Troppo - A Python Framework for the Reconstruction of Context-Specific Metabolic Models

Jorge Ferreira(✉), Vítor Vieira, Jorge Gomes, Sara Correia, and Miguel Rocha

Centre of Biological Engineering, Campus de Gualtar, University of Minho, Braga, Portugal
{jorge.ferreira,vvieira}@ceb.uminho.pt, jorge.gomes12@gmail.com, sarag.correia@gmail.com, mrocha@di.uminho.pt

Abstract. The surge in high-throughput technology availability for molecular biology has enabled the development of powerful predictive tools for use in many applications, including (but not limited to) the diagnosis and treatment of human diseases such as cancer. Genome-scale metabolic models have shown some promise in clearing a path towards precise and personalized medicine, although some challenges still persist. The integration of omics data and subsequent creation of context-specific models for specific cells/tissues still poses a significant hurdle, and most current tools for this purpose have been implemented using proprietary software. Here, we present a new software tool developed in Python, *troppo* - Tissue-specific RecOnstruction and Phenotype Prediction using Omics data, implementing a large variety of context-specific reconstruction algorithms. Our framework and workflow are modular, which facilitates the development of newer algorithms or omics data sources.

Keywords: Context-specific model reconstruction · Tissue specific models · Genome-scale metabolic models · Omics data integration

1 Introduction

Over the past years, the relationship between biology and informatics has proven to work to unveil the mysteries of the cell, from sequencing genomes to the reconstruction of the metabolism for a human cell. With ongoing advances on high-throughput technologies, the scientific community has been able to widen its scope of research, being able to analyze the cell as a complex layered system of interactions [5].

Genome-scale metabolic models (GSMMs) are the result of the integration of genome information into Constraint-Based Models (CBMs), connecting the

F. Fdez-Riverola et al. (Eds.): PACBB 2019, AISC 1005, pp. 146–153, 2020.
https://doi.org/10.1007/978-3-030-23873-5_18

genotype with cell metabolism. GSMMs are normally used to predict phenotypes which can be associated with consumption/ production rates of one or several metabolites [14].

The first reconstructions to emerge, through extensive manual curation, were Recon1 [8] and Edinburgh Human Metabolic Network (EHMN) [13]. To date, the most recent and complete reconstruction is the third version of Recon, Recon3D, integrating most of all the previous reconstructed models and including information related to the human microbiome and metabolism of dietary compounds [6]. However, these models represent the metabolism of a generalized human cell, eliciting the need for tissue specific model reconstruction algorithms. One of the main advantages of GSMMs is the ability to easily integrate omics data, i.e. to generate tissue specific metabolic models, with several algorithms already developed by the scientific community in the last couple of decades [5,16].

These allow the tuning of generic models to specific tissues or cell types, allowing to perform context-specific phenotype predictions and analyses. Although there are diverse algorithms proposed towards this aim, these are implemented in different platforms. Some are available in COBRA toolbox, which is dependent on a commercial platform (MatLab) [10].

In this work, we aim to provide a novel software platform, in open-source software, developed in Python, *Troppo - Tissue-specific RecOnstruction and Phenotype Prediction using Omics data*, implementing a wide range of context-specific algorithms, that can take as input a generic model and different types of omics data to provide context-specific models and/ or phenotype predictions (flux distributions).

2 Methods

The evolution of omics technologies led to the birth of the algorithms for reconstruction of tissue specific metabolic models since 2010, with the development of the MBA algorithm [12]. Although these algorithms have a similar goal, they can be divided into three different groups according to some of its characteristics: the definition of a core (Model-Building Algorithm (MBA)-like algorithms), testing Required Metabolic Functions (RMF) (Gene Inactivation Moderated by Metabolism and Expression (GIMME)-like algorithms) and creation of threshold based on the gene/protein expression (integrative Metabolic Analysis Tool (iMAT)-like family) [7,15].

2.1 GIMME-Like Family

This family contains the original GIMME algorithm [3], GIMMEp (allowing the incorporation of proteomics data) [4] and its extension GIM^{3}E [17]. The basis of this family is to perform a reconstruction by first optimizing the objective function (RMF) through a Linear Programming (LP) formulation. A second LP minimizing a penalty function (related to differences of the values obtained from the flux analysis to the respective transcript ones) is also performed, adding a constraint requiring a RMF value above a certain lower bound.

