Bioinformatics, 31(18), 2015, 3072–3074
doi: 10.1093/bioinformatics/btv323
Advance Access Publication Date: 21 May 2015
Applications Note

OXFORD

Systems biology

pwOmics: an R package for pathway-based integration of time-series omics data using public database knowledge

Astrid Wachter* and Tim Beißbarth

Department of Medical Statistics, Georg-August-University Göttingen, Germany

*To whom correspondence should be addressed

Associate Editor: Alfonso Valencia

Received on February 27, 2015; revised on May 5, 2015; accepted on May 17, 2015

Abstract

Summary: Characterization of biological processes is progressively enabled with the increased generation of omics data on different signaling levels. Here we present a straightforward approach for the integrative analysis of data from different high-throughput technologies based on pathway and interaction models from public databases. *pwOmics* performs pathway-based level-specific data comparison of coupled human proteomic and genomic/transcriptomic datasets based on their log fold changes. Separate downstream and upstream analyses results on the functional levels of pathways, transcription factors and genes/transcripts are performed in the cross-platform consensus analysis. These provide a basis for the combined interpretation of regulatory effects over time. Via network reconstruction and inference methods (Steiner tree, dynamic Bayesian network inference) consensus graphical networks can be generated for further analyses and visualization.
Availability and implementation: The R package *pwOmics* is freely available on Bioconductor (http://www.bioconductor.org/).
Contact: astrid.wachter@med.uni-goettingen.de

1 Introduction

High-throughput technologies applied in systems biology research generate large amounts of molecular information nowadays. Interpretation of genome- and proteome-wide data is dependent on current analysis tools. As each technique shows a certain bias and has natural limitations in identifying full signaling responses (Yeger-Lotem *et al.*, 2009), cross-platform analysis is an up-to-date approach in order to connect biological implications on different signaling levels. Usage of diverse data types provides a deeper understanding of global biological functions and the underlying processes (Kholodenko *et al.*, 2012). Thus, development of integrative software solutions for data from different high-throughput techniques is a current major challenge for bioinformatic analysis. Existing widely used commercial software solutions such as QIAGEN's Ingenuity® Pathway Analysis (IPA®, QIAGEN Redwood City, www.qiagen.com/ingenuity) or MetaCore (GeneGo, Inc., St. Joseph, MI) and also open-source software, such as Cytoscape (Shannon *et al.*, 2003), often handle proteomic and genomic/transcriptomic data as if coming from the same functional level. More specific integration tools which are considering these levels include, e.g. the web tool IMPaLA (Kamburov *et al.*, 2011), which provides knowledge based data integration on transcriptomics or proteomics data combined with metabolomics data, and the webserver SteinerNet (Tuncbag *et al.*, 2012), which enables integration of transcriptional, proteomic and interactome data utilizing Steiner trees. However, *pwOmics* combines these distinct omics levels of evidence in order to refine the understanding of molecular mechanisms including the biologically important time effect. Thereby, it joins tools used for network analysis (Kristensen *et al.*, 2014), but adds a level of complexity by attributing weight to the different functional levels of measurement in the first place and the dimension of time in the second place. We implemented *pwOmics* as open-source package for R, a free software environment for statistical computing commonly used for bioinformatic analyses.

2 Approach

pwOmics provides analyses functionalities and comparative integration features for coupled human proteome and genome/transcriptome datasets. The analysis workflow is adapted to account for the biological control mechanisms occurring on the different regulation levels such as transcriptional control on gene level, mRNA processing on transcript level and post-translational modifications on protein level, as illustrated in Figure 1. The two datasets are initially analyzed separately enabling a level-specific interpretation of up- and downstream changes of regulatory molecules. The protein based downstream analysis comprises the pathway-based identification of transcription factors (TF) of differentially abundant proteins and their target genes. The gene/transcription based upstream analysis identifies TFs and proteomic regulators based on differentially expressed transcripts or genes. As high-throughput data are increasingly used to follow time-dependent biological regulation after perturbation, the main benefit of *pwOmics* is the cross-platform time series analysis functionality, but consensus analysis can be performed also on single time point measurements.

