The Journal of Open Source Software

emba: R package for analysis and visualization of biomarkers in boolean model ensembles

John Zobolas[1, 2], **Martin Kuiper**[1], **and Åsmund Flobak**[2, 3]

1 Department of Biology, Norwegian University of Science and Technology (NTNU), Trondheim, Norway **2** Department of Clinical and Molecular Medicine, Norwegian University of Science and Technology (NTNU), Trondheim, Norway **3** The Cancer Clinic, St. Olav's Hospital, Trondheim, Norway

DOI: 10.21105/joss.02534

Software

- Review
- Repository
- Archive

Editor: Pending Editor

Submitted: 28 July 2020
Published: 28 July 2020

Introduction

Computational modeling of cellular systems has been one of the most powerful tools used to build interpretable knowledge of biological processes and help identify molecular mechanisms that drive diseases such as cancer (Aldridge, Burke, Lauffenburger, & Sorger, 2006). In particular, the use of logical modeling has proven to be a substantially useful approach, since it allows the easy construction, simulation and analysis of predictive models, capable of providing a qualitative and insightful view on the extremely complex landscape of biological systems (Abou-Jaoudé et al., 2016; Morris, Saez-Rodriguez, Sorger, & Lauffenburger, 2010; Wang, Saadatpour, & Albert, 2012). Such mechanistic models, with the systematic integration of prior knowledge and experimental data, have been extensively used to better understand what drives deregulation of signal transduction, the outcome of which is the manifestation of diseases (Traynard, Tobalina, Eduati, Calzone, & Saez-Rodriguez, 2017). Furthermore, their explanatory power has been used to provide insights into a drug's mode of action, investigate the mechanisms of resistance to drugs (Eduati et al., 2017) and suggest new therapeutic combination candidates, among others (Flobak et al., 2015).

One of the major challenges in systems medicine, has been the identification of scientifically validated, predictive biomarkers that correlate with patient response to given therapies. The analysis of biological predictive markers of pharmacologic response can not only further our understanding of the systemic processes involved in diseases but can also help to classify patients into groups with similar responses to specific therapeutic interventions, advancing personalized medicine (Senft, Leiserson, Ruppin, & Ronai, 2017). In addition, the identification of biomarkers in tumor cells (e.g. mutations) has enabled the discovery of drug targets which are utilized in combinatorial molecular-targeted therapies - a strategy which aims to treat specific patient subgroups and has shown larger overall survival rates and reduced side-effects than monotherapy (Al-Lazikani, Banerji, & Workman, 2012). Despite the huge advancements towards drug combination therapy, genetic heterogeneity, drug resistance and drug combination synergy mechanisms still pose fundamental challenges to clinicians, modelers and lab researchers.

To help bridge the model simulation results with the (clinical) laboratory observations, several optimization methods have been used, such as model calibration, parameter estimation and sensitivity analysis. These methods also allow us to determine which model parameters have the biggest influence in the overall behaviour of the system (Aldridge et al., 2006). For example, in (Fröhlich et al., 2018), a computational framework that allowed for the efficient parameterization and contextualization of a large-scale cancer signaling network, was used to predict combination treatment outcome from single drug data. This model was calibrated to fit and accurately describe specific cell-line experimental data, while enabling the identification

of biomarkers of drug sensitivity as well as molecular mechanisms that affect drug resistance. Furthermore, in (Dorier et al., 2016), a network optimization approach which topologically parameterized boolean models according to a genetic algorithm was used, in order to best match the experimentally observed behaviour. This method resulted in an ensemble of boolean models which can be used to simulate response under drug perturbations in order to assess the underlying mechanisms and to generate new testable hypotheses. Such an aggregation of best-fit models (wisdom of the crowds) has been shown to be quite robust and effective for model prediction performance (Marbach et al., 2012).

Statement of need

There is a plethora of software tools devoted to the qualitative modeling and analysis of biological networks. The Consortium for the development of Logical Models and Tools (CoLoMoTo) is a community effort which aims to standardize the representation of logical networks and provide a common repository of methods and tools to analyze these networks (Naldi et al., 2015). Furthermore, to facilitate the access to several software logical modeling tools and enable reproducible computational workflows, the CoLoMoTo Interactive Notebook was introduced as a unified computational framework (Aurélien Naldi, Hernandez, Levy, et al., 2018). The incorporated tools are accessed via a common programming interface (though originally implemented in different programming languages e.g. Java, Python, C++ and R) and offer a collection of features like accessing online model repositories (Helikar et al., 2012), model editing (Aurélien Naldi, Hernandez, Abou-Jaoudé, et al., 2018), dynamical analysis (finding attractors, stochastic simulations, reachability properties, model-checking techniques) (Klarner, Streck, Siebert, & Sahinalp, 2016; Müssel, Hopfensitz, & Kestler, 2010; Aurélien Naldi, 2018; Paulevé, 2017; Stoll et al., 2017) and model parameterization/optimization to fit perturbation signaling data (Gjerga et al., 2020; Terfve et al., 2012). Despite the diverse and multi-purpose logical modeling tools that exist, there is still a lack of data analysis-oriented software that assists with the discovery of predictive biomarkers in ensembles of parameterized boolean networks that have been subject to drug combination perturbations.

