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

Adriano Silva[1,2], Emanuel Cunha[1,2], and João Capela[1,2]

[1] Centro de Engenharia Biológica, Universidade do Minho, 4710-057 Braga, Portugal
[2] LABBELS – Laboratório Associado, Braga, Guimarães, Portugal

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. Interestingly structurally defined lipidic representation in models don't have cross references that corroborate the structure. In this project we aim to fill this gap using lipidic synonyms and abbreviation to access the lipid structure and further annotation.
– Resolution

Keywords: Genome-Scale Metabolic Models · Lipids representation · Lipids synonyms.

1 Introduction

1.1 Context and motivation

In the past two decades, Systems Biology has emerged as a discipline capable of integrating molecular knowledge into an understanding at a system level. Biological systems are complex and oftentimes present non-linear relationships between their components. Accordingly, it is often resorted to mathematical models to understand and contextualize both each component separately and the system as a whole. Genome-scale metabolic (GSM) models are useful tools that integrate genomic, biochemical, and physiological knowledge to better understand living organisms' metabolic behaviour [1, 2]. These approaches can guide strain optimization and the production of a compound with industrial interest, such as lipidic biofuels produced by optimized yeasts and microalgae [3].

The reconstruction of GSM models is gaining importance, mainly impulsed by the advances and cost-effectiveness of high-throughput technologies. Over 6000 GSM models were reconstructed [4] since the first model was published in 1999 [5]. Nonetheless, the pace of reconstruction cannot keep up with the advancements of high-throughput technologies and, consequently, the generation of *omics* data [6]. 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 to the biochemical data existent in the available databases. More precisely, complex macromolecules such as lipids and carbohydrates are often represented in their generic version, not providing the whole 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 [7].

Molecular structures are important to reliably integrate and annotate models' information regarding the different structurally defined lipid species. The integration of lipid molecular information can be performed by taking advantage of the *de facto* lipid databases such as SWISS LIPIDS [8] and LIPID MAPS [9]. A tool capable of annotating and linking the different lipid species represented in GSM models with those databases could improve models' interpretability and accuracy. Consequently, such improvement could leverage the yield optimisation of lipidic biofuels [3].

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 GSM models' lipids with their molecular structures.

2 State of art

2.1 Genome Scale Metabolic Models

GSM models are computational tools that conjugate biochemical and genomic data of an organism, with the capacity to perform *in silico* predictions of its phenotype under specific environmental and genetic conditions [10, 11].

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, 12]. Therefore, it is important to integrate reliable biochemical data into the reconstruction of these models to ensure their accuracy and further unequivocal interpretation [13, 14].

2.2 Lipid computational representation

Lipids are macromolecules grouped into different classes according to their structural composition. They are composed of two biochemical components, the *backbone*, usually composed of polar groups, and the *sidechains* composed of apolar carbon linked chains. The differences in lipid structural polarity confer amphipathic characteristics to this macromolecule [15]. This means that in a hydrophilic environment the polar part of the molecule is attracted and the apolar

one repelled. This allows the generation of micelles, which are important for lipids' biological roles such as energy storage, signalling molecules, and being the main cell membrane components [16].

As represented in Fig.1, lipid structures can be split into two different parts: the backbone and the sidechains. The former is not variable for the whole class, remaining the same to the whole structurally defined lipids of the same class. As for the sidechains, their structure can vary in the number of double bonds, stereochemistry and length.

According to Fahy and collaborators [17], lipids can be divided into eight main classes: Fatty acyls, Glycerolipids, Glycerophospholipids, Sphingolipids, Sterol Lipids, Prenol Lipids, Saccharolipids, and Polyketides. Due to the myriad of sidechains combinations, it is not possible to estimate how many distinct lipids can occur in nature [18, 7].
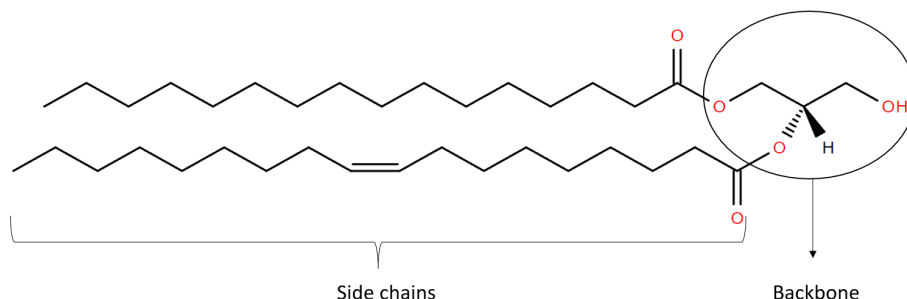
Fig. 1. Representation of 1-hexadecanoyl-2-(9Z-octadecenoyl)-sn-glycerol structure extracted from LIPID MAPS. The backbone is the hydrophilic part of the lipid while the sidechains confer hydrophobicity.

