生物审编组织管理模式及证据源质量控制和评价*

苏　燕

（中国科学院上海生命科学信息中心/
中国科学院上海营养与健康研究所/
中国科学院上海生命科学研究院　上海 200031）

凌鋆超

（中国科学院计算生物学重点实验室/中国科学院马普学会计算生物学伙伴研究所生物医学大数据中心/
中国科学院上海营养与健康研究所/中国科学院
上海生命科学研究院　上海 200031）

李祯祺

（中国科学院上海生命科学信息中心/
中国科学院上海营养与健康研究所/
中国科学院上海生命科学研究院　上海 200031）

张国庆

（中国科学院计算生物学重点实验室/中国科学院马普学会计算生物学伙伴研究所生物医学大数据中心/
中国科学院上海营养与健康研究所/中国科学院
上海生命科学研究院　上海 200031）

徐　萍

（中国科学院上海生命科学信息中心/中国科学院上海营养与健康研究所/
中国科学院上海生命科学研究院　上海 200031）

〔摘要〕　阐述国内外生物审编工作开展现状，从审编组织管理模式、证据源质量控制和评价两方面分析生物审编工作当前面临的主要挑战，提出我国发展生物审编相关建议，推进生物审编工作科学开展。

〔关键词〕　生物审编；组织模式；数据库；效率；质量
〔中图分类号〕R－056　　〔文献标识码〕A　　〔DOI〕10.3969/j.issn.1673－6036.2020.01.007

Biocuration Organization and Management Mode and Evidence Source Quality Control and Evaluation　*SU Yan, Shanghai Information Center for Life Sciences/Shanghai Institute of Nutrition and Health/Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China; LING Yunchao, CAS－Key Laboratory of Computational Biology/Bio－Med Big Data Center, CAS－MPG Partner Institute for Computational Biology/Shanghai Institute of Nutrition and Health/Shanghai Institutes for Biological Sci-*

〔修回日期〕　2019－07－22
〔作者简介〕　苏燕，馆员，发表论文 10 余篇；通讯作者：徐萍，研究馆员。
〔基金项目〕　国家重点研发计划“精准医学知识自动化注释与人工审编”（项目编号：2016YFC0901904）；中国科学院战略研究与决策支持系统建设专项“我国生物大数据发展的技术挑战分析”（项目编号：GHJ－ZLZX－2019－05－3）。

ences, Chinese Academy of Sciences, Shanghai 200031, China; LI Zhenqi, Shanghai Information Center for Life Sciences/Shanghai Institute of Nutrition and Health/Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China; ZHANG Guoqing, CAS – Key Laboratory of Computational Biology/Bio – Med Big Data Center, CAS – MPG Partner Institute for Computational Biology/Shanghai Institute of Nutrition and Health/Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China; XU Ping, Shanghai Information Center for Life Sciences/Shanghai Institute of Nutrition and Health/Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China

〔**Abstract**〕 The paper expounds on the current status of the development of biocuration at home and abroad, analyzes the current major challenges of biocuration from two aspects of organization and management mode, quality control and evaluation of evidence sources, puts forward relevant suggestions for the development of biocuration in China, to promote the scientific implementation of biocuration.

〔**Keywords**〕 biocuration; organization mode; database; efficiency; quality

1 引言

随着生物医学研究的飞速发展，生物医学数据呈现指数级增长。如何将海量异构数据收集整理成规范统一、高质量、可高效利用的知识成为亟待解决的问题。生物审编作为提升数据价值的重要方式，已经被国际大型公共和商业生物医学数据库广泛使用。国际生物审编学会将生物审编定义为：将生物相关信息转化集成到数据库或数据资源中，同时整合科学文献和大数据集，准确全面地描述生物知识，便于科研人员获取相关信息以及利用计算机进行数据分析，包括从杂乱数据集中提取和组织生物学和临床数据，录入到一个用户友好的数据库中。生物审编的发展历程，见图 1。经审编的数据一方面成为科研人员重要的参考资源，另一方面已被大规模用于文本挖掘或信息学分析，如生物信息学、神经信息学、卫生信息学等[1]。