3 Package features

3.1 Databases

Existing knowledge stored in public databases is a key element for data integration in the approach outlined above (Kramer *et al.*, 2014). Databases used here are pathway databases, TF-target databases and a protein–protein interaction database. Pathway databases can be selected individually or as combination of KEGG (Kanehisa *et al.*, 2014), Reactome (Croft *et al.*, 2014), Pathway Interaction Database (Schaefer *et al.*, 2009) and Biocarta (Nishimura, 2001). The information is used as gene sets in the downstream analysis and combined with topological information in upstream analysis. Prior knowledge for network reconstruction is based on the connected graph from protein–protein-interaction (PPI) database STRING (Franceschini *et al.*, 2013). For TF-target gene identification processes the user can choose from databases ChEA (Lachmann *et al.*, 2010) and/or Pazar (Portales-Casamar *et al.*, 2009) or specify an own file, e.g. containing commercial database information.

3.2 Individual comparative analysis

In the individual analysis database information is used to identify signaling molecules of the different functional levels for a level-specific comparison. Identification of pathways containing differentially abundant proteins is performed via a Biopax model generated by the R package *rBiopaxParser* (Kramer *et al.*, 2013) on basis of the selected pathway databases. Enrichment of pathways in downstream analysis and TFs in upstream analysis is optional. Upstream regulators of TFs are identified via their pathways, but only those pathways are considered further which contain a user-specified number of TFs. Overlapping proteins found as neighbors of a certain order of those TFs are assumed to be proteomic regulators. Easy access to the individual level results is provided.

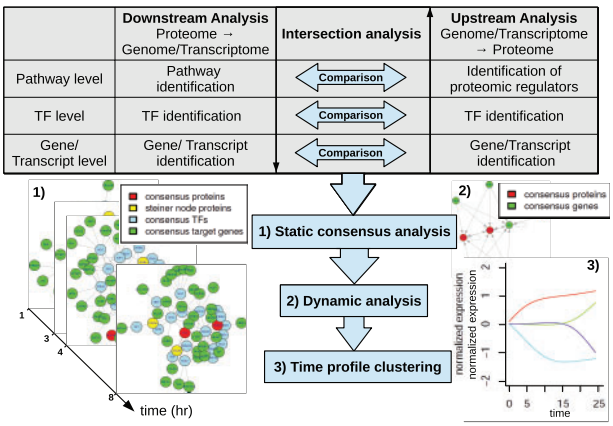

Fig. 1. *pwOmics* downstream and upstream analysis. Exemplarily shown are results of a static consensus analysis, a dynamic analysis and the time profile clustering

3.3 Consensus analysis

In the consensus analysis the intersection of signaling molecules on each functional level is identified and used for building consensus nets. For each matching time point a Steiner tree (Sadeghi and Fröhlich, 2013) is generated (implemented via the shortest paths based approximation algorithm) on the basis of intersecting proteins and TFs from up- and downstream analysis and the connected PPI STRING network. For this network reconstruction method intersecting molecules regarded as 'terminal nodes' are mapped to the PPI-network and those pathway components on shortest interconnecting paths are included which provide the shortest length of the overall network. Subsequently intersecting TF-target relations are included to contribute to the static consensus graphs for each matching time point. The dynamic consensus analysis additionally considers signaling changes over time by applying dynamic Bayesian network inference via the R package *ebdbNet* (Rau *et al.*, 2010). Nodes considered in this step are those identified in all static consensus graphs. With smoothing splines an appropriate number of time points are generated under the simplifying assumption of a gradual change of signaling over time. This longitudinal dataset is then used for the inference step. The result allows a significance level-based visualization of the dynamic Bayesian network.

