例如,在半诚实模型下,任何一方都不偏离协议执行,但是尝试推断关于另一方输入的一些信息,基于同态加密可以实现 DNA 搜索、查询、比较和匹配的隐私保护.

（2）代理重加密

对于代理重加密（Proxy Re-Encryption）[66],如图16所示,发送方 A 欲与接收方 B 共享数据,A 使用公钥加密数据并将密文发送给云计算服务提供商,同时结合自己的私钥和 B 的公钥生成转换密钥发送给云计算服务提供商,云计算服务提供商使用转换密钥将密文转化为 B 的私钥能够解密的密文,其中云计算服务提供商只提供计算服务,无法获得明文.将密文发送给 B,解密后获得 A 想要秘密共享其的明文 m.例如,发送方向多个接收方共享数据,在不可信第三方存在的情况下,第三方可能滥用用户的数据,可以通过代理重加密将数据发送给第三方来解决此问题,也就是原本使用发送方的公钥加密后的密文,只有发送方的相应私钥才能解密,允许代理转化为接收方的私钥也能够解密,发送方希望其愿意共享秘密的多个接收方获得密文的明文消息.

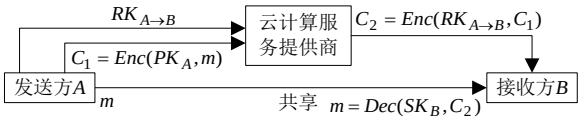

图 16 代理重加密

（3）基因组数据测序与存储

在基因组数据存储中,Teruya 等[72]指出目前的标准隐私保护密码协议可能不足以保护基因组数据隐私,这主要是由于基因组信息的典型特征,它是不变的,并且个体的基因组与个体后代的基因组相

关,因此 Teruya 等结合一次一密、消息认证码和同态加密提出具有永久安全性的密码协议来保护基因组隐私,该协议提供计算安全性和信息理论安全性的组合安全.基因组技术的快速发展导致基因组数据的指数增长,由于基因组数据的敏感性,研究机构需要考虑安全问题,提高基因组应用可扩展性和性能的方法以便能够处理大量数据和繁重的计算,于是 Kang 等[73]结合哈希、同态加密和保序加密提出完整的公共云基因组数据处理框架,该框架不仅可以保护基因组序列,而且可以保护公共云处理时的中间和最终计算结果,并且该框架将所有重量级计算委托给具有大规模存储和计算资源的公共云,隐私基础设施继续用于轻量级计算,包括明文数据,以及数据加密和解密,以便在发送到公共云进行分析之前保护隐私,使得该框架实现并行处理时具有较高的效率.基因组数据所有者可以将其数据库外包到集中的云服务器上,以方便访问数据库,然而数据所有者不愿意采用这种模式,因为将数据外包给不可信的云服务提供商（CSP）可能会导致数据泄露,为此 Ghasemi 等[106]提出将基因组数据外包到云的隐私保护模型,模型中基因组数据库的隐私性通过使用语义安全的同态加密方案对每个记录进行加密来保证,该模型在提供基因组数据库隐私保护的同时实现查询处理,通过在数据库中置换和添加假基因组记录来保证个体的隐私.

（4）基因组数据共享与聚集

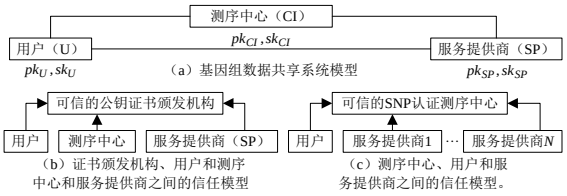

图 17 基因组数据共享系统模型与可信模型

个体与服务提供商共享其基因组数据或从其基因组数据中获得的结果,服务提供商希望确保接收到的基因组数据或结果事实上属于相应的个体且计算正确,个体希望提供数字许可以及指定是否允许服务提供商进一步共享其数据,如果未经其同意共享其数据,则个体希望确定对此泄漏负责的服务提供商.因同态签名具有不可伪造性和机密性的特点,聚合签名也可以实现不可伪造性,在图 17 中 Ayday 等[107]提出两种基于同态签名和聚合签名的方案来共享基因组数据和基因检测结果,利用所提出的方案服务提供商可以检查其从数据所有者接收的基因组数据的合法性,个体通过数字许可,可以确保服务提供商在未经其许可的情况下不会进一步共享其数据.所提出的方案将数据的合法性信息与个体许可和表型联系起来,因此为了验证数据,各方还需要使用数据拥有者的正确许可和表型.

（5）基因组数据研究与分析

目前,许多数据库都拥有大量的基因组数据,这些数据对于进行各种基因组研究的研究人员来说是非常宝贵的.然而,自愿提供基因组数据的患者有隐私泄露的风险,因此 Lauter 等[108]采用在 GWAS 中常用的基本基因组算法,例如 Pearson 检验、Cochran-Armitage 趋势检验等,并基于同态加密使它们在加密的基因型和表型数据上进行隐私计算,以此保护患者隐私.在 GWAS 的隐私保护研究中,逻辑回归是大多数 GWAS 中选择的方法,duVerle 等[109]使用同态加密提出基于采样的安全协议计算精确的统计量,它仅需要恒定数量的通信轮数和更少的计算量,目标是对患者数据在临床和基因组之间产生有用的统计检测,而不向另一方透露任何一方的信息或以任何方式将它们关联起来.GWAS 已被广泛应用于发现基因型与表型之间的关联研究,人类基因组数据包含有价值但高度敏感的信息,未保护地透露这些信息可能会危及个体隐私,因此保护人类基因组数据是非常重要的,精确逻辑回归是用于 GWAS 研究发现与疾病易感性相关罕见变异的惩罚可能性的偏差减少方法,Wang 等[110]基于同态加密提出 HEALER 框架,允许在云服务器中存储和计算加密的罕见疾病变异,以促进小样本安全罕见变异分析,旨在降低计算和存储成本,研究人员可以访问最终的同态加密精确逻辑回归模型评估结果 p 值,其中电路深度与数据大小的对数成比例.基因组数据的不断增加推动个性化治疗和精准医疗领域的大量研究,公共云服务提供灵活的方式来降低 GWAS 中的存储和计算开销,然而在云环境下共享敏感信息时,数据隐私受到广泛关注,Zhang 等[111]提出基于同态加密的安全外包 GWAS 框架 FORESEE,所提出的框架允许对加密数据进行安全划分,FORESEE 框架支持完全外包到云,并输出最终的加密 χ^2 统计.全基因组关联研究旨在发现与特定疾病相关的遗传变异,Lu 等[112]基于环 LWE 的同态加密提出 GWAS 安全外包计算方法,该方法的工作原理在于将基因型整数向量加密为单个密文,并且整数向量的标量积可以使用单个同态乘法求值.为了支持大规模的生物医学研究项目,组织需要在不侵犯数据对象隐私的情况下共享特定个体的基因组序列,于是 Kantarcioglu 等[113]提出使组织能够支持基因组数据挖掘而不公开原始基因组序列的密码框架,组织将同态加密的基因组序列记录到集中式存储库中,管理员可以在其中执行频率计数而不需要解密数据,该框架可以在生物医学环境中现有的信息和网络技术的基础上实现.

（6）基因组数据法律与取证

通过有限自动机的遗忘评估来处理容错 DNA 搜索的问题,其中客户端有DNA序列,服务提供商有与基因检测相对应的模式,通过将模式表示为有限自动机并在 DNA 序列上对其进行评估,实现容错搜索,其中必须保护模式和 DNA 序列的隐私.因此,Blanton 和 Aliasgari[114]提出将有限自动机评估安全外包给计算服务器的技术,因使用同态加密服务器不能获得任何信息,该技术适用于任何类型的有限自动机,其优化适合于 DNA 搜索.人类 DNA 序列提供丰富的信息,揭示各种疾病的易感性和亲子关系等,这些信息的广度和个性化本质突出隐私保护协议的必要性.因此,Troncoso-Pastoriza 等[115]结合同态加密和不经意传输提出适合于 DNA 查询的容错隐私保护字符串搜索协议,该协议检查一方所知的短模板是否存在于另一方拥有的 DNA 序列内,以解释可能的错误,并且不向一方泄露另一方的输入,每个查询在有限字母表上形成正则表达式,并作为自动机实现,复杂度在状态数和输入字母表的大小上是线性的.个体基因组具有固有的隐私风险,保护其隐私是重大的社会和技术挑战,考虑用户在存储大型基因组数据库的服务器上搜索遗传信息,例如等位基因,目的是接收与等位基因关联的信息,并且用户希望保护查询和结果的隐私,Shimizu 等[116]结合 Burrows-Wheeler 变换高效字符串数据结构与加性同态加密实现用户查询和查询结果的隐私保护,使用不经意传输技术基于加法同态加密来隐藏位置查询中的序列查询和感兴趣的基因组区域,使用 Burrows-Wheeler 转换使得序列数据在大型索引字典上的高效迭代查询操作中是可搜索的.基因组数据共享应该保证攻击者不能了解任何有关数据或各方对最终计算输出贡献的信息,因此 Aziz 等[117]提出实用的基因组数据安全共享和计算方案,采用 Paillier 密码体制和保序加密方法加密基因组数据并保持顺序,安全地执行计数查询和排序查询.将基因组序列或临床资料数据外包给第三方会导致参与者的隐私受到侵犯的潜在风险,为此 Hasan 等[118]提出在半诚实云服务器上实现基因组数据安全共享和计算的方法,共享数据的机密性是通过同态加密来保证的,同时基于混淆电路的整个基因组数据计算过程对于前沿的生物医学应用来说是高效的和可扩展的,该方法可以处理包含基因型和表型的生物医学数据,同时该方法通过索引树方法减少执行安全计数查询的计算开销.

（7）基因组数据直接面向消费者服务

在基因组数据直接面向消费者服务中,由于基因组信息的极度敏感性和唯一性,基因检测引发严重的隐私和伦理问题,Cristofaro 等[119]基于同态加密提出大小和位置隐藏隐私子字符串匹配协议,允许拥有数字化的基因组和拥有一组 DNA 标记的两方进行检测,只由前者获知结果,而任何一方都不了解其他信息,基因组所有者甚至不知道标记的大小或位置.因数据隐私和不能向医疗服务从业人员提供解释结果,McLarenet 等[120]使用同态加密和代理重加密对大量遗传标记进行加密,在隐私保护条件下推断患者祖先、计算单基因和多基因特征风险以及报告结果.对于亲子鉴定、个性化医学和遗传兼容性检测中的基因组隐私问题,Baldi 等[121]结合同态加密和隐私集合运算提出全序列人类基因组的高效安全检测方法.针对个性化医学场景,Djatmiko 等[122]提出使用基因组数据的线性组合来计算基因检测的安全评估算法,并以 Warfarin 剂量算法作为代表性例子,该方案结合部分同态 Paillier 加密和隐私信息检索以保护个体的隐私数据和检测细节.面对个性化医学检测中保护个体遗传数据的复杂挑战,在医疗中心是恶意但隐蔽（Malicious-but-Covert）的攻击模型下,在没有被抓到的情况下进行主动地欺骗,Barman 等[123]分析不同的隐私威胁,并基于同态加密提出新的实用解决方案,以防止恶意医疗中心试图主动推断患者的原始遗传信息的严重攻击.Ayday 等[67]提出利用患者基因组数据进行医学检测和个性化医学方法的隐私增强技术,在图 18 中通过利用同态加密和代理重加密提出疾病易感性检测的隐私保护框架.整个基因组测序是由测序中心（CI）完成的,在存储处理单元（SPU）中存储由患者的公钥加密的基因组数据.考虑存储处理单元是半诚实的,而恶意医疗单位可以被认为是入侵医疗中心系统的攻击者或不满员工访问医疗单位的数据库.所提出的模型允许存储处理单元或医疗单位处理用于医学检测的加密基因组数据和个性化医学方法,同时保护患者的基因组数据的隐私.在文献[67]的基础上,当加密 SNP 的位置及其内容时,Ayday 等[66]使用改进的 Paillier 密码系统和代理重加密提出疾病易感性检测的隐私保护协议.DNA 测序技术的进步使大规模基因检测越来越接近现实,然而遗传数据的敏感性表明需要小心保护患者的隐私,此外隐藏检测的细节是至关重要的,这常常构成制药公司的商业秘密,针对 Ayday 等[66]的体系

结构,存在泄露检测的 SNP 及其数量的问题,Danezis 和 Cristofaro[124]提出疾病易感性隐私保护协议计算患者的遗传标记 SNP 对疾病易感性的加权平均值,进而隐藏制药公司的商业机密的检测细节.专注于疾病风险检测,Ayday 等[125]提出利用同态加密和隐私保护整数比较来存储和处理基因组、临床和环境数据的隐私保护系统,并通过复杂度评价分析,表明系统的实用性.基因组研究领域的发展引起重要的隐私泄露问题,由于 DNA 信息在家庭成员之间是独特的、相互关联的,因此不能仅仅将其视为个人隐私问题,为此 Namazi 等[126]基于属性的同态加密提出隐私保护基因组易感性检测方法,将患者的基因组数据外包给不可信的平台,当医疗单位的属性满足定义的访问结构时,使得多个医疗单位只能访问患者医疗记录的授权部分,该方法使用同态加密处理遗传数据进而保护个体隐私,并向医疗服务提供者提供临床报告.

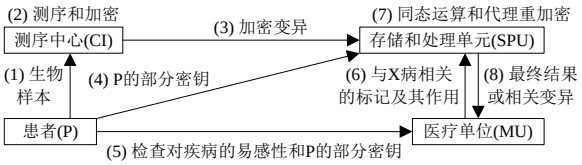

图 18 疾病易感性检测隐私保护框架

在表 14 中,同态加密用于基因组数据测序与存储、共享与聚集、研究与分析、法律与取证、直接面向消费者服务的安全和隐私保护,但因同态加密的计算复杂度高不适合于基因组数据医疗服务的安全和隐私保护.面向大规模、高维的基因组数据及其应用中的安全和隐私保护,设计高效的同态加密方法是关键的挑战.

表 14 基于同态加密的基因组数据安全和隐私保护

方法		特点	保护效果	面向场景	面向场景的保护效果
同态加密		允许对加密基因型数据进行计算,同态加密基因型数据计算开销大.	对基因型数据的运算转化为对加密基因型数据的运算,无需解密.	外包基因组数据库的查询处理	文献[106]提出将基因组数据外包到云上的隐私保护模型,在提供基因组数据库隐私保护的同时实现查询处理,并保证个体隐私.
				GWAS	文献[108]提出遗传关联研究的密码解决方案,对数据库中所有基因组数据进行同态加密,以此对加密数据进行有意义的计算和保护患者隐私.
					文献[109]基于抽样的方法在隐私保护环境中对 GWAS 的逻辑回归模型执行精确的统计检验.
					文献[110]提出用于 GWAS 罕见疾病变异分析的同态加密数据的精确逻辑回归参数 p -值估计框架 HEALER,且支持安全外包降低在不可信的云环境中分析敏感基因组数据的风险.
				GWAS 外包计算	文献[111]提出在公共云中实现安全且完全外包的 χ^2 统计计算的 FORESEE 框架.
					文献[112]提出安全外包 GWAS 计算,在保证基因型数据安全性的前提下有效地进行假设检验.
				基因组数据挖掘	文献[113]提出在加密的环境中存储和查询特定个体的基因组序列数据的密码框架.
				DNA 搜索	文献[114]通过有限自动机对 DNA 搜索进行外包计算,计算服务器不能获得任何信息.
				基因组数据查询	文献[116]结合有效的递归搜索数据结构和递归不经意传输提出以隐私保护方式搜索基因组序列.
				基因检测	文献[119]提出大小和位置隐藏的隐私子串匹配协议用于类似于个性化医学的隐私保护基因检测.
					文献[121]对亲子鉴定、个性化医疗和基因兼容性检测三个重要应用解决基因组隐私问题.
				个性化医学检测	文献[122]提出用于个性化医学的安全计算协议保护个体的敏感基因组数据和基因检测细节.
					文献[123]基于同态加密应对恶意医疗中心试图主动推断患者原始基因信息的严重攻击.
同态加密	一次一密	同表 12	同表 12	基因组数据长期存储	文献[72]提出具有永久安全性的密码协议来保护基因组隐私,并提供计算和信息论的组合安全.
	消息认证码	可认为是固定长度的、带密钥的 Hash 函数.因收发双方共享密钥则密钥分发困难,而且不能进行第三方证明、不能防止否认.	用于确认基因型数据的完整性,以及作为基因型数据来源的身份验证.		
	不经意传输	同表 13	同表 13		
	带密钥的哈希	同表 12	同表 12	公共云上的基因组数据处理	文献[73]提出完整的基因组数据处理安全框架,利用公共云资源来解决基因组数据的指数增长问题.
	保序加密	同表 12	同表 12		
	不经意传输	同表 13	同表 13	DNA 搜索	文献[115]提出适用于运行隐私 DNA 查询的具有容