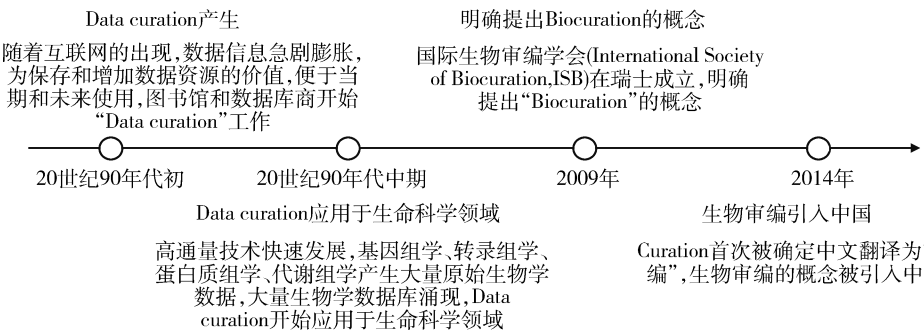

图 1 生物审编发展历程

2017 年 3 月欧洲药品监管机构负责人组织（Heads of Medicines Agencies，HMA）和药品管理局（European Medicines Agencies，EMA）成立大数据联合工作组，针对基因组学、蛋白质组学、临床试验等数据的标准化、质量、共享、连接、分析、监管等提出建议。2018 年 6 月美国国立卫生研究院（National Institutes of Health，NIH）发布《数据科学战略计划》，旨在对生物医药研究产生的海量数据进行存储和管理并进行标准化建设和数据公开。

2 国内外生物审编工作开展现状

2.1 国际数据库广泛开展

生物审编已在国际大型公共和商业生物医学数据库广泛使用，成为数据/知识抽取和标准化管理重要方式。公共大型生物医学数据库，如美国国立生物技术信息中心（National Center of Biotechnology Information，NCBI）开发的基因变异数据库 ClinGen，欧洲生物信息研究所（European Bioinformatics Institute，

EBI）开发的蛋白质相互作用数据库 IntAct，生物学通路数据库 Reactome，生物相关的化学实体数据库 ChEBI，基因本体、线虫模式生物数据库 WormBase，西班牙国家生物技术中心开发的蛋白质相互作用数据库 iHOP 等均采用审编方式提升数据附加值。同时经过审编的高质量数据的商业价值已经引起企业关注，如 GeneGo、IPA 和 Pathway Studio 等商业数据库通过自然语言处理技术从文档中提取信息和知识，聘请专业人士进行判读，提升数据价值，保证知识的可靠性。但这些商业软件核心数据保密且使用价格昂贵，在生物医学数据方面形成垄断。

2.2 国内刚刚起步

我国在生物审编领域研究几乎处于空白状态，目前仅有中国科学院北京基因组研究所[2]、中国医学科学院医学信息研究所[3]等机构对生物审编进行了探索性研究。随着大数据时代的到来以及我国自主开发意识的增强，国内各类生物医学知识库构建相继开展。尤其是 2016 年国家开始密集布局重大慢病、精准医学等领域的重点研发计划，依托这些计划搭建国家级生物医学知识库，旨在打破国际垄断，保护数据安全。

3 生物审编工作当前面临的主要挑战与建议

3.1 概述

虽然国际上已研发出 PubTator 等自动审编工具，在一定程度上提高了生物审编效率[4]，但受限于生物知识和文本的复杂性，目前审编工作仍以专业人员人工解读为主。生物审编是一项大体量、高耗时工作[5]，需要投入大量的人力和财力资源[6]，其开展和持续目前主要面临两方面挑战。一是国际尚无统一、规范的生物审编组织管理模式，以保障审编工作的质量和效率，审编人员的能力素养和个人偏好等因素容易引发审编质量和效率的差异。二是生物审编的数据来源于不同文本，其质量参差不齐，直接影响审编后生成知识的可信度，同时也给其开展造成一定困难。如何解决上述问题，实现审编过程的高质和高效管理成为其开展和持续运行的关键。

