Bull. Aust. Math. Soc. **96** (2017), 348–349
doi:10.1017/S0004972717000557

NOVEL PREPROCESSING APPROACHES FOR OMICS DATA TYPES AND THEIR PERFORMANCE EVALUATION

DARIO STRBENAC

(Received 21 April 2017; first published online 5 July 2017)

2010 *Mathematics subject classification*: primary 62E17; secondary 62N03, 68N01.
Keywords and phrases: signal processing, bioinformatics, Latin squares.

A diverse range of high-dimensional datasets has recently become available to help elucidate the functioning of biological systems and defects within those systems leading to disease. This improved understanding will aid our knowledge of fundamental biology as well as increasing our comprehension of the processes that are altered in complex disease. All of these new technologies come with the challenges of determining how the raw data should be efficiently processed or normalised and, subsequently, how can the data best be summarised for more complex downstream analysis. There are many approaches to summarising and normalising omics data, with new methods frequently being developed. Different kinds of omics data may also be integrated, in order to provide more confidence in predictions. To date, there has not been a comprehensive evaluation of existing methods for many omics data types. This thesis focuses on systematically evaluating existing methods for three different types of omics data and, having identified limitations in the current methods, also proposes new approaches to improve its quality.

Firstly, CAGE-seq data are considered. This type of data has unique characteristics such that regional summarisation algorithms developed for similar experiments, such as ChIP-seq, are not directly applicable. Additionally, the raw data also contain artefactual measurements from confounding biological processes, and a comprehensive evaluation of region-classification algorithms has not previously been carried out. A two-stage method based on a novel region-finding algorithm followed by a classifier that integrates sequence patterns surrounding the identified regions is shown to possess superior performance to two existing methods. Similarly, a novel data summarisation approach to gene expression data, which integrates changes

Thesis submitted to the University of Sydney in April 2016; degree awarded on 5 December 2016; supervisor Jean Yang.

in location and scale into a unified metric, demonstrates benefits in two-class classification problems. The error rates are found to be competitive with existing methods, and the feature selection has higher stability and increased biological relevance. Finally, in the proteomics setting, there are many choices for how to summarise peptides to proteins, as well as issues relating to batch effects and whether internal controls are necessary. By developing a broad variety of performance metrics that assess bias or variance, and an accompanying web-based framework for reproducible research, novel recommendations about peptide to protein summaries and batch correction algorithms are made, and a surprising result regarding the necessity of internal standards is revealed. The development and evaluation of novel dataset preprocessing approaches and the comprehensive evaluation of existing methods for three data types demonstrate the importance of systematic performance evaluation of statistical bioinformatics methods for more accurate and precise knowledge generation in modern biology.

DARIO STRBENAC, School of Mathematics and Statistics,
University of Sydney, New South Wales 2006, Australia
e-mail: dario.strbenac@sydney.edu.au

https://doi.org/10.1017/S0004972717000557 Published online by Cambridge University Press