Dissertation Proposal:
Methods for the Analysis of Compositonal RNA Sequence Data

Dominic D LaRoche

December 1, 2015

Contents

Chapter 1

Background and Introduction to the Problem

1.1 RNA Sequencing Data

1.1.1 Process of Collecting RNA Sequence Data

1.1.2 General Properties of RNA Sequence Data

1.1.3 Current Methodology Associated with RNA Sequencing Data

1.2 Principles of Compositional Data Analysis

Compositional data are non-negative data which are subject to a sum constraint, i.e. all the elements must sum to unity. This simple constraint has some important consequences for many standard statistical methodologies including correlation and regression. Compositional data contain only relative information, i.e. the information about any individual component, or group of componenets, is relative to the other components and no absolute information about the absolute value of the component. For example, if we know that 20% of the food in a refigerator is composed of fruit we do not know how much total fruit there is. If the refigerator is full then there will be substantially more fruit than if the refigerator is nearly empty.

Potential problems associated with compositional data were identified as early as 1897 by Pearson who noted that spurious correlations can be induced through ratios of independent variables, e.g. if X, Y, and Z are uncorrelated then X/Z and Y/Z will be correlated. Despite the fact that compositional data naturally arises in a wide variety of scientific disciplines, a general method for analysis of compositional data was not developed until John Aitchison published his seminal book in 1986. Aitchison outlines some basic principles for compostional data analysis (section 1.2.1) and provides some analysis tools for compositional data which conform to these principles (section 1.2.2). Additional methodology has been developed by a number of authors in the 29 years since the publication of Aitchison's book, although a number of problems remain.

1.2.1 Fundamental Principles

Aitchison outlined a set of fundamental principles to which all methods for compostional data should adhere (**Aitchison1986**)

Scale Invariance

Scale invariance requires that the results of a statistical procedure should not depend on the scale used.

Sub-compositional Coherance

Sub-compositional coherance requires that the results of a statistical procedure on a subset of componenets from a composition should depend only on the data contained in that subset.

Permutation Invariance

Permutation invariance requires that the results of a statistical procedure should not depend on the ordering of the of the components.

1.2.2 Statistical Methods for Compositional Data

1.3 Current Methodology with Respect to Compositional Data Analysis

1.3.1 Normalization

Previous authors have identified the compositional nature of RNA sequencing data (Robinson and Oshlack 2010). As stated previously, RNA sequence data are likely subject to two sum constraints: 1) the number of RNA sequences that can fit into the finite sample collected, and 2) the number of available reads of those sequences for the given sequencing technology.

Trimmed Mean of M-values Normalization Method

Robinson and Oshlack (2010) primarily focused on the latter when accounting for the compositional nature of RNA sequence data.

They, like many others (Anders and Huber 2010), also assume that the majority of genes in an assay are not differentially expressed.

Chapter 2

Normalization of Compositional RNA Sequence Data

Chapter 3

An Alternative to Correlation for Evaluation of Reproducibility and Repeatability of Compositional Data

Bibliography

Anders, Simon and W Huber (2010). "Differential expression analysis for sequence count data". In: *Genome Biol* 11.10, R106. ISSN: 1465-6906. DOI: 10.1186/gb-2010-11-10-r106. URL: http://www.biomedcentral.com/content/pdf/gb-2010-11-10-r106.pdf (cit. on p. 6).

Robinson, Mark D and Alicia Oshlack (2010). "A scaling normalization method for differential expression analysis of RNA-seq data." In: *Genome biology* 11.3, R25. ISSN: 1465-6906. DOI: 10.1186/gb-2010-11-3-r25 (cit. on p. 6).