医疗信息化领域中基因数据隐私安全的现状与对策

王乐子 母健康 郭 昊 王思圆 弓孟春

（神州数码医疗科技股份有限公司 北京 100000）

〔摘要〕 介绍数据安全相关法案以及现有的基因数据安全算法，阐述国内外基因数据安全使用与共享模式，分析基因医疗数据安全存在的问题并提出相应建议，包括尽快建立和完善相关法律法规及监管机制、建立基因精准医疗数据共享平台、推动基因组数据安全共享算法研究等方面。

〔关键词〕 基因数据；隐私；信息安全

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2019.01.003

Status Quo and Corresponding Solution of Privacy Security of Genetic Data in the Medical Informatization Field *WANG Lezi, MU Jiankang, GUO Hao, WANG Siyuan, GONG Mengchun, Digital China Health Technologies Corporation Limited, Beijing 100000, China*

〔**Abstract**〕 The paper introduces several bills related to data security and existing genetic data security algorithm, elaborates on the modes of safe use and sharing mode of genetic data home and abroad, analyzes existing problems of genetic medical data security and comes up with corresponding suggestions, including establishment and improvement of relevant laws and regulations and supervision mechanism as soon as possible, building of sharing platform of genetic precision medicine data, promotion of study on secure sharing of genetic data algorithm, etc.

〔**Keywords**〕 genetic data; privacy; information security

1 引言

针对人类基因数据的研究发展已久，从达尔文的《物种起源》到孟德尔遗传定律，德国科学家米歇尔发现 DNA，剑桥大学的詹姆斯·沃森和弗朗西斯·克里克发现 DNA 的双螺旋结构，再到 20 世纪与曼哈顿原子弹计划、阿波罗登月计划并称为人类自然科学史上 3 个重点计划的人类基因组计划[1-3]，以及近年来基因相关的精准医疗，漫长的发展过程也使得基因数据更加重要。随着生物医学和计算机相关领域的发展，基因组学的研究必将为人类医疗健康发挥重要效用。然而基因数据隐私问题也成为应用研究过程中重要的环节，因为这些隐私数据可能包含个人背景资料、生活习惯、生理和精神情况等极为敏感的信息，另外这些医疗数据还蕴含着极大的商业价值[4-6]，所以在最大限度使用基因数据的同时确保隐私数据安全的研究势在必行，主要从法律和技术两个方面进行。

〔收稿日期〕 2019－01－15

〔作者简介〕 王乐子，硕士；通讯作者：弓孟春，博士。

2 基因数据安全研究

2.1 数据安全相关法案

个人数据的安全是数据在使用过程中的根本问题。美国在健康卫生领域颁布的《健康保险流通与责任法案》(Health Insurance Portability and Accountability Act, HIPAA)[7-10]以及欧盟颁布的《通用数据保护条例》(General Data Protection Regulation, GDPR)[11-15]都体现出发达国家对于个人隐私保护的重视。我国现阶段虽然没有针对个人隐私信息保护的立法，但在多部法律法规中均有对个人信息数据保护的规定，也在不断向各界征求新的立法意见。基因数据是极为隐私的个人数据，通过基因测序后以数据库形式存在，当研究人员在使用这些数据进行基因疾病筛选、研究患者发病率、寻找疾病基因靶点时均会涉及个人基因库隐私安全的问题[16-18]。目前我国正在开展的千人基因计划，以后可能会涉及百万人群的基因组研究，其内容牵涉国人基因机构的组成、功能、演化等极为敏感的数据，一旦泄露会给国家及人民带来难以估计的损失和危害，所以在能够满足科研需求的情形下保证基因组数据的安全是迫切需要攻克的难关。