3.4 Time profile clustering

To identify similar co-regulation patterns over time *pwOmics* provides an integrated time profile clustering, based on the soft clustering fuzzy c-means algorithm implemented in the R package *Mfuzz* (Kumar *et al.*, 2007). The soft-clustering approach has the advantage of assigning several clusters to one signaling molecule based on similarity of log-fold change dynamics to several clusters. Thus it enables an adequate clustering of complex expression time profiles, which are characterized by fine-tuned transcriptional mechanisms.

3.5 Data visualization

For easier biological interpretation users can visualize following results: (i) Static consensus nets—based on matching time point comparisons of the two datasets. (ii) Dynamic consensus net—based on dynamic Bayesian network inference. (iii) Time profile clustering—based on softly clustered log-fold changes with a combined visualization of proteins and genes/transcripts.

4 Summary

We developed an R package as integrative pathway-based level-specific tool for the analysis and interpretation of signaling measured in parallel on different platforms. The presented approach enables the reduction of results to a very reliable set of regulatory signaling components, time profile clustering and the interpretation of static and dynamic consensus results. Further details and examples are provided in the package documentation.

Funding

This work was supported by the German Federal Ministry of Education and Research via the projects MetastaSys [0316173A] and MMML-Demonstrators [031A428B].

Conflict of Interest: none declared.

References

Croft,D. *et al*. (2014) The Reactome pathway knowledgebase. *Nucleic Acids Res.*, **42**, D472–D477.

Franceschini,A. *et al*. (2013) STRING v9.1: protein–protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.*, **41**, D808–D815.

Kamburov,A. *et al*. (2011) Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA. *Bioinformatics*, **27**, 2917–2918.

Kanehisa,M. *et al*. (2014) Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.*, **42**, D199–D205.

Kholodenko,B. *et al*. (2012) Computational approaches for analyzing information flow in biological networks. *Sci. Signal*, **5**, re1-re1.

Kramer,F. *et al*. (2013) rBiopaxParser—an R package to parse, modify and visualize BioPAX data. *Bioinformatics*, **29**, 520–522.

Kramer,F. *et al*. (2014) R-based software for the integration of pathway data into bioinformatic algorithms. *Biology*, **3**, 85–100.

Kristensen,V.B. *et al*. (2014) Principles and methods of integrative genomic analyses in cancer. *Nat. Rev. Cancer*, **14**, 299–313.

Kumar,L. and Futschik,M.E. (2007) Mfuzz: a software package for soft clustering of microarray data. *Bioinformation*, **2**, 5–7.

Lachmann,A. *et al*. (2010) ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics*, **26**, 2438–2444.

Nishimura,D. (2001) BioCarta. *Biotech Software & Internet Report*, **2**, 117–120.

Portales-Casamar,E. *et al*. (2009) The PAZAR database of gene regulatory information coupled to the ORCA toolkit for the study of regulatory sequences. *Nucleic Acids Res.*, **37**, D54–D60.

Rau,A. *et al*. (2010) An empirical Bayesian method for estimating biological networks from temporal microarray data. *Statistical Applications in Genetics and Molecular Biology*, **9**, 1544–6115.

Sadeghi,A. and Fröhlich,H. (2013) Steiner tree methods for optimal sub-network identification: an empirical study. *BMC Bioinformatics*, **14**, 144.

Schaefer,C.F. *et al*. (2009) PID: the pathway interaction database. *Nucleic Acids Res.*, **37**, D674–D679.

Shannon,P. *et al*. (2003) Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.

Tuncbag,N. *et al*. (2012) SteinerNet: a web server for integrating 'omic' data to discover hidden components of response pathways. *Nucleic Acids Res.*, **40**, W505–W509.

Wang,X. and Zhang,B. (2013) customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search. *Bioinformatics*, **29**, 3235–3237.

Yeger-Lotem,E. *et al*. (2009) Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. *Nat. Genet.*, **41**, 316–323.

Yosef,N. and Regev,A. (2011) Impulse control: temporal dynamics in gene transcription. *Cell*, **144**, 886–896.