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

Dominic D LaRoche

December 4, 2015

Contents

1 Background **5**

1.1 RNA Sequencing Data . . . 5

1.1.1 Process of Collecting RNA Sequence Data . . . 5

1.1.2 General Properties of RNA Sequence Data . . . 5

1.1.3 Current Methodology Associated with RNA Sequencing Data . . . 5

1.2 Principles of Compositional Data Analysis . . . 5

1.2.1 Fundamental Principles . . . 5

1.2.2 Statistical Methods for Compositional Data . . . 6

1.3 Current Methodology with Respect to Compositional Data Analysis . . . 6

1.3.1 Normalization . . . 6

2 Normalization **9**

3 Correlaton **11**

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**). These principles are outlined below.

Scale Invariance

Scale invariance requires that the results of a statistical procedure should not depend on the scale used. Any meaningful function x of a composition w must satisfy:

$$f(pw) = f(w), \text{ for every } p > 0.$$

Aitchison notes that any meaningful (scale-invariant) function of a composition can be expressed in terms of ratios of the components of the composition. For example, any method used for compositional data should not give different results whether the composition is given as proportions, percentages, parts per million, or any other scale.

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. A subcomposition is defined as the (1, 2, ..., C) parts of a D-part composition $[x_1, ..., x_D]$:

$$[s_1, ..., s_c] = \frac{[x_1, ..., x_c]}{(x_1 + ... + x_c)}$$

Any changes in components $[x_{c+1}, ..., x_D]$ should not have an impact on the inference from the subcomposition. For example, if we measure the number of reads of 14 mRNA sequences from a sample that contains 4,000 unique mRNA sequences the inference we obtain from those 14 sequences should not be affected by the expression level of any of the other 4,000 sequences.

Permutation Invariance

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

Median Normalization

Trimmed Mean of M-values Normalization Method

Robinson and Oshlack (2010) primarily focused on the mapped read constraint when developing their Trimmed-Mean of M-values (TMM) normalization method for RNA sequence data. Like many others (Anders and Huber 2010), they also assume that the majority of genes in an assay are not differentially expressed. For sequencing data Robinson and Oshlack first define the gene-wise log-fold-changes as:

$$M_g = log_2 \frac{Y_{gk}/N_k}{Y_{gk\prime}/N_{k\prime}},$$

where Y_{gk} is the read count for gene g in sample k (here we assume 1 library for each sample) and N_k as the total number of reads for sample k. They then define the absolute expression levels as:

$$A_g = \frac{1}{2} log_2 \left(Y_{gk}/N_k \cdot Y_{gk\prime}/N_{k\prime}\right) \text{ for } Y_{g\bullet} \neq 0,$$

assuming the absolute expression level is the same for the two samples. Both the M-values and A-values are timmed removing the upper and lower 30% of the M-values and the upper and lower 5% of the A-values (although these values are defaults and can be tailored for a given experiment). The resulting normalization factor for sample k is then a weighted average of the M-values:

$$log2\left(TMM_k^r\right) = \frac{\sum_{g \in G} w_{gk}^r M_{gk}^r}{\sum_{g \in G} w_{gk}^r}$$

where

$$M_{gk}^r = log_2 \frac{Y_{gk}/N_k}{Y_{gr}/N_r}, \; Y_{gk}, Y_{gr} > 0, \text{ for reference sample } r,$$

and the weights, w_{gk}^r, are defined as:

$$w_{gk}^r = \frac{N_k - Y_{gk}}{N_k Y_{gk}} + \frac{N_r - Y_{gr}}{N_r Y_{gr}}, \text{ for } Y_{gk}, Y_{gr} > 0.$$

The weights are a result of taking the inverse of the approximate asymptotic variance using the delta method **Casella2002** Note that these calculations omit any genes for which either observed read count is 0; these observations are generally removed by the trimming procedure.

The TMM normalization method attempts to derive a single value, f, such that $S_k = f \cdot S_{k\prime}$, where S_k is the (unobserved) total count for sample k. Unfortunately, this does not adequately account for the compositional nature of the data. In particular, the authors assume that the expression levels will be the same among samples for the majority of genes (probes). However, if we view the data as a composition, even if only a small number of genes differ with respect to expression level then all of the probes will differ with respect to what we actually observe, their *proportion* of the sample. The authors correctly identify that some probes will be under-represented, even if their absolute expression remains unchanged, in samples with a large total RNA output as compared to samples with a smaller total RNA output. This is because these probes will constitute a smaller proportion of the total number of reads, even if the absolute number will not have changed.

TMM normalization is not guaranteed to satisfy subcompositional coherance. This is because the selection of genes to be measured, the subcomposition, can determine the final normalization factor calculated.

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).