Biophysical Reviews (2019) 11:1–3
https://doi.org/10.1007/s12551-019-00500-x

EDITORIAL

Big data: the elements of good questions, open data, and powerful software

Joshua W. K. Ho[1,2,3] · Eleni Giannoulatou[2,3]

Received: 7 January 2019 / Accepted: 14 January 2019 / Published online: 25 January 2019

The field of bioinformatics has evolved in the last 15 years. Back in the days when we were students, we were taught the principle of hypothesis-driven scientific method—a process which involves formulating a testable hypothesis about a natural phenomenon, designing and carrying out an experiment with sufficient statistical power, and carefully analysing the experimental data to confirm or reject the initial hypothesis. This process is the bedrock of modern science, including the fields of biology and medicine. Good scientific findings always go hand-in-hand with good questions and well-designed experiments. Importantly, experimental data are generated to test a specific hypothesis or to address a specific question. In the field of molecular biology and genetics, genes are experimentally knocked out to assess their potential function. In the field of structural biology, X-ray crystallography is used to determine the structure of a protein. In clinical research, randomised controlled clinical trials are conducted to rigorously test the effect of a drug. In all these examples, data are generated experimentally to test a predefined hypothesis. Historically, bioinformatics largely plays a supportive role in this process by providing useful computational tools to analyse biological data. Bioinformatics software tools such as those designed for sequence alignment, phylogenetic tree inferences, molecular dynamics and statistical hypothesis testing for high throughput assays were all useful in helping scientists interpret the data generated from their experiments, often well-designed experiments. Nonetheless, the process largely still follows the standard hypothesis-driven scientific method in which experimental data are generated, collected and analysed for a specific purpose. The early development of biological databases, such as NCBI's GenBank and Gene Expression Omnibus (GEO), are some early predecessors of big data bioinformatics, but they were primarily designed to be repositories of completed experiments and their data.

This relatively linear scientific method seems to be slowly shifting in the big data era. We are increasingly witnessing consortium-based data collection projects which systematically generate a wide range of genome-wide data associated with cell lines, cultured cells, tissues and tumour samples. These data were not designed to answer one specific question by an individual researcher or research group; they were designed to act as reference data sets for all scientists such that they can use these data to address their specific questions. Using these open reference data, it is possible to explore specific hypotheses without individually conducting new experiments. For example, if we want to predict the effect of a DNA mutation in a particular gene in the human genome, we can now access data about this gene from various genetic databases online, and we can already make some reasonable predictions about the effect of this mutation without performing any experiment. This process significantly reduces the number of new experiments, allowing us to focus on formulating and testing more complex and interesting hypotheses.

Big data can also be collected from other unconventional sources, such as unstructured data from short text messages in social media, photographs and data from commercially available wearable devices. These data may not be initially generated for any scientific project, but they can be repurposed for scientific studies as they tend to be abundant and widely available. Nonetheless, these data also tend to have variable

This article is part of a Special Issue on 'Big Data' edited by Joshua WK Ho and Eleni Giannoulatou

✉ Joshua W. K. Ho
jwkho@hku.hk

[1] School of Biomedical Sciences, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Pokfulam, Hong Kong

[2] Victor Chang Cardiac Research Institute, Darlinghurst, NSW 2010, Australia

[3] St. Vincent's Clinical School, The University of New South Wales, Darlinghurst, NSW 2010, Australia

quality, such as high sampling bias, a high proportion of missing data and significant batch effects.

The use of big data in scientific studies not only necessitates new analytical tools, but also requires a major swift in mindset. In traditional biology classes, we learned experimental design and planning. In the big data era, data are often generated without a specific predefined hypothesis. As a consequence, the role of experimental design is partially replaced by the formulation of specific questions, data annotation and cleaning, feature selection, and retrospective data analysis. To fully harness the power of big data in biology and medicine, we believe there are three important elements:

1. Good questions: Biologically meaningful questions are necessary for any good scientific enquiry, regardless of whether big data are used in a study. A well-crafted question will determine what data sets need to be included or excluded, and what type of patterns will be explored and tested. In the traditional hypothesis-driven approach, a good question informs good experimental design. In the big data era, a good question informs the selection of data, expected patterns to observe, the analysis to be performed and the type of software to be used.
2. Open data: It is important to obtain good quality data that are openly accessible. It is especially important to understand the characteristics of the data, including how the data were generated, what are their statistical properties, and what the inherent artefacts in the data are.
3. Powerful software: It is a software that can store, process, analyse and visualise large-scale data that are necessary in big data projects. Moreover, advanced machine learning tools such as deep neural networks (deep learning) are used to build predictive models using a large collection of data. These artificial intelligence (AI)-based models are especially useful in translational medical applications. The validity of these software programs is especially important as key scientific conclusions are drawn based on their analytical results.