The growing importance of this macromolecule in health research and industrial applications brings 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 [19].

Computational approaches and, particularly, GSM models are helping to disentangle these issues [19], however, lipid representation in such models is still 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 [20]. Nevertheless, a growing number of models with structurally defined lipids start to appear (see http://bigg.ucsd.edu/models), however, they still lack 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 interpretation and integration into other databases.

2.3 Generic Representation in GSM models

The metabolites and reactions present in GSM models 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, as can be seen in Fig.2. This representation neglects the fact that sidechains are important components in the lipid metabolic network [19–21].

Fig. 2. Lipidic generic representation highlighted in *Tgondii* model [22].

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 [20, 7].

Interestingly, we can still see the presence of cross-references to databases with the structure of these macromolecules. However, the same structure represents a multitude of structurally defined lipids of the class being represented. In these models, the name of the lipids is defined as the name of the class being represented, which, in most cases, only includes the name of the backbone, as can be seen in Fig.2.

2.4 Structurally defined representation in GSM models

In the GSM models that comprise structurally defined lipids, the lipid name includes both the sidechains and the backbone. The sidechains' name is defined by the number of carbons followed by the number and the location of double bonds as well as their stereochemistry (Fig. 3).
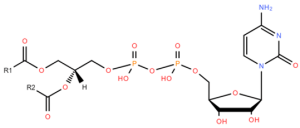

Fig. 3. Lipidic defined representation highlighted in *iLB1027-lipid* model [23].

These approaches allow the generation of individual reactions with structurally defined lipids, in contrast with generic representations. GSM models

with defined lipids are could be more accurate, 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 [20, 7]. Besides that, a lack of cross-references to *de facto* lipid databases can be witnessed, which is not ideal for lipid structural confirmation, integration and interpretability.

2.5   Lack of lipid annotations in GSM models

Oppositely to GSM with lipidic generic representation, defined ones do not have an annotation to structure only to metabolites. Overall, the use of lipid generic representations might not be appropriate in models aiming at lipid production optimization. However, defined representations usually lack in structural annotations, hindering their interpretation and curation.

Fig. 4. Comparison between generic and structurally defined representations. S generic version of the CDP-glycerol is represented on the left, extracted from the *Toxoplasma gondii* model [22]. A structurally defined representation of 1,2-Diacyl-sn-glycerol is represented, extracted from the *Phaeodactylum tricornutum* model [23]. The structure from the generic lipid was extrated from LIPID MAPS.

A possible approach to fill this gap would be the integration of structural information in models. Such a solution would be achieved by the integration of synonyms and abbreviations, extracted from *de facto* databases, as a link to the structural information. Those synonyms and abbreviations can be used to identify structurally defined lipids in the GSM models, as most of them provide a name or synonym, though not providing any cross-reference to other databases.

2.6   Computational tools

Neo4j The database used for this project is *BOIMMG*, a database formulated in graphs using the Neo4j program. Graphs are data structures composed of nodes and edges. Such an arrangement can emphasize the relationships between entities giving an edge over relational databases [24]. This formalism is often helpful for capturing biological entities and the relationships between their components, namely lipids and their structural components (e.g., backbones and side chains).

Neo4j, a platform for graph database management, provides a Application Programming Interface (API) that allows developers to connect and work on the database using Python programming language.

3   Methodology

The first step in this project was to access LIPID MAPS and Swiss Lipids online databases and extract the corresponding dataset. Afterwards, it was necessary to transform the raw data into a usable and trusted resource and upload it to our database to achieve the proposed objective.

An Extract-Transform-Load (ETL) pipeline was developed to facilitate accessing, treating, and loading lipid data from Lipid Maps or Swiss Lipids. ETL is an integration data tool that gathers, processes, and integrates data. Fig.5 showcases a representation of an ETL pipeline. Such a method is particularly useful in cleaning and organising raw data into a reliable and controlled version.

Fig. 5. Representation scheme of an ETL pipeline. First the ETL tool extracts (Extract) data from multiple databases then processes all that data (Transform) and finally integrates the data into the database (Load).

