Software and Application Note **Open Access**

Michael G. Schimek*, Eva Budinská, Karl G. Kugler, Vendula Švendová, Jie Ding and Shili Lin

TopKLists: a comprehensive R package for statistical inference, stochastic aggregation, and visualization of multiple omics ranked lists

Abstract: High-throughput sequencing techniques are increasingly affordable and produce massive amounts of data. Together with other high-throughput technologies, such as microarrays, there are an enormous amount of resources in databases. The collection of these valuable data has been routine for more than a decade. Despite different technologies, many experiments share the same goal. For instance, the aims of RNA-seq studies often coincide with those of differential gene expression experiments based on microarrays. As such, it would be logical to utilize all available data. However, there is a lack of biostatistical tools for the integration of results obtained from different technologies. Although diverse technological platforms produce different raw data, one commonality for experiments with the same goal is that all the outcomes can be transformed into a platform-independent data format – rankings – for the same set of items. Here we present the R package TopKLists, which allows for statistical inference on the lengths of informative (top-*k*) partial lists, for stochastic aggregation of full or partial lists, and for graphical exploration of the input and consolidated output. A graphical user interface has also been implemented for providing access to the underlying algorithms. To illustrate the applicability and usefulness of the package, we integrated microRNA data of non-small cell lung cancer across different measurement techniques and draw conclusions. The package can be obtained from CRAN under a LGPL-3 license.

Keywords: 62G99; 65K10; 68N01; 65C60; 62F07.

DOI 10.1515/sagmb-2014-0093

1 Introduction

Several high-throughput technologies have emerged in the past decade, most notably next generation sequencing, but also methods that estimate abundance levels of proteins and small molecules. Together, these methods are contributing to an enormous collection of experimental data. However, current research in molecular science is typically based on rather small studies in terms of sample size, many of them addressing the same disease or target. The findings obtained across platforms and studies are often quite diverse and an

***Corresponding author: Michael G. Schimek,** Statistical Bioinformatics, IMI, Medical University of Graz, Auenbruggerplatz 2/V, 8036 Graz, Austria, e-mail: michael.schimek@medunigraz.at
Eva Budinská: Bioinformatics in Translational Research, IBA, Masaryk University, Kotlarska 2, 61137 Brno, Czech Republic
Karl G. Kugler: Institute for Bioinformatics and Systems Biology, Helmholtz Centre Munich, Ingolstädter Landstrasse 1, 85764 Neuherberg, Germany
Vendula Švendová: Statistical Bioinformatics, IMI, Medical University of Graz, Auenbruggerplatz 2/V, 8036 Graz, Austria
Jie Ding: Stanford Cancer Institute, Stanford University, 265 Campus Drive, Stanford, CA 94305-5456, USA
Shili Lin: Department of Statistics, The Ohio State University, 1958 Neil Avenue, Columbus, OH 43210, USA

increasingly important task is to strengthen the evidence of these findings. Hence, there is a strong demand for statistical methods that integrate such findings, for example for combining microarray-based expression measurements with RNA-seq results.

A central task is the integration of such data, which differ in important aspects such as laboratory technology, quantification, scale, and study size. When several studies are combined, the involved sets of genes or of other omics entities usually do not match and missing observations are likely to occur. Moreover, often only subsets of unknown size of these data are relevant or informative. In almost all situations the original metric measurements from the involved studies can be transformed into rank data. Until recently, most integration tools for rank data have been heuristic in nature and could not meet all the above mentioned demands. The few statistical integration approaches in use are limited to microarray results (Yang et al., 2006; Plaisier et al., 2010). A general methodology allowing for the integration of other high-throughput technologies, as well as allowing for a platform and technology mix, even when ranked lists are incomplete, had been lacking until the work of Lin and Ding (2009) and Hall and Schimek (2012). Schimek et al. (2012) combined these approaches and extended them with the goal of processing arbitrarily long multiple ranked lists. To turn such novel statistical methods into practical tools, we have implemented them in the TopKLists R package. It focuses on the nonparametric estimation of the top-*k* list length and on the stochastic aggregation of the identified top-*k* lists. In addition, it also includes conventional aggregation techniques and visual aids for the analysis of ranked lists and the interpretation of aggregation results. In the following, we give an overview of the package and its statistical background, and we apply it to microRNA lung cancer data obtained from a number of different platforms.