These algorithms differ in the way the penalty function is defined. GIMME penalizes flux values of reactions which are below a defined threshold when compared to the associated expression values. As for the GIM^{3}E, the penalties are calculated for each reaction associated with expression levels (instead of a set in GIMME), allowing all the reactions to have a penalty score. Also, since GIM^{3}E includes thermodynamics constraints, it is formulated as a Mixed Integer Linear Programming (MILP) problem.

2.2 iMAT-Like Family

In this family, we find the algorithms iMAT [23], INIT [1] and tINIT [2], which also reconstruct a model based on experimental data. The main difference regarding the previous family is that they do not need a RMF to work, so they try to match the maximum possible number of reaction states (active) to related data expression. Their formulation is a MILP and despite they share a similar strategy, iMAT tries to incorporate data in the constraints of the model, while INIT and tINIT try do so on the objective function.

iMAT is based on a previously defined threshold for the expression data which separates the reactions in high and low expression. Next, after defining a minimum flux value for each group, it tries to match the maximum number of reactions to each group, and specifically for the highly expressed group, so it has to overcome the minimum flux by solving a MILP. However, due to the fact that there are several possible flux distributions which have the same objective function value, iMAT uses an adapted version of Flux Variability Analysis [23].

The INIT algorithm maximizes the matches between reaction states (active or inactive) and data regarding expression of genes/proteins, returning flux values and a tissue-specific model. The method solves a MILP, where binary variables represent the presence of each reaction from the template model in the resulting model. In the definition of the objective function, positive weights are given to reactions with a higher evidence from the input, and negative to the ones which have low or no expression. If there is supportive information (usually metabolomics) that corroborate the presence of a certain metabolite, the necessary reactions may be included in the final model to produce [1].

The tINIT is an extension of the previous algorithm [2]. The improvement is based on the possibility to define a set of metabolic tasks in agreement with the context of the reconstruction. These may be the consumption or production of a metabolite or activation of the reactions of a particular pathway for the tissue.

2.3 MBA-Like Family

The last family is constituted by MBA [12], mCADRE [21] and FASTCORE [20]. Unlike the previously described algorithms, these only return the resulting context-specific model. Their inputs are sets with predefined categorized reactions as core and non-core sets. Normally, the core is defined by the reactions which have higher evidence to be considered active (high-throughput data or curated biochemical information). When the core is built, the methods try to

eliminate or include the non-core reactions, while ensuring that the model is consistent and has no blocked reactions when reconstructed.

For MBA, both cores are created according to the provided data. In an iterative way, all the non-core reactions are removed in a random order, while the model is tested for consistency. The iteration ends when all the reactions have been submitted for the removal test in the final model. Since the order by which reactions are tested for removal matters, there is the need to repeat this algorithm several times to obtain a set of models. The final one should be a model based on the ranking of the frequency of the reactions in the set, adding them to the high evidence core until a coherent model is found [12].

The mCADRE algorithm is quite similar to the MBA, but only requires the reconstruction of a single model. It is initialized by ranking the reactions on the original model using three distinct scores: confidence, expression and connectivity. With the help of a threshold value for the scores, a core of reactions and the order of removal of the non-core ones is established [21].

FASTCORE algorithm uses another strategy by solving two LPs. The first maximizes the number of reactions in the core, comparing the values of a reaction with a constant, while the other decreases the number of reactions that are absent in the core by minimizing the L_1-norm of the flux vector. Until the core is coherent (the whole set of core reactions is activated with the smallest number of non-core reactions), both problems are solved alternatively and in a repeated way. For reversible reactions, the algorithm analyses both directions [20].

CORDA is a relatively new algorithm and due to its nature for the predefinition of a core of reactions, it might be classified MBA-like algorithm. One of its main features is the fact of only needing a LP, providing a faster reconstruction in comparison to other algorithms. As a novel approach, the developers created the *dependency assessment* as a new way to identify the importance of desirable reactions (with higher evidence) in contrast to the one with less information [18].

An overview of all the algorithm families and how to choose the better algorithm for a certain situation is depicted in the Fig. 1.