2.2 现有基因数据安全算法

2.2.1 *K*-匿名方法 由 Sweeny[19]等提出，主要是用来解决链接攻击个人数据隐私问题。基因数据隐私保护需要迫切，*K*-匿名方法不能完全保证将数据库中的 DNA 序列数据信息与这些数据信息提供者的个人身份信息之间的联系切断，于是 DNALA 被开发出来。DNALA 是 *K*-匿名方法应用到 DNA 数据隐私保护的一种方法，主要是对 DNA 数据模糊化处理，使得在数据集中的每个序列都至少有 *K*-1 个完全相同的序列，通过这种方法来防止攻击者的路径攻击，为保证数据的安全性降低数据的精度。另外该方法在数据预处理时用的是多序列对比，这个过程需要运算时间较长，在后面的数据处理中对序列利用贪心算法分组时精度不高。针对该问题的改进策略是在数据预处理阶段将多序列对比改为两两序列对比，这样就可以减少预处理阶段所用时间。研究人员为减少该方法对数据精度的影响，在原来的基础上提出随机爬山法，即以随机爬山法替代贪心算法，得到新的算法——Savior。经实验表明 Savior 对数据的变动程度远远小于 DNALA，可以通过爬山次数这个参数来影响进程中的数据精度。因此通过对 *K*-匿名算法进一步研究也成为保护基因数据安全的一个方向。

2.2.2 差分隐私方法 由计算机密码领域的专家 Bonnie Berger 和 Sean Simmons 提出[20]，可以用来保护基因组的数据库，从而防止个人基因组数据被泄露。以往的隐私模型存在两个主要缺点。其一，面对新型的攻击模式，如背景知识、合成式、deFinetti 等，分组的隐私保护模型难以提供有效的安全防护，攻击者掌握的知识背景与这类模型的安全性相关，而完全定义所有的知识背景极为困难。只有和背景知识无关的隐私安全保护模型在面临新型攻击时才能对数据形成有效的防护。其二，以往的模型在参数变化时不能对数据隐私水平进行定量计算分析，而这将极为影响此类模型处理后的数据可信度。差分隐私模型的出现能够克服以上两个缺点，具有较好的鲁棒性，能够抵挡攻击者各种攻击方式。差分隐私模型就是确保任意一个元素不管是否存在数据集中，其对最后的结果查询影响极小。这是由于该算法不需要知道攻击者掌握多少隐私数据相关的情况背景，对数据库进行随机变化、增加噪声，即在不影响整体的前提下对个人信息进行遮掩，这种输出的信息存在允许范围内的错误，从而达到保护个人数据隐私的目的。另外差分隐私模型建立在严格数学逻辑理论之上，不仅对数据隐私保护进行严密的定义，还提供评估的量化方法，使得模型在不同参数下输出的数据集的隐私保护水平具有可比性。隐私保护模型的可靠性使其逐渐成为数据隐私防护方面的研究热点。

2.2.3 区块链技术[21] 这是一种按照时间的顺序将数据块组合起来的链式数据结构，也是一种以密码学为基础的分布式账本数据库。由于区块链具有数据库的属性，可以对输入的数据信息进行保存和读取。另外只要有需求都可以通过构建服务器的方

式加入区块链网络结构，成为整个区块链网络中众多节点中的一个节点。庞大的网络中所有节点都是平等的，没有中心节点，所以区块链起到信任中介的作用，通过严密的数学逻辑算法保证基因数据的安全传输。区块链技术在基因隐私保护方面的特点是个人可以通过设置访问权限的方式使基因数据研究者得到授权，其只能得到公布的共享信息，也可以依据区块链的特性捕捉到个人数据的使用者。这个过程使用非对称加密——公钥加密，区块链用户通过加密其链上数据以确保隐私性。当基因数据被用于出售或捐赠时，数据的购买方或接收方通过被授予的私钥来解密数据信息，以保证数据不被两者之外的人或机构访问。可以看出区块链在保护用户隐私的同时还为研究机构深入研究特定人群的遗传规律提供一个安全平台。区块链中加密块的使用使得个人数据的修改及被恶性篡改的风险大大降低，从而为研究人员确保数据库的真实性。此外区块链技术还可以用于基因数据的管理，相关研究机构和企业通过获得准许证到基因链上存储其拥有的基因数据，这将能够避免伦理方面的问题。总之，随着基因技术的日趋成熟以及基因学临床数据的不断积累，在基因数据安全保护和应用方面会涌现更多深入的研究和全新方向。