2 Structure and availability of the R package

The TopKLists package comprises three modules: (i) TopKInference offers exploratory nonparametric inference for the estimation of the top-*k* list length of paired rankings; (ii) TopKSpace provides various rank aggregation techniques; (iii) TopKGraphics comprises a collection of graphical tools for the exploration of data and for the visualization of aggregation results. The analysis pipeline is to estimate the top-*k* consensus list length first, which also works for more than two ranked lists comprising tens of thousands of items, and to then aggregate the already obtained truncated lists. A new graphical concept, the aggregation map, has been implemented to visualize this graphically. It displays the selected top items with quality measures indicating their relevance with respect to the full ranked lists. Venn-type representations and a summary list form the end of the pipeline. The obtained formal results can then be used in succeeding downstream analysis and experimental validation. The modules can be used as stand-alone R libraries or via a graphical user interface (GUI) for ease of use (see Figure 1 for an example of the GUI interface).

TopKLists is available under the LGPL-3 license from CRAN for all major operating systems. Its R-Forge Web page http://TopKLists.r-forge.r-project.org/ offers the latest development version of the package, detailed vignette-based information about the methods and the package, and instructions on how to analyze the application data described in the example of this paper.

3 Implementation and performance of the R package

The TopKLists package has been designed and implemented for usage on standard desktop computers. To increase the computational speed and performance, parts of the sampling methods have been implemented in C. Therefore, when locally building the package from the source code these methods will be compiled. The graphical user interface, which provides interactive access to several TopKLists' procedures, has been implemented using the gWidgets2 package (Verzani, 2014).

The time needed for computation in the modules TopKInference and TopKSpace depends strongly on the choice of tuning parameters (see next section). Typically, when these parameters are chosen

Figure 1: GUI window with the final aggregation map of the NSCLC application.

appropriately, for lists with thousands of items, the runtime should be in the range of several seconds for inference as well as for aggregation. Most of the implemented aggregation techniques are computationally more demanding than the inference approach. For this reason, aggregation is usually performed on partial lists obtained from the inference procedure. For stochastic aggregation techniques runtime can amount to five or more seconds in a typical scenario.

4 Brief description of the statistical methods

The purpose of the `TopKInference` module is inference on the concordant top length of several rankings comprising the same set of items. The assumptions are: the reliability of rankings breaks down after the first k items due to lack of discriminatory information, irregular or even missing assessments, and substantially more ranked items than assessors exist. The index j_0 is the rank position where the consensus information of two lists degenerates into noise. The estimation of $\hat{j}_0 - 1 = \hat{k}$ is achieved via a moderate deviation-based method developed by Hall and Schimek (2012).

For a given set of items, the input is the overlap of rank positions represented by a sequence of indicators, where $I_j = 1$ if the ranking, given by the second assessor to the item ranked j by the first assessor, is not more than δ index positions distant from j, otherwise $I_j = 0$. The assumption that the variables I_j follow a Bernoulli random distribution can be relaxed. There is theoretical and simulation evidence that dependencies among the ranked lists do not impair the estimates (Hall and Schimek, 2012). As well as the *distance* δ, the

inter-assessor or inter-platform variability, there is another tuning parameter, the *pilot sample size* ν, which is a smoothing parameter controlling the irregularity of assessments or expression measurements. A graphical method called Δ-plot is implemented in `TopKGraphics` which helps to select δ. The parameter ν can be chosen interactively via the GUI.

The overall estimate $\hat{k}_{max}$ for ℓ multiple lists is calculated in the following way: The inference method is applied to all pairwise list combinations $\mathscr{L}=(\ell^2-\ell)$ of the lists, thus we obtain $\mathscr{L}$ values $\hat{k}_j$ (j=1, 2, ..., $\mathscr{L}$). The overall top-k list length is then defined by $\hat{k}_{max}=\max_j(\hat{k}_j)$ (note, other criteria could be chosen as well). The ℓ full lists truncated after $\hat{k}_{max}$ form the input to `TopKSpace`. As the reader will see from the description below, `TopKSpace` is more general, and this specific scenario constitutes a special case that `TopKSpace` is applicable to.

