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.

Keywords: Genome-Scale Metabolic Models · Lipids.

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 represent a huge amount of data, with the need to be treated and contextualized where this discipline comes to the aid. Whether in the construction of stochiometric models or the reconstruction of genome scale metabolic models (GSM), with means to understand the genomic, biochemical, and physiological knowledge gathered [1, 2]. These approaches can guide to strain optimization and the production of a compound with industrial interest, such as lipidic biofuel produced by optimized yeasts and microalgae [3].

Impulsed by the advances and cost-effectiveness in technologies that led to high-throughput biological data (Big Data), Systems Biology, and more precisely the reconstruction of GSM models is gaining importance. The reconstruction of GSM models is taking advantage of the high-quality data generated to create better simulations and predictions. In total, since the reconstruction of the first GSM model in 1999 [4], 6239 GSM models were reconstructed until 2019 [5]. Nonetheless, the pace of reconstruction of GSM models can't keep up with the growth of Big Data. The lack of integration of new data in 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 structural data incorporated. Complex macromolecules are often represented in their generic version not giving any biochemical and structural information [5]. Particularly in the case of lipids, only a small

chunk of GSM models reconstructed have structurally defined lipids. These models neglect the fact that each class is constituted of a countless number of combinations between the different components of the lipid. Thus the GSM models in these conditions are not able to capture the integrity of the lipid biosynthesis network. Therefore it is important the integration of such information into lipidic models, for better interpretability, handling, and predictions.

Integration of the structural information can be done by taking advantage of the *de facto* tools such as SWISS LIPIDS [6] and LIPID MAPS [7]. As mentioned above this is important to the reliability of the model allowing credible predictions and flexibility in the management of the model. This can turn into a major advancement in lipid models with better application in industry, such as in the case of lipidic biofuels [3].

1.2 Objective

The main objective of this project is to integrate structural data, from *de facto* tools SWISS LIPIDS and LIPID MAPS, into a graph-based database BOIMMG. For that, it will be done through the integration of the synonyms and abbreviations into a new label using ETL pipelines.

2 State of art

2.1 Genome Scale Metabolic Models

The use of computational tools brings to the science new tools to face the challenges in the scientific scope. Among them are GSM models, a computational tool that conjugate biochemical and genomic data from an organism, with the capacity to do *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[5, 10]. It is therefore important to integrate fresh and reliable biochemical data in the reconstruction of these models to ensure their accuracy and further actualization [11, 12].

2.2 Lipid computational representation

Lipids are a unique macromolecule grouped into different classes accordingly to their structural composition. They are composed of two distinct components, distinction in their composition confers to lipids the characteristic of amphipathic molecules. This characteristic allows the generation of micelles in a hydrophilic environment, which is important for their biological roles such as being the principal cell membrane components, energy storage, and signaling molecules.

Structurally it is impossible to say how many distinct lipids are due to the multitude of arrangements that can make a different structure [13]. The definition of the lipid class is well-established, existing already eight main classes are defined [14]. These can be split into two different parts the backbone, and the side chain represented in Fig. 1. The first one gives the lipid the name of the class, remaining the same to the whole lipids in the same class. In the case of side chains, their structure can vary in the same class, giving rise to new subclasses.

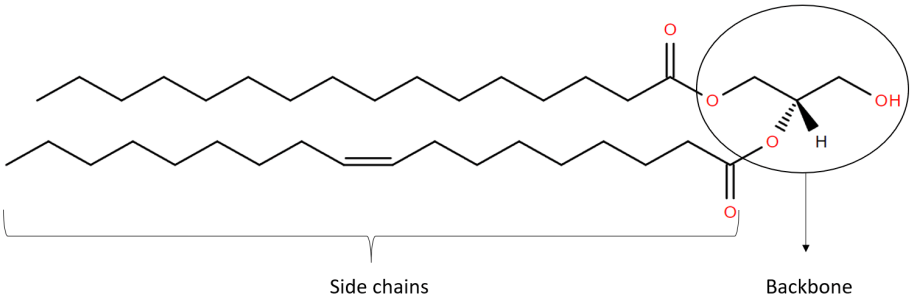

Fig. 1. Representation of 1-hexadecanoyl-2-(9Z-octadecenoyl)-sn-glycerol structure extracted from LIPID MAPS.The backbone is the hydrophilic part of the lipid while the side chains confer hydrophobicity.

The grown importance of this macromolecule in health research and industrial applications brings more than ever the need to characterize their metabolic network and roles in cells. Due to the immense amount of reactions, and complex biosynthetic pathways in that lipids play a role, it is almost impossible to study them in classical means [15].

Computational approaches can help disentangle these complications however, lipid representation is not always well implemented. Despite the existence of lipid databases with defined structures, computational representation is often done generically [16]. This means that the representation in computational approaches is only given by the backbone name, remaining the side chains as simple R groups. Given the importance of the structure in metabolic interactions, this hides information not allowing to achieve the full potential of these approaches.

2.3 Generic Representation in GSM models

In GSM models the representation of lipids is often generic (figura com a representação generica de um modelo?). This representation neglects the fact that side chains are important components in the lipid metabolic network. Based on this representation, we cannot access the specific lipid used in a hypothetic metabolic network due to the number of lipids with the same backbone name. Besides that, is almost impossible to say how accurate are the predictions to the successor lipids, or even to the one that ours came from [16].

The capacity to interchange side chains and form new lipids, with probable new metabolic interactions, turns this type of representation not helpful in acknowledging the lipidic metabolic network. Interestingly, we can still see the presence of cross-references to databases with the structure of these macromolecules. However, the same structure serves many lipids due to its generic input.

2.4 Structurally defined representation in GSM models

Contrary to the generic representation seeing defined ones in GSM models is not usual. Only a few chunks of the reconstructed GSM models have lipids structurally defined. Here, after the backbone name, is defined the constitution of the side chains is a second part of the lipid structure. This constitution is defined using the number of carbons in the chain followed by the number of double bonds and their conformation for each side chain (figura a demonstrar esta representação??). Starting from the idea that GSM models with defined lipids are more reliable, this is a better way to represent their structure in these models. Besides that, we can witness a lack of references, which is not ideal for lipid structural confirmation.

2.5 Lack in lipid annotation in GSM models

- Merge the two previous logics - Show that one is annotated and the other not. - Possível Resolução - integração dos sinonimos e abreviaturas das bases de dados com ETL (podes por uma frase para cada uma delas e um paragrafo sobre ETL) Anotação de modelos: - match direto - decomposição do nome em dois (backbone e side chain) - queries à base de dados

Fig. 2. Escheme of the process done with ETL pipeline. First gather data from conventional databases then processes 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. 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.
5. 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.
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. 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.
14. E. Fahy, D. Cotter, M. Sud, and S. Subramaniam, "Lipid classification, structures and tools," *Biochimica et biophysica acta*, vol. 1811, p. 637, 11 2011.
15. 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.
16. 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.