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. The abstract should briefly summarize the contents of the paper in 15–250 words.

Keywords: First keyword · Second keyword · Another keyword.

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. Particularly in the case of lipids[3], only a small chunk of GSM models reconstructed have defined lipids. These models neglect the fact that each subclass 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 lipid biosynthesis network. Therefore it is important the integration of such information in lipidic models, for a better interpertability, 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.

1.2 Objective

The main objetive 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 an itegration 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 this models to ensure their accuracy and further atualization [11, 12].

2.2 Lipid computational representation

- Generic/structurally defined duality – o que é um lipido, como é a sua estrutura – importancia dos lipidos – importancia de saber as estruturas dos lipidos – como são geralmente representados nos modelos, dualidade generica ou estrutural

2.3 Generic Representation in GSM models

- Lipid only with the backbone; - Name of the backbone; - Databases cross-references;

2.4 Structurally defined representation in GSM models

- Lipid with the whole structure - Name composed of backbone and side chains - No databases cross-references

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

Sample Heading (Third Level) Only two levels of headings should be numbered. Lower level headings remain unnumbered; they are formatted as run-in headings.

Sample Heading (Fourth Level) The contribution should contain no more than four levels of headings. Table 1 gives a summary of all heading levels.

Table 1. Table captions should be placed above the tables.

Heading level	Example	Font size and style
Title (centered)	**Lecture Notes**	14 point, bold
1st-level heading	**1 Introduction**	12 point, bold
2nd-level heading	**2.1 Printing Area**	10 point, bold
3rd-level heading	**Run-in Heading in Bold.** Text follows	10 point, bold
4th-level heading	*Lowest Level Heading.* Text follows	10 point, italic

Displayed equations are centered and set on a separate line.

$$x + y = z \tag{1}$$

Please try to avoid rasterized images for line-art diagrams and schemas. Whenever possible, use vector graphics instead (see Fig. 1).

Theorem 1. *This is a sample theorem. The run-in heading is set in bold, while the following text appears in italics. Definitions, lemmas, propositions, and corollaries are styled the same way.*

Proof. Proofs, examples, and remarks have the initial word in italics, while the following text appears in normal font.

For citations of references, we prefer the use of square brackets and consecutive numbers. Citations using labels or the author/year convention are also acceptable. The following bibliography provides a sample reference list with entries for journal

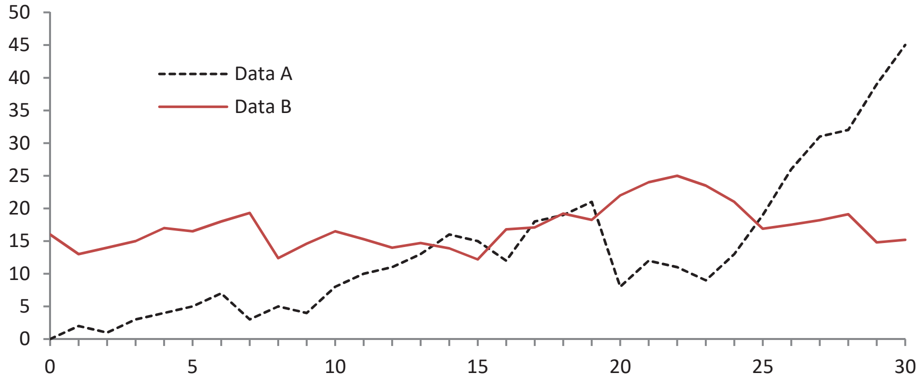

Fig. 1. A figure caption is always placed below the illustration. Please note that short captions are centered, while long ones are justified by the macro package automatically.

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. 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.
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.