The principle of the `TopKSpace` module is to consolidate information from the ℓ top-k lists to arrive at an aggregate list, *AL*. The top-k lists (L_1, L_2, ..., L_ℓ) may not only be of different lengths, they may also come from studies or assessments that consider different sets of items, hence the underlying spaces (S_1, S_2, ..., S_ℓ) from which the top-k lists are derived may actually be different. The goal therefore is to find the top-k list, *AL*, from the aggregate new space $(\cup_{i=1}^{\ell} L_i)$, such that the weighted sum of distances between each of the input lists and *AL* will be the minimum among lists of the same length. Two distance measures, Kendall's τ and Spearman's footrule, are available in the package. Both take the differences in the underlying spaces into account (Lin, 2010). There are three classes of algorithms implemented in `TopKSpace`, namely Borda's method, Markov chain (MC) algorithms (Lin, 2010), and a cross entropy Monte Carlo (CEMC) method taking advantage of the new order explicit algorithm (OEA) as described by Lin and Ding (2009). The Borda and MC methods consist of heuristic algorithms that do not directly optimize the objective function (i.e., minimizing the weighted distances), whereas the CEMC method employs a Monte Carlo search procedure for achieving this optimization. Borda and MC algorithms run substantially faster than the CEMC algorithm, however the latter usually achieves better results. Nevertheless, simulation studies indicate that taking the underlying space into consideration has a much greater impact than using different algorithms.

5 Application to cross platform microRNA profiles

Stimulated by the methodological discussion of microRNA profiling in Baker (2010), we compared non-small cell lung cancer (NSCLC) cell lines grown in vitro and in vivo as xenograft models across platforms. From the NCBI GEO database we retrieved data (Tam et al., 2014) of five in vitro and five in vivo samples from three different platforms: (i) GSE51501, Illumina Human v2 MicroRNA Expression BeadChip; (ii) GSE51504, NanoString nCounter Human v1 miRNA Expression Assay; (iii) GSE51507, Illumina HiSeq 2500 (High Throughput Sequencing, abb. HTS). Data (i) and (ii) were normalized using Bioconductor's `normalize.quantiles` and analysed with the `R`-package `samr` (Tibshirani et al., 2011) (cell line vs. xenograft). The next generation sequencing data (iii) were processed with Bioconductor's `DESeq2` (Love et al., 2014). The resulting miRNA expression values (items) from each platform were ranked according to their FDR-adjusted p-values. Those items common to all three lists were the input to the package `TopKLists` and comprise N=531 miRNAs. The thus obtained ranked lists and the corresponding code for the data analysis can be accessed and downloaded from the `TopKLists` Web page.

Data exploration led to the choice of δ=40 and ν=22 for the inference procedure (for details please refer to the show case instructions on the Web page). The obtained result was $\hat{k}_{max}=12$ and the three lists were truncated at this index position. The associated aggregation map is displayed in Figure 1. Its left group (*NanoString-HTS-BeadChip*) represents the aggregation result when all three platforms are integrated, and the right group (*HTS-BeadChip*) when *NanoString* is excluded. A group comparison allows us to identify platform differences ('white' denotes that an item is top-listed in only one of the concerned lists, 'gray' otherwise). NanoString had the strongest impact on the selection of top-ranking miRNAs and forms, with the other two platforms, a highly conforming group of six items. hsa-miR-107, on rank 7 in NanoString, is of special interest, as it was

shown to suppress growth of NSCLC cell lines and induced a G1 cell cycle arrest in H1299 cells (Takahashi et al., 2009). It was ranked 83 and 53 index positions away in HTS and BeadChip, and therefore is represented as *close* in the map by a 'red' color as opposed to more *distant* ranks, presenting themselves in 'yellow'.