3 基因数据安全使用与共享模式

3.1 国内

3.1.1 国家基因库 目前全球基因数据医疗领域的资金规模已超过 600 亿美元，其中基因精准诊断和基因精准治疗所占的资金规模分别约 100 亿美元和约 500 亿美元。全球精准医疗领域的增长速度达到 15%。我国“十三五”计划指出在 2030 年之前对精准医疗市场的投入资金将达到 600 亿元，这些资金由中央财政、地方财政、企业机构共同支付。在如此巨大的财政支持下，国内外对基因数据的使用分析能力与数据共享需求都在迅猛增长。面对如此庞大的市场，基因数据的隐私安全问题显得尤为重要。我国最具代表性的基因使用与共享的尝试是国家基因库（China National Genbank，CNGB）[22]。CNGB 于 2016 年 9 月 22 日正式对外运行，是目前我国首个获批筹建的国家级基因库，也是继美国的 GenBank[23]、日本的 DDBJ[24] 及欧盟的 EBI[25] 之后建成的战略级基因库。CNGB 管理用于研发的样本和数据共享，采取设置无限制和受控数据访问机制的方式，结合身份验证、分层访问控制和可审计的备案记录等技术手段。CNGB 只接受出于科研目的的访问请求，数据权限的管理和控制在数据提交者手中，数据提交者在提交数据时必须确定数据的受控范围，如果被设定为受控数据，则研究者必须向数据分析师协会（Certified Data Analyst Institute，CDA）提交数据权限申请，经 CDA 审批并授权后才可以下载和使用。CNGB 的监管体系采用大型国际数据库常规办法，CNGB 同意机构审查委员会（Institution Review Board，IRB）定期检查其已经获得批准的、涉及数据访问的项目。IRB 有权调查其中任何的负面事件并可以暂停或者终止违反访问条款或道德条例的项目。

3.1.2 推进行业规范的发展 2017 年 4 月华中科技大学与 CNGB 联合起草的《生物样本库样本/数据共享理论指南与管理规范》（征求意见稿）并对外发布，该指南明确界定样本或数据从收集、管理（存储安全、传输安全、使用安全和出境管理等）、国际研究合作、知识产权以及相关利益分配等的管理过程和规范。该指南参考包括国际生物与环境样本库协会（The International Society for Biological and Environmental Repositories，ISBER）相关实践及英国生物库（UK Biobank）伦理与治理框架在内的国际上遗传资源数据库和生物样本库的经验，同时还整合梳理国内有关管理部门的管理规定，为各种生物样本库的规范化管理奠定强有力的基础。另外指南中规定数据安全和隐患保护是处理数据时的安全准则。所有涉及人类样本或数据的相关项目均需要强制接受 IRB 的审查。同时该规范进一步规定跨境样本和数据共享的规则，数据的使用仅限于科学研究。尽管人类基因组计划完成多年，但人类基因组数据的医疗资源储存方式仍然是相互隔离的，为解决这一现状对精准医疗发展的制约，全球很多组织都在尝试打破隔离。在中国，尽管包括 CNGB 在内

的很多组织都在积极推进基因的共享和使用，其自身的规范也参考大量国际公认数据共享标准，对于涉及的跨境背景拥有一套完整的安全保护规范，然而除非得到中国人类遗传资源管理办的许可，目前所有共享仅限于中国境内使用。中国基因数据在国际间的共享仍处于初级阶段，而在国际上很多组织尝试不同的办法进行数据的共享和使用。