					错能力的隐私保护字符串搜索协议.
					文献[116]结合有效的递归搜索数据结构和递归不经意传输提出以隐私保护方式搜索基因组序列.
	保序加密	同表 12	同表 12	基因组数据共享和计算	文献[117]提出安全、高效的基因组数据共享和计算的方法.
	混淆电路	同表 13	同表 13	基因组数据外包	文献[118]提出安全、高效的基因组数据外包方法,并对其执行计数查询.
	代理重加密	云计算服务提供商将 *A* 用公钥加密基因型数据的密文转换为 *B* 的公钥加密的密文.	云计算服务提供商不能获得基因型数据的明文信息,没有数据泄漏风险,也不存在安全隐患.	疾病易感性检测	文献[66]提出使用患者基因组数据的医学检测和个性化医学检测的隐私增强技术.
					文献[67]提出在医学检测中利用基因组数据的隐私保护方案,患者的加密基因组数据存储在存储和处理单元中,并使用同态加密进行医学检测.
					文献[124]隐私地评估患者对特定疾病的易感性,除了检测结果外没有透露任何其他信息.
					文献[125]使用基因组数据以及临床和环境数据进行特定的疾病风险检测,同时保护个体隐私.
				基因检测	文献[120]在隐私保护框架下处理基因数据,向医疗服务提供者提交描述潜在可操作结果的临床报告.
	同态签名	类似于对加密基因型数据进行计算的同态加密方案,能够对签名基因型数据进行计算,确保基因型数据的不可伪造性和机密性.	使个体能够诚实地共享任何经过认证的基因型数据子集或检测结果,而无需与权威机构交互,还保证在共享检测结果时个体不会泄露不必要的信息.	个体与服务提供商共享基因组数据和基因检测结果	文献[107]提出两种基于同态签名和聚合签名的方案来共享基因组数据和基因检测结果,使得服务提供商可以检查其从数据所有者接收的基因组数据的合法性,个体通过数字许可确保服务提供商在未经其许可的情况下不会进一步共享其数据.
	聚合签名	每个用户对基因型数据进行签名,并共享加密基因型数据和签名,任意用户对收到的加密基因型数据和签名进行聚合,并验证聚合签名是否有效,确保基因型数据的不可伪造性.	有效防止服务提供商非法或未经授权共享基因组数据,接收数据的服务提供商未经数据所有者同意共享基因组数据,或者未经数据所有者同意服务提供商泄漏其基因组数据,因此要求该服务提供商对泄漏负责.		
	基于属性的同态加密	在基于密文策略属性的加密方案中, *A* 在加密其基因型数据时描述策略,并且可信第三方为属性提供解密密钥并在各方之间分发它们,如果 *B* 的属性满足 *A* 定义的策略, *B* 就可以解密这个密文.	对加密基因型数据进行计算,还能够实现细粒度的访问控制.	疾病易感性检测	文献[126]提出有效且实用的隐私保护易感性检测方法,该方法描述将患者基因组数据委托给不可信服务器的存储和处理方式,服务器在不损害个体隐私的情况下对加密的基因组数据执行所需的检测,只有具有授权属性的参与方可以获得检测的结果.

4.2.4 基于模糊加密的基因隐私保护

（1）模糊加密

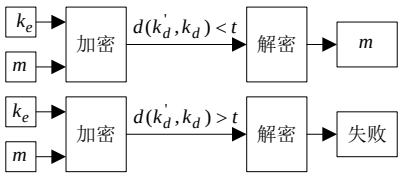

图 19 模糊加密

如图 19 所示,模糊加密（Fuzzy Encryption）[127]也包含公私钥对 (k_e,k_d),其中公钥 k_e 是公开的,而私钥 k_d 是秘密的,不同于非对称加密解密密钥与 k_d 相同才能解密,而是当解密密钥 $k_d^{'}$ 和 k_d 之间的汉明距离 $d(k_d^{'},k_d)$ 小于预定义的阈值 t 时才能解密.如果解密密钥 $k_d^{'}$ 和 k_d 之间的汉明距离为 t,则解密成功并终止,否则解密失败.例如,个体可以通过下载所有其他个体的可用公钥来检测遗传近亲,并将公钥与其私钥进行比较,如果两个个体有基因关联,可以在不泄露任何信息的情况下检测到遗传近亲,而与他人无遗传近亲的个体不会获得任何信息.

（2）基因组数据法律与取证

高通量测序技术已经影响到遗传研究的许多领域,其中一个领域是从遗传数据中识别亲属,鉴定遗传亲缘关系的标准方法是收集所有个体的基因组数据并将其存储在数据库中,并对每对个体进行比较,以确定遗传亲缘关系,但在确定亲属身份时产生固有的隐私问题.为此,He 等[128]提出使个体能够在不泄露任何有关其基因组信息的情况下识别基因亲属的方法,该方法使用个体基因组信息作为模糊加密密钥,并对个体基因组数据加密和公开发布密文,以此其他个体通过使用自己的基因组信息解密密文,如果两个个体有亲缘关系,那么基因组足够接近,因此解密将检测他们是相关的,否则不相关,在加密时指定基因组相似性的阈值,并调整到方案可以识别的亲属级别,例如,将阈值设置为表兄妹之间

的相似程度,则只有表兄妹或近亲的个体才能识别其亲属,而较远的亲属则不会彼此识别,在 HapMap 和 1000 Genomes 基因组数据库中模拟表明,该方法可以恢复一代和二代的遗传关系,并且可以在保护隐私的同时识别出与第三代近亲一样远的关系.遗传亲缘关系鉴定的主要缺点是需要与可信的第三方共享遗传数据,以执行遗传亲属关系检测,而且文献[128]的局限性是只有先前已知的常见变异才能在该方法中使用,在识别亲属中常见的变异不如通常只与近亲共享的罕见变异信息量大,因此 Hormozdiari 等[127]将每个个体的两个单倍型编码成一个集合,使得个体之间的对称差对应于两个个体之间的遗传相似性,类似于文献[128]从测序数据中检测出遗传亲缘关系,而不暴露任何利用普通和罕见变异的基因组信息,并且通过模拟证明可以检测到多达 5 代的近亲,还从 1000 Genomes 基因组数据库中发现包含隐秘关系的两个群体,并且可以检测出这些个体,使用构建的基因组编码可以用来检测未来的亲属.

模糊加密不同于非对称加密,解密加密基因型数据的私钥必须接近原始私钥,而不必需要相同的密钥.模糊加密可以用来实现固定大小集合的安全比较,这是进行隐私保护的亲属识别的基础.例如,面向遗传亲缘关系的检测,文献[128]使个体能够在不泄露任何有关其基因组信息的情况下发现基因亲属.在面向利用罕见变异检测遗传亲缘关系中,文献[127]从测序数据中检测出遗传亲缘关系,而不暴露任何利用普通和稀有变异的基因组信息.

4.2.5 基于蜂蜜加密的基因隐私保护

（1）蜂蜜加密

对于蜂蜜加密（Honey Encryption）[129],假设根据分布 π_m 从消息空间 M 中随机采样消息 m,并使用密钥 $k \in K$ 进行加密,以此产生密文 $c \in C$.使用不正确密钥 $k'(k' \neq k)$ 解密密文 c 也可以从分布 π_m 产生不正确的消息 m'.在传统密码学中,当使用错误的密钥解密密文时,通常会获得无效的消息,使得攻击者可以通过暴力攻击轻易地排除错误的密钥.然而,在蜂蜜加密中,使用错误的密钥 k' 解密密文 c 输出服从相同分布 π_m 的随机采样消息 m',所以攻击者不能通过暴力攻击获得明文消息.例如,正因蜂蜜加密为加密数据提供信息论上的安全保证,结合基因组数据的性质,将蜂蜜加密用于基因组数据的安全存储,使其对潜在的数据泄露和无界攻击者具有鲁棒性,并具有有效性.

（2）基因组数据测序与存储

基因组数据的安全存储具有重要的意义,在应用遗传数据时普遍使用口令来生成加密密钥会带来特别严重的问题,弱口令可能在短期内危害遗传数据,但考虑到遗传数据长期的使用寿命,即使使用强口令和常规加密也可能导致泄露.因此,Huang 等[129]基于蜂蜜加密提出 GenoGuard 为今天和长期的基因组数据提供强有力的保护,可以为加密基因组数据提供信息论的机密性保证,基于三叉树高效地编码基因重组和变异敏感的基因组序列,从而捕获基因组数据的高度不均匀概率分布和特殊结构,在患者口令下加密编码的基因组数据,并存储在基因库中,以此具有无限计算能力攻击者在不正确的密钥下对基因密文进行解密会产生基因组序列,该序列在统计上也似乎是可信的,还考虑到攻击者根据个体身体特征的附加信息,GenoGuard 可防止利用基因型-表型关联来确定个体的真实基因组,GenoGuard 高效的用于服务提供商提供直接面向消费者服务,以安全地存储用户的基因组,医疗单位也可以使用它来安全地存储患者的基因组,并在以后检索中供临床使用.同态加密方案可以应用于来自非常受限的概率分布集合的消息,而 GenoGuard 解决同态加密技术应用于遗传数据序列的高度非均匀概率分布的问题.可以证明在任何密钥下的解密都会产生可信的基因组序列,并且 GenoGuard 提供防止消息恢复攻击的信息理论安全保证.在基因组数据存储和检索中,图 20 是基于蜂蜜加密的 GenoGuard 协议,用于基因组数据的安全存储和检索.患者将其生物样本提供给测序中心,同时选择蜂蜜加密方案.测序中心进行排序、编码和加密,然后将密文发送到基因库.在检索期间,用户（患者或其医生）请求加密文本,对其进行解密,最后对其进行解码以获得原始序列.

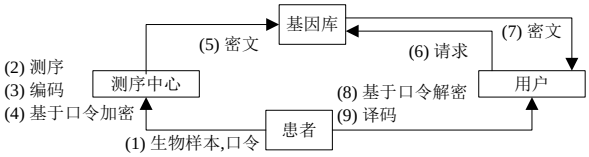

图 20 GenoGuard 协议

蜂蜜加密的特性是当密文用不正确的密钥解密时,结果是看似合理但却是不正确的明文.因此,蜂蜜加密通过响应对加密密钥或密码的每个错误猜测而提供假基因型数据,为加密基因型数据提供额外的保护层.从长远来看,蜂蜜加密提供防止暴力解密的方法,使其在基因组环境中具有特殊的价值.面向基因组数据长期存储,文献[129]提出 GenoGuard

为基因组数据提供长期保护,即使是计算能力无限的攻击者在不正确的密钥下对基因密文进行解密的基因组序列在统计上也似乎是可信的.

4.2.6 基于 SGX 的基因隐私保护

（1）SGX

SGX（Software Guard Extension）通过安全硬件和软件相结合来高效实现敏感数据的任何安全计算[130].SGX 允许应用程序创建安全区域（Secure Enclave）,以保证敏感数据的完整性和机密性,并实现潜在特权软件保护下的计算.SGX 是 Intel 处理体系结构的安全扩展.如图 21 所示,基于 SGX 的应用程序由数据持有者、不可信的云服务提供商和安全区域组成.首先,数据持有者通过远程证明（Remote Attestation）过程建立不可信 CSP 托管的安全区域.然后,数据持有者可以安全地将数据上传到 CSP.在 SGX 中,所有解密的秘密只能由授权代码访问,授权代码也位于安全区域内.硬件支持的访问控制代理保证代码和数据不能被安全区域之外的软件访问或修改.SGX 提供硬件级的安全保障,可以降低计算复杂度.例如,基于 SGX 模型的安全协作框架用于分析分布在不同大陆的罕见疾病遗传数据,首先远程认证协议允许多个数据持有者和安全区域相互验证对方的真实性和完整性,其次当接收到加密数据时,数据处理区域解密每个段并解压压缩数据以恢复原始数据段,然后对每个片段进行批量评估,同时为区域中的前 K 个 SNP 保持更新的全局队列,区域内的所有操作都由 SGX 保护,最后以安全的方式将结果返回数据持有者,以确保其机密性和完整性[71].

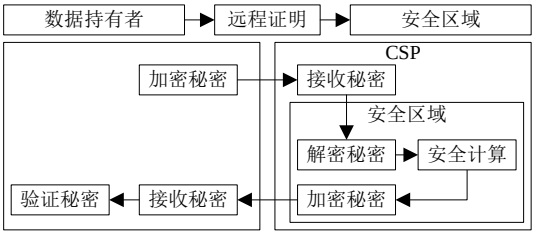

图 21 SGX 模型

（2）基因组数据测序与存储

在基因组变异搜索中,Mandal 等[131]使用 Intel SGX 构建实用的、隐私的、不经意的基因组变异搜索算法,更准确地说,考虑 2017 年 iDash 隐私与安全研讨会（iDash Privacy and Security Workshop 2017）竞赛的第 2 个问题,即在两个群体的某些遗传数据中搜索具有高 χ^2 统计的变异.竞赛的获胜解决方案非常有效,但不是内存遗忘（Memory Oblivious）,这可能使 SGX 易受基于内存和缓存的侧信道攻击.Mandal 等准确地量化这种泄漏,以一定的效率为代价,提供具有合理信息泄漏的内存遗忘的实现.该方案大约比非内存遗忘的实现慢 1 个数量级,但是仍然是实用的.为此,提出新的遗忘词典合并（Oblivious Dictionary Merging）的定义和模型,具有独立的理论意义.

（3）基因组数据研究与分析

在 GWAS 中,通常需要将不同来源的数据汇集,以揭示统计模式以及遗传变异与疾病之间的关系,主要的挑战是以隐私保护的方式访问多个基因组数据存储库进行协作研究.由于基因组数据的隐私问题,不同国家之间实施跨国界基因组数据共享的管辖法律和政策.为此,如图 22 所示,Sadat 等[132]提出混合框架 SAFETY,它可以使用同态加密在联邦基因组数据集上安全地执行 GWAS,并且引入 SGX 的安全硬件组件,以确保高效性和隐私性.Chen 等[71]提出隐私保护国际合作框架 PRINCESS,用于分析分布在不同大洲的罕见疾病遗传数据.PRINCESS 利用 SGX 和硬件进行可靠的计算,与传统的国际协作模型不同,PRINCESS 将个体级别的患者 DNA 物理地集中于单个站点,在使用 AES-GCM 加密的数据上执行安全分布式计算,从而满足保护健康信息的政策和法规.为了促进有效的协作或远程基因组分析,在不可信的云环境中,Kockan 等[133]基于安全区域与查询的草图算法（Sketching Algorithm）,对多个机构的基因组数据单独加密,以此在多个数据库上执行数据分析和查询的计算框架.与其他用于安全和隐私保护基因组数据分析的技术不同,该框架利用 SGX 支持的安全区域,并且该框架使用草图数据结构可以适合安全区域中可用的有限内存.此外,该框架能够在快速准确地识别出用户指定的 SNP 子集中与 χ^2 统计相关的显著 SNP,表明通过 SGX 对基因组医学进行安全和隐私保护计算的可行性.

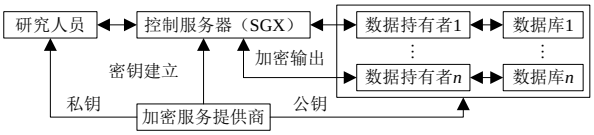

图 22 SAFETY 联邦体系结构

（4）基因组数据直接面向消费者服务

DNA 测序技术的进步促进基因组学在改善医疗服务和促进生物医学研究方面的广泛应用,然而,隐私和安全问题已经成为利用云计算处理敏感基因组数据的挑战.Chen 等[130]提出基于 SGX 的安全外包基因检测框架 PRESAGE,该框架利用对称加密、数字签名、最小完美哈希和消息认证码实现高效、安全的数据存储和外包计算.提出的方案为不可信云中安全高效的基因组数据外包提供另一种解决方案.

在表 15 中,SGX 可以用于基因组数据测序与存

储、研究与分析、直接面向消费者服务的安全和隐私保护.通过使用 SGX 展示基因组数据安全和隐私保护计算的可行性,利用 SGX 安全高效的计算能力来分析基因组数据有利于医疗服务和医学研究.