Finally, we calculated an optimized aggregate list $\widehat{AL}$ for the three lists truncated after $\hat{k}_{max}=12$ via CEMC under Kendall's τ and Spearman's footrule. In Table 1 (columns 1 and 2) the final items are displayed in their new rank order. For comparison, the 12 top-ranked miRNAs based on Fisher's method for combining p-values (Fisher, 1925) are listed in the third column of the same table. We have used the function `fisher.method` from the `R` package `MADAM` (Kugler et al., 2010) with Benjamini-Hochberg p-value correction.

The CEMC stochastic search algorithm may select items that are top-ranked only in one of the lists (here *BeadChip*). This applies to the following items in Table 1: hsa-miR-576-5p, hsa-miR-490-5p, hsa-miR-139-5p, hsa-miR-1233, hsa-miR-1284, and hsa-miR-505. In contrast, Fisher's method tends to select 'consensus' items, thus having greater agreements with the aggregation map results. Within the top-5 positions the same items are selected by all methods. Only the orders are permuted. However, apart from this rather limited set of overlapping miRNAs, both aggregate lists from CEMC, as well as the aggregation map discussed before, clearly point at substantial platform differences.

Using the miRSystem (Lu et al., 2012) we found the final lists (one for Kendall, one for Spearman, and one for Fisher's method) of ranked miRNAs to be highly enriched for the JAK-STAT signaling pathway and the Hedgehog signaling pathway both of which were suggested to play an important role in NSCLC. The interesting candidates comprise hsa-miR-143, which is among a set of 43 miRNAs that were found to be differentially expressed between noncancerous lung tissues and lung cancer tissues (Yanaihara et al., 2006) and has also been suggested as a putative biomarker for NSCLC (Gao et al., 2010). Finally, on rank 1 and on rank 2, respectively, we have the RAB14 targeting tumor suppressor hsa-miR-451 (Wang et al., 2011).

6 Discussion

A major advantage over ground truth-based and other ad hoc methods is `TopKLists`'s ability to provide an objective data-driven top-list length estimate and a consolidated as well as optimized aggregate ranking based on multiple input lists. In the described NSCLC application it allowed us to efficiently select those miRNAs which are supported by all three or at least by two platforms. In addition, a consolidated set of miRNAs under different aggregation criteria (distance measures) could be obtained. The aggregation map

Table 1: Aggregate list results of the NSCLC application.

Rank	$\widehat{AL}$ (Kendall)	$\widehat{AL}$ (Spearman)	Fisher's method
1	**hsa-miR-451**	**hsa-miR-143**	**hsa-miR-143**
2	**hsa-miR-223**	**hsa-miR-451**	**hsa-miR-451**
3	**hsa-miR-199a-5p**	**hsa-miR-223**	**hsa-miR-223**
4	**hsa-miR-143**	**hsa-miR-144**	**hsa-miR-144**
5	**hsa-miR-144**	**hsa-miR-199a-5p**	**hsa-miR-199a-5p**
6	**hsa-miR-150**	hsa-miR-1284	**hsa-miR-145**
7	hsa-miR-576-5p	hsa-miR-139-5p	**hsa-miR-133a**
8	hsa-miR-490-5p	**hsa-miR-150**	**hsa-miR-195**
9	hsa-miR-139-5p	**hsa-miR-195**	**hsa-miR-214**
10	**hsa-miR-107**	**hsa-miR-145**	**hsa-miR-150**
11	hsa-miR-1233	hsa-miR-505	**hsa-miR-1246**
12	**hsa-miR-133a**	**hsa-miR-1246**	**hsa-miR-142-5p**

First and second columns: CEMC consolidated list results under the distance measures Kendall's τ and Spearman's footrule. Third column: consolidated list using Fisher's method for combining p-values (miR-symbols in **bold** coincide with the aggregation map result in Figure 1).

as well as the stochastic CEMC aggregation method also aid in giving an answer to the problem raised in Baker (2010): Although there is high conformity among the top-5 items across all (graphical and stochastic) aggregation techniques, our results support the observation that substantial platform differences exist with respect to all other miRNA measurements. As has been demonstrated in this paper, TopKLists offers a variety of highly useful and computationally efficient state-of-the-art methods for omics data integration, most of them implemented in R for the first time.

