Genomics Proteomics Bioinformatics 12 (2014) 153–155

HOSTED BY

ELSEVIER

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

PERSPECTIVE

Bringing Biocuration to China

Zhang Zhang [1,*], Weimin Zhu [2,3], Jingchu Luo [4]

[1] *CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China*
[2] *Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences, Beijing 100730, China*
[3] *Taicang Institute of Life Sciences Information, Taicang 215400, China*
[4] *College of Life Sciences and Center for Bioinformatics, Peking University, Beijing 100871, China*

Received 30 June 2014; accepted 10 July 2014
Available online 17 July 2014

KEYWORDS

Biocuration;
Integration;
Big data;
Database

Abstract Biocuration involves adding value to biomedical data by the processes of standardization, quality control and information transferring (also known as data annotation). It enhances data interoperability and consistency, and is critical in translating biomedical data into scientific discovery. Although China is becoming a leading scientific data producer, biocuration is still very new to the Chinese biomedical data community. In fact, there currently lacks an equivalent acknowledged word in Chinese for the word “curation”. Here we propose its Chinese translation as “审编” (Pinyin: *shěn biān*), based on its implied meanings taken by biomedical data community. The 8th International Biocuration Conference to be held in China (http://biocuration2015.tilsi.org) next year bears the potential to raise the general awareness in China of the significant role of biocuration in scientific discovery. However, challenges are ahead in its implementation.

Introduction

With the rapid advancement of high-throughput sequencing technologies and the resulting accumulation of large-scale data, biomedical databases have integrated a wealth of highly-diversified information from omics studies, such as genomics, transcriptomics, proteomics, epigenomics and phenomics, which provide essential reference information on genes, proteins, and their sequence, structure and expression. According to the 2014 database collection issue of Nucleic Acids Research, a total of 1552 published biomedical databases are available online [1], which represent only a fraction of all biomedical databases. Together, these databases are the ensemble of research results in life sciences, serving as an information pool for wet-lab biologists to validate their research results and develop new hypotheses and for bioinformaticians to mine new biomedical knowledge. As researchers become increasingly dependent on databases, the proper management of these data becomes a crucial task.

What is biocuration?

In order to make biomedical databases relevant and useful, great efforts are required to keep data in these databases accurate, comprehensive and up-to-date. A biocurator is a data specialist whose entrance level normally requires PhD training in biomedical fields with at least 2–3 years research experience. A typical biocuration process involves the following steps: (1) the administration of raw and meta data from a submitter to ensure the data integrity and syntactic interoperability;

[*] Corresponding author.
E-mail: zhangzhang@big.ac.cn (Zhang Z).

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

http://dx.doi.org/10.1016/j.gpb.2014.07.001

(2) the exhaustive search of literature to extract relevant information associated with the submitted data; (3) the formal representation of the terms and their relationship using ontology and controlled vocabularies to ensure their semantic interoperability with other databases; (4) the identifier and/or term mapping between databases to annotate and enrich new data with known knowledge; and finally, (5) the loading of the information into a database to make them ready for targeted users. As biocuration is an indispensable process to use and reuse biomedical data, it has become an integral part of med/bioinformatics and is critical in translating big biomedical data into big scientific discovery.

The 2015 International Biocuration Conference in China and its significance

The International Society for Biocuration (ISB; http://www.biocurator.org) is a not-for-profit organization with more than 300 registered members. These members come from a wide range of biomedical databases including those from 3 member databases of International Nucleotide Sequence Database Collaboration (INSDC; http://www.insdc.org): GenBank (http://www.ncbi.nlm.nih.gov/genbank), European Nucleotide Archive (ENA; http://www.ebi.ac.uk/ena/) and DNA Data Bank of Japan (DDBJ; http://www.ddbj.nig.ac.jp). Since 2005, seven International Biocuration Conferences have been organized aiming to provide a forum for knowledge exchange, project promotion and collaboration fostering among the biocuration community. It was recently announced that the 8th International Biocuration Conference will be held in Beijing, April 23–26, 2015. As of July 10, 2014, four distinguished data scientists, Dr. David Lipman, Director of the National Center for Biotechnology Information (NCBI), Dr. Alex Bateman, the Head of Protein Sequence Resources at the European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI), Dr. Amos Bairoch, the Head of neXtProt and Director of the Department of Human Protein Science at University of Geneva, and Dr. Takashi Gojobori, Distinguished Professor of Bioscience in King Abdullah University of Science and Technology, have confirmed to give keynote speeches. Further details of the conference can be found at http://biocuration2015.tilsi.org and http://www.biocurator.org.

Although the value of biocuration is evident among mature biomedical databases, biocuration is still very new in China, as testified by the lack of an acknowledged equivalent word in Chinese for "curation" and the fact that biocurator as a profession is still unknown to many. Here we suggest that the Chinese translation of curation is "审编" (Pinyin: *shěn biān*), based on the implied meanings taken by biomedical data community. It is our strong belief that bringing the International Biocuration Conference to China will raise the general awareness of the important role of biocuration among data scientists and bioinformaticians, as well as funding agencies [2].