表 15 基于 SGX 的基因组数据安全和隐私保护

方法		特点	保护效果	面向场景	面向场景的保护效果
SGX		Intel SGX 是 Intel 体系结构的扩展,SGX 可以在解密密文后高效地执行任何安全计算.但是 SGX 易受到侧信道攻击,而且 SGX 安全区域提供的内存有限,可以通过使用草图数据结构在有限的范围内处理基因组数据.	SGX 允许应用程序在 CPU 的受保护执行区域内运行,包括特权软件内核、管理程序等在内的不可信实体无法访问安全区域.SGX 确保不能从安全区域外部读取或修改安全区域内的代码和基因组数据,以此提供受保护区域内敏感基因组数据分析.	基因变异搜索	文献[131]使用 SGX 构建实用的、隐私的、不经意的基因组变异搜索算法.
				国际合作罕见疾病分析	文献[71]提出 PRINCESS 框架实现隐私保护国际合作罕见疾病分析的高安全级别.
				协同基因组数据分析	文献[133]提出基于安全区域的基因组数据分析草图算法,用于不可信的云平台上执行安全的协同基因组分析（特别是 GWAS）和在线查询.
SGX	同态加密	同表 14	同表 14	联邦环境中的 GWAS	文献[132]提出混合框架 SAFETY 用于在联邦环境中安全地执行 GWAS 的统计检测.
	对称加密	同表 12	同表 12	基因检测	文献[130]提出基因检测的混合框架 PRESAGE 用于不可信云中安全高效的基因组数据外包计算,能够有效地防御恶意攻击.
	数字签名	同表 12	同表 12		
	最小完美哈希	对于任意基因型值 x_i 和 x_j,当 $i=j$ 时哈希函数 $H(x)$ 满足 $H(x_i)=H(x_j)$,且 $H(x)$ 一一映射输入基因型值到唯一整数.	对输入的基因型值集合进行哈希时不可能产生任何碰撞,两个不同基因型值在相同哈希值相互碰撞的可能性越小.		
	消息认证码	同表 14	同表 14		

4.2.7 基因隐私保护的密码学方法比较与分析

对称加密相比于非对称加密计算较快,因此对称加密广泛用于实现基因组数据测序与存储、共享与聚集,以及基因组数据应用中的安全和隐私保护.结合对称加密和保序加密可以实现基因组数据存储、检索和处理的安全和隐私保护.不过,对于大规模分布、高维的基因组数据,使用对称加密存在密钥协商和管理困难的问题.因此,可以使用非对称加密来进行密钥协商,结合对称加密和非对称加密实现基因组数据的安全和隐私保护.将非对称加密和隐私信息检索结合可以实现基因组数据共享的安全和隐私保护.通过结合承诺方案、先应秘密共享、量子密钥分发和一次一密,以及结合数字签名、承诺方案、Merkle 哈希树等密码技术可以实现基因组数据存储的永久安全.使用带密钥的哈希可以实现基因组数据研究与分析中的安全和隐私保护.在基因检测中,通过使用受控函数加密可以实现个性化医学、患者相似性检测、亲子鉴定和血缘关系检测中的隐私保护.不过,直接使用这些密码技术不便于进行基因组数据医疗服务的安全计算和隐私保护.

安全多方计算不泄露各自输入基因组数据的隐私,各方只是获得联合计算结果.因此,基于安全多方计算可以实现基因组数据研究与分析、医疗服务、法律与取证和直接面向消费者服务应用中的安全和隐私保护.在基因组数据研究与分析中,基于混淆电路可以实现基因组数据的第三方安全计算、联邦基因组数据集的安全 GWAS 分析、微生物测序研究的安全和隐私保护,以及相似患者查询的隐私保护.使用秘密共享实现大规模 GWAS 的安全和隐私保护.在直接面向消费者服务中,基于不经意传输实现亲子鉴定和祖先鉴定中的安全和隐私保护.不过,安全多方计算存在计算和通信开销的瓶颈.

而同态加密可以对密文进行计算,不需要进行复杂的通信.因此,同态加密广泛用于基因组数据测序与存储、共享与聚集,以及应用中的安全和隐私保护.此外,结合同态加密、一次一密、消息认证码和不经意传输可以实现基因组数据存储的永久安全.结合同态加密、带密钥的哈希和保序加密实现公共云上的基因组数据处理的安全和隐私保护.以同态加密为基础,使用不经意传输可以实现 DNA 搜索的安全和隐私保护,使用保序加密可以实现基因组数据共享和计算的安全和隐私保护,使用混淆电路实现基因组数据的安全外包,结合代理重加密实现疾病易感性检测的隐私保护.结合同态签名和聚合签名实现个体与服务提供商共享基因组数据和基因检测结果的可信性和可追责性.使用基于属性的同态加密可以实现基因组数据的细粒度访问控制和疾病易感性检测的隐私保护.不过,使用同态加密的计算复杂度高,加解密速度慢延迟基因组数据分析,影响医疗诊断和治疗,因此同态加密不能用于基因组数据医疗服务中实现安全和隐私保护.

此外,模糊加密可以实现固定大小集合的安全

比较,常用于基因组数据法律与取证中遗传亲缘关系检测的安全和隐私保护.蜂蜜加密使用不正确的密钥解密时,解密结果是看似合理但又不正确的明文,因此可以防止穷举攻击,常用于基因组数据存储中的永久安全保护.因为 SGX 在安全区域可以高效地执行任何计算,因此常用于基因组数据测序与存储、研究与分析、直接面向消费者服务中的安全和隐私保护.在基因组数据测序与存储中,基于 SGX 实现基因变异搜索的安全和隐私保护.在基因组数据研究与分析中,使用 SGX 实现国际合作罕见疾病分析、协同基因组数据分析的安全和隐私保护.结合 SGX 和同态加密可以实现联邦环境中 GWAS 的安全和隐私保护.在直接面向消费者服务中,结合 SGX、对称加密、数字签名、最小完美哈希和消息认证码实现基因检测的安全和隐私保护.

4.3 基于匿名的基因组数据隐私保护

在敏感数据的隐私保护中,密码学方法只适用于最终和处理的序列数据,在不可信第三方存在的情况下,解密后获得明文会导致数据保护级别不足.此外,计算复杂度和通信开销是密码学方法的瓶颈,使用密码学方法因计算复杂使得在进一步的分析工作中会出现相当大的时间延迟,不利于医疗单位对基因组数据的快速分析.而匿名方法通过有效的泛化敏感数据实现个体隐私保护,而且不会导致延迟分析的问题.为此,匿名方法已被用于敏感基因组数据的隐私保护,下面将主要介绍实现匿名的相关方法,及其用于基因组数据隐私保护的已有工作.

(1) 匿名

k-匿名(k-Anonymity)要求数据集中每条记录与至少 $k-1$ 条其他记录具有相同的准标识(Quasi-Identifier)属性值[10],其中准标识符是识别个体信息的一组属性.k-匿名将基于准标识属性的个体链接到相应的记录的概率限制为 $1/k$,所以可以防止身份泄露.参数 k 控制提供的隐私保护程度,并由数据发布者来设置.为了陈述 k-匿名,首先介绍准标识符的定义.

定义 1. 准标识符.假设 $T(A_1,A_2,\ldots,A_n)$ 是二维表,T 的准标识符是个体的某些属性集合 $(A_1,A_2,\ldots,A_j)\subseteq(A_1,A_2,\ldots,A_n)$.

定义 2. k-匿名.假设 $T(A_1,A_2,\ldots,A_n)$ 是二维表,$Q|_T$ 是与之关联的准标识符.T 满足 k-匿名当且仅当对于每个准标识符 $QI\in Q|_T$,$T[QI]$ 中值的至少出现 k 次.

例如,将腺嘌呤和鸟嘌呤进行 2-匿名泛化为嘌呤族,将胞嘧啶、胸腺嘧啶和尿嘧啶进行 3-匿名泛化为嘧啶族,再将嘌呤族和嘧啶族进一步 2-匿名泛化为核苷酸,以此碱基可以 5-匿名泛化为核苷酸.不过 k-匿名不能防止属性泄露攻击,为此,隐私保护模型 ℓ-多样性(ℓ-Diversity)被提出,要求数据集中的每个匿名组至少包含 ℓ 个敏感属性值[134]. t-封闭性(t-Closeness)是另一种保护数据免受属性泄露攻击的隐私保护模型,旨在限制匿名组中敏感属性值的概率分布和整个数据集中敏感属性值的概率分布之间的距离,防止攻击者学习关于数据集中不可用的个体敏感属性值的信息[135]. k-匿名和 ℓ-多样性都不支持在插入和删除后重新发布微数据,但 m-不变性(m-Invariance)有效地限制重新发布微数据的隐私泄露风险[136].

(2) 基因组数据测序与存储

为了能够在不损害个体隐私的情况下分析数据,要求对基因组数据进匿名保护.事实上,基因测序是获得基因组数据最重要的手段,而且很容易带来个体所关心的隐私泄露问题.在下一代测序中,即使是匿名数据,易遭到重新识别个体和其他隐私泄露风险,而且也适用于人类 DNA 作为副产品获得的应用,例如来自人类宿主的病毒或亚基因组样本,于是 Loka 等[137]提出 PriLive,在测序机运行时自动删除敏感数据,基于此实现基因组测序过程中的隐私保护,以此人类序列敏感信息可以在完全产生之前被检测和移除,这有助于遵守严格的数据保护法规,且对于数据保护以外的应用来说,导致几乎不延迟进行进一步分析的独特特性也是明显的优势.为了能够在不损害研究对象隐私的情况下分析数据,必须开发从医学和基因组数据中去除标识信息的方法,因此 Lin 等[138]建立装箱(Binning)数据库记录更难追溯到个体,按照层次结构表示符号数据和数字数据,并通过泛化记录来进行装箱,以此可以将数据装箱到不同的精度级别,并使用装箱的大小来控制隐私和数据完整性之间的权衡.越来越多个体特定 DNA 序列的收集、储存和分析对保护这些序列所对应的身份提出严峻的挑战,通过重识别来破坏 DNA 隐私,推断 DNA 来源的个体的明确身份,取决于从 DNA 序列推断出的独特特征,于是 Malina[139]基于 k-匿名提出特定个体的 DNA 数据库序列集合匿名的计算方法 DNALA(DNA Lattice Anonymization),使用该方法不可能观察或了解将一个遗传序列记录与 $k-1$ 个其他记录区分开来的特征,使用概念泛化格来确定单个核苷酸区域中两个残基之间的距离,为两个残基提供最相似的泛化概念,例如腺嘌呤

和鸟嘌呤都是嘌呤.DNALA 方法选择要匿名化的序列对,使之成为序列对之间最小距离的序列,并对其进行相应的泛化.在研究个体 DNA 数据库时,必须有适当的匿名性保证数据不能与个体相关,DNALA 是 DNA 序列匿名化的成功方法,然而它使用耗时的多序列比对和低精度贪心聚类算法,并且 DNALA 不是在线算法,当数据库被更新时,它不能快速返回结果,为此 Li 等[140]改进 DNALA 方法,通过更换多序列比对为全局双序列比对比 DNALA 节省时间,并且设计由最大权匹配（Maximum Weight Matching,MWM）算法和在线算法组成的混合聚类算法,在 DNALA 中基于 MWM 的算法比贪心算法更精确,并且具有相同的时间复杂度.健康信息技术有助于收集大量的患者数据,可以支持新的、大规模的生物医学研究,虽然鼓励医疗机构以未标识的形式共享这些数据,但仍对是否允许重识别相应的患者存在担心,目前提出的匿名化临床数据的技术可能会对接收者确定患者身份的能力做出不切实际的假设,于是基于 k-匿名 Heatherly 等[141]表明更实用的假设使得匿名算法的设计能够在可证明的保护下传播详细的临床资料.基因组的快速和低成本测序使得基因组数据在研究和个性化需求中得到广泛应用,并且基因组数据在公共数据库中共享.尽管参与者的身份在这些数据库中是匿名的,但有关个体的敏感信息仍然可以被推断出来,其中有些信息就是亲属关系,为此 Kale 等[142]研究基因型相似和异常等位基因对计数会导致亲属隐私泄露的途径,并在图 23 中提出最大限度地利用共享数据的同时保护亲属隐私免受泄露风险的方法.该方法系统地识别最小部分的基因组数据,以隐藏新的参与者被添加到数据库中,选择合适的隐藏位置被转化为优化问题,在该问题中要隐藏的位置数量在隐私限制的约束下最小化,以确保家庭关系不被暴露.研究结果表明同时共享父母和子女的数据会导致亲属隐私泄露的高风险,而共享来自其他亲属的数据通常更安全.此外,家庭成员的到达顺序对隐私风险水平和共享数据的效用有很大影响.目前的基因组数据隐私保护依赖于去标识符技术,其中隐私保护与数据效用成为零和博弈,取而代之的是 Erlich 等[143]使用信任启用技术来构建研究人员和参与者都双赢的解决方案,通过引入透明创造信任、加强控制增强信任和互惠维系信任三个原则,促进对基因组数据研究的信任,并构造出建立在这些原则基础上的框架,这种以信任为中心的框架提供可持续的解决方案,使遗传隐私与数据共享相协调,并促进遗传研究.

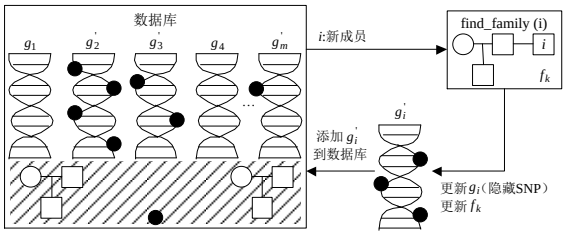

图 23 亲属隐私保护框架

（3）基因组数据共享与聚集

在基因组数据共享中,为了实现隐私保护和发布有用的输出,隐私保护机制发布区间作为输出,而不是真正的输出值.可以通过图 24（a）中的输出进行概率推断来唯一地标识隐私输入,则发布输出不具有隐私保护.攻击者使用后验分布的概率推断输入,假定攻击者持有先验概率 $\Pr[X]$,并从发布者获得 y,则攻击者通过估计后验概率 $\Pr[X_i = a \mid Y = y]$ 来推断输入,其中后验概率为

$$\Pr[X_i = a \mid Y = y] = \frac{\Pr[X_i = a, Y = y]}{\Pr[Y = y]}$$

因此,为了在隐私保护下发布有用的输出,在图 24（b）中 Kusano 等[70]将发布区间作为输出,而不是输出值.定义该机制为映射 $M: R \to I$,其中 I 是连续的实数区间.假设攻击者使用发布的输出执行概率推断,以增强目标属性的后验分布.机制 M 具有两个重要性质,一是当机制 M 提供输出时,攻击者在任何输入属性上的后验值的增加规定在一定水平的上限,二是在这种隐私约束下,机制 M 可以提供包括真实输出的最优区间.不过,该机制计算复杂度为 d 的指数,其中 d 为输入的维数.

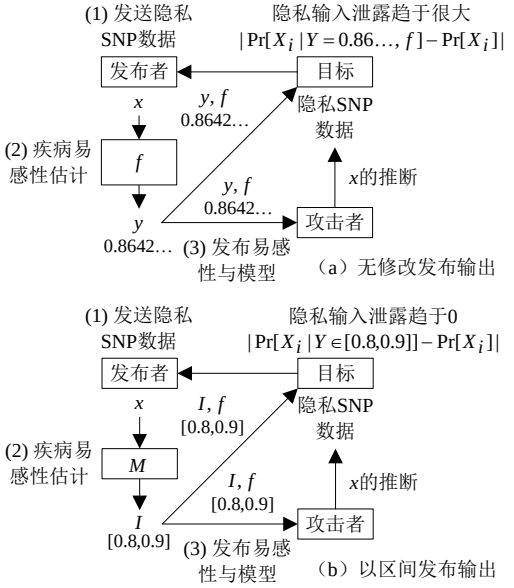

图 24 无修改发布和区间发布

文献[70]考虑单个函数的区间发布,而多个函数的区间发布通常是实际应用所必需的,例如对于几

种常见疾病的易感性.在这种情况下,该机制需要输出确保隐私的长方体.然而,最优长方体发布问题比最优区间发布问题困难得多.

（4）基因组数据直接面向消费者服务

直接对消费者进行基因检测使每个人都有可能了解自己的基因组序列,为了对医学研究做出贡献,越来越多的人在网络上发布他们的基因组数据,有时是以他们的真实身份发布的.然而,这不仅导致他们自己的隐私泄露,也会造成他们亲属的隐私泄露.因此,由于亲属的基因组高度相关,一些家庭成员可能反对透露任何家庭的基因组数据.因此,Humbert 等[144]研究基因组学中效用与隐私的权衡问题,关注最相关的SNP变异类型,如图25所示提出基因组隐私保护框架.考虑到个体的 SNP 包含有关其家庭成员的 SNP 的信息,并且 SNP 相互关联.此外,假设 SNP 在医学研究中有不同的用途,并且对个体的敏感程度也不同.Humbert 等提出混淆机制,该机制依赖于组合优化和概率图模型来优化效用和满足隐私要求,以此最大限度地发挥研究效用,同时保护个体和亲属的基因组和健康隐私.此外,通过扩展的优化算法,以此处理由 SNP 之间的相关性引起的非线性约束问题.该机制最大限度地用于基因组研究,并满足家庭成员的隐私限制.

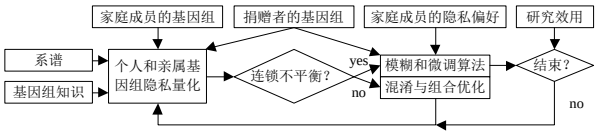

图 25 基因组隐私保护框架

考虑匿名方法的特点、保护效果,面向场景及其保护效果,表 16 对基因隐私保护的匿名方法进行比较,使用各种匿名方法可以实现基因组数据测序与存储、共享与聚集、直接面向消费者服务中的隐私保护.不过,匿名方法没有严格的数学定义和形式化证明隐私保护效果,易遭到链接攻击.因此,需要使用或构建形式化方法实现基因组数据的隐私保护.

