Front. Comput. Sci.

LETTER

Detecting differential transcript usage across multiple conditions for RNA-seq data based on the smoothed LDA model

Jing Li[1,2], Xuejun Liu(✉)[1,2], DaoQiang Zhang[1,2]

[1] College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, MIIT Key Laboratory of Pattern Analysis and Machine Intelligence, Nanjing 211106, China.

[2] Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing 210023, China.

1 Introduction and main contributions

Differential transcript usage (DTU), which refers to the event that the relative transcript abundance within a gene changes between conditions. To detect DTU, various methods have been proposed, which can be classified into exon-based models and gene-based models. These approaches either cannot estimate the relative transcript abundance, or they cannot deal properly with the multi-source mapping problems of reads. Besides, few methods currently consider sample-to-sample variability under multiple conditions [1]. In light of the similarity in structures between text data and RNA-Seq data, we previously proposed a statistic model, NLDMseq [2], based on LDA [3] to accurately estimate gene and isoform expression under a single condition, which has showed the applicability of the topic model of text data on RNA-Seq data.

In this paper, we further propose a new statistical approach, MLDA for detecting DTU across multi-condition based on smoothed LDA [3]. MLDA can be considered as a hybrid of exon-based and gene-based methods, taking the pre-processed exon counts as input data, and modeling relative transcript abundances as random variables. Meanwhile, MLDA models the random process of the generation of reads for multiple conditions, which accounts for the uncertainty caused by multi-source mapping. MLDA combines with LRT to detect DTU under multiple conditions.

The main contributions of this letter are the following:

- We propose MLDA to detect DTU under multiple conditions.
- MLDA not only avoids the uncertainty caused by multi-source mapping, but also obtains the relative transcript abundance.
- MLDA uses Python to extract exon-based read counts from the alignment file without relying on other quantization software, and the software package is MLDA under PUGEA users on GitHub.

More details about MLDA are given in online resource.

2 Design and implement of MLDA

The structures of text data and RNA-seq data are

Received September 13, 2019; accepted December 30, 2019

E-mail: xuejun.liu@nuaa.edu.cn

shown in Fig. 1(a)(b) where the analogies between items can be found, the collection of exons (represented by reads) and the document, the isoforms and the topics, and the exons and the words. Based on the similarity in structures between them, we propose MLDA driven from the smoothed LDA [3] to model the stochastic process of the generation of RNA-seq data. To detect DTU, we establish two models, as shown in Fig. 1(c)(d), a null model LR0 based on H0 hypothesis, in which the relative transcript abundances θ_c under all conditions are the same, and an alternative model LR1 based on H1 hypothesis, in which the relative transcript abundances θ_c under different conditions varies. *C, N* and *K* are the number of conditions, the number of reads and the number of isoforms.

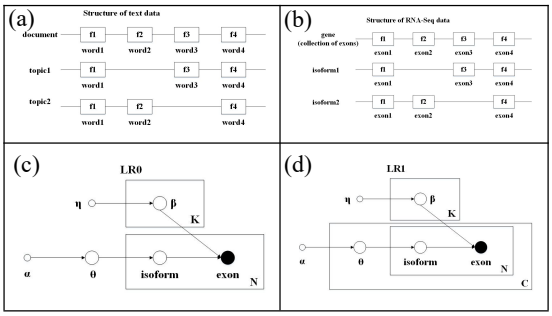

Fig. 1 Structure diagrams of data and model. Structures of (a) text data and (b) RNA-Seq data, f_i represents the frequency of the corresponding item. The probabilistic graphical model of (c) LR0 and (d) LR1.

In the assumption of MLDA, the reads (corresponding to exons) for a given gene are sampled from the multinomial distribution with parameters β given an *isoform*, where *isoform* and β are related to the hidden isoform and the isoform- and exon-specific read sequencing biases, respectively. The hidden isoform *isoform* is sampled from the multinomial distribution with parameters θ which represents the relative transcript abundance of the isoforms for this given gene. The θ and β are sampled from the Dirichlet prior with hyperparameter α and η, respectively.

We assume that each read is associated with a hidden isoform. The model assumption, $exon \sim Multinomial(\beta_k \mid isoform)$ and $isoform \sim Multinomial(\theta)$, mean that reads can be generated by the mixture of the hidden isoforms, so MLDA accounts for the uncertainty caused by multi-source mapping. What's more, since the isoform- and exon- specific read sequencing bias is an inherent nature of a particular gene, we share β across all conditions, and this enables MLDA to model the generation of reads for multiple conditions.

There are three major differences between the MLDA and NLDMseq. First, MLDA places a Dirichlet prior over the isoform- and exon- specific read sequencing bias, β, to avoid the data sparsity problems. Second, MLDA deals with multi-condition data, while NLDMseq can only deal with single condition data. Finally, MLDA is designed for DTU detection, while NLDMseq is for transcript expression estimation.