3.2 审编组织管理模式

3.2.1 国际 目前尚无统一、规范的生物审编组织管理模式。国际上的数据库建设机构主要采用成立联盟、与协会合作、与期刊合作、开放式群体审编等组织模式，根据自身资源和优势设计适应性管理模式。国内外生物审编组织管理模式，见图 2。国际分子交换联盟（International Molecular Exchange，IMEx）成员包括 UniProt、IntAct 等 16 家数据库，其审编模式是联盟成员承诺认领数量不等的期刊进行审编。为减少审编人员个体差异造成的审编质量和效率差异，IMEx 联盟安排审编人员进行跨库交叉培训，促进不同机构间的审编人员交流，缩小机构以及人员间的审编差异。IMEx 联盟编制了面向蛋白质互作审编的 IMEx 审编规则（IMEx Curation Rules），对证据源、审编字段、更新方式等作了详细规定[7]。线虫模式生物数据库 WormBase 通过与美国遗传学会（Genetics Society of America，GSA）和达特茅斯期刊服务机构合作获取全文资源，引导作者自行审编。2010 年作者反馈率约 40%，其中 75% 的作者进行了较为详细的注释。IBM 人工智能系统 Watson 的审编过程管理中设计了 4 类职位：团队领导、数据专家、内容审编人员和领域专家。团队领导负责管理和监督 Watson 审编项目，启动新的任务工作；数据专家负责查找与收集相关符合目的、范围、标准的内容，利用计算机技术进行内容分类；内容审编人员负责指导数据专家收集内容、审编收集的内容、指导领域专家对审编内容进行分类和改进；领域专家利用专业知识对收集的内容进行质量控制，评估内容的有效性、准确性和价值。通路数据库 IPA 聘请 500 名博士开展审编工作，通过多轮审编进行质控。

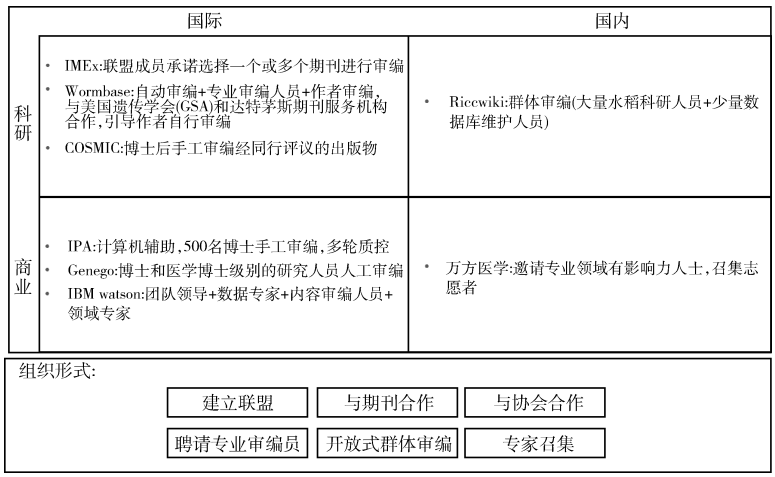

图 2 国内外生物审编组织管理模式

3.2.2 国内 中科院北京基因组所开发的 RiceWiki 数据库采用开放式群体审编的模式。开放式群体审编不同于作者审编和聘用专业审编员审编，其审编员范围广泛，线上协作审编的方式不受时间、地域限制，为审编工作的大规模开展提供可能。但也存在审编员个体差异大以致于审编质量和效率参差不齐的问题。RiceWiki 利用 AuthorReward 评分明确审编人员的贡献程度，该系统从一定程度上对审编人员进行评价但并未从根本上解决人员造成的审编质量和效率差异问题[8]。此外国内的商业数据库也开始尝试通过审编构建知识库，如万方基于其文献资源优势构建万方医学知识库，其组织模式为邀请专业领域内有影响力人士，由其召集志愿者进行审编。国家重点研发计划“疾病研究精准医学知识库”项目设计了基于文本挖掘的自动与人工结合、数量与质量并重的审编模式，见图 3。

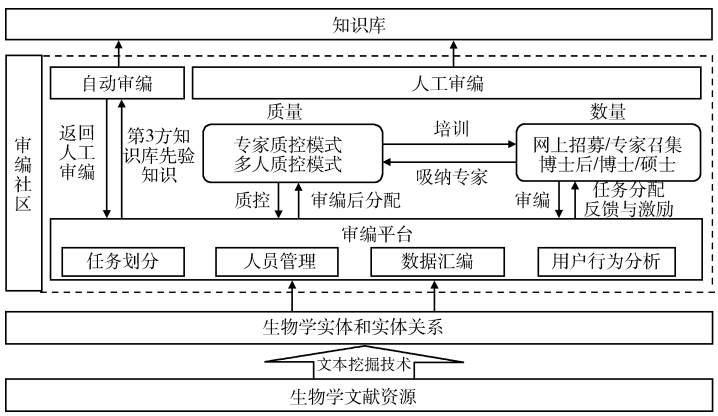

图 3 自动与人工结合、数量与质量并重的审编模式

文本挖掘生成的生物学实体和实体关系数据汇集到审编平台，首先比照第 3 方数据库的先验知识进行自动审编，与第 3 方数据库匹配的直接输出到知识库，不匹配的返回审编平台进行人工审编。在审编数量控制上采用网上招募或者专家召集博士后、博硕士研究生的模式，保障对审编员数量的需求，同时对审编结果进行一定的反馈和激励。在质量控制上通过专家质控、多人质控模式，不合格的审编条目返回审编平台重新进行任务分配。同时组织专家对审编人员进行线上/线下培训，将审编质量较高的人员吸纳为审编专家。

