BIOINFORMATICS APPLICATIONS NOTE Vol. 31 no. 1 2015, pages 128–130
doi:10.1093/bioinformatics/btu610

Gene expression

Advance Access publication September 18, 2014

RAMONA: a Web application for gene set analysis on multilevel omics data

Steffen Sass[1], Florian Buettner[1], Nikola S. Mueller[1] and Fabian J. Theis[1,2,*]

[1]Institute of Computational Biology, Helmholtz Zentrum München, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany and [2]Department of Mathematics, Technische Universität München, Boltzmannstraße 3, 85747 Garching, Germany

Associate Editor: Ziv Bar-Joseph

ABSTRACT

Summary: Decreasing costs of modern high-throughput experiments allow for the simultaneous analysis of altered gene activity on various molecular levels. However, these multi-omics approaches lead to a large amount of data, which is hard to interpret for a non-bioinformatician. Here, we present the remotely accessible multilevel ontology analysis (RAMONA). It offers an easy-to-use interface for the simultaneous gene set analysis of combined omics datasets and is an extension of the previously introduced MONA approach. RAMONA is based on a Bayesian enrichment method for the inference of overrepresented biological processes among given gene sets. Overrepresentation is quantified by interpretable term probabilities. It is able to handle data from various molecular levels, while in parallel coping with redundancies arising from gene set overlaps and related multiple testing problems. The comprehensive output of RAMONA is easy to interpret and thus allows for functional insight into the affected biological processes. With RAMONA, we provide an efficient implementation of the Bayesian inference problem such that ontologies consisting of thousands of terms can be processed in the order of seconds.

Availability and implementation: RAMONA is implemented as ASP.NET Web application and publicly available at http://icb.helmholtz-muenchen.de/ramona.

Contact: fabian.theis@helmholtz-muenchen.de

Received on March 19, 2014; revised on September 9, 2014; accepted on September 10, 2014

1 INTRODUCTION

Decreasing costs of large-scale molecular profiling studies, such as transcriptomics or proteomics, allow for the joint analysis of several molecular levels in parallel. The crucial step in the analysis of such diverse data is to combine the different levels such that a comprehensive insight in the response of genes to these conditions can be assessed. This in turn can be directly linked to the underlying biological processes affecting the activity of genes on several molecular levels. However, these kinds of analyses are not straightforward, and often the molecular levels are treated as independent to allow the use of single-omics analysis techniques.

In practice, gene response is initially determined by using statistical methods. Among the resulting set of altered genes, one usually searches for overrepresented biological processes by applying common gene set enrichment methods (Boyle *et al.*, 2004; Subramanian *et al.*, 2005) that incorporate functional annotations from databases like Gene Ontology (GO) (Ashburner *et al.*, 2000) or the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa *et al.*, 2011). Even though there exists a multitude of easy-to-use Web-based enrichment tools (Huang *et al.*, 2007; Zhang *et al.*, 2005), they are only capable of analyzing a single molecular level. Furthermore, no Web tool is available that properly deals with term redundancies appearing frequently, e.g. due to the tree structure of GO.

To provide a powerful method to integrate multilevel gene response data for the determination of altered biological processes, we recently introduced the *multilevel ontology analysis* (*MONA*) (Sass *et al.*, 2013). MONA is a model-based Bayesian method, which is able to integrate datasets from multiple molecular levels by simultaneously dealing with term redundancies and related multiple testing problems.

However, the usage of the standalone MONA application can be a cumbersome process, as the user has to specify the data structure of the activated genes and their term annotations by himself/herself. Furthermore, it lacks a comprehensive visualization of the results and can be run only on Windows machines, as it depends on the .NET library.

Here we introduce a Web-based implementation of MONA, called *remotely accessible MONA* (*RAMONA*), which is designed with the focus on practical usability for any applied researcher. It offers three models to analyze most common experimental setups. The Web interface is capable of processing many given gene identifiers as well as of automatically mapping them to widely used ontologies derived from GO and KEGG. The detailed output of RAMONA includes an interactive visualization of the inferred active terms in the context of their respective pathways or ontology hierarchy. This provides functional insight into the activity of biological processes and the role of associated genes responding to the given conditions by providing relevant details on the resulting processes.