3 Software

We have developed *troppo*, a modular framework implemented using the Python programming language and providing routines for the reconstruction of context-specific metabolic models (Fig. 1). *troppo* does not natively depend or provide any model reading or manipulation capabilities. This is intended, since external wrappers for cobrapy [9] and framed (https://github.com/cdanielmachado/framed) are available, allowing the user to load and manipulate previously validated models with these tools, and use them as inputs for the reconstruction algorithms provided in *troppo*. Additionally, the framework is built such that these operations are performed with short and simple commands.

Troppo depends on the cobamp library, an open-source tool implementing constraint-based pathway analysis methods, as the underlying framework

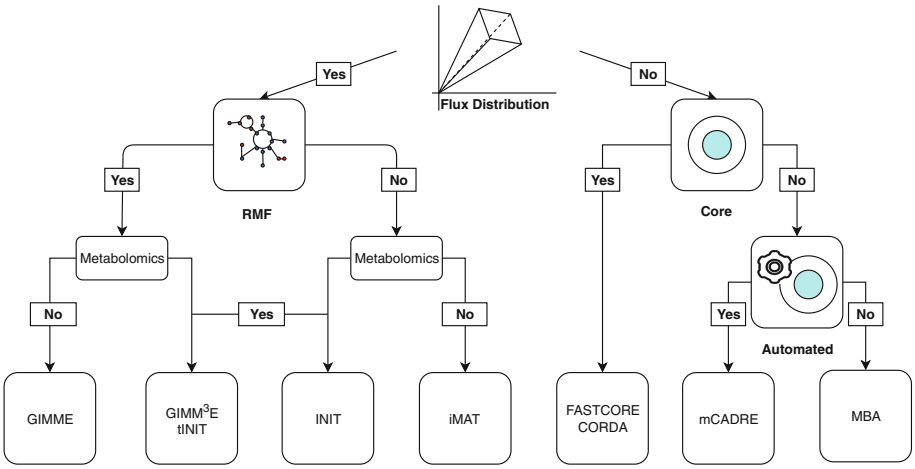

Fig. 1. Overview of the developed algorithms for context-specific model reconstruction and their properties, including the ability to obtain a flux distribution, as well as possible inputs such as metabolic functions (RMFs), metabolomics data or a core of curated reactions

providing methods to connect with external model loading frameworks and for building and solving LP/MILP problems. The latter is also dependent on *optlang* [11], allowing a wide selection of commercial and non-commercial solvers to be used.

The software is comprised of two main components, namely an omics data processing layer capable of handling transcriptomics and proteomics data, and a modular reconstruction layer implementing context-specific reconstruction methods. The general workflow implemented in this software is depicted on Fig. 2.

3.1 Omics Layer

Omics data is obtained using one of the provided readers for proteomics data from HPA [19] and as microarray experiments from any source. Using mappings from the HUGO Gene Nomenclature Committee [22], *troppo* is capable of mapping genes encoding enzymes with their respective reactions. This integration relies on gene-protein-reaction rules (GPRs) usually present on GSMMs, which can be used to modulate gene expression and adjust flux predictions accordingly. The preprocessed and mapped data is then passed as an algorithm property in the form of either weights for each reaction or sets of reactions, depending on the algorithm.

3.2 Reconstruction Layer

In the reconstruction layer, two key inputs are needed for any of the implemented algorithms. A metabolic model must be supplied **(2a)**, which at a bare minimum, implies a stoichiometric matrix encoding the relations between metabolites and

reactions, as well as thermodynamic constraints for the latter (lower and upper bounds). All algorithms also require a matching set properties, individual to each method, where specific execution options and omics inputs pre-processed using the omics layer described in the previous section **(1)** are included **(2b)**.

After defining these inputs, the selected algorithm will be run **(3)** and the expected output is a set of reactions to exclude from the generic model **(4a)**. Some methods are also capable of returning a flux distribution constrained by omics data **(4b)**, which can later be used for analysis or to manually extract the inactive reactions from it. Finally, these reactions are excluded from the model, alongside with their respective genes **(5)**, completing the reconstruction process.