2.1 Variational EM algorithm

For each gene, the joint probability of the model for θ, z, w, β given hyperparameters α, η is $p(\theta, z, w, \beta \mid \alpha, \eta)$. Due to the coupling of β and θ, we apply variational inference, which uses variational distribution, $q(\theta, z, \beta \mid \gamma, \varphi, \lambda)$ with parameters γ, φ and λ, to approximate the posterior distributions of the latent variables. By assuming the independence of the hidden variables in the variational distributions, we use the variational EM algorithm to work out the model. Please refer to the supplementary file for the details of the EM algorithm.

In MLDA, θ represents the relative transcript abundance, and its posterior distribution is approximated by a Dirichlet distribution with variational parameter γ. Therefore, the relative transcript abundance θ can be estimated in a principled way by considering all stochastic factors in the data generation procedure.

2.2 LRT for DTU detection

Once the two models, LR0 and LR1, are worked out, LRT is further used to identify DTU. The test statistic is twice the difference between the log-likelihood of the two models $LR = 2\log\frac{L(D \mid LR0)}{L(D \mid LR1)}$, and LR follows the Chi-square distribution with degrees of freedom equal $(C-1)*K$. For error control we utilize FDR (False Discovery Rate) by Benjamini to determine a significant level to limit the FDR under a certain level.

3 Results

Datasets We verify our method using a simulated dataset and four real datasets, the number of conditions and replicates of all datasets are shown in Table 1. For the purpose of evaluating our method on the detection of DTU,

we compare MLDA to three existing DTU analysis methods, DEXSeq[4], SUPPA2[5] and Cuffdiff[6] on a simulated dataset and three real datasets. To access the accuracy of the calculated relative transcript abundance, we compare with RSEM, Kallisto and Cufflinks on the simulated dataset and a real dataset.

Table 1 The number of conditions and replicates of datasets.

Datasets	# of conditions	# of replicates
Simulated	4	3
TH17(GSE52260)	9	3
PSC(GSE90053)	8	2
HHD(GSE90053)	3	4
SEQC(GSE47792)	4	8

Results For the real dataset, we consider the common DTU genes detected by multiple methods as the ground truth when comparing different methods. The comparison results of AUC on a simulated dataset and three real datasets are summarized in Table 2. It can be seen that MLDA obtains competitive results on these datasets. Besides the DTU detection analysis, we also further verify the accuracy of the relative transcript abundance obtained from MLDA on real dataset. For this purpose, we utilize the well predefined benchmark SEQC dataset, where sample C consists of 3/4 of sample A and 1/4 of sample B and sample D consists of 1/4 of sample A and 3/4 of sample B. Consistent estimates of transcript expression in A, B, C and D should agree with these mixing proportions. The consistence of the four methods under conditions C and D are shown in Table 3.

Table 2 The AUC results of different methods on datasets.

Datasets	MLDA	SUPPA2	DEXSeq	Cuffdiff
Simulated	**0.9842**	0.9556	0.9759	0.8021
TH17	**0.8177**	0.7649	0.7923	0.6489
PSC	**0.7509**	0.7003	0.6217	0.7104
HHD	**0.8185**	0.7984	0.8055	0.7896

Table 3 Consistence between transcript-level estimates and predefined results for the mixed SEQC samples C, D.

Condition	MLDA	RSEM	Kallisto	Cufflinks
C	**0.9939**	0.8679	0.8687	0.9681
D	**0.9927**	0.8634	0.8644	0.9655

4 Conclusion

In this paper, we propose a statistical framework MLDA, based on the smoothed LDA to detect the differential usage of transcript isoforms for RNA-seq data. The experimental results show that our method performs competitively on the detection of DTU and obtain more accurate relative transcript abundance compared with other alternatives on both simulated and real data.

Acknowledgements This work was supported by the National Key R&D Program of China (Grant Nos. 2018YFC2001600, 2018YFC2001602).

Supporting information The supporting information is available online at journal.hep.com.cn and link.springer.com.

References

[1] Hooper J E. A survey of software for genome-wide discovery of differential splicing in RNA-Seq data. Human genomics, 2014, 8(1): 3.

[2] Liu X, Shi X, Chen C, Zhang L. Improving RNA-Seq expression estimation by modeling isoform-and exon-specific read sequencing rate. BMC bioinformatics, 2015, 16(1): 332.

[3] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation. Journal of machine Learning research, 2003, 3(Jan): 993-1022.

[4] Anders S, Reyes A, Huber W. Detecting differential usage of exons from RNA-seq data. Genome research, 2012, 22(10): 2008-2017.

[5] Trincado J L, Entizne J C, Hysenaj G, Singh B, Skalic M, Elliott D J., Eyras E. SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. Genome biology, 2018, 19(1): 40.

[6] Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley D R, Pimentel H, Salzberg S L, Rinn J L, Pachter L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature protocols, 2012, 7(3): 562.