ABSTRACT

THESIS: ZERO-INFLATED MODELS FOR RNA-Seq COUNT DATA

STUDENT: Morshed Alam

DEGREE: Master of Science

COLLEGE: Sciences and Humanities

DATE: May 2015

PAGES: 82

Motivation: Next Generation Sequencing (NGS) methods for RNA-Seq result millions of short sequences, called reads that provide fundamental information in the fields of genomics, epigenetics and transcriptomes. One of the main objectives of many biological studies is gene expression profiling between samples. Gene expression profiling studies involve mapping of short reads to reference genome, if available, summarizing, normalizing, and finally performing downstream analysis such as making a list of differentially expressed (DE) genes. One of the common assumptions of RNA-Seq data is that, all gene counts follow an overdispersed Poisson or Negative Binomial (NB) distribution which is sometimes misleading because some genes may have stable transcription levels with no overdispersion and some of them may have excessive number of zero counts. Thus, a more realistic assumption in RNA-Seq data is to consider four sets of genes: overdispersed with limited number of zeros and excessive number of zeros, and

non-overdispersed with limited number of zeros and excessive number of zeros. Our objective is to apply zero inflated models to the data with excessive number of zero counts and to evaluate their performance.

Method: Available methods can handle read counts data with limited number of zero counts for both overdispersed and non-overdispersed data. With excessive number of zeros in the data, we adopt a new approach and apply it to the real RNA-Seq data obtained from Gilad et al.[1] to detect DE genes. Our approach is to consider Zero Inflated Poisson (ZIP) mixed model for non-overdispersed genes and Zero Inflated Negative Binomial (ZINB) mixed model for overdispersed genes. This is an integrated approach because this method can be combined with any other Poisson and NB based methods for detecting DE genes. We also evaluate the performance of the models by conducting a simulation study.

Results: Heat maps for DE genes obtained by ZIP and ZINB mixed models demonstrate the notable performance of the models for the real data. Area under receiver operating characteristics curve (AUC) and Receiver operating characteristics (ROC) curve depict that the models perform well for simulated data. However, ZIP performs better in identifying DE genes from both real and simulated data with excessive zeros.