3.1 ETL Pipeline Execution

Extract Phase Firstly the online database from the *de facto* databases was accessed using the package *requests* that accessed and downloaded a zip file with the data from Swiss Lipids in a comma-separated values (CSV) format and Lipid Maps in a Structure-Data File (SDF) format. Then, the file is used to create a *pandas DataFrame.*

The process was done separately for Swiss Lipids and Lipid Maps, using Python Classes to create a data frame for each. In the data frames, we observe a large amount of information that can be converted to a desirable data structure.

Transform Phase The original data frame contains numerous columns, each having information about the lipids represented in each row. However, only three are necessary for the scope of this project. They are the columns for synonyms, abbreviations, and the column with the identifier (ID) needed to match the lipids in the database. Another step is necessary to treat the cases where a lipid has more than one synonym or abbreviation.

The Python class developed for this phase creates a new data frame with only two columns, one for lipid ID and another for synonyms and abbreviations. Then, successive iterations over the rows of the original dataset are performed, adding the lipid ID into the first column and its synonym and abbreviations into the second. A new line is created with the same ID for each synonym and abbreviation in cases where one or more synonyms and abbreviations are present. Since the data frames are relatively large, multiprocessing was used to help reduce the final time by nearly ten times for this phase.

Load Phase Having obtained the transformed data, the information was loaded into the database. For this purpose, it was necessary to connect to the database and formulate queries that create the nodes corresponding to synonyms and abbreviations as well as to the relations between these and the database lipids.

For instance, one synonym node could be linked to many lipids, especially synonyms of the backbone that are equal for many lipids. The lipid nodes, on the other hand, are restricted to a certain number of synonyms that depends on their composition (backbone plus the number of side chains).

In the implementation first the synonym node is created and only then the relations with the corresponding lipid nodes are formulated. As in the transformation phase, multiprocessing was here used to speed up loading into the database.

3.2 Model testing

The final step was the testing in a model case with structurally defined lipids and a lack of cross-references to databases; for this, the choice fell on the *iLB1027-lipid* model [23].

Regular expressions were used to search for characteristic patterns of lipid names so that they could be separated from the remaining metabolites. Next, the same process was used to split the lipid names into backbone and side chains; these components were used to get the corresponding database IDs. As expected, not all synonyms are represented in the database, so it is necessary to limit the selection of those that get the ID for the backbone and at least one side chain.

The last step consists in the gathering of information for statistic purpose. Primarily all lipid compounds containing a specific backbone/side-chain combination are searched in the database. They represent the connection with a *de facto* database essential to annotate the model. At the same time the occurrences in the model for each backbone (lipid class) are counted. Similarly for each lipid in the model is created an entry in a dictionary where hits are stored in the form of database id. In this case lipids with more than one hit may appear in the database.

Using the information mentioned above, means and standard deviations are calculated for the number of hits for each model lipid in the database. A comparison is also made between the lipids annotated in the model and those that could be annotated following this process.

4 Results and discussion

This project had as main objectives the integration of lipid synonyms and abbreviations into a graph database. It also sought to corubrate the work by testing in a model case where these would serve as connectors to the molecular structures of the lipids in the model. These goals were successfully achieved.

5 Conclusion

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. T. Y. Kim, S. B. Sohn, Y. B. Kim, W. J. Kim, and S. Y. Lee, "Recent advances in reconstruction and applications of genome-scale metabolic models," *Current Opinion in Biotechnology*, vol. 23, pp. 617–623, 8 2012.

7. "Universidade do minho: Revising lipid chemical structures in genome-wide metabolic models with boimmg."

8. 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.

9. 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.

10. 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.

11. 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.

12. 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.

13. 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.

14. 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.

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

16. P. R. Cullis, M. J. Hope, and C. P. Tilcock, "Lipid polymorphism and the roles of lipids in membranes," *Chemistry and Physics of Lipids*, vol. 40, pp. 127–144, 6 1986.

17. 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.

18. 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.

19. 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.

20. 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.

21. 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.

22. S. Tymoshenko, R. D. Oppenheim, R. Agren, J. Nielsen, D. Soldati-Favre, and V. Hatzimanikatis, "Metabolic needs and capabilities of toxoplasma gondii through combined computational and experimental analysis," *PLoS computational biology*, vol. 11, 5 2015.

23. J. Levering, J. Broddrick, C. L. Dupont, G. Peers, K. Beeri, J. Mayers, A. A. Gallina, A. E. Allen, B. O. Palsson, and K. Zengler, "Genome-scale model reveals metabolic basis of biomass partitioning in a model diatom," *PloS one*, vol. 11, 5 2016.

24. J. J. Miller, "Graph database applications and concepts with neo4j,"