4:40 pm – 6:00 pm

Lecture Hall H 23-3

Statistical Methods in Bioinformatics II

Plasmode data sets: a real-life data simulation with application on RIP-Seq experiments

Tichy, Diana[1]; Benner, Axel[1]

[1]German Cancer Research Center, Division of Biostatistics, Heidelberg, Germany

Abstract. RNA immunoprecipitation combined with high-throughput sequencing (RIP-Seq) is a newly developed method to identify micro-RNA (miRNA) target genes. Analysis of RIP-Seq data is a challenge due inherent noise in the underlying experiments. This noise is caused by unspecific background bindings and artificially created overexpression of specific miRNAs necessary to detect these targets. In other words, Rip-Seq data is similar to RNA-data with respect to overdispersion in the distribution of the genewise read counts, but differs with respect to the underlying noise distribution. To investigate different approaches for the analysis of RIP-Seq data we conduct a simulation study. A central aspect of this study is to avoid distributional assumptions of the RIP-Seq read counts and to preserve the real-life structure of genewise read counts for the underlying RIP-Seq experiment. This can be achieved by using plasmode methods, which generate simulated data sets via resampling from the original data. In doing so a genewise permutation of read counts is applied in a specific manner. Based on the idea of plasmode data generation, we develop a real-life data simulation approach for the RIP-Seq scenario. In this talk we focus on the presentation of the simulation workflow within our published studies on miRNA target identification.

References

1. Tichy, D., Pickl, J., Benner A., Sueltmann, H. (2017) Experimental design and data analysis of Ago-RIP-Seq experiments for the identification of microRNA targets. Briefings in Bioinformatics.
2. Pickl. J.M., Tichy, D., Kuryshev, V., Tolstov, Y., Falkenstein, M., Schueler, J., Reidenbach, D., Hotz-Wagenblatt, A., Kristiansen, G., Roth, W. et al (2016) Ago-RIP-Seq identifies the Polycomb repressive complex I member CBX7 as a major target of miR-375 in prostate cancer progression. Oncotarget, 7(37): 59589-59603.

Demonstrating an association between a clinical covariate and multiple analytes, partly with detecting limits, in cross-sectional metabolomics data

Hothorn, Ludwig A.[1]; Ferrario, Paola G.[2]

[1]Retired from Leibniz University Hannover
[2]Max Rubner-Institut, Department of Physiology and Biochemistry of Nutrition, Karlsruhe

A recent problem is to demonstrate the association between a selected clinical covariate and multiple analytes from diverse metabolomic platforms, e. g. between glomerular filtration rate (to characterize allograft dysfunction) and analytes from both liquid chromatography-mass spectrometry and two dimensional correlated spectroscopy (Bassi, 2017).

A first issue is the style of modeling the clinical covariate: either quantitatively, i.e. in a regression context, or qualitatively, i.e. multiple comparisons by design, or post-hoc classification. In the regression context, a maximum test on three regression models for the arithmetic, ordinal, and logarithmic-linear dose metameters is used (Tukey et al., 1985), whereas the distribution is available

via multiple marginal model approach (mmm) (Pipper, 2012).

The second issue is the different distribution of the many analytes. Some publications ignore this problem (and assume implicitly normal distribution throughout) or use the same transformation for all analytes, such as log-transformation. But it is not a realistic assumption that all analytes follow the same distribution.

The third issue is the measurement of complete analytes and those with detection limits (Snowden, 2017). Analytes with detection limits are considered as left-censored variables assuming some data points are below a certain limit but it is unknown how many.

The fourth issue is the joint consideration of analytes without or with detection limits and following different distributions. Therefore, the concept of most likely distribution (mlt) (Hothorn, 2015) is used, where models for the unconditional or conditional distribution function of any univariate response variable can be estimated by choosing an appropriate transformation function and related parametrisation. Fortunately, most left-censored variables belong to the class of distributions within the mlt-framework.

Using the KarMeN cross-sectional data (Rist, 2017) and the CRAN packages multcomp (for mmm), tukeytrend and mlt, the association between age and selected analytes without and with detection limits are shown in detail.

References

1. Bassi, R., et al.: Metabolomic proling in individuals with a failing kidney allograft. Plos One 12(1), 0169077 (2017).
2. J. W. Tukey, et al. Testing the statistical certainty of a response to increasing doses of a drug. Biometrics, 41(1):295{301, 1985. Hothorn, T., Moest, L., Buehlmann, P.: Most likely transformations. eprint arXiv:1508.06749 (2015)
3. C. B. Pipper, et al. A versatile method for conrmatory evaluation of the eects of a covariate in multiple models. Journal of the Royal Statistical Society Series C-applied Statistics, 61:315{326, 2012.
4. Snowden, S.G., et al. Association between fatty acid metabolism in the brain and alzheimer disease neuropathology and cognitive performance: A nontargeted metabolomic study. Plos Medicine 14(3), 1002266 (2017).
5. Rist, M. Et al.. Metabolite patterns predicting sex and age in participants of the Karlsruhe Metabolomics and Nutrition (KarMeN) study PLOS ONE 12;8 ; e0183228 2017

64. BIOMETRISCHES KOLLOQUIUM
BIOMETRIE: GELEBTE VIELFALT

25.-28. MÄRZ 2018 an der Goethe-Universität Frankfurt

ABSTRACTS

IBS

FACHBEREICH MEDIZIN FRANKFURT
GOETHE-UNIVERSITÄT