3.2 国外

全球基因组学与健康联盟（GA4GH）[26]是由生命科学研究机构、医疗机构以及研究型大学等组合成的联盟组织，主持发起制定基因组学和健康数据的共享框架。目的是为所有机构或个人提供、存储、访问、管理或使用基因组及健康相关数据。研究人员向指定医院发送数据查询指令，该指定医院来决定数据共享程度以及共享对象，通过该方法避免隐私方面的问题。其中 Beacon 项目是 GA4GH 在基因数据共享方面具有代表性的一个项目，重点在于联合全球具有数据共享意向的各大企业和研究机构，从而分享使用其基因数据库，建立具有信息安全性、使用简便的国际信息共享数据库。Beacon 项目设计一个简单的网络平台，任何使用者都可以在不违反隐私规则的条件下提出其他实验室所掌握的基因组数据的相关问题，使用者可以发出类似“你是否有一个基因包含‘A’在 3 号染色体的位点 100，735 处?”的问题，得到“Yes”或“No”的答复。每愿意提供类似这种平台服务的机构都被称作 Beacon。针对难以收集数据的罕见病或者有强遗传倾向的家族疾病的研究，由于此类研究涉及的基因往往具有极强的特异性，需通过重复询问，可以唯一定位某个持有罕见基因的人在该平台中的风险是否存在。GA4GH 目前也在推行所有者同意书，该同意书对基因组数据提供者所享有的权利做出明确规定，与其他大多数同意书相比，该同意书允许全球范围内的研究人员进行受控访问。如果某个机构查询的问题多次涉及同一个人，则认为该机构在有意探寻该人的隐私，将封锁该机构的查询权限。同时一些相关隐私算法的研究也在进行中，通过变更阈值，随机反转，加密交换的方式来保障个人隐私安全。

4 基因医疗数据安全领域存在的问题和建议

4.1 尽快建立和完善相关法律法规及监管机制

基因精准医疗的核心是基因数据库的建立，然而在建立基因数据库的过程中，涉及个人基因数据隐私、伦理的相关问题也会随之产生。由于基因精准医疗处于起步阶段，相应的技术标准、共享平台、法律法规还没有建立起来，在使用、保存、传输基因数据时有极大的泄露风险。现今各国都在积极探索相应的法律条文，2016 年美国食品药物管理局（Food and Drug Administration，FDA）颁布基于下一代测序（Next Generation Sequencing，NGS）技术的设计、开发及检测结果诊断标准指南，规定相关研究机构要严格遵守 FDA 标准分析检测结果的有效性，尽量减少错误结果。我国针对基因检测方面也颁布相关法规条文，如《药物代谢酶和药物作用靶点基因检测技术指南（试行）》、《肿瘤个体化治疗检测技术指南（试行）》等。然而法律规定只是在技术层面的规范指导，缺乏确切的法律方面的监管与规范。为使基因精准医疗有条不紊的发展，国家政府应明确国家卫健委和食品药品管理局在精准医疗领域的相应监督职责并进一步细化相关法案。

4.2 建立基因精准医疗数据共享平台

精准医疗的基础是数据的累积，在数据安全的前提下应建立精准医疗基因数据共享平台。我国在建立平台时可以参考美国 FDA 与 DNAnexus 生物信息公司构建的精准医疗 FDA 平台[27]，该平台为新型的基因测序研究提供云工具，可以帮助研究者上传临床验证成果和共享基因数据信息，其他研究机构也可以在该平台上调用、验证、分享其他人或机构的研究成果。构建平台时会涉及数据整合标准、信息安全构架以及规范、平台基础构架技术体系、大数据分析技术。国家层面相应标准规范以及技术发展支持应建立在精准医疗基因方面，另外通过现有的电子病历系统，共同加入基因测序数据信息，为建立精准医疗基因数据共享平台奠定基础。也可