3.3 证据源质量控制和评价

生物医学领域的文本数量飞速增长，文本中

研究结果的可信度不高，低质量的文本往往造成审编信息冗余甚至错误。同时在文本中信息的呈现形式极大影响生物数据的识别和审编速度[9]。生物审编的证据来源于不同文本，其质量参差不齐，从根本上影响了审编后生成知识的可信度。因此一方面需要对证据源加以一定控制，保障基于审编的数据库质量，另一方面需要对审编生成的数据进行可信度评价，帮助数据库用户快速识别入库数据的可信度。目前许多数据库都采取一定措施从证据源角度进行审编质量控制。肿瘤突变信息数据库 COSMIC 采用经同行评议的出版物为证据源。线虫模式生物数据库 WormBase 证据源筛选主要依赖审编人员的主观判断，后台系统首先利用 Perl 脚本在 PubMed 数据库中对关键词“elegans”进行检索，检索获得的文献再通过人工判读摘要或全文，分析文献信息数量和质量进而决定是否采纳该证据源。WormBase 数据库每年审编文献量约 1 200 篇[10]。此外 ClinGen 数据库针对提交数据的实验室发布数据质量标准，仅收录满足特定要求的实验室数据[11]。在可信度评价方面 ClinGen、CIViC 等都已建立证据分级评价模型，为数据库用户提供直观的可信度展示。ClinGen 根据试验方法、试验材料、分析方法等因素对证据进行分级[12]。肿瘤基因变异数据库 CIViC 数据库分别建立证据评级和可信度评级两种评价模式。证据评级赋予共识/指南、临床试验、个案报道、试验模型等证据由高到低的分级，可信度评级则根据期刊影响、研究规模、再现性等因素进行分级[13]。国际分子交换联盟面向蛋白质互作关系设计证据评分模型，利用 MIscore 工具对审编获得的蛋白之间的相互作用进行可靠性评价，评分因素包括文献数量、实验方法类型和相互作用类型[14]。

3.4 我国发展生物审编相关建议

在审编工作组织管理方面我国已发展一批具有国际影响力的生物医学期刊，包括《细胞研究》（*Cell Research*）、《分子细胞生物学报》（*Journal of Molecular Cell Biology*）等，发挥政府、学术团体、期刊的联动能力，组织引导作者在投稿过程中共享数据和开展审编是发展我国自有生物医学数据/知识库的直接、有效途径。同时通过重大专项等形式持续支持生物医学数据/知识库建设，招募和培育专业审编人才，规模化开展审编工作是发展我国自有生物医学数据/知识库的必由之路。在审编工作的质量和效率控制方面，利用计算机技术辅助审编已成为提高审编效率的重要方式，应积极推动生物医学本体构建、异构数据整合、文本挖掘等研究，支持和引导国内生物医学信息系统采用统一通用的数据标准，打破数据孤岛，通过自动或半自动审编技术和工具提高审编效率。探索生物审编过程管理机制，制定标准、规范的审编流程，保障生物审编工作的科学开展，推进审编质量和效率的双重提升。

4 结语

生物数据已被视为重要的国家战略资源，美、欧、日等国家在 20 世纪就已经布局数据的收集、存储、审编和利用，美国 NCBI、欧洲 EBI 等数据中心以及 GeneGo、IPA 和 Pathway Studio 等商业数据库已对生物数据形成垄断。近年来我国高度重视生物数据资源的开发和利用，中科院北京基因组研究所生命与健康大数据中心、中科院上海生科院生物医学大数据中心、北科生物国家生物医学大数据产业园等多个生物大数据中心、集群、平台也相继建成。但我国在生物审编方面刚刚起步，生物数据缺乏有效积累与管理，面临严峻的数据资源风险，亟需探索生物审编工作的支持和激励机制，充分挖掘数据资源，打破数据垄断，保障科研、临床对数据资源的需求。

参考文献

1 Mitchell C S, Cates A, Kim R B, et al. Undergraduate Biocuration: developing tomorrow's researchers while mining today's data [J]. Journal of Undergraduate Neuroscience Education, 2015, 14 (1): A56 - A65.

