ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ & ΤΕΧΝΟΛΟΓΙΑΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ
SCHOOL OF INFORMATION SCIENCES & TECHNOLOGY

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
DEPARTMENT OF STATISTICS

ΚΥΚΛΟΣ ΣΕΜΙΝΑΡΙΩΝ ΣΤΑΤΙΣΤΙΚΗΣ ΙΟΥΝΙΟΣ 2018

Panagiotis Papastamoulis

The University of Manchester, Division of Informatics, Imaging & Data Sciences

Bayesian identification of differentially expressed transcripts using RNA-Seq data

ΠΕΜΠΤΗ 14/6/2018
13:00 – 15:00

ΑΙΘΟΥΣΑ Τ103, 1ος ΟΡΟΦΟΣ,
ΝΕΟ ΚΤΙΡΙΟ ΟΠΑ
(ΤΡΟΙΑΣ 2)

ΠΕΡΙΛΗΨΗ

The transcriptome is the set of all transcripts in a studied organism, at a given developmental stage or biological condition. Quantification and characterization of the transcriptome is an essential task in molecular biology. RNA-seq is a next generation sequencing technology resulting in large datasets of short reads, consisting of nucleotide sequences. These reads are aligned to the reference genome or transcriptome and the task of inferring transcript abundances given the aligned short reads would be straightforward if these alignments were unique. However, during the process of transcription, most genes can be alternatively spliced into various transcripts which share specific parts of their sequence. Consequently, many reads map to several transcripts of interest and the estimation of transcript abundance has to be treated probabilistically, since the origin of each read is unknown.

In statistical terms, the problem of inferring relative transcript expression reduces to estimating the weights of a mixture model, assuming that the transcriptome is known. A fundamental task is to discover differentially expressed transcripts between two biological conditions. Most inference methods deal with this question applying a two-stage procedure: a primary model is used to estimate expression and its output is post-processed using a differential expression model. In this talk, both issues are simultaneously addressed by presenting the cjBitSeq model [1, 2] which jointly estimates expression levels and differential expression: the unknown relative abundance of each transcript can either be equal or not between two conditions. It is shown that the proposed model enjoys conjugacy for fixed dimension variables, thus the full conditional distributions are analytically derived. Two samplers are constructed, a reversible jump MCMC sampler and a collapsed Gibbs sampler, and the latter is found to perform best. The method is benchmarked using synthetic RNA-Seq data and compared against other popular approaches.

References

[1] Papastamoulis P. and Rattray M. (2018). A Bayesian model selection approach for identifying differentially expressed transcripts from RNA-sequencing data. Journal of the Royal Statistical Society Series C (Applied Statistics), 67(1): 3-23.

[2] Papastamoulis P. and Rattray M. (2017). Bayesian estimation of Differential Transcript Usage from RNA-seq data. Statistical Applications in Genetics and Molecular Biology 16(5-6): 387-405.

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ & ΤΕΧΝΟΛΟΓΙΑΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ
SCHOOL OF INFORMATION SCIENCES & TECHNOLOGY

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
DEPARTMENT OF STATISTICS

AUEB STATISTICS SEMINAR SERIES JUNE 2018

Panagiotis Papastamoulis
The University of Manchester, Division of Informatics, Imaging & Data Sciences

Bayesian identification of differentially expressed transcripts using RNA-Seq data

THURSDAY 14/6/2018
13:00 – 15:00

ROOM T103, 1st FLOOR,
NEW AUEB BUILDING
(2 TROIAS STR.)

ABSTRACT

The transcriptome is the set of all transcripts in a studied organism, at a given developmental stage or biological condition. Quantification and characterization of the transcriptome is an essential task in molecular biology. RNA-seq is a next generation sequencing technology resulting in large datasets of short reads, consisting of nucleotide sequences. These reads are aligned to the reference genome or transcriptome and the task of inferring transcript abundances given the aligned short reads would be straightforward if these alignments were unique. However, during the process of transcription, most genes can be alternatively spliced into various transcripts which share specific parts of their sequence. Consequently, many reads map to several transcripts of interest and the estimation of transcript abundance has to be treated probabilistically, since the origin of each read is unknown.

In statistical terms, the problem of inferring relative transcript expression reduces to estimating the weights of a mixture model, assuming that the transcriptome is known. A fundamental task is to discover differentially expressed transcripts between two biological conditions. Most inference methods deal with this question applying a two-stage procedure: a primary model is used to estimate expression and its output is post-processed using a differential expression model. In this talk, both issues are simultaneously addressed by presenting the cjBitSeq model [1, 2] which jointly estimates expression levels and differential expression: the unknown relative abundance of each transcript can either be equal or not between two conditions. It is shown that the proposed model enjoys conjugacy for fixed dimension variables, thus the full conditional distributions are analytically derived. Two samplers are constructed, a reversible jump MCMC sampler and a collapsed Gibbs sampler, and the latter is found to perform best. The method is benchmarked using synthetic RNA-Seq data and compared against other popular approaches.

References

[1] Papastamoulis P. and Rattray M. (2018). A Bayesian model selection approach for identifying differentially expressed transcripts from RNA-sequencing data. Journal of the Royal Statistical Society Series C (Applied Statistics), 67(1): 3-23.

[2] Papastamoulis P. and Rattray M. (2017). Bayesian estimation of Differential Transcript Usage from RNA-seq data. Statistical Applications in Genetics and Molecular Biology 16(5-6): 387-405.