Development of text-mining solutions to facilitate lipid metabolism interpretation in Genome-Scale Metabolic Models

Adriano Silva[1], João Ribeiro[1], and Emanuel Cunha[1]

University of Minho

Abstract. Systems Biology is gaining importance in the unveil of cellular secrets. More precisely GSM models allow the contextualization of omic data and the progress of genomic engineering. Still, the lack of macromolecule structural defined representation, such as lipids, is glaring.
– Generic and defined duality
– Resolution

Keywords: Genome-Scale Metabolic Models · Lipids representation.

1 Introduction

1.1 Context and motivation

In the past two decades, Systems Biology has emerged as a discipline capable of integrating molecular biological knowledge into an understanding at a system level, from a complete, precise, and efficient perspective. Biological systems are complex and oftentimes present non-linear relationships between their components. Accordingly, there is the need to understand and contextualize them. Genome-scale metabolic models (GSM) are useful tools that integrate genomic, biochemical, and physiological knowledge for the better understanding of living organisms metabolism [1, 2]. These approaches can guide strain optimization and the production of a compound with industrial interest, such as lipidic biofuel produced by optimized yeasts and microalgae [3].

The reconstruction of GSM models is gaining importance, mainly impulsed by the advances and cost-effectiveness in technologies that led to high-throughput biological data. Over 6000 GSM models were reconstructed in total [4] since the reconstruction of the first GSM model in 1999 [5]. Nonetheless, the pace of reconstruction of GSM models cannot keep up with the advancements of high-throughput technologies and thus *omics* data. The lack of integration of new data into GSM models is a problem inherent to this growth discrepancy.

Besides the usefulness of these models, their reconstruction is limited due to the lack of biochemical and incorporated structural data. More precisely, complex macromolecules are often represented in their generic version not giving any biochemical and structural information [4]. Particularly in the case of lipids,

only a small number of reconstructed GSM models have structurally defined lipids with no or few relevant cross-references.

The integration of molecular information can be done by taking advantage of the *de facto* databases such as SWISS LIPIDS [6] and LIPID MAPS [7]. Molecular information is important to the reliably integrate and annotate models' information regarding the different structurally defined lipids. A tool capable of annotating and linking the different lipid species represented in GSM models with relevant databases could improve models' interpretability. Such improvement could leverage the yield optimisation of lipidic biofuels Sawangkeaw2013.

1.2 Objective

The main objective of this project is to integrate synonyms and abbreviations of lipids from SWISS LIPIDS and LIPID MAPS into a graph-based database. Then, those synonyms and abbreviations will be used to link lipids present in GSM models and their molecular structures.

2 State of art

2.1 Genome Scale Metabolic Models

GSM models are computational tools that conjugate biochemical and genomic data from an organism, with the capacity to perform *in silico* predictions of a given organism phenotype in specific environmental and genetic conditions [8, 9].

Thus, these models are key to the contextualization of high-throughput data and helpful in many other applications such as metabolic engineering, production of biochemicals and bio-materials, prediction of enzyme functions, or even in the discovery of drug targets [4, 10]. Therefore, it is important to integrate reliable biochemical data into the reconstruction of these models to ensure their accuracy and further interpretability [11, 12].

2.2 Lipid computational representation

Lipids are macromolecules grouped into different classes according with their structural composition. They are composed by two biochemical different components, the head usually composed by polar groups and the tail composed by apolar carbon linked chains. The differences in lipid structural polarity conferes amphipathic characteristic to this macromolecule [13]. This means that in an hidrophilic environment the polar part of the molecule is atracted and the apolar one repelled. This allows the generation of micelles, which is important for their biological roles such as being the principal cell membrane components, energy storage, and signaling molecules.

As represented in Fig.1, lipid structures can be split into two different parts: the backbone and the side chains. The former is not variable for the whole class,

remaining the same to the whole structurally defined lipids in the same class. As for the side chains, their structure can vary in the same class in the number of double bonds, stereochemistry and length.

According to Fahy and collaborators [14], lipids can be divided into eight main classes:Fatty acyls,Glycerolipids,Glycerophospholipids,Sphingolipids,Sterol Lipids,Prenol Lipids,Saccharolipids,Polyketides. Due to the myriad of side chain combinations, it is not possible to estimate how many distinct lipids can occur, both naturally and synthetically [15].
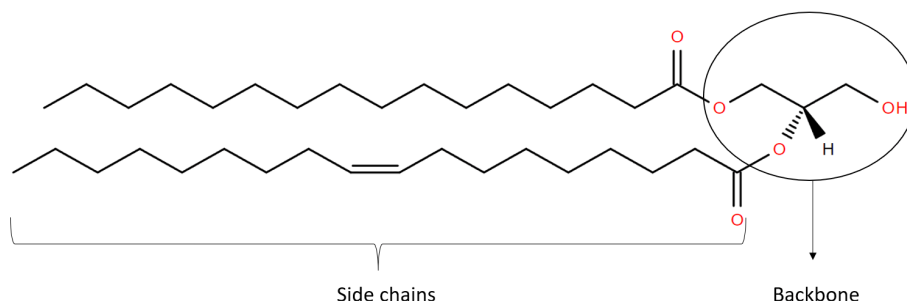

Fig. 1. Representation of 1-hexadecanoyl-2-(9Z-octadecenoyl)-sn-glycerol structure extracted from LIPID MAPS. The backbone is the hydrophilic part of the lipid while the side chains confer hydrophobicity.