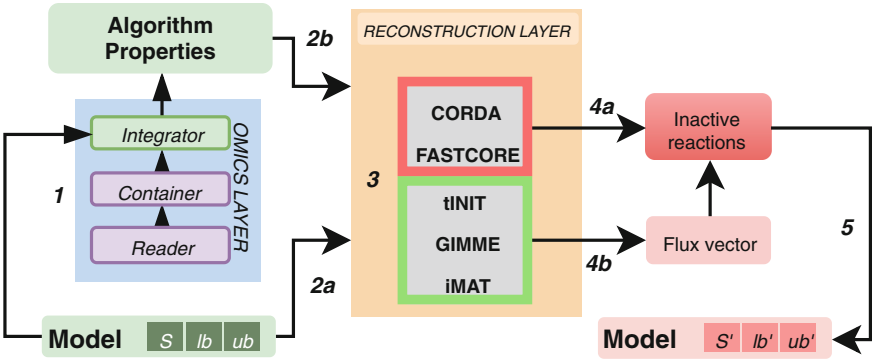

Fig. 2. Overview of the generic pipeline implemented within *troppo*. The process starts with a generic metabolic model - stoichiometric matrix (**S**), lower and upper bounds (**lb**, **ub**) - with pre-processed omics data associated with its identifiers **(1)**. This model as well as specific algorithm parameters, including those obtained using omics data, are the inputs for the reconstruction algorithms **(2)**, whose result is ultimately a set of inactive reactions **(3, 4)**. This final set is then removed from the model **(5)**

3.3 Availability

Troppo is available in the Python programming language, preferably from version 3.6 onwards. It is licensed under the GNU Public License (version 3.0), with source-code available on GitHub (https://github.com/BioSystemsUM/troppo). A LP/MILP solver is required to run the algorithms and our framework is currently compatible with CPLEX, Gurobi and GNU Linear Programming Kit (GLPK).

4 Results

The software featured in this work was used to reconstruct tissue-specific metabolic models of tumors arising from glial cells (glioma). To validate the implementations, we also used similar MATLAB routines based on the COBRA

Toolbox to run the algorithms which were implemented in *troppo* package using CPLEX as the undelying LP solver. These are then compared with solutions from *troppo*. These analyses are included in the GitHub repository as Jupyter notebooks which users can download and run (https://github.com/BioSystemsUM/troppo/tests/validation/).

5 Conclusion

In this work, a Python open-source library for context-specific model reconstruction was presented. The software currently integrates a large and growing variety of algorithms for integration of omics data within a generic framework that is also modular, and thus, facilitates the development of newer algorithms. Unlike other implementations, this one does not depend on proprietary software, enabling easy access for the whole community.

Acknowledgements. This study was supported by the Portuguese Foundation for Science and Technology (FCT) under the scope of the strategic funding of UID/BIO/04469/2019 unit and BioTecNorte operation (NORTE-01-0145-FEDER-000004) funded by the European Regional Development Fund under the scope of Norte2020 - Programa Operacional Regional do Norte. The authors also thank the PhD scholarships funded by national funds through Fundação para a Ciência e Tecnologia, with references: SFRH/BD/133248/2017 (J.F.), SFRH/BD/118657/2016 (V.V.).

References

1. Agren, R., Bordel, S., Mardinoglu, A., et al.: Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT. PLoS Comput. Biol. **8**(5), e1002518 (2012). https://doi.org/10.1371/journal.pcbi.1002518
2. Agren, R., Mardinoglu, A., Asplund, A., et al.: Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling. Mol. Syst. Biol. (2014). https://doi.org/10.1002/msb.145122
3. Becker, S.A., Palsson, B.O.: Context-specific metabolic networks are consistent with experiments. PLoS Comput. Biol. **4**(5), e1000082 (2008). https://doi.org/10.1371/journal.pcbi.1000082
4. Bordbar, A., Mo, M.L., Nakayasu, E.S., et al.: Model-driven multi-omic data analysis elucidates metabolic immunomodulators of macrophage activation. Mol. Syst. Biol. (2012). https://doi.org/10.1038/msb.2012.21
5. Bordbar, A., Monk, J.M., King, Z.A., Palsson, B.O.: Constraint-based models predict metabolic and associated cellular functions (2014). https://doi.org/10.1038/nrg3643
6. Brunk, E., Sahoo, S., Zielinski, D.C., et al.: Recon3D enables a three-dimensional view of gene variation in human metabolism. Nat. Biotechnol. **36**(3), 272–281 (2018). https://doi.org/10.1038/nbt.4072
7. Correia, S., Costa, B., Rocha, M.: Reconstruction of consensus tissue-specific metabolic models. bioRxiv (2018). https://doi.org/10.1101/327262