2 Zhang Z, Zhu W, Luo J. Bringing Biocuration to China [J]. Genomics, Proteomics & Bioinformatics, 2014, 12 (4): 153 - 155.

3 杨林, 李姣, 侯丽, 等. 跨领域数据审编（Curation）流程研究——以环境健康数据为例 [J]. 现代图书情报技术, 2016, 31 (12): 80 - 88.
4 Poux S, Arighi C N, Magrane M, et al. On Expert Curation and Scalability: UniProtKB/Swiss - Prot as a case study [J]. Bioinformatics, 2017, 33 (21): 3454 - 3460.
5 Dauga D. Biocuration: a new challenge for the tunicate community [J]. Genesis, 2015, 53 (1): 132 - 142.
6 International Society for Biocuration. Biocuration: distilling data into knowledge [J]. PLoS Biology, 2018, 16 (4): e2002846.
7 Orchard S, Kerrien S, Abbani S, et al. Protein Interaction Data Curation: the International Molecular Exchange (IMEx) consortium [J]. Nature Methods, 2012, 9 (4): 345.
8 Zhang Z, Sang J, Ma L, et al. RiceWiki: a wiki - based database for community curation of rice genes [J]. Nucleic Acids Research, 2014, 42 (D1): 1222 - 1228.
9 Howe D, Costanzo M, Fey P, et al. Big data: the future of biocuration [J]. Nature, 2008, 455 (7209): 47.
10 Van Auken K, Fey P, Berardini T Z, et al. Text Mining in the Biocuration Workflow: applications for literature curation at WormBase, dictyBase and TAIR [J]. Database, 2012 (2012): 1 - 13.
11 ClinGen. Clinical Laboratories Meeting Minimum Requirements for Data Sharing to Support Quality Assurance [EB/OL]. [2019 - 01 - 10]. https://www.clinicalgenome.org/lablist/.
12 Strande N T, Riggs E R, Buchanan A H, et al. Evaluating the Clinical Validity of Gene - disease Associations: an evidence - based framework developed by the Clinical Genome Resource [J]. The American Journal of Human Genetics, 2017, 100 (6): 895 - 906.
13 Griffith M, Spies N C, Krysiak K, et al. CIViC is a Community Knowledgebase for Expert Crowdsourcing the Clinical Interpretation of Variants in Cancer [J]. Nature Genetics, 2017, 49 (2): 170.
14 Villaveces J M, Jimenez R C, Porras P, et al. Merging and Scoring Molecular Interactions Utilising Existing Community Standards: tools, use - cases and a case study [J]. Database, 2015 (2015): 1 - 12.

（上接第 33 页）

级。医学生应尽可能避免使用本人姓名的拼音、缩写或生日等方式来申请和设置社交网络账户密码，减少用同一密码申请多个 APP 应用的情况，同时还应避免传送个人网络账号密码等。最后主动接受信息安全教育。研究表明接受过信息安全培训的医学生具有信较强的信息安全意识，也有较好的个人信息保护能力[5]。

5.3 大数据服务商应改善移动社交网络的安全性

良好的行业规范和自律公约是一个行业蓬勃发展的重要条件。大数据服务商应加强行业自律，规范企业行为，提高准入标准，做好成员资质审核。此外要积极创新，优化网络结构，做好信息保护，规范数据采集行为，保证用户个人信息的安全和自由。

6 结语

大数据是当前信息产业发展的基础，也是其未来发展的重要方向，做好大数据的技术研究对于提高信息技术质量具有十分重要的作用。而在这一过程中怎样应对医学生个人隐私信息安全所面临的挑战成为重要研究内容。这一研究的开展不但有助于大数据背景下医学生网络信息安全性的提高，也为大数据技术提供更为广阔的发展空间。

参考文献

1 吴静梅. 大数据下大学生移动社交网络个人信息保护研究 [J]. 智库时代, 2018 (45): 208 - 209.
2 封二英, 齐晓娜, 张宇敬. 大数据下大学生移动社交网络个人信息保护研究 [J]. 产业与科技论坛, 2018, 17 (1): 39 - 40.
3 殷国伟. 网络环境下高校学生个人信息资料的法律保护 [J]. 湖州师范学院学校, 2009, 31 (3): 73 - 77.
4 维克托·迈尔·舍恩伯格, 肯尼思·库克耶. 大数据时代：生活, 工作与思维的大变革 [M]. 杭州：浙江人民出版社, 2012.
5 王来兵. 大数据时代社交网络个人信息安全问题研究 [J]. 赤峰学院学报（自然版）, 2017, 33 (13): 19 - 20.