The growing importance of this macromolecule in health research and industrial applications bring the need to characterize their metabolic network and roles in cells. Due to the immense amount of reactions, complex lipid biosynthetic pathways, and their inherent combinatorial complexity, it is almost impossible to study them by means of classical molecular biology [16].

Computational approaches and, particularly, GSM models are helping to disentangle these issues [16], however, lipid representation in such models is not as trivial as desirable. Despite the existence of lipid databases with defined structures, GSM models still fail to represent lipids with their structure completely defined [17]. Nevertheless, a growing number of models with structurally defined lipids start to appear (see http://bigg.ucsd.edu/models), however, they still lack of cross-references for lipid-specific databases, such as LIPID MAPS and SwissLipids. Such fact creates a gap between GSM models and *de facto* lipid databases, hindering their integration into other databases and interpretability.

2.3 Generic Representation in GSM models

The metabolites and reactions present in a GSM model highly depend on their source. As most databases (e.g., KEGG and MetaCyc) do not represent lipids as structurally defined, the absence of completely defined structures is propagated

to those models see Fig.2. This representation neglects the fact that side chains are important components in the lipid metabolic network [16–18].

Fig. 2. Generic representation highlighted.

Accordingly, biosynthetic pathways are represented in generic terms, with generic lipids as reactants and products. Based on this, we cannot access the exact lipids used in an hypothetic biosynthetic network. Besides that, an abstract representation is linked to the loss of specificity of individual reactions. The use of a generic representation will impose the utilization of many lipids in the reactions, and does not allow the transformation of a generic compound into a well defined one [17].

Interestingly, we can still see the presence of cross-references to databases with the structure of these macromolecules. However, the same structure represent a multitude of structurally defined lipids of the class under representation. In these models, the name of the lipids is defined as the name of the class under representation, which, in most cases, only includes the name of the backbone.

2.4 Structurally defined representation in GSM models

Contrary to the generic representation, GSM models do not usually include structurally defined lipids [16]. In those that include, the lipid name includes both the side chains and the backbone. The side chains' name is defined by the number of carbons followed by the number and the location of double bonds as well as their stereochemistry fig3.

Fig. 3. Defined representation highlighted.

These approaches allow the generation of individual reactions with structurally defined lipids, in contrast with generic representations [17]. GSM models with defined lipids are more reliable, allowing the improvement of the flexibility,

accuracy, and level of detail of these models. On the other hand, the inclusion of structurally defined versions of lipids can significantly increase the number of reactions in the model, which can be a drawback for some users. Besides that, we can witness a lack of cross-references to *de facto* lipid databases, which is not ideal for lipid structural confirmation, integration and interpretability.

2.5   Lack in lipid annotation in GSM models

GSM models with lipidic generic representation have few annotations to metabolites, and contrary defined ones have almost no annotation to the structure, only to the metabolites. All things considered, the use of lipidic generic representation is not appropriate in models aiming lipid production optimization. But defined representation is lacking in structural annotation as well.
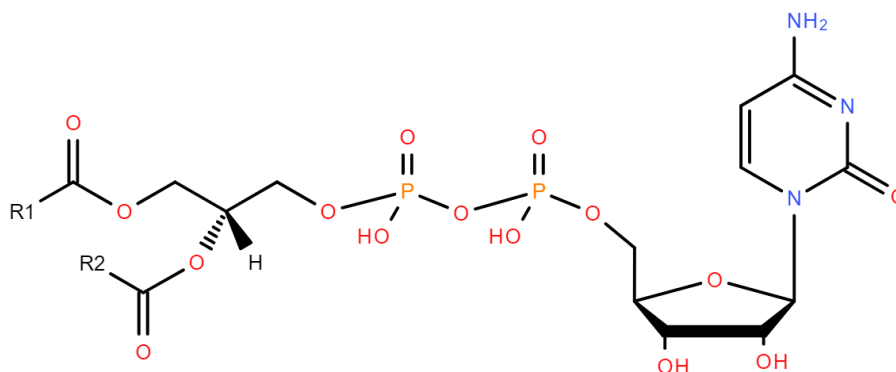

Fig. 4. Comparation between the representations aborded above.

A possible aproche to fill this gap would be the integration of structural information in models. It would be achieved by the integration of synonyms and abbreviations, extracted from *de facto* databases, as a link to the structural information.

ETL is a integration data tool, that allows the gathering, processing and integration of data, see Fig.5. This is extremely usefull in data cleaning and organization providing the bases to data analytics and machine learning aproaches. Based on this, ETL pipelines could make more easier the integration of structural information in our database.