8. Duarte, N.C., Becker, S.A., Jamshidi, N., et al.: Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proc. Nat. Acad. Sci. USA **104**(6), 1777–82 (2007). https://doi.org/10.1073/pnas.0610772104
9. Ebrahim, A., Lerman, J.A., Palsson, B.O., Hyduke, D.R.: COBRApy: constraints-based reconstruction and analysis for Python. BMC Syst. Biol. (2013). https://doi.org/10.1186/1752-0509-7-74
10. Heirendt, L., Arreckx, S., Pfau, T., et al.: Creation and analysis of biochemical constraint-based models: the COBRA toolbox v3.0. Protocol Exchange (2017). https://doi.org/10.1038/protex.2011.234
11. Jensen, K., G.R. Cardoso, J., Sonnenschein, N.: Optlang: an algebraic modeling language for mathematical optimization. J. Open Sour. Softw. (2017). https://doi.org/10.21105/joss.00139
12. Jerby, L., Shlomi, T., Ruppin, E.: Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism. Mol. Syst. Biol. **6**, 401 (2010). https://doi.org/10.1038/msb.2010.56
13. Ma, H., Sorokin, A., Mazein, A., et al.: The Edinburgh human metabolic network reconstruction and its functional analysis. Mol. Syst. Biol. **3**(1), 135 (2007). https://doi.org/10.1038/msb4100177
14. Raškevičius, V., Mikalayeva, V., Antanavičiūtė, I., et al.: Genome scale metabolic models as tools for drug design and personalized medicine. PLOS ONE **13**(1), e0190636 (2018). https://doi.org/10.1371/journal.pone.0190636
15. Robaina Estévez, S., Nikoloski, Z.: Generalized framework for context-specific metabolic model extraction methods. Front. Plant Sci. **5**, 491 (2014). https://doi.org/10.3389/fpls.2014.00491
16. Ryu, J.Y., Kim, H.U., Lee, S.Y.: Reconstruction of genome-scale human metabolic models using omics data. Integr. Biol. (2015). https://doi.org/10.1039/C5IB00002E
17. Schmidt, B.J., Ebrahim, A., Metz, T.O., et al.: GIM^{3}E: condition-specific models of cellular metabolism developed from metabolomics and expression data. Bioinformatics **29**(22), 2900–8 (2013). https://doi.org/10.1093/bioinformatics/btt493. (Oxford, England)
18. Schultz, A., Qutub, A.A.: Reconstruction of tissue-specific metabolic networks using CORDA. PLoS Comput. Biol. (2016). https://doi.org/10.1371/journal.pcbi.1004808
19. Uhlen, M., Oksvold, P., Fagerberg, L., et al.: Towards a knowledge-based human protein Atlas. Nat. Biotechnol. (2010). https://doi.org/10.1038/nbt1210-1248
20. Vlassis, N., Pacheco, M.P., Sauter, T.: Fast reconstruction of compact context-specific metabolic network models. PLoS Comput. Biol. **10**(1), e1003424 (2014). https://doi.org/10.1371/journal.pcbi.1003424
21. Wang, Y., Eddy, J.A., Price, N.D.: Reconstruction of genome-scale metabolic models for 126 human tissues using mCADRE. BMC Syst. Biol. (2012). https://doi.org/10.1186/1752-0509-6-153
22. Yates, B., Bruford, E., Gray, K., et al.: Genenames.org: the HGNC and VGNC resources in 2019. Nucleic Acids Res. **47**(D1), D786–D792 (2018). https://doi.org/10.1093/nar/gky930
23. Zur, H., Ruppin, E., Shlomi, T.: iMAT: an integrative metabolic analysis tool. Bioinformatics (2010). https://doi.org/10.1093/bioinformatics/btq602