表 16 基因组数据隐私保护的匿名方法比较

方法	特点	保护效果	面向场景	面向场景的保护效果	关键问题
自动删除敏感数据	在测序机运行时自动删除敏感数据,几乎不延迟进一步分析,还实现至少与密码学、后置过滤基因型数据保护策略同样精确的过滤结果.	隐私保护实时过滤只比测序机晚几分钟完成,比密码学、后置过滤工具提供更高水平的数据保护,遵守严格的数据保护法规.	下一代测序	文献[137]提出下一代测序的隐私保护实时过滤工具PriLive,在测序机运行时自动删除敏感数据.	量化和识别敏感基因组数据
去标识符	将基因型数据映射到层次结构中,并通过将特定基因型数据表示为层次结构中更一般的节点来匿名化它们,确保没有唯一的基因型可供用户使用.	在层次结构中向上泛化数据,直到基因型的值由用户指定数量的基因型共享为止,泛化层次控制隐私和数据效用之间的权衡,可以使用泛化层次表明不同的匿名级别.	基因组数据存储	文献[138]通过装箱泛化基因组数据库记录更难追溯到个体,可以改变装箱的大小参数来控制隐私保护和数据效用之间的权衡.	定义标识符
k-匿名	选择要匿名化的序列对,使之成为序列对之间最小距离的序列,并对其进行相应的泛化.	不可能观察或学习将一个遗传序列记录与其他 $k-1$ 条记录区分开来的特征.	DNA 序列	文献[139]提出基于 k-匿名的 DNALA 方法,该方法通过保证个体的 DNA 序列与采集机构发布的数据中另一个体的序列完全相同来保护隐私.	构建泛化的等价类
			DNA 序列	文献[140]改进个体 DNA 序列匿名算法 DNALA,在计算距离矩阵时使用成对序列比对以提高效率,使用基于最大权匹配和在线算法的混合算法获得更好的聚类结果.	
			基因组数据传播	文献[141]表明更实用的假设使得匿名算法的设计能够在可证明的保护下传播详细的临床资料.	
在隐私约束下确保家庭关系不被泄露来隐藏最小化的等位基因位点数量	当基因型为 g_i 的个体 i 被添加到数据库中时,会在数据库中查找其亲属,并确定其所属的家庭,所属家庭的隐私通过选择性地隐藏 g_i 的一部分而得到保护,然后 i 的基因型被部分共享,部分共享的基因型表示为 $g_i^{'}$.假设所有的基因位置都会对亲属关系产生同样的影响,但是罕见变异在推断亲属关系时会具有影响力.	基因序列某些位置与疾病状态或易感性有关而揭示更多的信息,根据基因位置所能显示的信息级别可以为其分配重要性权重,使得优先地隐藏重要基因位置,考虑基因位置之间的统计相关性,根据应用的不同对效用函数进行修改,使某些基因位置减小权重或增加权重,重新定义优化模型的目标.	基因组数据共享	文献[142]提出基因组数据库中保护亲缘关系的效用最大化隐私保护方法,在隐私约束下通过最小化合适的基因组数据的隐藏位置,以隐藏新的参与者被添加到数据库中,确保家庭关系不被暴露.	选择合适隐藏位置的优化问题
信任原则	以信任为中心的框架能够提供可持续的解决方案,允许参与者和研究人员从基因组数据共享中获益,使遗传隐私与	奖励行为适当的研究人员,以及惩罚违反其信任的研究人员的机制为持续的双赢行为提供激励,以信	基因组数据共享	文献[143]建议使用以信任为中心的框架能够提供可持续的解决方案,使遗传隐	构建参与者和研究人员之间

	数据共享相协调,并促进遗传研究.	任为中心的框架将创建奖励良好行为、阻止恶意行为和惩罚不遵从行为的系统.		私与数据共享相协调,并促进遗传研究.	的信任框架
发布区间	通过发布包含真实输出值的区间,而不是真实输出值,攻击者关于任何输入属性的后验概率的增加都具有指定的上界,在这种隐私约束下,可以提供包含真实输出的最优区间,在保护隐私的情况下发布有用的输出.	在满足隐私约束下将输出修改为区间,输入就不能被高置信度地推断出来,输出必须尽可能精确,以保证输出区间总是包含真实输出.	基因组数据共享	文献[70]在保护隐私的情况下发布包含真实输出的区间,使得输入不能被高置信度地推断出来.	构建包含真实输出值的最优区间映射
混淆机制	依赖于基因组数据的概率图模型,使用组合优化通过隐藏 SNP 来最大限度地发挥研究效用,同时保护自己和亲属的基因组和健康隐私.	使基因组数据能够公开用于研究,同时保护家庭中个体的基因组隐私,将混淆机制扩展到关联的 SNP,同样满足隐私约束且提供最大化效用.	基因检测	文献[144]提出混淆机制使基因组数据能够公开用于研究,同时保护家庭中个体和亲属的基因组隐私.	在关联公开数据集和 SNP 连锁不平衡下构建混淆机制

4.4 基于差分隐私的基因组数据隐私保护

虽然使用匿名方法可以实现敏感数据的隐私保护,但是不能对隐私保护效果进行严格的数学证明.为此,针对统计推断攻击的具有严格数学证明的差分隐私被提出[11],下面介绍差分隐私的定义及其实现机制,以及目前差分隐私用于基因组数据隐私保护的研究工作.

（1）差分隐私

在本论文中使用 X 表示所有可能数据库记录的全集,x 表示多条数据记录的数据库.具有 k 条记录的数据库 $x=(x_1,x_2,\ldots,x_k)\in \mathrm{X}^k$,其中 $k\in \mathrm{N}$,$x_i\in x$ 是数据库 x 的第 i 条记录或第 i 个元素.数据库 x 和 y 具有同样的大小,并且除了某条数据记录外其他都是相同的,那么数据库 x 和 y 是邻近数据库.因此,邻近数据库 x 和 y 的汉明距离 $d(x,y)=1$.基于邻近数据库的定义,差分隐私的定义及其实现机制如下,其中随机机制 M 的输出空间表示为 $Range(M)$.

定义 3. 差分隐私.如果对于所有 $S\subseteq Range(M)$,$x,y\in \mathrm{X}^k$,且 $d(x,y)=1$,使得

$$P(M(x)\in S)\le e^{\varepsilon}P(M(y)\in S)+\delta$$

那么关于输入空间 X^k 的随机机制 M 是 (ε,δ)-差分隐私.如果 $\delta=0$,那么 M 是 $(\varepsilon,0)$-差分隐私.

差分隐私独立于任何个体是否存在于数据库中,从而保证响应查询的数据是等可能的.根据差分隐私的定义,对于所有邻近数据库 x 和 y,随机机制 M 以至少 $1-\delta$ 的概率满足 ε-差分隐私.差分隐私具有下面的性质[145],包括后处理（Post-Processing）、群组隐私（Group Privacy）和序列组合（Sequential Composition）,以及具有并行组合（Parallel Composition）[146]的性质.在本论文的后续部分,符号 R 表示所有实数的集合.

定理 1. 后处理.如果 $M:\mathrm{X}^k\to \mathrm{R}$ 满足 (ε,δ)-差分隐私,且 $f:\mathrm{R}\to \mathrm{R}'$ 是任意随机映射,那么 $f\circ M:\mathrm{X}^k\to \mathrm{R}'$ 是 (ε,δ)-差分隐私.

定理 2. 群组隐私.如果 M 满足 (ε,δ)-差分隐私,对于所有 $S\subseteq Range(M)$,$x,y\in \mathrm{X}^k$,且 $d(x,y)=t$,使得

$$P(M(x)\in S)\le e^{t\varepsilon}P(M(y)\in S)+te^{(t-1)\varepsilon}\delta$$

那么对于群组大小为 t 的随机机制 M 是 $(t\varepsilon,te^{(t-1)\varepsilon}\delta)$-差分隐私.如果 $\delta=0$,对于群组大小为 t 的随机机制 M 是 $(t\varepsilon,0)$-差分隐私.

定理 3. 序列组合.如果 M_i 满足 $(\varepsilon_i,0)$-差分隐私,那么序列 $M(x)=(M_1(x),M_2(x),\ldots,M_t(x))$ 是 $(\sum_{i=1}^{t}\varepsilon_i,0)$-差分隐私.

定理 4. 并行组合.如果 M_i 满足 $(\varepsilon_i,0)$-差分隐私,且 x_i 是数据库 x 的任意不相交的子集,那么序列 $M(x)=(M_1(x_1),M_2(x_2),\ldots,M_t(x_t))$ 是 $(\max\{\varepsilon_i\},0)$-差分隐私.

在差分隐私机制中,概率分布都是对称的指数分布.对于数值数据,使用 Laplace 机制、Gaussian 机制和离散 Laplace 机制[147]可以实现差分隐私.对于分类数据,使用指数机制可以实现差分隐私[148].

针对本地数据管理和使用的隐私泄露问题,可以使用本地化差分隐私实现隐私保护[149],本地化差分隐私的定义及其实现机制如下.

定义 4. 本地化差分隐私.如果对于任意可能输出 $\beta\in Range(M)$ 和任意两个输入 b_1,b_2,使得

$$P(M(b_1)=\beta)\le e^{\varepsilon}P(M(b_2)=\beta)+\delta \quad (5)$$

那么随机机制 M 是 (ε,δ)-本地化差分隐私.如果 $\delta=0$,那么 M 是 $(\varepsilon,0)$-本地化差分隐私.

随机响应是实现本地化差分隐私的主要扰动机制[150].二元随机响应可以实现 (ε,δ)-本地化差分隐私[151].当 $\delta=0$ 时,二元随机响应是最优机制,可以实现 $(\varepsilon,0)$-本地化差分隐私[151].四元随机响应也可以实现 (ε,δ)-本地化差分隐私[151].多元随机响应可以实现 $(\varepsilon,0)$-本地化差分隐私[151].

例如,在 GWAS 中,通过输入扰动次要等位基因频率,以及输出扰动 χ^2 统计和 p-值,然后再发布差分隐私 GWAS 数据而不泄个体隐私.

（2）基因组数据共享与聚集

在基因组数据共享与聚集的隐私保护研究中,由于匿名方法没有严格的形式化证明隐私保护效果,而且由于链接到其他公开数据集易遭受到隐私泄露问题[44].因此,具有严格数学证明的差分隐私成为基因组数据隐私保护的热点方法.针对 GWAS 中的个体重识别攻击,Li 等[152]引入成员隐私框架,该框架包括正成员隐私（Positive Membership Privacy）,防止攻击者显著增加其判断实体在输入数据集中的能力,以及负成员隐私（Negative Membership Privacy）防止非成员泄露,该框架由一系列分布参数化,这些分布捕捉到攻击者的先验知识,能够选择不同的分布族来实例化成员隐私,在隐私保护和数据效用之间提供更好的权衡,该框架还提供原理性的方法来发展新的隐私概念,以此可以实现更好的效用.对于合理的隐私预算,差分隐私显著影响期望效用,文献[152]引入成员隐私框架目的是防止具有概率分布族先验知识的攻击者泄露集合成员身份.在成员隐私框架下 Tramèr 等[153]通过考虑先验分布来捕获更合理的背景知识,研究松弛的差分隐私,在不同的隐私预算下差分隐私可用于实现各种攻击环境的成员隐私,从而在隐私保护和数据效用之间实现有趣的权衡,并且重新评估在全基因组关联研究中发布差分隐私 χ^2 统计的方法,通过添加 Laplace 噪声到 SNP 值为 0 的案例组患者数,并使用具有特定距离分数的指数机制输出显著重要的 SNP,表明可以达到更高的效用,同时在相关的攻击环境中仍然保证成员隐私.隐私问题阻碍有效的基因组数据共享,因此 Wang 等[154]提出新的传播基因组数据的方法,同时满足差分隐私,该方法将原始基因组序列分割成块,自上而下对块进行细分,最后在计数中加入噪声以实现隐私保护,所提出的算法具有较高的敏感度,能够保持一定的数据利用率.生物医学研究中测序技术的普及引起许多新的隐私问题,包括在基因组尺度上发布聚合数据,例如次要等位基因频率、回归系数,差分隐私方法可以通过提供强有力的隐私保护来克服这些问题,但代价是极大地干扰感兴趣分析的结果,因此 Simmons 等[155]研究在不牺牲差分隐私准确性的情况下,在每个 SNP 的次要等位基因频率中添加 Laplace 噪声,通过将最小扰动量与贝叶斯统计和马尔可夫链蒙特卡罗（Markov Chain Monte Carlo）技术相结合来实现隐私保护的聚合基因组数据共享.SNP 连锁不平衡容易导致患者的隐私信息泄露,为此结合 SNP 连锁不平衡相关系数,在图 26 中刘海等[156]使用差分隐私机制对基因组数据进行随机扰动,以此实现基因组数据共享的隐私保护,该模型可以实现 SNP 连锁不平衡下基因组数据隐私与效用的权衡,并对 SNP 连锁不平衡下的基因差分隐私保护研究具有促进作用.进一步,在关联序列下考虑高阶 SNP 连锁不平衡,通过标准化熵差（Normalized Entropy Difference）度量高阶 SNP 连锁不平衡,使用 Pearson 相关系数度量序列之间的关联性,Liu 等[157]基于差分隐私实现基因组数据隐私保护与数据效用之间的权衡.为了促进基因组数据共享,GA4GH 建立 Beacon 系统,旨在帮助研究人员找到感兴趣数据集的搜索引擎.虽然目前的 Beacon 系统只支持基因组数据,但其他类型的生物医学数据,如 DNA 甲基化,对于促进生物信息学领域的研究和发展也是必不可少的.由于目前的基因组 Beacon 易受成员推断攻击,且 DNA 甲基化数据高度敏感,因此 Hagestedt 等[158]构建 DNA 甲基化数据共享 Beacon 系统 MBeacon,MBeacon 的核心组件是满足差分隐私的双稀疏向量技术（Double Sparse Vector Technique）,稀疏向量技术在与阈值比较之前向所有查询添加噪声确保差分隐私,以此能够在不显著损害效用的情况下成功地降低成员隐私泄露风险.基于差分隐私,Fienberg 等[159]提出发布聚集 GWAS 数据而不泄露个体隐私的新方法,该方法通过添加 Laplace 噪声到次要等位基因频率、χ^2 统计和 p -值,发布这些扰动统计实现个体的隐私保护,并且还提出基于惩罚逻辑回归（Penalized Logistic Regression）的差分隐私方法进行全基因组关联研究.在 Homer 等[20]发表对 GWAS 数据的攻击之后,在 GWAS 数据库中保护个体信息的隐私一直是研究人员关注的主要问题,于是 Yu 等[160]扩展 Fienberg 等的方法,用于任意数量的案例和对照组研究发布差分隐私 χ^2 统计信息,以及用于发布差分隐私等位基因检验统计信息,而且不泄露个体的隐私,实现隐私预算和统计效用之间的权衡,从而有助于为发布的数据确定适当的隐私保护级别.由于基因组数据聚集能揭示个体的敏感信息,Simmons 和 Berger[161]通过对次要等位基因频率使用差分隐私机制进行随机扰动后,为从基因组研究中获得的聚集数据提供可证明的隐私保护.但是,差分隐私的隐私保护与数据效用之间的权衡导致效用灾难或隐私泄露的问题,为此 Liu 等[162]提出字符型自适应差分隐私机制,并将其用于基因组数据共享中实现期望隐私保护和期望数据效用.

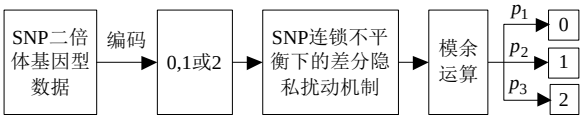

图 26 SNP 连锁不平衡下的差分隐私随机扰动模型

（3）基因组数据研究与分析

GWAS 已经成为分析 DNA 序列以发现遗传疾病的流行方法,不过作为 GWAS 的结果发布的统计数据可以用来识别参与研究的个体.为了防止隐私威胁,甚至先前发布的结果也被从公共数据库中删除,阻碍研究人员对数据的访问,并阻碍协作研究.现有的隐私保护 GWAS 技术侧重于回答特定的问题,

例如给定 SNP 对之间的相关性,这不符合典型的 GWAS 过程,分析者可能事先不知道要考虑哪些 SNP,要使用哪些统计检测,对于给定的数据集有多少 SNP 是重要的等等.因此,Johnson 和 Shmatikov[163]提出实用的、隐私保护的 GWAS 数据挖掘算法,支持探索性数据分析,其中分析者事先不知道考虑多少和哪些 SNP,基于差分隐私计算与疾病显著相关的 SNP 的数量和位置、SNP 与疾病之间任何统计检验的显著性、SNP 之间相关性的任何度量以及相关性的块结构,在保证差分隐私的同时产生更准确的结果.由于隐私问题对基因组数据的访问仅限于少数可信的个体,降低对生物医学研究的影响,于是 Simmons 等[164]提出差分隐私 GWAS 的计算框架,使用差分隐私保护敏感表型信息,同时校正种群分层,该框架基于最常用的 EIGENSTRAT 和线性混合模型（Linear Mixed Model）统计信息生成隐私保护 GWAS 结果,以此能够在返回有意义的 GWAS 结果的同时保护隐私.通过隐私保护数据选择方法,帮助数据分析者在不侵犯患者隐私的情况下有效地访问人类基因组数据集,Zhao 等[165]的主要思想是让每个数据所有者发布一组不同的差分隐私试验数据,数据用户可以在其上检测运行任意的关联检测算法,包括数据所有者事先不知道的那些算法,在试验数据生成过程中利用人类基因组中的连锁不平衡来保护数据的效用和患者的隐私,以及通过效用值帮助用户以高度自信评估试验版本真实数据的价值,尽管试验数据不能直接用于科学发现,但它表明哪些数据集更有可能对数据用户有用,因此数据用户可以适当与数据所有者联系,以获得对数据的访问.到目前为止,以差分隐私寻找高分 SNP 的所有方法在准确性或计算效率方面都存在重大缺陷,因此 Simmons 和 Berger[69]基于邻近距离（Neighbor Distance）方法改进差分隐私 GWAS 来克服这些限制,从而能够产生更可行的机制.具体地,使用输入扰动和自适应边界方法来克服准确度问题,并且设计和实现基于凸分析的算法,在固定时间内计算每个 SNP 的邻近距离,克服邻近距离法的主要计算瓶颈.对于给定的 SNP,设 s_0、s_1 和 s_2 分别为对照组中具有 0、1 或 2 个次要等位基因的个体数,r_0、r_1 和 r_2 分别为案例组中的具有 0、1 或 2 个次要等位基因的个体数.设 $x=2r_0+r_1$ 和 $y=2s_0+s_1$,如果 x' 和 y' 是 x 和 y 邻近数据库,那么敏感度为 $|x-x'|+|y-y'|\leq 2$.让 $x_{dp}=x+Lap(2/\varepsilon)$ 和 $y_{dp}=y+Lap(2/\varepsilon)$,则 (x_{dp},y_{dp}) 和 (x,y) 满足 ε-差分隐私,则使用差分隐私的等位基因检验统计量为

$$Y=\frac{2N(x_{dp}S-y_{dp}R)^2}{RS(x_{dp}+y_{dp})(2N-x_{dp}-y_{dp})}$$

其中,S 为案例数,R 为对照数,N 为参与者总数.

