Guest Editorial

Ming Chen[1] / Andrew Harrison[2] / Hugh Shanahan[3] / Yuriy Orlov[4,5]

Biological Big Bytes: Integrative Analysis of Large Biological Datasets

[1] Department of Bioinformatics, College of Life Sciences, Zhejiang University, Hangzhou, P.R. China, E-mail: mchen@zju.edu.cn
[2] Department of Mathematical Sciences, University of Essex, Essex, England, UK
[3] Department of Computer Science, Royal Holloway, University of London, Egham, England, UK
[4] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
[5] Novosibirsk State University, Novosibirsk, Russia

DOI: 10.1515/jib-2017-0052

This current special issue aims to provide a forum for researchers to present the latest advances and state-of-the-art techniques, tools and applications in biological big data analysis. The past decade has witnessed tremendous growth in massive and complex omic data sets that are generated continuously by high-throughput experimental technologies such as Next Generation Sequencing (NGS). Today, we are surrounded by a world of big data and we are struggling to make sense of it. The increasing availability of omics data represents an unprecedented opportunity for bioinformaticians, but also a major challenge because of diverse heterogeneous factors. Uncovering hidden patterns from such a huge and heterogeneous amount of omics data allows the creation of predictive models for real-life applications. Research in bioinformatics is moving from a hypothesis-driven to a data-driven approach. The main issue is the need of improved bioinformatics solutions. For example, how to build an appropriate architecture for specific omics data system to manage, access and interpret data. Therefore, new paradigms are needed to deal with omics data, for its annotation and integration and finally for inferring knowledge, answering biological questions and making it available to biologists.

For this special issue, we selected one review and six original articles:

Benstead-Hume et al. describe how genetic interactions are being therapeutically exploited to identify novel targeted treatments for cancer. They discuss current approaches – both experimental and computational – that use big data to identify genetic interactions both in humans and model organisms [1]. Detecting sources of bias in transcriptomic data is essential to determine signals of biological significance. Alnasir and Shanahan [2] outline a novel method to detect sequence specific bias in short read NGS data, based on determining intra-exon correlations between specific motifs using the big data analysis platform Spark. A Miniature Inverted-repeat Transposable Element (MITE) is a short transposable element, carrying no protein-coding regions. Ge et al. [3] developed MUSTv2 to represent an accurate detection program of recently active MITE copies. Orlov et al. [4] present an analysis of alternative splicing events on an example of glioblastoma cell culture samples using a set of computer tools and integration of the databases. Based on RNA-seq analysis, Babenko et al. revealed a highly statistically significant level increase of ApoE expression in the hypothalamus in chronically aggressive and defeated mice compared to the control. Correlation analysis revealed a close association of ApoE expression profile and proopiomelanocortin gene in the hypothalamus, implying the putative neuroendocrine stress response background of ApoE expression elevation therein [5]. Wang et al. present a generalized approach for measuring relationships between any pairs of genes, which is based on statistical prediction. They derived two particular versions of the generalized approach, least squares estimation (LSE) and nearest neighbors prediction (NNP). The approach can be extended from two-genes relationships to multi-genes relationships [6]. Phenomics is a fast emerging field wherein high-throughput phenotyping images can be obtained. Rahaman et al. propose an approach to biomass estimation based on image derived phenotypic traits. They modeled plant volume as a function of plant area, plant compactness, and plant age to generalize the linear biomass model. The obtained results confirmed that the proposed model can explain most of the observed variance during image derived biomass estimation [7].

Overall, this issue presents a small set of approaches for big data analysis in bioinformatics. The development and application of computational algorithms, databases and tools are crucial for the efficient processing, management and visualization of large-scale omics data. However, analyzing such big data and deriving biological knowledge, applying it for predictions and further experimentation is becoming a challenging task.

Ming Chen is the corresponding author.

Integrative bioinformatics analyses are important to comprehensively understand how a system works by combining different methods to investigate multiple datasets. The annual international symposium on Integrative Bioinformatics will be of interest to bioinformaticians. The 2018 event will be held in Rothamsted Research, UK.

References

[1] Benstead-Hume G, Wooller S, Pearl F. Computational approaches to identify genetic interactions for cancer therapeutics. J Integr Bioinform. 2017;14:20170027.
[2] Alnasir J, Shanahan HP. A novel method to detect bias in Short Read NGS RNA-seq data. J Integr Bioinform. 2017;14:20170025.
[3] Ge R, Mai G, Zhang R, Wu Q, Wu X, Zhou F. MUSTv2: an improved de novo detection program for recently active miniature inverted repeat transposable elements (MITEs). J Integr Bioinform. 2017;14:20170029.
[4] Orlov Y, Babenko VN, Gubanova N, Bragin A, Chadaeva I, Vasiliev G, et al. Computer analysis of glioma transcriptome profiling: alternative splicing events. J Integr Bioinform. 2017;14:20170022.
[5] Babenko VN, Smagin DA, Kudryavtseva NN. RNA-Seq mouse brain regions expression data analysis: focus on ApoE functional network. J Integr Bioinform. 2017;14:20170024.
[6] Wang L, Ahsan A, Chen M. A generalized approach of measuring relationships among genes. J Integr Bioinform. 2017;14:20170026.
[7] Rahaman M, Ahsan A, Gillani Z, Chen M. Digital biomass accumulation of cereal plants using high-throughput phenotype image 2 analysis. J Integr Bioinform. 2017;14:20170028.