BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 14 2013, pages 1823–1824
doi:10.1093/bioinformatics/btt291

Systems biology

Advance Access publication June 5, 2013

opm: an R package for analysing OmniLog® phenotype microarray data

Lea A. I. Vaas[1], Johannes Sikorski[2,*], Benjamin Hofner[3], Anne Fiebig[2], Nora Buddruhs[2], Hans-Peter Klenk[2] and Markus Göker[2]

[1]CBS-KNAW Fungal Biodiversity Centre, Bioinformatics, Uppsalalaan 8. 3584 CT Utrecht, The Netherlands, [2]Leibniz Institute DSMZ–German Collection of Microorganisms and Cell Cultures, Inhoffenstraße 7B, 38124 Braunschweig, Germany and [3]Institut für Medizininformatik, Biometrie und Epidemiologie, Universität Erlangen-Nürnberg, Waldstr. 6, 91054 Erlangen, Germany

Associate Editor: Olga Troyanskaya

ABSTRACT

Summary: opm is an R package designed to analyse multidimensional OmniLog® phenotype microarray (PM) data. opm provides management, visualization and statistical analysis of PM data, including curve-parameter estimation and discretization, dedicated and customizable plots, metadata management, automated generation of textual and tabular reports, mapping of substrates to databases, batch conversion of files and export to phylogenetic software in the YAML markup language.

Availability: opm is distributed under the GPL through the Comprehensive R Archive Network (http://cran.r-project.org/package=opm) along with a comprehensive manual and a user-friendly tutorial. Further information may be found at http://www.dsmz.de/research/microorganisms/projects/.

Contact: johannes.sikorski@dsmz.de

Received on October 26, 2012; revised on March 4, 2013; accepted on May 17, 2013

1 INTRODUCTION

High-throughput phenotypic testing is increasingly important for exploring the biology of bacteria, fungi, yeasts and animal cell lines, e.g. human cancer cells (Bochner, 2009), and offers a great potential for testing gene function and improving genome annotation (Bochner *et al.*, 2001). The OmniLog® PM system monitors simultaneously, on a longitudinal time scale, the phenotypic reaction to up to 2000 environmental challenges spotted on sets of 96-well microtiter plates (Bochner, 2009) as respiration kinetics with an often sigmoidal shape.

As we discussed previously, by using a combination of existing R packages, *ad hoc* code and even manual manipulations (Vaas *et al.*, 2012), there is an increasing demand to explore OmniLog® PM data not only qualitatively but also quantitatively, taking into account associated metadata on the organisms and experimental settings. As a result, we present here the broad and flexible design of the R (R Development Core Team, 2012) package opm, which allows users to analyse OmniLog® PM data within a wide frame of research tasks such as -omics approaches, systems biology, ecology and taxonomy.

*To whom correspondence should be addressed.

2 FEATURES

2.1 Data input and storage

The raw kinetic values can be imported from CSV [from the OmniLog® or MicroStation™ reader (BiOLOG Inc., 2009)], or YAML (http://www.yaml.org/; used by opm itself), yielding S4 objects (Chambers, 1998). These containers comprise one object per input plate (with the measurements and the few metadata output by the OmniLog® software) and optionally estimated curve parameters together with the estimate settings, and/or additional metainformation on the experiment. Batch conversion of large numbers of files (optionally non-interactively via `Rscript`) and parallel computation is supported. R functions and experimental examples are provided and described in great detail in the manual and the user-friendly tutorial. The main opm workflow as described later in the text is summarized in Figure 1.

2.2 Data enrichment

Important biological information stored in the raw curve kinetics can be summarized in the *curve parameters* lag phase (λ), steepness of slope (μ), maximum curve height (A) and area under the curve (AUC), using a fast and approximate method (only A and AUC) or spline-fit algorithms (Eilers *et al.*, 1996; Kahm *et al.*, 2010; Wood, 2006). Fitting splines rather than growth models is more robust for these data (Vaas *et al.*, 2012). Confidence intervals can be obtained via bootstrapping or aggregating predefined groups. To enable PM data analysis with respect to such organismal or experimental features, the user may include to the objects any *metadata* of interest for each 96-well plate.

2.3 Data manipulation and export

Objects can be indexed for specific plates, time points of measurements or wells. The *stored metainformation* can be modified and queried with specific functions or infix operators. This allows for easy and flexible subsetting of objects based on any combination of metadata keys and values and on positive versus negative reactions. The *aggregated* or *discretized curve parameters* can be extracted into a matrix or data frame for any further analysis outside opm but within R. For storage in files or

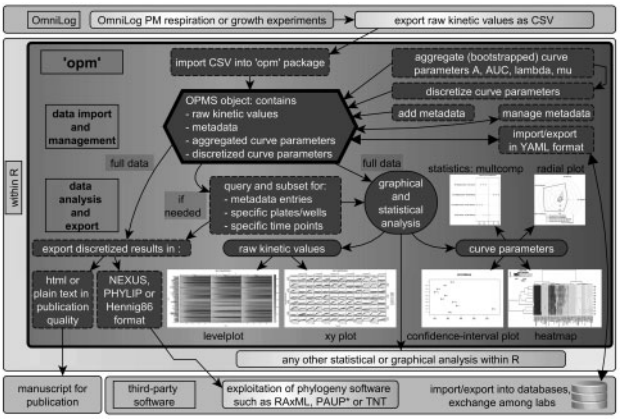

Fig. 1. General workflow of `opm` and its interplay with R and third-party software

databases and exchange between laboratories self-describing YAML files containing all stored information can be generated.