以在现有的电子病历系统基础上加入基因测序数据信息，建立标准化、结构化和统一编码的电子病历数据共享系统。

4.3 推动基因组数据安全共享算法研究

基因组学领域发展面临的问题在于已收集的大量数据难以共享，其中一个关键因素是数据所占的存储空间，基因自身大小导致很多问题，如单人的全基因数据大小可达 100G 左右，即使是原始数据也有 10G 左右。然而 DNA 序列具有不同于其他数据的序列特征，导致目前通用的数据压缩算法[28]难以进行有效压缩，其时间和空间代价很大，因此研究基因序列压缩算法对于基因数据的使用和共享具有重要意义。此外患者的隐私保护也是基因数据共享过程中无法回避的问题，因为个人基因组数据所含有的信息与个人和其家庭密切相关，除在法律和安全共享平台方面进行规范外，在数据共享安全算法方面也应展开深入研究。除区块链研究方向外，可搜索加密技术也是保护用户隐私的方向[29-30]。传统的搜索算法是基于明文的技术，这个过程中不论是查询者提交的查询字段，还是服务器数据库中的信息数据均是以明文的形式出现的，这种情况极容易造成信息泄露，从而侵害个人数据信息安全。可搜索加密技术是用密码学技术在密文的形式下进行搜索查询，但该技术在大规模应用方面需要深入研究。

5 结语

总的来说基因数据的安全既要国家政府在法律层面进行规范化，也需要在技术层面深入研究。目前我国政府虽然对基因数据隐私保护进行规定，但是现阶段还没有建立起完整的基因数据隐私安全立法系统，涉及的基因隐私法律分散于法律及行政规范中，缺少层次性、针对性及统一性。另外在安全技术层面的研究也有待深入。

参考文献

1 Sawicki M P, Samara G, Hurwitz M, et al. Human Genome Project [J]. The American Journal of Surgery, 1993, 165 (2): 258-264.

2 Collins F S, Patrinos A, Jordan E, et al. New Goals for the US Human Genome Project: 1998-2003 [J]. Science, 1998, 282 (5389): 682-689.

3 Goodman L. The Human Genome Project Aims for 2003 [J]. Genome research, 1998, 8 (10): 997-999.

4 Wang H J, Quigley G J, Kolpak F J, et al. Molecular Structure of a Left-handed Double Helical DNA Fragment at Atomic Resolution [J]. Nature, 1979, 282 (5740): 680.

5 Sorenson J R, Cheuvront B. The Human Genome Project and Health Behavior and Health Education Research [J]. Health Education Research, 1993, 8 (4): 589-593.

6 Vasemägi A, Primmer C R. Challenges for Identifying Functionally Important Genetic Variation: the promise of combining complementary research strategies [J]. Molecular Ecology, 2005, 14 (12): 3623-3642.

7 Lloyd A L, Rasko D A, Mobley H L T. Defining Genomic Islands and Uropathogen-specific Genes in Uropathogenic Escherichia Coli [J]. Journal of Bacteriology, 2007, 189 (9): 3532-3546.

8 Feil E J, Li B C, Aanensen D M, et al. eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data [J]. Journal of Bacteriology, 2004, 186 (5): 1518-1530.

9 Stranger B E, Forrest M S, Dunning M, et al. Relative Impact of Nucleotide and Copy Number Variation on Gene Expression Phenotypes [J]. Science, 2007, 315 (5813): 848-853.

10 Joyner M J. Precision Medicine, Cardiovascular Disease and Hunting Elephants [J]. Progress in Cardiovascular Diseases, 2016, 58 (6): 651-660.

11 Ahrendt S A, Decker P A, Alawi E A, et al. Cigarette Smoking is Strongly Associated with Mutation of the K-ras Gene in Patients with Primary Adenocarcinoma of the Lung [J]. Cancer, 2001, 92 (6): 1525-1530.

12 Ahrendt S A, Decker P A, Doffek K, et al. Microsatellite Instability at Selected Tetranucleotide Repeats is Associated with p53 Mutations in Non-small Cell Lung Cancer [J]. Cancer Research, 2000, 60 (9): 2488-2491.

