


Class Discovery, Comparison, and Prediction Methods for RNA–Seq Data

Ahu Cephe

Erciyes University, Turkey

Necla Koçhan

 <https://orcid.org/0000-0003-2355-4826>

İzmir Biomedicine and Genome Center, Turkey

Gözde Ertürk Zararsız

Erciyes University, Turkey

Vahap Eldem

İstanbul University, Turkey

Gökmen Zararsız

Erciyes University, Turkey

INTRODUCTION

Measuring gene-expression plays a vital role in life sciences such as cancer genomics. It enables us to quantify the level at which a particular gene is expressed within a cell, tissue or organism, thereby providing a tremendous amount of information (Alberts et al., 2002). There are different technologies (i.e., microarray and next-generation technologies) that can measure gene-expression levels. Microarray technology is an outdated technology with some limitations and lost its popularity with the advent of next-generation technologies. On the other hand, RNA-seq is one of the next-generation technologies capable of coping with these limitations, using the capabilities of next generation sequencing technologies, and performing operations quickly and cheaply based on the principle of high-throughput sequencing technology. Moreover, compared to microarrays, RNA-seq offers several advantages: (i) having less noisy data, (ii) being able to detect new transcripts and coding regions, (iii) not requiring pre-determination of the transcriptomes of interest.

RNA-seq technology allows measuring the expression levels of thousands of genes in cells simultaneously, leading to high dimensional data to be further analyzed. The information stored in these high dimensional data can be used for different purposes: (i) identifying “biomarker” genes that can characterize different disease subclasses, that is, class comparison; (ii) identifying new subclasses for a particular disease, that is, class discovery and (iii) assigning samples into known disease classes, that is, class prediction (Dudoit et al., 2002; Weigelt et al., 2010).

Class comparison is known as differential analysis or analysis of differential-expression. In these studies, gene-expression profiles of samples, which are predefined groups, are compared to identify differentially expressed genes between groups. Differentially expressed genes are identified in cells from different tissues, different patients, or cells exposed to different experimental conditions. For example,

DOI: 10.4018/978-1-7998-9220-5.ch123

comparing treated and untreated cells to detect the effect of a new drug on gene-expression levels; comparisons between healthy tissue and diseased tissue to identify genes with altered expression; comparing gene-expression in tumor tissue for patients responding to a particular treatment versus gene-expression in patients with the same cancer diagnosis who do not respond to treatment. Such studies yield lists of genes that were significantly altered between groups. The aim is to provide insight into the underlying biological mechanisms and perhaps identify potential therapeutic targets.

In class prediction studies, as in class comparison studies, genes that differ between predefined classes are tried to be determined. However, in class prediction studies, gene-expression values are explanatory variables rather than outcome variables. Moreover, the purpose of the analysis of class prediction studies is to identify a small set of genes that can accurately distinguish between different classes rather than identify all genes that differ. Classes are defined beforehand in class predictions, and the aim is to create a classifier that can distinguish between these classes based on the gene-expression profiles of the samples and can be applied to the expression profiles of a new sample. For example, a classifier that distinguishes between 2 different disease states; a classifier that distinguishes short-term survivors from long-term survivors; a classifier can be created that predicts whether a patient will respond to a particular drug. In class comparison studies, whether a new patient will react to treatment can be predicted based on gene-expression profiles.

Class discovery differs from class comparison and class prediction studies in that classes are not predefined. The purpose of these studies is to determine whether subsets of samples with apparently homogeneous phenotypes can be distinguished based on differences in gene-expression profiles. For example, there are many diseases in which individuals with apparently similar phenotypes have significant variability in outcomes such as survival. This variability is due to differences at the molecular level. Class discovery studies are used to identify molecular differences that define subgroups for new diseases or known diseases. Class discovery studies need to analyze a set of gene-expression profiles in order to discover subgroups that share common characteristics. For example, subgroups of patients with similar expression profiles are classified. It can also describe different stages of disease severity or identify groups of genes that may behave similarly in a disease state.

Although there are many studies and methods using microarray data for class comparison, class prediction and class discovery design model in the literature, these methods cannot be directly applied to RNA-seq data, which has a discrete, skewed and over-dispersed structure, quite different from the continuous data structure of microarrays. In this case, the first way to analyze the RNA-seq data is to transform and get closer data to a normal distribution and use the methods developed for microarrays. The second way is to work directly with count data using methods based on discrete probability distributions such as Poisson and Negative Binomial (Anders & Huber, 2010; Love et al., 2014; Zheng et al., 2014).

This chapter provides an overview of algorithms/methods used for class discovery, comparison and prediction of RNA-seq data. It also covers the difficulties and challenges encountered by these algorithms. Finally, the algorithms proposed and developed for RNA-seq data analysis are discussed with their pros and cons.

23 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:

www.igi-global.com/chapter/class-discovery-comparison-and-prediction-methods-for-rna-seq-data/317607

Related Content

Identifying Patterns in Fresh Produce Purchases: The Application of Machine Learning Techniques

Timofei Bogomolov, Malgorzata W. Korolkiewicz and Svetlana Bogomolova (2020). *Handbook of Research on Big Data Clustering and Machine Learning* (pp. 378-408).

www.irma-international.org/chapter/identifying-patterns-in-fresh-produce-purchases/241384

Malware Analysis and Classification Using Machine Learning Models

Swadeep Swadeep, Karmel Arockiasamy and Karthika Perumal (2024). *Machine Learning Algorithms Using Scikit and TensorFlow Environments* (pp. 209-220).

www.irma-international.org/chapter/malware-analysis-and-classification-using-machine-learning-models/335190

Generating an Artificial Nest Building Pufferfish in a Cellular Automaton Through Behavior Decomposition

Thomas E. Portegys (2019). *International Journal of Artificial Intelligence and Machine Learning* (pp. 1-12).

www.irma-international.org/article/generating-an-artificial-nest-building-pufferfish-in-a-cellular-automaton-through-behavior-decomposition/233887

Robust Output Only Health Monitoring of Steel Railway Bridges: Analysis of Applicability of Different Sensors

Ahmed Rageh, Daniel Linzell, Samantha Lopez and Saeed Eftekhari Azam (2020). *Handbook of Research on Engineering Innovations and Technology Management in Organizations* (pp. 24-41).

www.irma-international.org/chapter/robust-output-only-health-monitoring-of-steel-railway-bridges/256668

Balanced Scorecard as a Tool to Evaluate Digital Marketing Activities

Tasnia Fatin, Mahmud Ullah and Nayem Rahman (2023). *Encyclopedia of Data Science and Machine Learning* (pp. 2368-2383).

www.irma-international.org/chapter/balanced-scorecard-as-a-tool-to-evaluate-digital-marketing-activities/317676