Acknowledgments: Gratefully acknowledged are financial support of the Medical University of Graz (MGS, VS) as well as funding by the Deutsche Forschungsgemeinschaft (DFG) to SFB 924 (KGK) and by the US National Science Foundation grant DMS-1220772 (SL).

References

Baker, M. (2010): "MicroRNA profiling: separating signal from noise," Nat. Methods, 7, 687–692.

Fisher, R. A. (1925): Statistical methods for research workers, Edinburgh: Oliver and Boyd, ISBN 0-05-002170-2.

Gao, W., Y. Yu, H. Cao, H. Shen, X. Li, S. Pan and Y. Shu (2010): "Deregulated expression of miR-21, miR-143 and miR-181a in non small cell lung cancer is related to clinicopathologic characteristics or patient prognosis," Biomed Pharmacother, 64, 399–408.

Hall, P. and M. G. Schimek (2012): "Moderate-deviation-based inference for random degeneration in paired rank lists," J. Am. Stat. Assoc., 107, 661–672.

Kugler, K. G., L. A. Mueller and A. Graber (2010): "MADAM: an open source meta-analysis toolbox for R and Bioconductor," Source Code Biol. Med., 5, 3.

Lin, S. (2010): "Space oriented rank-based data integration," Stat. Appl. Genet. Mol. Biol., 9:Article20. doi: 10.2202/1544-6115.1534. Epub 2010 Apr 9.

Lin, S. and J. Ding (2009): "Integration of ranked lists via Cross Entropy Monte Carlo with applications to mRNA and microRNA studies," Biometrics, 65, 9–18.

Love, M. I., W. Huber, and S. Anders (2014): "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," Genome. Biol., 15:550.

Lu, T.-P., C.-Y. Lee, M.-H. Tsai, Y.-C. Chiu, C. K. Hsiao, L.-C. Lai and E. Y. Chuang (2012): "miRSystem: an integrated system for characterizing enriched functions and pathways of microRNA targets," PloS One, 7, e42390.

Plaisier, S. B., R. Taschereau, J. A. Wong and T. G. Graeber (2010): "Rank-rank hypergeometric overlap: identification of statistically significant overlap between gene-expression signatures," Nucleic Acids Res., 38, e169.

Schimek, M. G., A. Myšicková and E. Budinská (2012): "An inference and integration approach for the consolidation of ranked lists," Commun. Stat. – Simul. C., 41, 1152–1166.

Takahashi, Y., A. R. Forrest, E. Maeno, T. Hashimoto, C. O. Daub and J. Yasuda (2009): "MiR-107 and MiR-185 can induce cell cycle arrest in human non small cell lung cancer cell lines," PloS One, 4, e6677.

Tam, S., R. de Borja, M.-S. Tsao, and J. D. McPherson (2014): "Robust global microRNA expression profiling using next-generation sequencing technologies," Lab. Invest., 94, 350–358.

Tibshirani, R., G. Chu, B. Narasimhan and J. Li (2011): "Significance analysis of microarrays – samr: R package version 2.0," URL http://CRAN.R-project.org/package=samr.

Verzani, J. (2014): "gWidgets: gWidgets API for building toolkit-independent, interactive GUIs – R package version 0.0-54," URL http://CRAN.R-project.org/package=gWidgets.

Wang, R., Z. Wang, J. Yang, X. Pan, W. De and L. Chen (2011): "MicroRNA-451 functions as a tumor suppressor in human non-small cell lung cancer by targeting ras-related protein 14 (RAB14)," Oncogene, 30, 2644–2658.

Yanaihara, N., N. Caplen, E. Bowman, M. Seike, K. Kumamoto, M. Yi, R. M. Stephens, A. Okamoto, J. Yokota, T. Tanaka, et al. (2006): "Unique microRNA molecular profiles in lung cancer diagnosis and prognosis," Cancer Cell, 9, 189–198.

Yang, X., S. Bentink, S. Scheid and R. Spang (2006): "Similarities of ordered gene lists," J Bioinform Comput Biol., 4, 693–708.