In this special issue on *Big Data*, we are fortunate to have fifteen reviews and letters that discuss all three of these elements of big data analysis in biology and medicine. We have articles that discuss how a large amount of open data are used to address specific biomedical questions, such as discovery of driver mutations in cancer (Nussinov et al. 2019; Poulos and Wong 2018); prediction of drug response (Ali and Aittokallio 2018); and discovery of cell type and their gene regulatory networks (Kabir and O'Connor 2019). We have articles that discuss the characteristics of a variety of big data in biology and medicine, including genomic data (Wong 2019); metagenomic data (Wang et al. 2019); Hi-C data (Pal et al. 2018); ChIP-seq data (Tan and Wong 2019); physiological data (Orphanidou 2019); and transcriptomic data (Mar 2019). Last but not least, we have articles that discuss the advances of software technology and analytical methods in big data analysis, including Bayesian statistical learning for big data biology (Yau and Campbell 2019); natural language processing and other curation approaches for GEO (Wang et al. 2018); machine learning of imaging data (Nichols et al. 2018), a software package for processing data from wearable wrist-based heart rate monitors (Djordjevic et al. 2019); and testing techniques for big data software (Zhang and Xie 2018).

In the big data era, bioinformaticians are no longer just playing a supportive role in biological and medical investigations by producing software and analytical methods; they are very much playing a leading role in using big data to conduct their own scientific inquiries. We hope this special issue in *Biophysical Reviews* will serve as a reference for many aspiring big data bioinformaticians for years to come.

Compliance with ethical standards

Funding information The work was supported in part by funds from the National Health and Medical Research Council (1105271 to JWKH); National Heart Foundation (100848 to JWKH, 101204 to EG); and a NSW Health Early-Mid Career Fellowship (to EG).

Conflict of interest Joshua W. K. Ho declares that he has no conflict of interest. Eleni Giannoulatou declares that she has no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Ali M, Aittokallio T (2018) Machine learning and feature selection for drug response prediction in precision oncology applications. Biophys Rev. https://doi.org/10.1007/s12551-018-0446-z (in this issue)

Djordjevic D, Cawood BK, Rispin S, Shah A, Yim LHH, Haywood CS, Ho JWK (2019) CardiacProfileR: an R package for extraction and visualisation of heart rate profiles from wearable fitness trackers. Biophys Rev. https://doi.org/10.1007/s12551-019-00498-2 (in this issue)

Kabir MH, O'Connor MD (2019) Stem cells, big data and compendium-based analyses for identifying cell types, signalling pathways and gene regulatory networks. Biophys Rev. https://doi.org/10.1007/s12551-018-0486-4 (in this issue)

Mar JC (2019) The rise of the distributions: why non-Normality is important for understanding the transcriptome and beyond. Biophys Rev. https://doi.org/10.1007/s12551-018-0494-4 (in this issue)

Nichols JA, Chan HWH, Baker MAB (2018) Machine learning: applications of artificial intelligence to imaging and diagnosis. Biophys Rev. https://doi.org/10.1007/s12551-018-0449-9 (in this issue)

Nussinov R, Jang H, Tsai C-J, Cheng F (2019) Precision medicine review: rare driver mutations and their biophysical classification.

Biophys Rev. https://doi.org/10.1007/s12551-018-0496-2 (in this issue)

Orphanidou C (2019) A review of big data applications of physiological signal data. Biophys Rev. https://doi.org/10.1007/s12551-018-0495-3 (in this issue)

Pal K, Forcato M, Ferrari F (2018) Hi-C analysis: from data generation to integration. Biophys Rev. https://doi.org/10.1007/s12551-018-0489-1 (in this issue)

Poulos RC, Wong JWH (2018) Finding cancer driver mutations in the era of big data research. Biophys Rev. https://doi.org/10.1007/s12551-018-0415-6 (in this issue)

Tan K, Wong KH (2019) RNA polymerase II ChIP-seq - a powerful and highly affordable method for studying fungal genomics and physiology. Biophys Rev. https://doi.org/10.1007/s12551-018-00497-9 (in this issue)

Wang Z, Lachmann A, Ma'ayan A (2018) Mining data and metadata from the gene expression omnibus. Biophys Rev. https://doi.org/10.1007/s12551-018-0490-8 (in this issue)

Wang Q, Wang K, Wu W, Giannoulatou E, Ho JWK, Li L (2019) Host and microbiome multi-omics integration: applications and methodologies. Biophys Rev. https://doi.org/10.1007/s12551-018-0491-7 (in this issue)

Wong K-C (2019) Big data challenges in genome informatics. Biophys Rev. https://doi.org/10.1007/s12551-018-0493-5 (in this issue)

Yau C, Campbell KR (2019) Bayesian statistical learning for big data biology. Biophys Rev. https://doi.org/10.1007/s12551-019-00499-1 (in this issue)

Zhang Z, Xie X (2018) Towards testing big data analytics software: the essential role of metamorphic testing. Biophys Rev. https://doi.org/10.1007/s12551-018-0492-6 (in this issue)