2.4 Graphical and statistical analysis

High-quality graphical visualizations of either the *raw kinetic measurements* or *curve parameters* are indispensable for quantitative analysis of PM data. Therefore, functions for level plots, *x*/*y* plots, confidence interval plots, heat maps and radial plots have been integrated. Plots can be easily annotated with the stored metainformation. Statistical comparison of multiple groups (Hothorn *et al.*, 2008) defined by stored metainformation is also straightforward. Data conversions allow for user-defined visualization strategies based on, e.g. `lattice` (Sarkar, 2008).

2.5 Data discretization and report generation

The continuous curve parameter data can be converted to discrete values in several ways for the export of character data to software for inferring character evolution or phylogenies (Goloboff *et al.*, 2008; Stamatakis *et al.*, 2005; Swofford, 2003). For a qualitative classification of the curves into negative and positive (optionally also weak/ambiguous) reactions relevant for assessing gene function and annotation (Bochner *et al.*, 2001) or for prokaryote taxonomy, several approaches are available. The results can be displayed as HTML tables or text suitable, e.g. for scientific journals (Fiebig *et al.*, 2013; Montero-Calasanz *et al.*, 2012).

2.6 Substrate information provided by opm

Almost all OmniLog® and MicroStation™ plates are supported regarding the mapping of wells to standardized substrate names. To further explore PM data regarding gene function and annotation, CAS, KEGG (http://www.genome.jp/kegg/), MeSH (http://www.ncbi.nlm.nih.gov/mesh) and Metacyc (http://metacyc.org/) IDs are supplied for each substrate as far as available from literature.

3 CONCLUSION

The unique strength of opm is the qualitative and quantitative analysis of raw kinetic OmniLog® PM data via direct visualization and the robust estimation of curve parameters and their confidence intervals. Thereby, the user retains full control on the directions of data analyses according to his research interests. Flexible ways to add and query metainformation enable the user to efficiently explore these data statistically, also across multiple groups. Export to matrices and data frames allows for the exploitation of all functionality available in the R environment, whereas standardized and easily readable output formats ease the interaction with external software.

ACKNOWLEDGEMENTS

The authors thank B. Bochner, A. Chouankam, J. Kirkish (BiOLOG Inc.), J. Meier-Kolthoff and S. Ehrentraut (DSMZ) for helpful advice.

Funding: Deutsche Forschungsgemeinschaft [grant SFB/TRR 51 to H.-P.K. and M.G., grant SI 1352/1-2 to J.S.]; and European Commission Microme project 222886 within the Framework 7 program to H.-P.K.

Conflict of Interest: none declared.

REFERENCES

BiOLOG Inc. (2009) *Converter, File Management Software, Parametric Software, Phenotype MicroArray, User Guide, Part 90333*. Biolog Inc., Hayward.

Bochner,B.R. (2009) Global phenotypic characterization of bacteria. *FEMS Microbiol. Rev.*, **33**, 191–205.

Bochner,B. *et al.* (2001) Phenotype microarrays for high throughput phenotypic testing and assay of gene function. *Genome Res.*, **11**, 1246–1255.

Chambers,J. (1998) *Programming with Data. Statistics and Computing*. Springer, New York.

Eilers,P.H.C. *et al.* (1996) Flexible smoothing with B-splines and penalties. *Stat. Sci.*, **11**, 89–121.

Fiebig,A. *et al.* (2013) Genome of the marine alphaproteobacterium *Hoeflea phototrophica* type strain (DFL-43^T). *Stand. Genomic Sci.*, **7**, 440–448.

Goloboff,P.A. *et al.* (2008) TNT, a free program for phylogenetic analysis. *Cladistics*, **24**, 774–786.

Hothorn,T. *et al.* (2008) Simultaneous inference in general parametric models. *Biometr. J.*, **50**, 346–363.

Kahm,M. *et al.* (2010) Grofit: fitting biological growth curves with R. *J. Stat. Soft.*, **33**, 1–21.

Montero-Calasanz,M.C. *et al.* (2012) *Geodermatophilus siccatus* sp. nov., isolated from arid sand of the Saharan desert in Chad. *A. van Leeuw. J. Microb.*, **103**, 449–456.

R Development Core Team. (2012) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

Sarkar,D. (2008) *Lattice: Multivariate Data Visualization with R*. Springer, New York.

Stamatakis,A. *et al.* (2005) RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics*, **21**, 456–463.

Swofford,D. (2003) *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4*. Sinauer Associates, Sunderland.

Vaas,L.A.I. *et al.* (2012) Visualization and curve-parameter estimation strategies for efficient exploration of phenotype microarray kinetics. *PLoS One*, **7**, e34846.

Wood,S.N. (2006) *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC Press, Boca Raton.