（4）基因组数据医疗服务

差分隐私也被广泛应用于医疗服务实现其隐私保护,Fredrikson 等[166]将差分隐私机器学习模型用于基于患者的基因型和背景指导医学治疗,通过添加 Laplace 噪声到线性回归模型的系数,以此实现药物基因组学研究中的隐私保护,例如在个性化 Warfarin 剂量研究中实现隐私保护,不过使用差分隐私模拟临床试验来分析会严重影响效用.在个性化医学中,现有方法的差分隐私学习并不能改善具有可行数据大小和维度的预测,Honkela 等[167]使用差分隐私线性回归模型实现药物敏感性预测的隐私保护,将噪声注入到从数据中计算出的统计量中,使用 Wishart 机制扰动输入协方差项,使用 Laplace 机制扰动目标项,从而实现差分隐私,该方法可以推广到其他的预测因子,并且是渐近一致和有效的隐私保护方法,同时它在有限数据上表现良好,通过降低维度限制隐私信息的共享,并通过投影异常值以适应更严格的边界,从而添加更少的噪声以获得相同的隐私,进而获得良好的有限数据效用.从低预测误差的高维生物数据中将个体分类为疾病或临床类别是生物信息学中统计学习的重要挑战,特征选择可以提高分类精度,但是必须仔细地结合交叉验证以避免过度拟合.近年来,学者们提出基于差分隐私的特征选择方法,如差分隐私随机森林（Random Forest）和可重用保持集（Reusable Holdout Set）.然而,在生物信息学领域,特征数量远大于观察数量,使用差分隐私方法容易导致过度拟合,为此 Le 等[168]通过差分隐私 Evaporative Cooling 实现个体疾病分类的隐私保护,使用指数机制随机输出属性,使用 Relief-F 进行特征选择,使用随机森林进行隐私保护分类,进而防止过度拟合,将隐私保护阈值机制与热力学 Maxwell-Boltzmann 分布联系起来,其中温度代表隐私阈值,利用原子气体 Evaporative Cooling 的热统计物理概念进行反向逐步隐私保护特征选择,在属性之间的相互作用中,差分隐私 Evaporative Cooling 提供更高的分类精度,而不会基于独立的验证集导致过度拟合.

差分隐私具有严格的数学定义和形式化证明的隐私保护效果,并且除了某条记录外,差分隐私考虑所有背景知识.通过添加噪声或随机扰动的方法,差分隐私独立于任何个体是否存在于数据库中,保证响应查询的输出是等可能的.在表 17 中,从面向场景及其保护效果两方面对比分析目前基于差分隐私的基因组数据隐私保护的研究工作.使用差分隐私进行随机扰动,在基因组数据共享与聚集、基因组

数据研究与分析、基因组数据医疗服务中可以实现隐私保护与数据效用之间的权衡,但可能因为添加噪声影响而带来基因组数据效用的偏差,进而严重影响共享与聚集、研究与分析、医疗服务的结果.在差分隐私中,较小的隐私预算导致效用灾难,而较大的隐私预算导致隐私泄露.因此,在基因组数据的差分隐私保护研究中,需要解决隐私保护与数据效用之间的均衡,不过这是关键的挑战.

表 17 基于差分隐私的基因组数据隐私保护

面向场景	面向场景的保护效果	关键问题
基因组数据共享	文献[152]提出成员隐私框架,其中正成员隐私防止攻击者显著提高其对实体在输入数据集中的信心,负成员隐私防止攻击者显著提高其对实体不在数据集中的信心.	隐私与效用权衡
	文献[153]在成员隐私框架下通过考虑捕获更合理数量背景知识的先验分布来研究差分隐私的松弛,对于不同的隐私预算差分隐私可以用于实现不同攻击环境下的成员隐私,从而实现隐私保护和数据效用之间的权衡.	
	文献[154]以严格的差分隐私保护的方式传播基因组数据,能以较高的敏感度保持数据的效用,还可用于保护由医学数据和基因组数据组成的记录等异构数据.	
	文献[155]在保证准确度的情况下基于差分隐私实现隐私保护的聚集基因组数据共享的替代方法.	
	文献[156]提出 SNP 连锁不平衡下的基因隐私保护模型,实现基因数据和单核苷酸多态性连锁不平衡的隐私保护与基因数据效用之间的权衡.	
	文献[157]在关联序列高阶 SNP 连锁不平衡下提出基因组隐私保护框架,该框架满足差分隐私和基因组数据效用之间的权衡.	
DNA 甲基化数据共享	文献[158]基于差分隐私提出共享 DNA 甲基化数据的 Beacon 系统 MBeacon,以此解决 DNA 甲基化数据共享带来的隐私泄露风险问题.	
发布 GWAS	文献[159]基于差分隐私发布聚集次要等位基因频率、 χ^2 -统计和 p 值.	
	文献[160]提出在不泄露个体隐私的情况下发布聚集 GWAS 数据的方法,实现隐私预算和统计效用之间的权衡,以此为发布的数据确定适当的隐私保护级别.	
基因组数据聚集	文献[161]提出基于模型的测量方法 PrivMAF,以此为从基因组研究中获得的聚集数据即次要等位基因频率提供可证明的隐私保护.	
GWAS	文献[163]在没有任何背景的情况下提出 GWAS 的隐私保护查询工具包能够进行数据探索分析,允许分析者查询与疾病显著相关的 SNP 的数量、位置和 p 值,以及在最感兴趣的区域中基因组的相关块结构.	
	文献[164]在执行隐私保护 GWAS 时,同时校正人口分层的影响,而不会显著增加运行时间.	
基因组数据选择	文献[165]提出差分隐私试验数据发布方法来解决具有价值和敏感信息的大量可用人类基因组数据与参与者重识别风险之间的矛盾,为数据所有者和研究人员共享人类基因组数据提供安全、及时的方法.	
发布频繁 SNP	文献[69]通过对邻近距离方法的改进来克服准确性和计算效率的缺陷,并执行差分隐私 GWAS.	
药物基因组学研究	文献[166]在医疗应用中使用差分隐私进行端到端案例研究,当差分隐私算法用于指导 Warfarin 治疗剂量水平,探索隐私和效用之间的权衡.	
药物敏感性预测	文献[167]使用稳健的差分隐私线性回归模型预测给定基因表达数据的药物敏感性.	
个体疾病分类	文献[168]提出差分隐私 Evaporative Cooling 算法,使用 Relie-F 进行特征选择,并利用随机森林进行隐私保护分类,进而防止过度拟合.	
基因组数据共享	文献[162]提出字符型自适应差分隐私机制用于实现基因组数据隐私保护与数据效用之间的均衡.	隐私与效用均衡

4.5 基于混合方法的基因组数据隐私保护

在基因组数据隐私保护中,如果仅使用密码学方法会导致很高的计算复杂度,不利于基因组数据分析,仅使用匿名技术会因为链接公开的生物医学数据导致隐私泄露,而差分隐私会因使用较小的隐私预算导致效用灾难,反之亦然.为此,通过结合密码学、匿名和差分隐私方法,利用各种方法的优势在基因组数据隐私保护中起到扬长避短的作用,以此更好地实现基因组数据的隐私保护.

(1) 基因组数据测序与存储

基因组数据是非常隐私敏感的,并且在亲属之间高度相关,因此个体决定如何管理和保护他们的基因组数据是至关重要的,在如何保护和是否揭示他们的基因组方面,同一个家庭的成员可能有不同的意见,Humbert 等[169]基于博弈论的方法来研究这种紧张关系,首先模型化两个纯粹自利的家庭成员之间的相互作用,还分析当亲属表现出利他行为时,博弈是如何演变的,并且定义不同情形下的闭式纳什均衡,多代理影响图将博弈扩展到 N 个参与者,能够有效地计算纳什均衡,利他主义并不总是导致基因组隐私博弈更有效的结果,而且如果参与者意识到的基因组共享利益之间的差异太大,将遵循相反的共享策略,这将对家庭效用产生负面影响.在隐私保护和数据共享之间寻求平衡是当今人类基因组数据管理的主要挑战之一,在不妨碍数据共享的情况下,需要新的隐私增强技术来解决已知的对个人敏感基因组数据的泄露威胁,为此 Cogo 等[170]利用已知的隐私敏感核酸和氨基酸序列的知识库实现隐私敏感基因组数据的检测,该方法使用已知的隐私敏感核酸和氨基酸序列的知识数据库作为参考,可以系统地检测直接来自输入流的隐私敏感 DNA 片段,将检测方法添加到标准安全技术中,可以提供稳健、高效的隐私保护解决方案,消除与最近发布的

基于短串联重复序列、疾病相关基因和基因组变异的基因组隐私攻击相关的威胁,目前关于人类基因组的全球知识表明,该方法可以立即获得全面的数据库,随着新的隐私敏感序列的识别,该数据库可以自动进化应对未来的攻击,此外,通过使用 Bloom 过滤器和扩展到更快的测序机,验证该检测方法可以与下一代测序生产周期相匹配.患者健康档案的再利用可以为临床研究提供巨大的效益,然而当研究人员需要访问基因组数据时,隐私和安全问题是主要障碍,因此在图 27 中 Raisaro 等[171]在 i2b2（Informatics for Integrating Biology and the Bedside）框架之上,利用同态加密和差分隐私实现 i2b2 基因组数据的安全和隐私保护,i2b2 负责管理数据存储、处理查询、计算密文同态加法,并根据用户的访问权限,在计算结果上添加噪声以满足差分隐私.

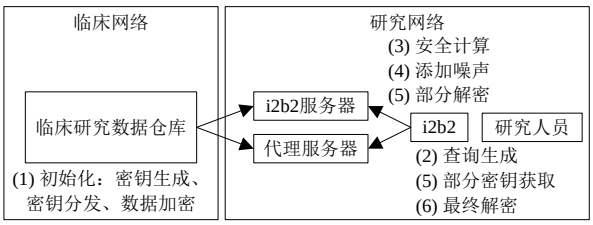

图 27 i2b2 基因组数据的隐私和安全系统

（2）基因组数据共享与聚集

个体研究的聚集数据会受到推断攻击,因此当研究人员在全基因关联元分析中共享研究数据时会出现隐私泄露问题,在图 28 中 Huang 等[172]结合安全多方计算和差分隐私提出安全质量控制（Secure Quality Control,SQC）协议,通过添加 Gaussian 噪声到安全多方计算的统计结果,该协议能够在不向潜在攻击者泄露敏感信息的情况下以隐私保护的方式检查数据质量,以此实现基因组数据质量控制的隐私保护,其中 $[X]$ 表示 X 的加密.人类基因组可以揭示敏感信息,并且具有潜在的可识别性,这就引起人们对大规模共享此类数据的隐私和安全问题的担忧,因此在去中心化网络（Decentralized Network）中 Zhang 等[173]基于同态加密、安全多方计算和差分隐私提出预防性的方法来实现基因组数据共享的隐私保护,以便于全基因组关联研究.在多个机构之间共享大量敏感的临床和基因组数据对于精准医学的扩展至关重要,但是现有的解决方案不能提供法规所要求的强有力的隐私和安全保障,因此考虑研究者是恶意但隐蔽的攻击者,Raisaro 等[174]结合同态加密和差分隐私提出 MEDCO 系统,通过添加 Laplace 噪声来混淆同态加密的患者计数,MEDCO 允许一组临床站点联合并集体保护它们的数据,以便与外部研究人员共享它们,而不必担心安全和隐私问题.考虑分布式环境下大规模基因组数据的共享和聚集,正由于区块链具有分布式、无需可信第三方和自主可控的特点,为基因组数据共享相关的技术和治理挑战提供解决方案,以及执行数据访问协议和拥有本地数据所有权提供有效的方法,因此 Shabani[175]基于区块链实现基因组数据共享的隐私保护,以应对基因组数据共享中的治理挑战,基于区块链的解决方案可以提供自动化数据访问控制过程的机会,提高基因组数据访问的透明度和公平性.

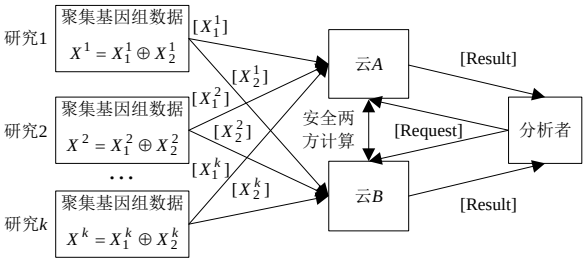

图 28 基因组数据安全质量控制系统模型

因为匿名方法不是严格形式化的,而且易受到链接攻击.因此,在表 18 中,根据博弈论、隐私敏感序列知识库、安全多方计算、同态加密、差分隐私和区块链等方法的各自特点,可以使用混合方法实现基因组数据测序与存储、基因组数据共享与聚集中的隐私保护.