13 Raphael J, Verma S, Hewitt P, et al. The Impact of Angelina Jolie (AJ)'s Story on Genetic Referral and Testing at an Academic Cancer Centre in Canada [J]. Journal of Genetic

Counseling, 2016, 25 (6): 1309 - 1316.

14 Cavallaro S, Paratore S, de Snoo F, et al. Genomic Analysis: toward a new approach in breast cancer management [J]. Critical Reviews in Oncology/Hematol - ogy, 2012, 81 (3): 207 - 223.

15 Insel T R. The NIMH Research Domain Criteria (RDoC) project: precision medicine for psychiatry [J]. American Journal of Psychiatry, 2014, 171 (4): 395 - 397.

16 彭霞. 基因专利与基因资源的法律保护 [J]. 科技管理研究, 2010, 30 (7): 191 - 193.

17 Tomes J P. The Health Insurance Portability and Accountability Act of 1996: understanding the anti - kickback laws [J]. Journal of Health Care Finance, 1998, 25 (2): 55 - 62.

18 Mantelero A. The EU Proposal for a General Data Protection Regulation and the Roots of the'right to be forgotten' [J]. Computer Law & Security Review, 2013, 29 (3): 229 - 235.

19 Sweeney L. K - anonymity: a model for protecting privacy [J]. International Journal of Uncertainty, Fuzziness and Knowledge - based Systems, 2002, 10 (5): 557 - 570.

20 Lin C, Wang P, Song H, et al. A Differential Privacy Protection Scheme for Sensitive Big Data in Body Sensor Networks [J]. Annals of Telecommunications, 2016, 71 (9 - 10): 465 - 475.

21 Yue X, Wang H, Jin D, et al. Healthcare Data Gateways: found healthcare intelligence on blockchain with novel privacy risk control [J]. Journal of Medical Systems, 2016, 40 (10): 218.

22 Jia G, Shi S, Wang C, et al. Molecular Diversity and Population Structure of Chinese Green Foxtail [Setaria viridis (L.) Beauv.] Revealed by Microsatellite Analysis [J]. Journal of Experimental Botany, 2013, 64 (12): 3645 - 3656.

23 Link M P, Hagerty K, Kantarjian H M. Chemotherapy Drug Shortages in the United States: genesis and potential solutions [J]. Journal of Clinical Oncology, 2012, 30 (7): 692 - 694.

24 Miyazaki S, Sugawara H, Gojobori T, et al. DNA Data Bank of Japan (DDBJ) in XML [J]. Nucleic Acids Research, 2003, 31 (1): 13 - 16.

25 Brazma A, Parkinson H, Sarkans U, et al. ArrayExpress—a public repository for microarray gene expression data at the EBI [J]. Nucleic Acids Research, 2003, 31 (1): 68 - 71.

26 Knoppers B M. International Ethics Harmonization and the Global Alliance for Genomics and Health [J]. Genome Medicine, 2014, 6 (2): 13.

27 Xu J, Gong B, Wu L, et al. Comprehensive Assessments of RNA - seq by the SEQC Consortium: FDA - led efforts advance precision medicine [J]. Pharmaceutics, 2016, 8 (1): 8.

28 Lee S J, Kim J, Lee M. A Real - time ECG Data Compression and Transmission Algorithm for an E - health Device [J]. IEEE Transactions on Biomedical Engineering, 2011, 58 (9): 2448 - 2455.

29 Abdalla M, Bellare M, Catalano D, et al. Searchable Encryption Revisited: consistency properties, relation to anonymous IBE, and extensions [J]. Journal of Cryptology, 2008, 21 (3): 350 - 391.

30 Curtmola R, Garay J, Kamara S, et al. Searchable Symmetric Encryption: improved definitions and efficient constructions [J]. Journal of Computer Security, 2011, 19 (5): 895 - 934.