2 OVERVIEW

RAMONA is a Web-based application whose interface is implemented in the Mono ASP.NET framework. The underlying MONA application is written in C# and is based on the Infer.NET framework (Minka *et al.*, 2012).

MONA currently provides three models of molecular interactions (Fig. 1). The single-level model can be used when

*To whom correspondence should be addressed.

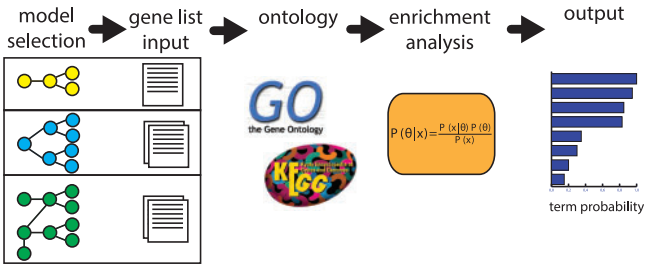

Fig. 1. RAMONA workflow. The user has to specify the input according to the selected model. He can choose between GO and KEGG as ontologies. Using a Bayesian modeling approach, the tool is able to infer non–redundant-enriched terms among the given gene lists

measurements are available only on a single level. This corresponds to the principle of the model-based gene set analysis (Bauer *et al.*, 2010). The cooperative model accounts for studies where measurements of two different levels are available, which may be regarded as independent noisy observations (e.g. mRNA and protein) of an underlying common gene response. The inhibitory model is applicable when two species are measured, but they could not be interpreted as independent measurements of the hidden gene function. A prominent example is the post-transcriptional modulation of an mRNA expression by miRNAs.

Given the user input, the MONA algorithm infers the marginal posterior probability of the term activity using a Bayesian network as described in Sass *et al.* (2013). The user has to specify the input of RAMONA according to the selected model. In general, this must be a set of genes that show a special behavior like the response to a certain condition and a set of measured genes, which is referred to as background. A typical example for an input would be two lists of differentially expressed genes between two conditions for both mRNA and protein level. For the cooperative model, two lists of differentially expressed genes together with a background of all measured genes have to be provided. The probabilistic nature of RAMONA allows for the analysis of experiments, where different numbers of genes are measured (e.g. usually the case for mRNA and protein data). For the inhibitory model, a set of inhibited genes has to be specified in addition to the responding genes and background. All these sets can be provided by the text field input or text file upload. The user can manipulate the shape of all priors via the expert settings to gain a sparser result; uniform priors are used as default settings for the single and cooperative case. In case of the inhibitory model, weakly informative priors are used as discussed previously (Sass *et al.*, 2013).

RAMONA supports a variety of common gene identifiers for several organisms that are mapped to specific terms. These terms include biological processes, molecular functions and cellular components from GO (Ashburner *et al.*, 2000) as well as pathways from KEGG (Kanehisa *et al.*, 2011).

The actual MONA process runs in a background thread on the Web server with runtime depending on the size of the input and the selected ontology. For common setups, RAMONA runs no longer than a minute. In addition to the term probabilities provided by the model-based enrichment analysis, *P*-values for enrichment of the individual terms are calculated by using Fisher's exact test on each molecular level separately when the cooperative model is chosen.

