Biophys Rev (2017) 9:297–298
DOI 10.1007/s12551-017-0282-6

LETTER TO THE EDITOR

Modelling, inference and big data in biophysics

Joshua W. K. Ho[1,2] · **Guy H. Grant**[3]

Received: 9 July 2017 / Accepted: 17 July 2017 / Published online: 30 July 2017

Abstract In recognition of the increasing importance of big data in biophysics, a new session called 'Modelling, inference, big data' is incorporated into the IUPAB/EBSA Congress on 18 July 2017 at Edinburgh, UK.

The term 'big data' is used to describe data that are large and complex, often many magnitudes larger than current data size, and likely containing much more noise. In the context of biology, big data are emerging due to technological advances such as next-generation sequencing, single-cell technology and high content imaging, to name just a few. For example, while a typical RNA sequencing (RNA-seq) study generates tens of genome-wide gene expression profiles, a single-cell RNA-seq (scRNA-seq) study can generate tens of thousands of gene expression profiles, each representing an individual cell. Currently, research in computational biophysics tends to be compute-intensive (e.g., molecular dynamic simulation, large-scale simulation of the physiological function of an organ, etc.) but uses relatively modest data size. Big data are transforming the field into a much more data-intensive discipline, requiring us to adapt to new challenges. If utilised properly, big data will open new opportunities to model gene regulation and cell-to-cell variability at a much finer scale.

In recognition of the increasing importance of big data in biophysics, a new session called 'Modelling, inference, big data' is incorporated into the IUPAB/EBSA Congress on 18 July 2017 at Edinburgh, UK. This session has two invited speakers: Dr. Joshua Ho will present a talk on 'Software scalability and validation in big data analysis', and Dr. Christopher Yau will present a talk on 'From single cells to populations: statistical models for heterogeneous biological systems'. Collectively, these two talks will examine three important aspects of modern big data analysis: on-demand scalability; validation of analytical software; and robust statistical inference of noisy data.

Dr. Ho's talk focuses on his recent work on using a commercial on-demand cloud computing platform (e.g., Amazon Web Services cloud) and big data programming frameworks (e.g., Apache Hadoop and Spark) to drastically speed up the processing of scRNA-seq data by parallelising the processing task on large cluster of virtual machines (Yang et al. 2017). He discusses how the modern cloud computing paradigm is being used to deal with the scalability problem in big data analysis. Besides the scalability issue, Dr. Ho's talk also touches on another important yet often ignored issue in big data analysis: validation of the correctness of big data analysis software. In particular, he explains why testing such big data software is difficult and how state-of-the-art software testing techniques can be used to implement effective software quality assurance strategies (Kamali et al. 2015; Troup et al. 2016).

Dr. Yau's talk focuses on his recent work on developing robust statistical methods and software to analyse scRNA-seq data. The analysis of scRNA-seq data is challenging because

This article is part of a Special Issue on 'IUPAB Edinburgh Congress' edited by Damien Hall

✉ Joshua W. K. Ho
j.ho@victorchang.edu.au

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

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

[3] School of Life Sciences, University of Bedfordshire, Park Square, Luton LU1 3JU, UK

they are much noisier compared to standard 'bulk' RNA-seq data, with an excessive amount of 'zeros' in the data due to signal dropouts. Dr. Yau's group specialises in developing statistical inference methods to capture properly that additional noise, and correctly account for the additional variability in downstream analysis, such as clustering (Pierson and Yau 2015) and inference of temporal transcriptional dynamics (Campbell and Yau 2016).

The IUPAB/EBSA big data session also features five short oral presentations selected from abstracts. They cover a wide range of topics: 'Transforming protein sequence and composition into numbers: a big data analysis tool for proteomics' by Rajaram Swaminathan, 'Colonisation dynamics of bacteria in mice' by Florence Bansept, 'Neuronal signalling pathways estimated from whole-brain imaging data of *C. elegans*' by Yuishi Iwasaki, 'MDbox: a cloud-based repository for molecular dynamics simulations' by Karman Condic-Jurkic, and 'Understanding cancer phenomena using a thermodynamic-based approach' by Nataly Kravchenko-Balasha.

Acknowledgments The author would like to thank Andrian Yang for critical review of the manuscript.

Compliance with ethical standards

Conflict of interest Joshua Joshua W.K. Ho declares that he has no conflict of interest. Guy H. Grant declares that he 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

Campbell KR, Yau C (2016) Order under uncertainty: robust differential expression analysis using probabilistic models for Pseudotime inference. PLoS Comput Biol 12:e1005212

Kamali AH, Giannoulatou E, Chen TY, Charleston MA, McEwan AL, Ho JWK (2015) How to test bioinformatics software? Biophys Rev 7:343–352

Pierson E, Yau C (2015) ZIFA: dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biol 16:241

Troup M, Yang A, Kamali AH, Giannoulatou E, Chen TY, and Ho JWK (2016) A Cloud-based Framework for Applying Metamorphic Testing to a Bioinformatics Pipeline. In Proceedings of the 1st International Workshop on Metamorphic Testing. p 33–36

Yang A, Troup M, Lin P, Ho JWK (2017) Falco: a quick and flexible single-cell RNA-seq processing framework on the cloud. Bioinformatics 33:767–769