Challenges in bringing biocuration to China

China is becoming a powerhouse in producing biological data. However, very few biocuration projects aiming at adding and bringing up values of these data have been funded in China. In addition, there is no existence of biocurator as a profession, for their work is behind the scene, and current academic system is not compatible in evaluating their work of "no authorship" [3,4]. A general awareness and long-term funding support are critical for the Chinese biocuration community to be born, grow, mature and flourish.

The interdisciplinary and collaborative feature of biocuration poses additional challenges for biocurators in coordinating between data producers from wet labs, bioinformaticians who analyze data, informatics specialists who make data professionally manageable and accessible and computational biologists who translate data into knowledge. It requires wide recognition of this crucial role of biocurators among research community in China to truly cash on what has been invested in biomedical research.

Considering a big gap between the deluge of biomedical data and a limited number of expert biocurators available,

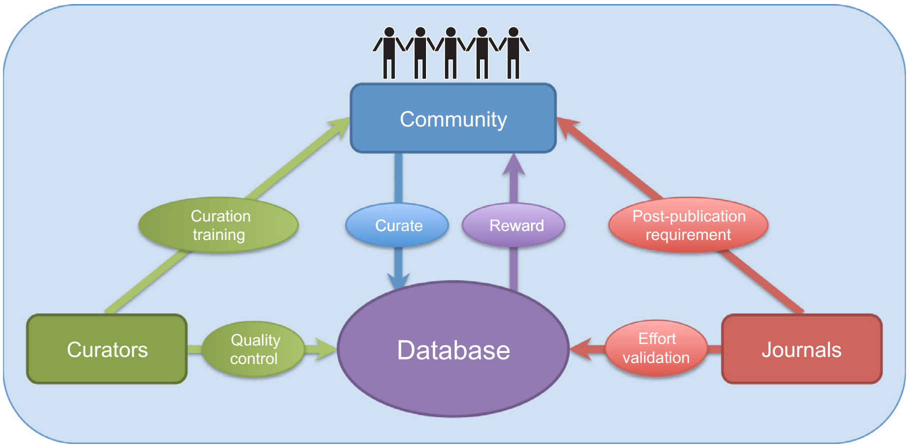

Figure 1 A community-contributed, contribution-rewarded, expert curator-validated and journal-involved model for big biological data curation

The community can perform curation and would be rewarded by contribution quantification and explicit authorship. Curators provide curation trainings for the community and conduct quality control for curated information. Journals, for one thing, require authors submitting data to a relevant database as a compulsory post-publication process, and for another, validate their efforts made to the database.

community curation is becoming a viable alternative complement to expert curation. Community and journals are also playing critical roles in facilitating biocuration on "big data" from omics studies [4]. A model of community curation is proposed (**Figure 1**) to curate big data [5,6]. In order to make this model practical and sustainable, community-curation efforts should be rewarded by giving explicit authorship based on quantified contributions [7]. This community-contributed, contribution-rewarded, expert curator-validated and journal-involved model would make it possible to accurately and effectively curate big omics data. "Union makes strength", as a Chinese saying has put. China has an advantage in this model with a large population of researchers and students available to make these joint efforts.

Acknowledgements

We thank Dr. Pascale Gaudet from the Swiss Institute of Bioinformatics and Dr. Alex Bateman from the European Bioinformatics Institute for constructive comments on this manuscript. This work was supported by grants from the National High Technology Research and Development Program (863 Program, Grant No. 2014AA021503 and 2012AA020409) of China and the "100-Talent Program" of Chinese Academy of Sciences of China.

References

[1] Fernández-Suárez XM, Rigden DJ, Galperin MY. The 2014 nucleic acids research database issue and an updated NAR online molecular biology database collection. Nucleic Acids Res 2014;42: D1–6.

[2] Baker M. Databases fight funding cuts. Nature 2012;489:19.

[3] Bateman A. Curators of the world unite: the International Society of Biocuration. Bioinformatics 2010;26:991.

[4] Howe D, Costanzo M, Fey P, Gojobori T, Hannick L, Hide W, et al. Big data: the future of biocuration. Nature 2008;455:47–50.

[5] Waldrop M. Big data: wikiomics. Nature 2008;455:22–5.

[6] Zhang Z, Sang J, Ma L, Wu G, Wu H, Huang D, et al. RiceWiki: a wiki-based database for community curation of rice genes. Nucleic Acids Res 2014;42:D1222–8.

[7] Dai L, Tian M, Wu J, Xiao J, Wang X, Townsend JP, et al. AuthorReward: increasing community curation in biological knowledge wikis through automated authorship quantification. Bioinformatics 2013;29:1837–9.