The `emba` R package aims to fill that gap and provide a first implementation of such a novel software. Initially, it was designed as a complementary software tool, to help the analysis of the parameterized boolean model ensembles which were produced by modules from the DrugLogics NTNU software pipeline (see respective documentation (Zobolas, 2020a)). Later, we generalized most of the functions in the package and modularized them to package-essential (that form the core of the `emba` package) and various general-purpose yet useful functions (that are now part of the dependency package `usefun` (Zobolas, 2020b)).

Summary

The main functionality of the `emba` R package is to find *performance* and *synergy* biomarkers. Performance biomarkers are nodes in the input boolean networks whose activity state and/or model parameterization affects the predictive performance of those models. The prediction performance can be assessed via the number of true positive predictions or the Matthews correlation coefficient score which is more robust to class imbalances (Chicco & Jurman, 2020). On the other hand, synergy biomarkers are nodes which provide hints for the mechanisms behind the complex process of synergy manifestation in drug combination datasets.

For more information, see our "Get started guide" and the reference manual in the package website (Zobolas, 2020c). Several analyses using the `emba` R package are available in a separate repository (Zobolas, 2020d). Future developments will include the implementation of a method for the identification of *topology* biomarkers, where we will be able to assess

which interactions in the network are important for the manifestation of synergies in specific cell-contexts.

Acknowledgements

This work was supported by ERACoSysMed grant *COLOSYS* (JZ, MK) and The NTNU Strategic Research Area *NTNU Health* (AF).

References

Abou-Jaoudé, W., Traynard, P., Monteiro, P. T., Saez-Rodriguez, J., Helikar, T., Thieffry, D., & Chaouiya, C. (2016). Logical Modeling and Dynamical Analysis of Cellular Networks. *Frontiers in genetics*, *7*, 94. doi:10.3389/fgene.2016.00094

Aldridge, B. B., Burke, J. M., Lauffenburger, D. A., & Sorger, P. K. (2006). Physicochemical modelling of cell signalling pathways. *Nature Cell Biology*, *8*(11), 1195–1203. doi:10.1038/ncb1497

Al-Lazikani, B., Banerji, U., & Workman, P. (2012). Combinatorial drug therapy for cancer in the post-genomic era. *Nature Biotechnology*, *30*(7), 679–692. doi:10.1038/nbt.2284

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1). doi:10.1186/s12864-019-6413-7

Dorier, J., Crespo, I., Niknejad, A., Liechti, R., Ebeling, M., & Xenarios, I. (2016). Boolean regulatory network reconstruction using literature based knowledge with a genetic algorithm optimization method. *BMC Bioinformatics*, *17*(1). doi:10.1186/s12859-016-1287-z

Eduati, F., Doldàn-Martelli, V., Klinger, B., Cokelaer, T., Sieber, A., Kogera, F., Dorel, M., et al. (2017). Drug Resistance Mechanisms in Colorectal Cancer Dissected with Cell Type-Specific Dynamic Logic Models. *Cancer research*, *77*(12), 3364–3375. doi:10.1158/0008-5472.CAN-17-0078

Flobak, Å., Baudot, A., Remy, E., Thommesen, L., Thieffry, D., Kuiper, M., & Lægreid, A. (2015). Discovery of Drug Synergies in Gastric Cancer Cells Predicted by Logical Modeling. (I. Xenarios, Ed.)*PLOS Computational Biology*, *11*(8), e1004426. doi:10.1371/journal.pcbi.1004426

Fröhlich, F., Kessler, T., Weindl, D., Shadrin, A., Schmiester, L., Hache, H., Muradyan, A., et al. (2018). Efficient Parameter Estimation Enables the Prediction of Drug Response Using a Mechanistic Pan-Cancer Pathway Model. *Cell Systems*, *7*(6), 567–579.e6. doi:10.1016/J.CELS.2018.10.013

Gjerga, E., Trairatphisan, P., Gabor, A., Koch, H., Chevalier, C., Ceccarelli, F., Dugourd, A., et al. (2020). Converting networks to predictive logic models from perturbation signalling data with CellNOpt. *Bioinformatics*. doi:10.1093/bioinformatics/btaa561

Helikar, T., Kowal, B., McClenathan, S., Bruckner, M., Rowley, T., Madrahimov, A., Wicks, B., et al. (2012). The Cell Collective: Toward an open and collaborative approach to systems biology. *BMC Systems Biology*, *6*(1), 96. doi:10.1186/1752-0509-6-96