表 18 基因组数据隐私保护的混合方法比较

方法	特点	保护效果	面向场景	面向场景的保护效果	关键问题
博弈论	博弈论能够模拟潜在利益冲突之间的相互作用,并预测策略行为,例如,在同一个家庭中,基于博弈论分析如何保护和是否揭示基因组数据的策略行为.	利他主义并不总是导致基因组隐私博弈更有效的结果,且如果参与者感知到的基因组共享利益之间的差异太大,将遵循相反的共享策略,这将对家庭效用产生负面影响.	基因组数据存储	文献[169]研究家庭成员在是否公开他们的基因组以及如何确保他们在个人设备上的存储安全方面的策略,运用博弈论的方法模型化不同激励的家庭成员之间的相互作用,并预测他们在均衡状态下的行为.	隐私保护与共享价值之间的均衡
已知的隐私敏感核酸和氨基酸序列的知识库	基于短串联重复序列、疾病关联基因和个体基因组变异,建立隐私敏感核酸和氨基酸序列的知识库,基于此系统地识别测序机中 DNA 序列的隐私敏感基因组	将隐私敏感序列检测方法添加到标准安全技术中,可以提供稳健、高效的隐私保护解决方案,消除基于短串联重复序列、疾病关联基因和基因组变异的基因组隐私攻击相关的敏感	基因组数据测序	文献[170]以隐私敏感序列的知识库为参考,提出从输入流中自动检测敏感 DNA 序列的方法.	建立隐私敏感序列知识库

	数据.	信息.随着新的隐私敏感序列的识别,可以获得全面的隐私敏感序列知识库以应对未来的攻击,该检测方法能够匹配下一代测序技术的周期.			
同态加密	同表 14	同表 14	基因组数据存储	文献[171]设计、实现和部署安全、高效的隐私保护解决方案,用于在实际操作场景中探索基因组队列.	机密性 完整性 可用性 可认证性 不可否认性 隐私-效用权衡
差分隐私	见 4.4 节	见 4.4 节			
安全多方计算	同表 13	同表 13	基因组数据共享	文献[172]面向全基因组关联元分析提出安全质量控制协议,能够以隐私保护的方式检查数据的质量,而不会泄露敏感信息.	
差分隐私	见 4.4 节	见 4.4 节			
安全多方计算	同表 13	同表 13	基因组数据共享	文献[173]在中心化网络中采用隐私保护共享协议和数据碎片化算法实现 GWAS 基因组数据的隐私保护共享.	
同态加密	同表 14	同表 14			
差分隐私	见 4.4 节	见 4.4 节			
同态加密	同表 14	同表 14	基因组数据共享	文献[174]提出 MEDCO 系统能够实现敏感医疗数据安全共享,以隐私意识和监管合规的方式促进医疗数据共享.	
差分隐私	见 4.4 节	见 4.4 节			
区块链	区块链提供分布式数据管理和访问控制的能力,以及执行数据访问协议和数据所有权的有效方式,以此为基因组数据共享相关的技术和治理挑战提供解决方案.	基于区块链解决基因组数据共享中的治理挑战,确保组织和参与者可以使用隐私保护算法共享基因组数据,从而促进遵守法律和道德标准.	基因组数据共享	文献[175]基于区块链解决基因组数据共享等方面的治理挑战,最终确保组织和参与者能够与隐私保护算法共享数据,从而促进遵守法律和道德标准.	机密性 完整性 可用性 可认证性 不可否认性

5 基因隐私保护方法分析与展望

使用密码学、匿名和博弈论可以实现基因组数据测序与存储中的隐私保护.但是,考虑基因组数据具有不随时间变化的特点,以及对基因组数据完整性和机密性的要求,而且匿名方法存在非严格形式化的特点,需要使用密码学方法,结合密码学方法与博弈论的策略实现基因组数据测序与存储中的长期安全和隐私保护.在基因组数据共享与聚集中,根据基因组数据的具体使用目的,可以通过使用密码学、匿名和差分隐私方法,以及混合方法实现基因组数据的隐私保护.考虑基因组数据的关联性容易导致隐私泄露,在基因型-疾病、基因型-表型和基因型-基因型关联的基因组数据研究与分析中,主要使用密码学和差分隐私方法,以及混合方法实现基因组数据的隐私保护.在相似患者查询、基因组诊断、药物基因组学研究、药物敏感性预测和疾病分类中,以及进行个性化医学治疗的基因组数据医疗服务中,需要使用精确的基因组数据,否则稍有差池会给患者带来灾难性的结果.由于差分隐私方法通过随机扰动而实现隐私保护,因此在基因组数据医疗服务中可以使用密码学方法来保持基因组数据的完整性,并且实现其隐私保护.对于亲子鉴定和刑事取证,需要对基因序列进行精确的搜索、匹配和查询.因此,在基因组数据法律与取证应用中,通过使用密码学方法不仅可以保证精确搜索、匹配和查询,而且可以实现基因组数据的隐私保护.在基因组数据直接面向消费者服务中,因个体希望准确地了解祖先、配偶遗传兼容性和自己的健康状况等,服务提供商要求个体提供精确的基因组数据.因此,使用密码学方法可以保证基因组数据的完整性和准确性,并且实现基因组数据直接面向消费者服务的隐私保护.

从基因组数据测序与存储、共享与聚集,到基因组数据的广泛应用,包括基因组数据研究与分析、医疗服务、法律与取证和直接面向消费者服务,在这个基因组数据的生态系统中,由于基因组数据固有的敏感性带来人类所担心的隐私泄露问题.为此,除了相关法律法规的监管外,目前的研究主要基于密码学、匿名、差分隐私和混合方法解决整个基因组数据生态系统中存在的隐私泄露问题.然而,使用密码学方法具有不可忽略的计算复杂度[176],并且在不可信第三方存在的情况下解密后容易导致隐私泄露,同时解密后通过关联公开的数据集容易遭到链接攻击.此外,使用没有严格数学证明的匿名技术也容易遭到链接攻击[44].虽然使用差分隐私可以实现严格形式化的隐私保护,但是差分隐私具有隐私-效用单调性的性质,使用较小的隐私预算容易导致效用灾难,而使用较大的隐私预算导致隐私泄露[177],并且考虑数据的关联性直接使用差分隐私容易导致隐私泄露[178].因此,考虑大规模的基因组数据,以及基因组数据固有的敏感性,根据存在的隐私威胁,在表 19 中,本论文分析基因组数据安全与隐私保护的研究挑战,包括具体场景、使用的方法、实现目标和

拟解决的关键科学问题.接下来,就其具体未来的研究挑战进行展望.

表 19 基因组数据安全与隐私保护研究挑战

基因组数据生态系统	场景	方法	目标	科学问题
基因组数据测序与存储	生物样本	法律法规	生物样本保护	公共安全政策研究
	测序 DNA	法律法规	敏感基因组数据保护	敏感基因数据量化和识别
		自动化敏感基因检测		
	外包 DNA 序列生物信息处理	密码技术	DNA 数据安全和隐私保护	机密性、完整性、可用性
	DNA 数据存储	抗量子计算密码	DNA 数据永久安全	
基因组数据共享与聚集	基因组数据共享	密码技术	基因组数据共享的安全和隐私保护	机密性、完整性、可用性、可认证性、不可否认性
		同态签名	基因组数据共享的可信性和可追责性	
		聚合签名		
	基因组数据统计信息发布	差分隐私	隐私与效用平衡	隐私与效用均衡模型
	基因组数据聚集	安全多方计算	基因组数据的安全处理	机密性、完整性、可用性
		同态加密		
基因组数据研究与分析	查询基因组数据库	密码学	查询及查询结果的隐私保护	机密性、完整性、可用性
	基因组数据查询结果			
	研究结果发布	差分隐私	隐私与效用平衡	隐私与效用均衡模型
基因组数据医疗服务	查询患者 DNA 序列	高效的安全多方计算	查询及查询结果的隐私保护	机密性、完整性、可用性
	DNA 序列查询结果			
基因组数据法律与取证	亲子鉴定	高效的密码方案	基因组数据法律与取证中参与者的隐私保护	机密性、完整性、可用性
	法医 DNA 数据库			
基因组数据直接面向消费者服务	查询基因组数据	高效的密码方案	基因组数据查询及查询结果、基因检测结果的隐私保护	机密性、完整性、可用性
	基因组数据查询结果			
	基因检测结果			

（1）基因组数据测序与存储的隐私保护

通过测序患者生物样本获得基因组数据,基因组数据唯一识别个体,并与疾病、血缘关系,以及任何其他敏感信息关联,因此需要保护生物样本和测序基因组数据的隐私信息.但因生物样本的物理特性,需要通过法律法规保护生物样本.对于测序基因数据,需要设计自动化敏感基因数据检测方法,结合法律法规保护测序基因数据的隐私.由于大规模、高维的基因组数据计算成本高,基于密码学方法设计外包到云服务提供商进行标准化生物信息处理的隐私保护机制.因为基因组数据具有稳定性,不随时间而变化,基因组数据中不敏感的部分在将来可能被认为是敏感的,已有工作在基因组数据永久安全方面使用标准密码技术存在不足,即使无限地增加密钥长度,但在实际中对于所有密码系统是不可行的,而且现在认为是安全的密码系统会因量子计算发现其存在安全缺陷.因此,需要研究抗量子计算的密码学方法保证基因组数据的永久安全.

（2）基因组数据共享与聚集的隐私保护

基因组数据共享促进生物医学研究,不过基因组数据共享关联亲属的隐私.即使个体没有公开自己的基因组数据,其家庭成员公开基因组数据也会泄露关于该个体的敏感信息.此外,共享基因组数据计数查询结果,例如等位基因频率,以及共享基因型-表型研究结果,例如 GWAS 统计信息,会导致个体重识别,并且获得个体及其亲属的疾病易感性隐私信息.因此,需要基于密码学方法实现基因组数据共享的隐私保护,将差分隐私用于基因组数据研究的统计发布信息实现可证明的隐私保护.不过,差分隐私因隐私预算过小而导致效用灾难,因此需要使用满足隐私与效用均衡的差分隐私机制共享基因组数据的敏感统计信息.在基因组数据共享中,存在数据源不可信和服务提供商在未获得授权的情况滥用基因组数据的问题,需要结合同态签名和聚合签名实现数据的可信性和服务提供商未授权行为的可追责性.此外,为了便于跨机构、跨地区和跨国家之间进行生物医学研究,例如国际罕见疾病分析,需要聚集大规模基因组数据.但因基因组数据聚集的服务提供商不可信,通过关联公开可用的表型数据导致个体重识别,进而获得其他任何敏感信息.因此,需要使用安全多方计算、同态加密实现聚集基因组数据的处理,并且实现基因组数据聚集的隐私保护.

（3）基因组数据研究与分析的隐私保护

在基因组数据与疾病关联研究中,科学家或制药公司研究人员利用隐私基因数据库进行分析.但数据库持有者担心个体敏感信息泄露,不愿向研究人员的查询请求提供查询结果,而且由于经济利益的竞争,研究人员都希望在发表或授予研究成果专利之前,对彼此的研究保密.在基因组数据研究与分析中,除了查询的机密性之外,基因数据库中的任何

隐私信息也应该受到保护.因此,需要使用密码学实现基因组数据研究与分析中查询及查询结果的隐私保护.此外,在研究人员发布的研究结果中,例如GWAS 统计量,该统计结果会带来参与者的隐私泄露风险.因此,基于差分隐私实现基因组数据研究与分析中参与者的隐私保护.但是,人类基因组的很大部分是相同的,而单个个体所特有的区域相对较小,需要大量的噪声来满足少量 SNP 的差分隐私保护是不切实际的.因此,需要设计满足隐私与效用均衡的差分隐私机制,用于基因组数据研究与分析中保护参与者的隐私同时维持数据效用.

(4)基因组数据医疗服务的隐私保护

在基因组数据医疗服务中,医疗服务提供者查询患者的 DNA 序列,以便于提供个性化医疗服务.医疗服务提供者对基因组数据的查询应该保密,以保护其商业机密,同时医疗服务提供者不能获得患者的敏感信息.即使在隐私保护下,医疗服务提供者可以通过重复查询患者的隐私 DNA 序列,根据查询的输出猜测患者的隐私基因组数据.为了提供更好的个性化医疗服务,医疗服务提供者希望获得更准确的基因组数据,应该在保护患者隐私的同时保持数据的效用.因此,可以使用安全多方计算实现基因组数据医疗服务中序列比较和相似序列搜索,而不泄露患者的隐私,同时保护医疗服务提供者的查询隐私.但是,因为安全多方计算存在计算和通信开销的瓶颈,考虑实际中高效的医疗服务,需要设计有效的安全多方计算协议用于基因组数据医疗服务的隐私保护.

(5)基因组数据法律与取证的隐私保护

基因组数据广泛用于亲子关系鉴定,参与检测的个体将自己的 DNA 发送给第三方可能会严重影响个体的隐私.此外,基因组也广泛用于执法机构确定犯罪嫌疑人,执法机构通常可以无限制访问 DNA 数据库,易导致DNA数据滥用,进而能够获得其他个体的全基因组序列,或者破坏数据库.因此,在基因组数据法律与取证中,基于安全多方计算可以实现参与亲子鉴定的个体隐私保护,而不像第三方透露任何信息.在执法机构检测犯罪嫌疑人时,使用安全多方计算保护 DNA 数据库中其他个体的隐私,保证 DNA 数据库的完整性.在基因组数据法律与取证的隐私保护中,不需要考虑查询隐私,使得问题更容易处理,但对于数百万条记录的大型DNA数据库,使用密码学方法效率较低.因此,需要设计高效的密码方案实现基因组数据法律与取证的隐私保护.

(6)基因组数据直接面向消费者服务的隐私保护

服务提供商查询个体的基因组数据,并提供直接面向消费者的基因组学服务,包括祖先、疾病易感性、血缘关系和配偶兼容性检测等.个体将自己的基因组数据发送到服务提供商,服务提供商进行祖先、疾病易感性、血缘关系或其他检测,检测结果可供个体下载或通过浏览器查看.但是,基因组数据包含有关个体生物学的基本和隐私信息,例如基因组信息包含个体是否有某种特定疾病的易感性,基因组数据泄露对个体亲属的隐私具有重大威胁.而且,服务提供商查询个体基因组数据易导致商业机密泄露.因为基因检测结果明确关联祖先、疾病易感性、血缘关系和配偶兼容性等敏感信息,更能直接导致个体的隐私泄露风险.因此,基于密码技术可以实现基因组数据直接面向消费者服务中查询及查询结果,还有基因检测结果的隐私保护.不过,因为密码学方法计算复杂度高,需要设计高效的密码方案用于实现基因组数据直接面向消费者服务的隐私保护.

6 总 结

本论文首先建立基因组数据生态系统,并分析基因组数据生态系统中的隐私泄露问题.其次,比较分析基因组数据存在的隐私威胁,以及基因组数据隐私和效用度量.然后,对比分析基因组数据的常用隐私保护方法.同时,根据基因组数据的生态系统,对比分析基因组数据测序与存储、共享与聚集、研究与分析、医疗服务、法律与取证,以及直接面向消费者服务中隐私保护的相关研究成果.最后,通过比较分析现有基因组数据隐私保护方法的不足,讨论基因组数据生态系统中隐私保护研究的挑战.希望该工作有益于激励基因组数据的安全和隐私保护研究,进而解决基因组数据的安全和隐私泄露问题.

参 考 文 献

[1] Christensen K D, Dukhovny D, Siebert U, et al. Assessing the costs and cost-effectiveness of genomic sequencing. Journal of Personalized Medicine, 2015, 5(4): 470-486

[2] Naveed M, Ayday E, Clayton E W, et al. Privacy in the genomic era. ACM Computing Surveys, 2015, 48(1): 6

[3] Ayday E, Cristofaro E, Hubaux J P, et al. Whole genome sequencing: Revolutionary medicine or privacy nightmare?. Computer, 2015, 48(2): 58-66

[4] Raisaro J L, Ayday E, Hubaux J P. Patient privacy in the genomic era. Praxis, 2014, 103(10): 579-86

[5] Akgün M, Bayrak A O, Ozer B, et al. Privacy preserving processing of genomic data: A survey. Journal of Biomedical Informatics, 2015, 56: 103-111

[6] Humbert M, Ayday E, Hubaux J P, et al. Quantifying interdependent risks in genomic privacy. ACM Transactions on Privacy and Security, 2017, 20(1): 3

[7] Hudson K L, Rothenberg K H, Andrews L B, et al. Genetic discrimination and health insurance: An urgent need for reform. Science, 1995, 270(5235): 391-393

[8] Stajano F, Bianchi L, Liò P, et al. Forensic genomics: Kin privacy, driftnets and other open questions//Proceedings of the 7th ACM Workshop on Privacy in the Electronic Society. Alexandria, USA, 2008: 15-22

[9] Katz J, Lindell Y. Introduction to Modern Cryptography. Second Edition. Boca Raon, USA: CRC Press, 2014

[10] Samarati P, Sweeney L. Generalizing data to provide anonymity when disclosing information (Abstract)//Proceedings of the 17th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems. Seattle, USA, 1998: 188

[11] Dwork C, McSherry F, Nissim K, et al. Calibrating noise to sensitivity in private data analysis//Proceedings of the 3rd Theory of Cryptography Conference. New York, USA, 2006: 265-284

[12] Dugan T, Zou X. A survey of secure multiparty computation protocols for privacy preserving genetic tests//Proceedings of the IEEE 1st International Conference on Connected Health: Applications, Systems and Engineering Technologies. Washington, USA, 2016: 173-182

[13] Shi X, Wu X. An overview of human genetic privacy. Annals of the New York Academy of Sciences, 2017, 1387(2017): 61–72

[14] Hasan Z, Mahdi M S R, Mohammed N. Secure count queries on encrypted genomic data: A survey. IEEE Internet Computing, 2018, 22(2): 71-82

[15] Al Aziz M M, Sadat M N, Alhadidi D, et al. Privacy-preserving techniques of genomic data—A survey. Briefings in Bioinformatics, 2019, 20(3): 887-895

[16] Mittos A, Malin B, Cristofaro E. Systematizing genome privacy research: A privacy-enhancing technologies perspective. Proceedings on Privacy Enhancing Technologies, 2019, 2019(1): 87-107

[17] Yakubu A M, Chen Y P P. Ensuring privacy and security of genomic data and functionalities. Briefings in Bioinformatics, 2020, 21(2): 511-526

[18] Weidman J, Aurite W, Grossklags J. On sharing intentions, and personal and interdependent privacy considerations for genetic data: A vignette study. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2019, 16(4): 1349-1361

[19] Lin Z, Owen A B, Altman R B. Genomic research and human subject privacy. Science, 2004, 305(5689): 183

[20] Homer N, Szelinger S, Redman M, et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genetics, 2008, 4(8): e1000167

[21] Jacobs K B, Yeager M, Wacholder S, et al. A new statistic and its power to infer membership in a genome-wide association study using genotype frequencies. Nature Genetics, 2009, 41(11): 1253-1257

[22] Wang R, Li Y F, Wang X F, et al. Learning your identity and disease from research papers: Information leaks in genome wide association study//Proceedings of the 16th ACM Conference on Computer and Communications Security. Chicago, USA, 2009: 534-544

[23] Lumley T, Rice K. Potential for revealing individual-level information in genome-wide association studies. The Journal of the American Medical Association, 2010, 303(7): 659-660

[24] Zhou X, Peng B, Li Y F, et al. To release or not to release: Evaluating information leaks in aggregate human-genome data//Proceedings of the 16th European Symposium on Research in Computer Security. Leuven, Belgium, 2011: 607-627