Anotação de modelos: - match direto - decomposição do nome em dois (backbone e side chain) - queries à base de dados

Fig. 5. Escheme of the process done with ETL pipeline. First the ETL tool extracte data from multiple databases then processes all that data and finally integrates the data in our database.

References

1. Y. Zou and M. D. Laubichler, "From systems to biology: A computational analysis of the research articles on systems biology from 1992 to 2013," *PLOS ONE*, vol. 13, p. e0200929, 7 2018.
2. I. Tavassoly, J. Goldfarb, and R. Iyengar, "Systems biology primer: The basic methods and approaches," *Essays in Biochemistry*, vol. 62, pp. 487–500, 10 2018.
3. R. Sawangkeaw and S. Ngamprasertsith, "A review of lipid-based biomasses as feedstocks for biofuels production," *Renewable and Sustainable Energy Reviews*, vol. 25, pp. 97–108, 9 2013.
4. C. Gu, G. B. Kim, W. J. Kim, H. U. Kim, and S. Y. Lee, "Current status and applications of genome-scale metabolic models," *Genome Biology 2019 20:1*, vol. 20, pp. 1–18, 6 2019.
5. J. S. Edwards and B. O. Palsson, "Systems properties of the haemophilus influenzae rd metabolic genotype," *The Journal of biological chemistry*, vol. 274, pp. 17410–17416, 6 1999.
6. L. Aimo, R. Liechti, N. Hyka-Nouspikel, A. Niknejad, A. Gleizes, L. Götz, D. Kuznetsov, F. P. David, F. G. V. D. Goot, H. Riezman, L. Bougueleret, I. Xenarios, and A. Bridge, "The swisslipids knowledgebase for lipid biology," *Bioinformatics*, vol. 31, pp. 2860–2866, 9 2015.
7. M. Sud, E. Fahy, D. Cotter, A. Brown, E. A. Dennis, C. K. Glass, A. H. Merrill, R. C. Murphy, C. R. Raetz, D. W. Russell, and S. Subramaniam, "Lmsd: Lipid maps structure database," *Nucleic Acids Research*, vol. 35, pp. D527–D532, 1 2007.
8. I. Rocha, J. Förster, and J. Nielsen, "Design and application of genome-scale reconstructed metabolic models," *Methods in Molecular Biology*, vol. 416, pp. 409–431, 12 2007.
9. J. Zhou, P. Liu, J. Xia, and Y. Zhuang, "Advances in the development of constraint-based genome-scale metabolic network models," *Shengwu Gongcheng Xuebao/Chinese Journal of Biotechnology*, vol. 37, pp. 1526–1540, 5 2021.

10. W. J. Kim, H. U. Kim, and S. Y. Lee, "Current state and applications of microbial genome-scale metabolic models," *Current Opinion in Systems Biology*, vol. 2, pp. 10–18, 4 2017.

11. B. Moseley, A. Passi, J. D. Tibocha-Bonilla, M. Kumar, D. Tec-Campos, K. Zengler, and C. Zuniga, "Genome-scale metabolic modeling enables in-depth understanding of big data," *Metabolites 2022*, vol. 12, p. 14, 2021.

12. A. Passi, J. D. Tibocha-Bonilla, M. Kumar, D. Tec-Campos, K. Zengler, and C. Zuniga, "Genome-scale metabolic modeling enables in-depth understanding of big data," *Metabolites*, vol. 12, 1 2021.

13. E. Fahy, D. Cotter, M. Sud, and S. Subramaniam, "Lipid classification, structures and tools," *Biochimica et biophysica acta*, vol. 1811, p. 637, 11 2011.

14. E. Fahy, S. Subramaniam, R. C. Murphy, M. Nishijima, C. R. Raetz, T. Shimizu, F. Spener, G. V. Meer, M. J. Wakelam, and E. A. Dennis, "Update of the lipid maps comprehensive classification system for lipids," *Journal of Lipid Research*, vol. 50, p. S9, 4 2009.

15. D. Gyamfi, E. O. Awuah, and S. Owusu, "Classes, nomenclature, and functions of lipids and lipid-related molecules and the dietary lipids," *The Molecular Nutrition of Fats*, pp. 3–16, 1 2018.

16. V. Schützhold, J. Hahn, K. Tummler, and E. Klipp, "Computational modeling of lipid metabolism in yeast," *Frontiers in Molecular Biosciences*, vol. 3, p. 57, 9 2016.

17. H. W. Aung, S. A. Henry, and L. P. Walker, "Revising the representation of fatty acid, glycerolipid, and glycerophospholipid metabolism in the consensus model of yeast metabolism," *Industrial Biotechnology*, vol. 9, p. 215, 8 2013.

18. B. J. Sánchez, F. Li, E. J. Kerkhoven, and J. Nielsen, "Slimer: Probing flexibility of lipid metabolism in yeast with an improved constraint-based modeling framework," *BMC Systems Biology*, vol. 13, pp. 1–9, 1 2019.