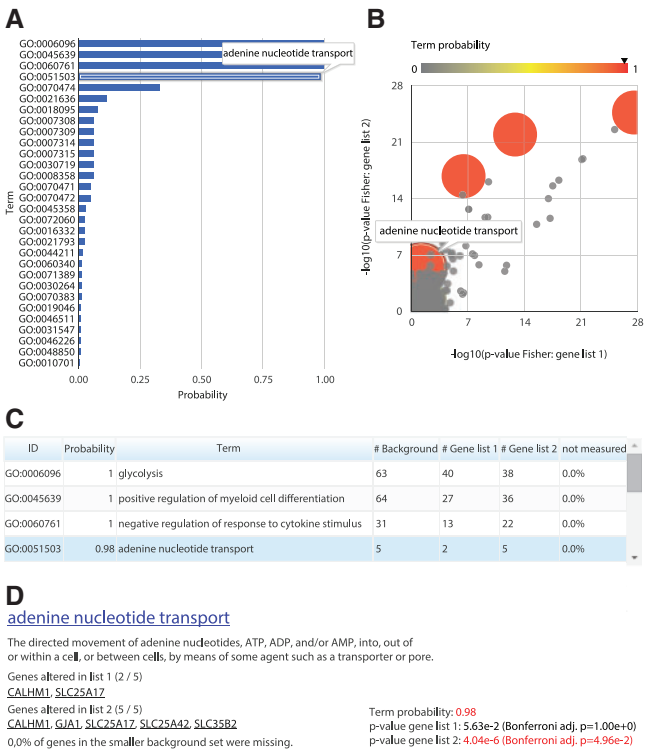

Fig. 2. The RAMONA output. (**A**) Resulting term probabilities are shown in a barplot. (**B**) If the cooperative model is chosen, a scatterplot can be displayed that shows the *P*-values for each term determined by the traditional gene set analysis (Fisher's exact test) on the two input gene lists individually. The color and size of the points correspond to the RAMONA term probability. (**C**) The tabular representation gives an overview of all relevant term information. (**D**) Additional information can be obtained by clicking on the terms in any of the three panels. This information includes the set of altered genes for each level as well as the decision whether a term is active (red) or not (black) for RAMONA or Fisher's exact test

The output of RAMONA consists of three parts: a plot panel, a table and a panel for further term information (Fig. 2). If the cooperative model was chosen, the user can switch between a barplot (Fig. 2A) and a scatterplot (Fig. 2B) to illustrate the results of RAMONA. Otherwise, only the barplot is shown, which displays the term probabilities for the top 30 terms. The scatterplot displays the *P*-values of the Fisher tests, which are performed on the two input lists individually, in comparison with the term probabilities. This representation allows the user to determine the effect of the two individual input gene lists on the RAMONA outcome. In addition, it exposes the redundancies that arise from the traditional gene set analyses and that do not appear in the RAMONA results. The table (Fig. 2C) provides an overview of all relevant information on the terms, namely, the number of assigned genes and the number of altered genes in the given gene list(s). Additionally, the percentage of assigned genes is shown, which were missing in the smaller background set.

By selecting a term, in the barplot, scatterplot or table, detailed information for the respective term can be displayed (Fig. 2D). This includes for each molecular level a list of regulated genes assigned to this term as well as the percentage of missing genes in

case of the cooperative model. Furthermore, a link to the term database is provided, which allows for a graphical mapping of the results. In case of KEGG, the respective pathway is displayed, and the regulated genes are marked by a color for each molecular level. If GO was selected, the GO tree will be shown illustrating the term hierarchy, including all active terms ($P > 0.5$).

3 CONCLUSION

The integration of data from multiple molecular levels for gene set analysis is becoming more and more important and therefore requires appropriate methods, which are easy to use for applied researchers. Important challenges we address with RAMONA include dependencies between terms with an ontology and interactions between molecular levels. We provide an easy-to-use Web application that can be used to infer non-redundant biological processes either from multiple molecular levels or from a single molecular level.

ACKNOWLEDGEMENT

The authors thank Benedikt Rauscher and Michael Schollerer for preliminary work on the Web interface.

Funding: This work was supported by The European Research Council [Latent Causes: 259294]; the Deutsche Forschungsgemeinschaft [InKoMBio: SPP 1395]; and the Federal Ministry of Education and Research [GerontoSys: FKZ 0315576C; LungSys: FKZ 0316042I; Virtual Liver: FKZ 0315752].

Conflict of interest: none declared.

REFERENCES

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Bauer,S. *et al.* (2010) GOing bayesian: model-based gene set analysis of genome-scale data. *Nucleic Acids Res.*, **38**, 3523–3532.

Boyle,E.I. *et al.* (2004) Go::termfinder–open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes. *Bioinformatics*, **20**, 3710–3715.

Huang,D.W. *et al.* (2007) David bioinformatics resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Res.*, **35**, W169–W175.

Kanehisa,M. *et al.* (2011) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, **40**, D109–D114.

Minka,T. *et al.* (2012) Infer.NET 2.5. Microsoft Research Cambridge. http://research.microsoft.com/infernet (22 September 2014, date last accessed).

Sass,S. *et al.* (2013) A modular framework for gene set analysis integrating multi-level omics data. *Nucleic Acids Res.*, **41**, 9622–9633.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.

Zhang,B. *et al.* (2005) Webgestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res.*, **33**, W741–W748.