[25] Cai R, Hao Z, Winslett M, et al. Deterministic identification of specific individuals from GWAS results. Bioinformatics, 2015, 31(11): 1701-1707

[26] Sankararaman S, Obozinski G, Jordan M I, et al. Genomic privacy and limits of individual detection in a pool. Nature Genetics, 2009, 41(9): 965-967

[27] Visscher P M, Hill W G. The limits of individual identification from sample allele frequencies: Theory and statistical analysis. PLoS Genetics, 2009, 5(10): e1000628

[28] Shringarpure S S, Bustamante C D. Privacy risks from genomic data-sharing beacons. The American Journal of Human Genetics, 2015, 97(5): 631-646

[29] Raisaro J L, Tramèr F, Ji Z, et al. Addressing Beacon re-identification attacks: Quantification and mitigation of privacy risks. Journal of the American Medical Informatics Association, 2017, 24(4): 799-805

[30] Backes M, Berrang P, Humbert M, et al. Membership privacy in MicroRNA-based studies//Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. Hofburg Palace, Austria, 2016: 319-330

[31] Im H K, Gamazon E R, Nicolae D L, et al. On sharing quantitative trait GWAS results in an era of multiple-omics data and the limits of genomic privacy. The American Journal of Human Genetics, 2012, 90(4): 591-598

[32] Backes M, Berrang P, Hecksteden A, et al. On epigenomic privacy: Tracking personal microrna expression profiles over time//Proceedings of the Workshop on Understanding and Enhancing Online Privacy, affiliated with NDSS. San Diego, USA, 2016

[33] Erlich Y, Shor T, Peer I, et al. Identity inference of genomic data

using long-range familial searches. Science, 2018, 362(6415): 690-694

[34] Lippert C, Sabatini R, Maher M C, et al. Identification of individuals by trait prediction using whole-genome sequencing data. Proceedings of the National Academy of Sciences, 2017, 114(38): 10166-10171

[35] von Thenen N, Ayday E, Cicek A E. Re-identification of individuals in genomic data-sharing beacons via allele inference. Bioinformatics, 2018, 35(3): 365-371

[36] Malin B, Sweeney L. Determining the identifiability of DNA database entries//Proceedings of the American Medical Informatics Association Annual Symposium. Los Angeles, USA, 2000: 537-541

[37] Malin B, Sweeney L. How (not) to protect genomic data privacy in a distributed network: Using trail re-identification to evaluate and design anonymity protection systems. Journal of Biomedical Informatics, 2004, 37(3): 179-192

[38] Malin B. Re-identification of familial database records//Proceedings of the American Medical Informatics Association Annual Symposium. Washington, USA, 2006: 524-528

[39] Malin B, Airoldi E. The effects of location access behavior on re-identification risk in a distributed environment//Proceedings of the International Workshop on Privacy Enhancing Technologies. Cambridge, UK, 2006: 413-429

[40] Nyholt D R, Yu C E, Visscher P M. On Jim Watson's APOE status: Genetic information is hard to hide. European Journal of Human Genetics, 2009, 17(2): 147-150

[41] Humbert M, Huguenin K, Hugonot J, et al. De-anonymizing genomic databases using phenotypic traits. Proceedings on Privacy Enhancing Technologies, 2015, 2015(2): 99-114

[42] Alser M, Almadhoun N, Nouri A, et al. Can you really anonymize the donors of genomic data in today's digital world?. Garcia-Alfaro J, Navarro-Arribas G, Aldini A, Martinelli F, Suri N. Data Privacy Management, and Security Assurance. Cham: Springer, 2015: 237-244

[43] Li S, Bandeira N, Wang X, et al. On the privacy risks of sharing clinical proteomics data. AMIA Summits on Translational Science Proceedings, 2016, 2016: 122-131

[44] Harmanci A, Gerstein M. Quantification of private information leakage from phenotype-genotype data: Linking attacks. Nature Methods, 2016, 13(3): 251-256

[45] Malin B A, Sweeney L A. Inferring genotype from clinical phenotype through a knowledge based algorithm//Proceedings of the Pacific Symposium on Biocomputing. Hawaii, USA, 2002: 41-52

[46] Cassa C A, Schmidt B, Kohane I S, et al. My sister's keeper?: Genomic research and the identifiability of siblings. BMC Medical Genomics, 2008, 1(1): 32

[47] Gitschier J. Inferential genotyping of Y chromosomes in Latter-Day Saints founders and comparison to Utah samples in the HapMap project. The American Journal of Human Genetics, 2009, 84(2): 251-258

[48] Gymrek M, McGuire A L, Golan D, et al. Identifying personal genomes by surname inference. Science, 2013, 339(6117): 321-324

[49] Samani S S, Huang Z, Ayday E, et al. Quantifying genomic privacy via inference attack with high-order SNV correlations//Proceedings of the 2015 IEEE Security and Privacy Workshops. San Jose, USA, 2015: 32-40

[50] Backes M, Berrang P, Bieg M, et al. Identifying personal DNA methylation profiles by genotype inference//Proceedings of the 2017 IEEE Symposium on Security and Privacy. San Jose, USA, 2017: 957-976

[51] Humbert M, Ayday E, Hubaux J P, et al. Addressing the concerns of the lacks family: Quantification of kin genomic privacy//Proceedings of the 2013 ACM SIGSAC Conference on Computer and Communications Security. Berlin, Germany, 2013: 1141-1152

[52] Deznabi I, Mobayen M, Jafari N, et al. An inference attack on genomic data using kinship, complex correlations, and phenotype information. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2018, 15(4): 1333-1343

[53] Lee S H, van der Werf J H J, Hayes B J, et al. Predicting unobserved phenotypes for complex traits from whole-genome SNP data. PLoS Genetics, 2008, 4(10): e1000231

[54] Schadt E E, Woo S, Hao K. Bayesian method to predict individual SNP genotypes from gene expression data. Nature Genetics, 2012, 44(5): 603-608

[55] Wang Y, Wen J, Wu X, et al. Infringement of individual privacy via mining differentially private GWAS statistics//Proceedings of the International Conference on Big Data Computing and Communications. Shenyang, China, 2016: 355-366

[56] Zhang L, Pan Q, Wang Y, et al. Bayesian network construction and genotype-phenotype inference using GWAS statistics. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2019, 16(2): 475-489

[57] Goodrich M T. The mastermind attack on genomic data//Proceedings of the 2009 30th IEEE Symposium on Security and Privacy. Oakland, USA, 2009: 204-218

[58] Wan Z, Vorobeychik Y, Xia W, et al. A game theoretic framework for analyzing re-identification risk. PloS One, 2015, 10(3): e0120592

[59] Cristofaro E, Faber S, Gasti P, et al. GenoDroid: Are privacy-preserving genomic tests ready for prime time?//Proceedings of the 11th ACM Workshop on Privacy in the Electronic Society. Raleigh, USA, 2012: 97-108

[60] Chen Y, Peng B, Wang X F, et al. Large-scale privacy-preserving mapping of human genomic sequences on hybrid clouds//Proceedings of the 19th Annual Network and Distributed System Security Symposium. San Diego, USA, 2012

[61] Jha S, Kruger L, Shmatikov V. Towards practical privacy for genomic computation//Proceedings of the 2008 IEEE Symposium on

Security and Privacy. Oakland, USA, 2008: 216-230

[62] Wang X S, Huang Y, Zhao Y, et al. Efficient genome-wide, privacy-preserving similar patient query based on private edit distance//Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. Denver, USA, 2015: 492-503

[63] Schneider T, Tkachenko O. EPISODE: Efficient privacy-preserving similar sequence queries on outsourced genomic databases//Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security. Auckland, New Zealand, 2019: 315-327

[64] Wagner J, Paulson J N, Wang X, et al. Privacy-preserving microbiome analysis using secure computation. Bioinformatics, 2016, 32(12): 1873-1879

[65] Wagner I. Evaluating the strength of genomic privacy metrics. ACM Transactions on Privacy and Security, 2017, 20(1): 2

[66] Ayday E, Raisaro J L, Rougemont J. Protecting and evaluating genomic privacy in medical tests and personalized medicine//Proceedings of the 12th ACM Workshop on Privacy in the Electronic Society. Berlin, Germany, 2013: 95-106

[67] Ayday E, Raisaro J L, Hubaux J P. Personal use of the genomic data: Privacy vs. storage cost//Proceedings of the 2013 IEEE Global Communications Conference. Atlanta, USA, 2013: 2723-2729

[68] Mohammed N, Wang S, Chen R, et al. Private genome data dissemination. Aris G-D, Loukides G. Medical Data Privacy Handbook. Cham: Springer, 2015: 443-461

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

[70] Kusano K, Takeuchi I, Sakuma J. Privacy-preserving and optimal interval release for disease susceptibility//Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security. Abu Dhabi, United Arab Emirates, 2017: 532-545

[71] Chen F, Wang S, Jiang X, et al. PRINCESS: Privacy-protecting rare disease international network collaboration via encryption through software guard extensions. Bioinformatics, 2016, 33(6): 871-878

[72] Teruya T, Nuida K, Shimizu K, et al. On limitations and alternatives of privacy-preserving cryptographic protocols for genomic data//Proceedings of the International Workshop on Security. Nara, Japan, 2015: 242-261

[73] Kang S, Aung K M M, Veeravalli B. Towards secure and fast mapping of genomic sequences on public clouds//Proceedings of the 4th ACM International Workshop on Security in Cloud Computing. Xi'an, China, 2016: 59-66

[74] Wang X, Zhang Y. E-SC: Collusion-resistant secure outsourcing of sequence comparison algorithm. IEEE Access, 2017, 6: 3358-3375

[75] Senf A. End-to-end security for local and remote human genetic data applications at the EGA. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2019, 16(4): 1324-1327

[76] Hosseini M, Pratas D, Pinho A J. Cryfa: A secure encryption tool for genomic data. Bioinformatics, 2018, 35(1): 146-148

[77] Canim M, Kantarcioglu M, Malin B. Secure management of biomedical data with cryptographic hardware. IEEE Transactions on Information Technology in Biomedicine, 2011, 16(1): 166-175

[78] Ayday E, Raisaro J L, Hengartner U, et al. Privacy-preserving processing of raw genomic data. Garcia-Alfaro J, Lioudakis G, Cuppens-Boulahia N, Foley S. Data Privacy Management and Autonomous Spontaneous Security. Heidelberg, Berlin: Springer, 2013: 133-147

[79] Huang Z, Ayday E, Lin H, et al. A privacy-preserving solution for compressed storage and selective retrieval of genomic data. Genome Research, 2016, 26(12): 1687-1696

[80] Braun J, Buchmann J A, Demirel D, et al. LINCOS - A storage system providing long-term integrity, authenticity, and confidentiality//Proceedings of the 12th ACM on Asia Conference on Computer and Communications Security. Abu Dhabi, United Arab Emirates, 2017: 461-468

[81] Buchmann J, Geihs M, Hamacher K, et al. Long-term integrity protection of genomic data. EURASIP Journal on Information Security, 2019, 2019(1): 1-14

[82] Cassa C A, Miller R A, Mandl K D. A novel, privacy-preserving cryptographic approach for sharing sequencing data. Journal of the American Medical Informatics Association, 2012, 20(1): 69-76

[83] Oprisanu B, Cristofaro E. AnoniMME: Bringing anonymity to the matchmaker exchange platform for rare disease gene discovery. Bioinformatics, 2018, 34(13): i160-i168

[84] Gulcher J R, Kristjánsson K, Gudbjartsson H, et al. Protection of privacy by third-party encryption in genetic research in Iceland. European Journal of Human Genetics, 2000, 8(10): 739-742

[85] Singh A P, Zafer S, Peer I. MetaSeq: Privacy preserving meta-analysis of sequencing-based association studies//Proceedings of the Pacific Symposium on Biocomputing. Hawaii, USA, 2013: 356-367

[86] Zhao Y, Wang X F, Tang H. Secure genomic computation through site-wise encryption. AMIA Summits on Translational Science Proceedings, 2015, 2015: 227-231

[87] Turchin M C, Hirschhorn J N. Gencrypt: One-way cryptographic hashes to detect overlapping individuals across samples. Bioinformatics, 2012, 28(6): 886-888

[88] Bohannon P, Jakobsson M, Srikwan S. Cryptographic approaches to privacy in forensic DNA databases//Proceedings of the International Workshop on Public Key Cryptography. Melbourne, Australia, 2000: 373-390

[89] Naveed M, Agrawal S, Prabhakaran M, et al. Controlled functional encryption//Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security. Scottsdale, USA, 2014: 1280-1291

[90] Kamm L, Bogdanov D, Laur S, et al. A new way to protect privacy in large-scale genome-wide association studies. Bioinformatics,

2013, 29(7): 886-893

[91] Deuber D, Egger C, Fech K, et al. My genome belongs to me: Controlling third party computation on genomic data. Proceedings on Privacy Enhancing Technologies, 2019, 2019(1): 108-132

[92] Constable S D, Tang Y, Wang S, et al. Privacy-preserving GWAS analysis on federated genomic datasets. BMC Medical Informatics and Decision Making, 2015, 15(5): S2

[93] Hamada K, Hasegawa S, Misawa K, et al. Privacy-preserving fisher's exact test for genome-wide association study//Proceedings of the International Workshop on Genome Privacy and Security. Orlando, United States, 2017: 99-102

[94] Bogdanov D, Kamm L, Laur S, et al. Implementation and evaluation of an algorithm for cryptographically private principal component analysis on genomic data. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2018, 15(5): 1427-1432

[95] Cho H, Wu D J, Berger B. Secure genome-wide association analysis using multiparty computation. Nature Biotechnology, 2018, 36(6): 547-551

[96] Xie W, Kantarcioglu M, Bush W S, et al. SecureMA: Protecting participant privacy in genetic association meta-analysis. Bioinformatics, 2014, 30(23): 3334-3341

[97] Asharov G, Halevi S, Lindell Y, et al. Privacy-preserving search of similar patients in genomic data. Proceedings on Privacy Enhancing Technologies, 2018, 2018(4): 104-124

[98] Mahdi M S R, Aziz M M A, Alhadidi D, et al. Secure similar patients query on encrypted genomic data. IEEE Journal of Biomedical and Health Informatics, 2019, 23(6): 2611-2618

[99] Wang R, Wang X F, Li Z, et al. Privacy-preserving genomic computation through program specialization//Proceedings of the 16th ACM Conference on Computer and Communications Security. Chicago, USA, 2009: 338-347

[100] Jagadeesh K A, Wu D J, Birgmeier J A, et al. Deriving genomic diagnoses without revealing patient genomes. Science, 2017, 357(6352): 692-695

[101] Katz J, Malka L. Secure text processing with applications to private DNA matching//Proceedings of the 17th ACM Conference on Computer and Communications Security. Chicago, USA, 2010: 485-492

[102] Karvelas N, Peter A, Katzenbeisser S, et al. Privacy-preserving whole genome sequence processing through proxy-aided ORAM//Proceedings of the 13th ACM Workshop on Privacy in the Electronic Society. Scottsdale, USA, 2014: 1-10

[103] Atallah M J, Li J. Secure outsourcing of sequence comparisons. International Journal of Information Security, 2005, 4(4): 277-287

[104] Shen L, Chen X, Wang D, et al. Efficient and private set intersection of human genomes//Proceedings of the 2018 IEEE International Conference on Bioinformatics and Biomedicine. Madrid, Spain, 2018: 761-764

[105] Gentry C. Fully homomorphic encryption using ideal lattices//Proceedings of the 41st Annual ACM Symposium on Theory of Computing. Bethesda, USA, 2009: 169-178

[106] Ghasemi R, Aziz M M A, Mohammed N, et al. Private and efficient query processing on outsourced genomic databases. IEEE Journal of Biomedical and Health Informatics, 2017, 21(5): 1466–1472

[107] Ayday E, Tang Q, Yilmaz A. Cryptographic solutions for credibility and liability issues of genomic data. IEEE Transactions on Dependable and Secure Computing, 2019, 16(1): 33-43

[108] Lauter K, López-Alt A, Naehrig M. Private computation on encrypted genomic data//Proceedings of the International Conference on Cryptology and Information Security in Latin America. Florianópolis, Brazil, 2014: 3-27

[109] duVerle D A, Kawasaki S, Yamada Y, et al. Privacy-preserving statistical analysis by exact logistic regression//Proceedings of the 2015 IEEE Security and Privacy Workshops. San Jose, USA, 2015: 7-16

[110] Wang S, Zhang Y, Dai W, et al. HEALER: Homomorphic computation of exact logistic regression for secure rare disease variants analysis in GWAS. Bioinformatics, 2016, 32(2): 211-218

[111] Zhang Y, Dai W, Jiang X, et al. FORESEE: Fully outsourced secure genome study based on homomorphic encryption. BMC Medical Informatics and Decision Making, 2015, 15(5): S5

[112] Lu W, Yamada Y, Sakuma J. Efficient secure outsourcing of genome-wide association studies//Proceedings of the 2015 IEEE Security and Privacy Workshops. San Jose, USA, 2015: 3-6

[113] Kantarcioglu M, Jiang W, Liu Y, et al. A cryptographic approach to securely share and query genomic sequences. IEEE Transactions on Information Technology in Biomedicine, 2008, 12(5): 606-617