Klarner, H., Streck, A., Siebert, H., & Sahinalp, C. (2016). PyBoolNet: a python package for the generation, analysis and visualization of boolean networks. *Bioinformatics*, *33*(5), btw682. doi:10.1093/bioinformatics/btw682

Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., Allison, K. R., et al. (2012). Wisdom of crowds for robust gene network inference. *Nature Methods*, *9*(8), 796–804. doi:10.1038/nmeth.2016

Morris, M. K., Saez-Rodriguez, J., Sorger, P. K., & Lauffenburger, D. A. (2010). Logic-based models for the analysis of cell signaling networks. *Biochemistry*, *49*(15), 3216–3224. doi:10.1021/bi902202q

Müssel, C., Hopfensitz, M., & Kestler, H. A. (2010). BoolNet—an R package for generation, reconstruction and analysis of Boolean networks. *Bioinformatics*, *26*(10), 1378–1380. doi:10.1093/bioinformatics/btq124

Naldi, A. (2018). BioLQM: A Java Toolkit for the Manipulation and Conversion of Logical Qualitative Models of Biological Networks. *Frontiers in Physiology*, *9*, 1605. doi:10.3389/fphys.2018.01605

Naldi, A., Hernandez, C., Abou-Jaoudé, W., Monteiro, P. T., Chaouiya, C., & Thieffry, D. (2018). Logical Modeling and Analysis of Cellular Regulatory Networks With GINsim 3.0. *Frontiers in Physiology*, *9*, 646. doi:10.3389/fphys.2018.00646

Naldi, A., Hernandez, C., Levy, N., Stoll, G., Monteiro, P. T., Chaouiya, C., Helikar, T., et al. (2018). The CoLoMoTo Interactive Notebook: Accessible and Reproducible Computational Analyses for Qualitative Biological Networks. *Frontiers in Physiology*, *9*, 680. doi:10.3389/fphys.2018.00680

Naldi, A., Monteiro, P. T., Mussel, C., Kestler, H. A., Thieffry, D., Xenarios, I., Saez-Rodriguez, J., et al. (2015). Cooperative development of logical modelling standards and tools with CoLoMoTo. *Bioinformatics*, *31*(7), 1154–1159. doi:10.1093/bioinformatics/btv013

Paulevé, L. (2017). Pint: A Static Analyzer for Transient Dynamics of Qualitative Networks with IPython Interface. In *CMSB 2017 - 15th conference on computational methods for systems biology*, Lecture notes in computer science (Vol. 10545, pp. 316–370). Springer. doi:10.1007/978-3-319-67471-1_20

Senft, D., Leiserson, M. D. M., Ruppin, E., & Ronai, Z. A. (2017). Precision Oncology: The Road Ahead. *Trends in Molecular Medicine*, *23*(10), 874–898. doi:10.1016/j.molmed.2017.08.003

Stoll, G., Caron, B., Viara, E., Dugourd, A., Zinovyev, A., Naldi, A., Kroemer, G., et al. (2017). MaBoSS 2.0: an environment for stochastic Boolean modeling. (J. Wren, Ed.)*Bioinformatics*, *33*(14), 2226–2228. doi:10.1093/bioinformatics/btx123

Terfve, C., Cokelaer, T., Henriques, D., MacNamara, A., Goncalves, E., Morris, M. K., Iersel, M. van, et al. (2012). CellNOptR: a flexible toolkit to train protein signaling networks to data using multiple logic formalisms. *BMC Systems Biology*, *6*(1), 133. doi:10.1186/1752-0509-6-133

Traynard, P., Tobalina, L., Eduati, F., Calzone, L., & Saez-Rodriguez, J. (2017). Logic Modeling in Quantitative Systems Pharmacology. *CPT: Pharmacometrics & Systems Pharmacology*, *6*(8), 499–511. doi:10.1002/psp4.12225

Wang, R.-S., Saadatpour, A., & Albert, R. (2012). Boolean modeling in systems biology: an overview of methodology and applications. *Physical Biology*, *9*(5), 55001. doi:10.1088/1478-3975/9/5/055001

Zobolas, J. (2020a). DrugLogics software documentation. GitHub Pages. Retrieved from https://druglogics.github.io/druglogics-doc/

Zobolas, J. (2020b). *Usefun: A collection of useful functions by john*. Retrieved from https://github.com/bblodfon/usefun

Zobolas, J. (2020c). Emba package website. GitHub Pages. Retrieved from https://druglogics.github.io/druglogics-doc/

Zobolas, J. (2020d). Ensemble boolean model analyses related to drug prediction performance. GitHub. Retrieved from https://github.com/bblodfon/gitsbe-model-analysis