Next-masigpro: dealing with RNA-SEQ time series

Ana Conesa[1], María José Nueda[2]✉
[1]Prince Felipe Research Centre, Valencia, Spain
[2]University, Alicante, Spain

Motivation and Objectives

During the last decades the development of specific statistics methods to deal with microarray data has been key in transcriptome study. Most of the developed statistics methods have become as reference or classic methods due to their ability to deal with transcriptomics data. However, recent advances in sequencing technologies have created alternatives to microarrays. These new type of data require an appropriate statistical treatment to get good results. New methods are needed but it is also important the study of the adequateness of the existing methods and the adaptation of them to the new type of data.

maSigPro (Conesa *et al.,* 2006) is a method to deal with time course microarray (TCM) data that has been applied in several biological scenarios. maSigPro is in Bionconductor since 2005 and it is implemented in several web-services (Nueda *et al.*, 2010 and Medina *et al,*. 2010). However, maSigPro has been designed to deal with normal microarray intensity signals, rather than with count data. In this work, we adapt maSigPro to RNA-Seq time series analysis.

Methods

maSigPro deals with regression linear models where the response is considered as normally distributed data, a continuous variable. Sequencing technologies give us counts data which distribution is discrete. Therefore, applying the original version of maSigPro to discrete data can not be appropriate and results can be wrong.

The statistical model for counts data may be Poisson or Binomial. However, there are studies (Lu *et al.,* 2005) that show overdispersion of the data and suggest the negative binomial (NB) distribution for being more flexible to estimate the variance of the data. Generalized linear models (GLMs) are an extension of linear models to non-normally distributed response data (McCullagh and Nelder 1989, Dobson 2002). We have modified maSigPro package replacing linear models functions by GLMs functions and giving them the appropriate statistical treatment.

To study the need of adapting the maSigPro package to RNA-Seq data several binomial negative time series datasets have been simulated in different scenarios with different number of replicates in each experimental condition (example in table 1). Linear regression models and GLMs have

Table 1: 4 RNA-Seq simulated datasets with 6000 genes, 300 differentially expressed genes, 6 time-points and different number of replicates in each one. FP: false positives. FN: false negatives. R2: model good of fit threshold for gene selection..

		LM MODEL					GLM MODEL				
replicates	**R2**	**selection**	**FP**	**FN**	**Sensitivity**	**Specificity**	**selection**	**FP**	**FN**	**Sensitivity**	**Specificity**
1	0.5	0	0	300	0.000	1.000	663	454	91	0.697	0.920
	0.6	0	0	300	0.000	1.000	657	453	96	0.680	0.921
	0.7	0	0	300	0.000	1.000	601	523	122	0.260	0.926
2	0.5	11	0	289	0.037	1.000	420	144	24	0.920	0.975
	0.6	11	0	289	0.037	1.000	330	75	45	0.850	0.996
	0.7	11	0	289	0.037	1.000	214	20	106	0.647	0.995
3	0.5	217	11	94	0.687	0.998	325	29	4	0.987	0.999
	0.6	171	5	134	0.553	0.999	273	6	33	0.890	1.000
	0.7	81	1	220	0.267	1.000	198	1	103	0.657	1.000
5	0.5	238	0	62	0.793	1.000	299	0	1	0.997	1.000
	0.6	120	0	180	0.400	1.000	280	0	20	0.933	1.000
	0.7	32	0	268	0.107	1.000	174	0	126	0.580	1.000

been applied to the simulated datasets to compare the results.

Results and Discussion

Results show an improved performance of maSigPro to deal with RNA-Seq when using generalized linear models. Therefore the maSigPro package has been updated to include RNA-Seq compatible statistical model. This new version is available in Bioconductor 2.12. The package main structure, analysis steps and visualization options are maintained, hence current maSigPro users can upgrade seamlessly to RNA-Seq time series analysis.

References

Conesa A, Nueda MJ, Ferrer A and Talón M (2006) maSigPro: a Method to Identify Significantly Differential Expression Profiles in Time-Course Microarray Experiments. *Bioinformatics*, **22**(9), 1096-1102. doi: 10.1093/bioinformatics/btl056

Dobson AJ (2002) An introduction to generalized linear models. Chapman & Hall/CRC, Boca Ratón, Florida, 2nd edition.

Lu J, Tomfohr JK and Kepler TB (2005) Identifying diferential expression in multiple SAGE libraries: an overdispersed log-linear model approach. BMC Bioinformatics, 6, 165.

McCullagh P and Nelder JA (1989) Generalized linear models. Chapman & Hall/CRC, Boca Ratón, Florida, 2nd edition.

Medina I, Carbonell J, Pulido L, Madeira SC, Goetz S, et al. Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling", (2010) Nucleic Acids Research (38) Web Server Issue, W210-213. doi: 10.1093/nar/gkq388

Nueda MJ, Carbonel J, Medina I, Dopazo J and Conesa A (2010) Serial Expression Analysis: a web tool for the analysis of serial gene expression data. Nucleic Acids Research (38) Web Server Issue, 239-245. doi:10.1093/nar/gkq488