[114] Blanton M, Aliasgari M. Secure outsourcing of DNA searching via finite automata//Proceedings of the IFIP Annual Conference on Data and Applications Security and Privacy. Rome, Italy, 2010: 49-64

[115] Troncoso-Pastoriza J R, Katzenbeisser S, Celik M. Privacy preserving error resilient DNA searching through oblivious automata//Proceedings of the 14th ACM Conference on Computer and Communications Security. Alexandria, USA, 2007: 519-528

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

[117] Aziz A, Momin M, Hasan M Z, et al. Secure and efficient multiparty computation on genomic data//Proceedings of the 20th International Database Engineering & Applications Symposium. Montreal, Canada, 2016: 278-283

[118] Hasan M Z, Mahdi M S R, Sadat M N, et al. Secure count query on encrypted genomic data. Journal of Biomedical Informatics, 2018, 81: 41-52

[119] Cristofaro E, Faber S, Tsudik G. Secure genomic testing with size- and position-hiding private substring matching//Proceedings of the 12th ACM Workshop on Privacy in the Electronic Society. Berlin, Germany, 2013: 107-118

[120] McLaren P J, Raisaro J L, Aouri M, et al. Privacy-preserving genomic testing in the clinic: A model using HIV treatment. Genetics in Medicine, 2016, 18(8): 814-822

[121] Baldi P, Baronio R, Cristofaro E, et al. Countering GATTACA: Efficient and secure testing of fully-sequenced human genomes//Proceedings of the 18th ACM Conference on Computer and Communications Security. Chicago, USA, 2011: 691-702

[122] Djatmiko M, Friedman A, Boreli R, et al. Secure evaluation protocol for personalized medicine//Proceedings of the 13th ACM Workshop on Privacy in the Electronic Society. Scottsdale, USA, 2014: 159-162

[123] Barman L, Elgraini M T, Raisaro J L, et al. Privacy threats and practical solutions for genetic risk tests//Proceedings of the 2015 IEEE Security and Privacy Workshops. San Jose, USA, 2015: 27-31

[124] Danezis G, Cristofaro E. Fast and private genomic testing for disease susceptibility//Proceedings of the 13th ACM Workshop on Privacy in the Electronic Society. Scottsdale, USA, 2014: 31-34

[125] Ayday E, Raisaro J L, McLaren P J, et al. Privacy-preserving computation of disease risk by using genomic, clinical, and environmental data//Proceedings of USENIX Security Workshop on Health Information Technologies. Washington, USA, 2013

[126] Namazi M, Eryonucu C, Ayday E, et al. Dynamic attribute-based privacy-preserving genomic susceptibility testing//Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing. Limassol, Cyprus, 2019: 1467-1474

[127] Hormozdiari F, Joo J W J, Wadia A, et al. Privacy preserving protocol for detecting genetic relatives using rare variants. Bioinformatics, 2014, 30(12): i204-i211

[128] He D, Furlotte N A, Hormozdiari F, et al. Identifying genetic relatives without compromising privacy. Genome Research, 2014, 24(4): 664-672

[129] Huang Z, Ayday E, Fellay J, et al. GenoGuard: Protecting genomic data against brute-force attacks//Proceedings of the 2015 IEEE Symposium on Security and Privacy. San Jose, USA, 2015: 447-462

[130] Chen F, Wang C, Dai W, et al. PRESAGE: Privacy-preserving genetic testing via software guard extension. BMC Medical Genomics, 2017, 10(2): 48

[131] Mandal A, Mitchell J C, Montgomery H, et al. Data oblivious genome variants search on Intel SGX. Garcia-Alfaro J, Herrera-Joancomartí J, Livraga G, Rios R. Data Privacy Management, Cryptocurrencies and Blockchain Technology. Cham: Springer, 2018: 296-310

[132] Sadat M N, Aziz A, Momin M, et al. SAFETY: Secure GWAS in federated environment through a hybrid solution. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2019, 16(1): 93-102

[133] Kockan C, Zhu K, Dokmai N, et al. Sketching algorithms for genomic data analysis and querying in a secure enclave//Proceedings of the 23rd Annual International Conference Research in Computational Molecular Biology. Washington, USA, 2019: 302-304

[134] Machanavajjhala A, Gehrke J, Kifer D, et al. l-diversity: Privacy beyond k-anonymity//Proceedings of the 22nd International Conference on Data Engineering. Atlanta, USA, 2006: 24

[135] Li N, Li T, Venkatasubramanian S. t-closeness: Privacy beyond k-anonymity and l-diversity//Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering. Istanbul, Turkey, 2007: 106-115

[136] Xiao X, Tao Y. m-invariance: Towards privacy preserving re-publication of dynamic datasets//Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data. Beijing, China, 2007: 689-700

[137] Loka T P, Tausch S H, Dabrowski P W, et al. PriLive: Privacy-preserving real-time filtering for next-generation sequencing. Bioinformatics, 2018, 34(14): 2376-2383

[138] Lin Z, Hewett M, Altman R B. Using binning to maintain confidentiality of medical data//Proceedings of the American Medical Informatics Association Annual Symposium. San Antonio, USA, 2002: 454-458

[139] Malin B A. Protecting genomic sequence anonymity with generalization lattices. Methods of Information in Medicine, 2005, 44(05): 687-692

[140] Li G, Wang Y, Su X. Improvements on a privacy-protection algorithm for DNA sequences with generalization lattices. Computer Methods and Programs in Biomedicine, 2012, 108(1): 1-9

[141] Heatherly R D, Loukides G, Denny J C, et al. Enabling genomic-phenomic association discovery without sacrificing anonymity. PloS One, 2013, 8(2): e53875

[142] Kale G, Ayday E, Tastan O. A utility maximizing and privacy preserving approach for protecting kinship in genomic databases. Bioinformatics, 2017, 34(2): 181-189

[143] Erlich Y, Williams J B, Glazer D, et al. Redefining genomic privacy: Trust and empowerment. PLoS Biology, 2014, 12(11): e1001983

[144] Humbert M, Ayday E, Hubaux J P, et al. Reconciling utility with privacy in genomics//Proceedings of the 13th ACM Workshop on Privacy in the Electronic Society. Scottsdale, USA, 2014: 11-20

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

[146] McSherry F D. Privacy integrated queries: An extensible platform for privacy-preserving data analysis//Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data. Providence, USA, 2009: 19-30

[147] Ghosh A, Roughgarden T, Sundararajan M. Universally utility-maximizing privacy mechanisms. SIAM Journal on Computing, 2012, 41(6): 1673-1693

[148] McSherry F, Talwar K. Mechanism design via differential privacy//Proceedings of the 48th Annual IEEE Symposium on

Foundations of Computer Science. Providence, USA, 2007: 94-103

[149] Duchi J C, Jordan M I, Wainwright M J. Local privacy and statistical minimax rates//Proceedings of the 2013 IEEE 54th Annual Symposium on Foundations of Computer Science. Berkeley, USA, 2013: 429-438

[150] Warner S L. Randomized response: A survey technique for eliminating evasive answer bias. Journal of the American Statistical Association, 1965, 60(309): 63-69

[151] Kairouz P, Oh S, Viswanath P. Extremal mechanisms for local differential privacy. Journal of Machine Learning Research, 2016, 17(1): 492-542

[152] Li N, Qardaji W, Su D, et al. Membership privacy: A unifying framework for privacy definitions//Proceedings of the 2013 ACM SIGSAC Conference on Computer and Communications Security. Berlin, Germany, 2013: 889-900

[153] Tramèr F, Huang Z, Hubaux J P, et al. Differential privacy with bounded priors: Reconciling utility and privacy in genome-wide association studies//Proceedings of the 2015 ACM SIGSAC Conference on Computer and Communications Security. Denver, USA, 2015: 1286-1297

[154] Wang S, Mohammed N, Chen R. Differentially private genome data dissemination through top-down specialization. BMC Medical Informatics and Decision Making, 2014, 14(1): S2

[155] Simmons S, Berger B, Sahinalp C S. Protecting genomic data privacy with probabilistic modeling//Proceedings of the Pacific Symposium on Biocomputing. Hawaii, USA, 2019: 403-414

[156] Liu H, Wu Z, Peng C, et al. Genomic privacy preserving framework for SNP linkage disequilibrium. Ruan Jian Xue Bao/Journal of Software, 2019, 30(4): 1094-1105(in Chinese)
(刘海, 吴振强, 彭长根等. 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)

[157] Liu H, Wu Z, Peng C, et al. Genomic privacy preserving framework for high-order SNPs linkage disequilibrium on correlated sequences//Proceedings of the 2017 International Conference on Networking and Network Applications. Kathmandu, Nepal, 2017: 125-132

[158] Hagestedt I, Zhang Y, Humbert M, et al. MBeacon: Privacy-preserving beacons for DNA methylation data//Proceedings of the 26th Annual Network and Distributed System Security Symposium. San Diego, USA, 2019

[159] Fienberg S E, Slavkovic A, Uhler C. Privacy preserving GWAS data sharing//Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops. Vancouver, Canada, 2011: 628-635

[160] Yu F, Fienberg S E, Slavković A B, et al. Scalable privacy-preserving data sharing methodology for genome-wide association studies. Journal of Biomedical Informatics, 2014, 50: 133-141

[161] Simmons S, Berger B. One size doesn't fit all: Measuring individual privacy in aggregate genomic data//Proceedings of the 2015 IEEE Security and Privacy Workshops. San Jose, USA, 2015: 41-49

[162] Liu H, Wu Z, Peng C, et al. Adaptive differential privacy of character and its application for genome data sharing//Proceedings of the 2019 International Conference on Networking and Network Applications. Daegu, Korea South, 2019: 429-436

[163] Johnson A, Shmatikov V. Privacy-preserving data exploration in genome-wide association studies//Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Chicago, USA, 2013: 1079-1087

[164] Simmons S, Sahinalp C, Berger B. Enabling privacy-preserving GWAS in heterogeneous human populations. Cell Systems, 2016, 3(1): 54-61

[165] Zhao Y, Wang X, Jiang X, et al. Choosing blindly but wisely: Differentially private solicitation of DNA datasets for disease marker discovery. Journal of the American Medical Informatics Association, 2014, 22(1): 100-108

[166] Fredrikson M, Lantz E, Jha S, et al. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing//Proceedings of the 23rd USENIX Security Symposium. San Diego, USA, 2014: 17-32

[167] Honkela A, Das M, Nieminen A, et al. Efficient differentially private learning improves drug sensitivity prediction. Biology Direct, 2018, 13(1): 1

[168] Le T T, Simmons W K, Misaki M, et al. Differential privacy-based evaporative cooling feature selection and classification with relief-F and random forests. Bioinformatics, 2017, 33(18): 2906-2913

[169] Humbert M, Ayday E, Hubaux J P, et al. On non-cooperative genomic privacy//Proceedings of the International Conference on Financial Cryptography and Data Security. San Juan, Puerto Rico, 2015: 407-426

[170] Cogo V V, Bessani A, Couto F M, et al. A high-throughput method to detect privacy-sensitive human genomic data//Proceedings of the 14th ACM Workshop on Privacy in the Electronic Society. Denver, USA, 2015: 101-110

[171] Raisaro J L, Choi G, Pradervand S, et al. Protecting privacy and security of genomic data in i2b2 with homomorphic encryption and differential privacy. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2018, 15(5): 1413-1426

[172] Huang Z, Lin H, Fellay J, et al. SQC: Secure quality control for meta-analysis of genome-wide association studies. Bioinformatics, 2017, 33(15): 2273-2280

[173] Zhang Y, Zhao X, Li X, et al. Enabling privacy-preserving sharing of genomic data for GWASs in decentralized networks//Proceedings of the 12th ACM International Conference on Web Search and Data Mining. Melbourne, Australia, 2019: 204-212

[174] Raisaro J L, Troncoso-Pastoriza J, Misbach M, et al. MEDCO: Enabling secure and privacy-preserving exploration of distributed clinical and genomic data. IEEE/ACM Transactions on

Computational Biology and Bioinformatics, 2019, 16(4): 1328-1341

[175] Shabani M. Blockchain-based platforms for genomic data sharing: A de-centralized approach in response to the governance problems?. Journal of the American Medical Informatics Association, 2019, 26(1): 76-80

[176] Wang S, Jiang X, Singh S, et al. Genome privacy: Challenges, technical approaches to mitigate risk, and ethical considerations in the United States. Annals of the New York Academy of Sciences, 2017, 1387(1): 73-83

[177] Liu H, Wu Z, Peng C, et al. Bounded privacy-utility monotonicity indicating bounded tradeoff of differential privacy mechanisms. Theoretical Computer Science, 2020, 816: 195-220

[178] Kifer D, Machanavajjhala A. No free lunch in data privacy//Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data. Athens, Greece, 2011: 193-204

LIU Hai, Ph. D. His research interest includes privacy preserving.

PENG Chang-Gen, Ph. D. , professor. His research interests include cryptography, information security, and privacy preserving.

WU Zhen-Qiang, Ph. D. , professor. His research interests include network security, privacy preserving, and trusted computing.

TIAN You-Liang, Ph. D. , professor. His research interests include game theory, cryptography, and secure protocol.

TIAN Feng, Ph. D. , associate professor. His research interests include cloud computing, network security, and privacy preserving.

Background

This paper surveys recent advances in the theories and methods of privacy preserving of genome data, and discusses some challenges to the future research. Genome data can uniquely identify an individual and closely associate with inheritance, health, phenotype, and kinship. Moreover, genome data are not change over time. Thus, genome data will bring about privacy concerns in a wide range of applications, such as scientific research, healthcare, legal and forensic, and direct-to-consumer. To solve this problem, in addition to the supervision of relevant laws and regulations, privacy preserving technologies are also used to achieve the privacy preserving of genome data, such as cryptography, anonymity, differential privacy, and hybrid approach. At present, privacy preserving of genome data has been developed based on these technologies of privacy preserving. Recently, some surveys have been studied on existing work of privacy preserving of genome data.

For example, “Patient Privacy in the Genomic Era” discusses the important privacy issues related to genome data, and discusses the methods of protecting the privacy of genome data from the perspective of cryptography; “Privacy Preserving Processing of Genomic Data: A Survey” is a survey of the privacy preserving of query processing of genome data; “Privacy in the Genomic Era” is a survey of privacy preserving of applications of genome data; “A Survey of Secure Multiparty Computation Protocols for Privacy Preserving Genetic Tests” reviews the research on the privacy preserving of genome data based on secure multiparty computation; “An Overview of Human Genetic Privacy” surveys the research on privacy preserving of genome data from three aspects, including access control, differential privacy, and cryptography; “Secure Count Query on Encrypted Genomic Data” summarizes the existing security solutions for outsourcing genome data to the cloud; “Privacy-Preserving Techniques of Genomic Data—A Survey” discusses differential privacy and security issues on genome data; “Systematizing Genome Privacy Research: A Privacy-Enhancing Technologies Perspective” systematically reviews the research on privacy preserving of genome data from the perspective of privacy enhancement technology; Considering the semi-honest and malicious model, “Ensuring Privacy and Security of Genomic Data and Functionalities” discusses and summarizes the privacy leakage of genome data and the relevant privacy attack methods, and classifies the latest privacy preserving schemes of genome data to mitigate the existing attacks, and also discusses the challenges and future research direction of privacy

preserving of genome data; “On Sharing Intentions, and Personal and Interdependent Privacy Considerations for Genetic Data: A Vignette Study” has shown that the institutional trust, concern for family and friend’s privacy, and the likelihood of sharing genetic data are closely associated with factors, such as carrying a genetic marker and the institutional type of the data requester, as well as the demographic factors, such as age and ethnicity, and this work is good for developing a comprehensive explanatory model for the intention to share genetic data.

However, these surveys do not systematically summarize and analyze the theories and methods of privacy preserving of genome data. Unlike these research surveys, this paper systematically summarizes and analyzes the theories and methods of privacy preserving of genome data following six aspects, specifically including the ecosystem of genome data, privacy concerns of genome data, privacy threat of genome data, privacy and utility metrics of genome data, privacy-preserving methods of cryptography, anonymity, differential privacy, and hybrid approach, privacy preserving of genome data in sequencing and storage, aggregation and sharing, research and analysis, healthcare, legal and forensic, and direct-to-consumer. Finally, this paper compares and analyzes the existing privacy-preserving methods of genome data, and discusses future research challenges to privacy preserving of genome data. This work provides the basis for solving the problem of privacy leakage of genome data. Furthermore, this work is helpful to inspire the research of security and privacy preserving of genome data.

As a part of our project, this work is supported by the National Natural Science Foundation of China (U1836205, 62002081, 61662009, 61772008, 61602290), the Project Funded by China Postdoctoral Science Foundation (2019M663907XB), the Foundation of Guizhou Provincial Key Laboratory of Public Big Data (2018BDKFJJ004), the Major Scientific and Technological Special Project of Guizhou Province (20183001). These projects focus on the exploration of theories and methods of security and privacy preserving of public big data. The results and findings of this paper can provide theoretical guidance and solutions for the security and privacy preserving of public big data. Therefore, our work is a key part of these projects.