Towards a multivariate analysis of Genome-Scale Metabolic Models Derived from the BiGG Models database

Alexandre Oliveira (✉), Emanuel Cunha, Fernando Cruz, João Capela, João Sequeira, Marta Sampaio, and Oscar Dias

Centre of Biological Engineering, University of Minho, Braga, Portugal
{alexandre.oliveira, ecunha, fernando.cruz, joao.capela, jsequeira, msampaio}@ceb.uminho.pt, odias@deb.uminho.pt

Abstract. Genome-Scale metabolic models (GEMs) are a relevant tool in systems biology for *in silico* strain optimisation and drug discovery. An easier way to reconstruct a model is to use available GEMs as templates to create the initial draft, which can be curated up until a simulation-ready model is obtained. This approach is implemented in *merlin*'s BiGG Integration Tool, which reconstructs models from existing GEMs present in the BiGG Models database. This study aims to assess draft models generated using models from BiGG as templates for three distinct organisms, namely, *Streptococcus thermophilus*, *Xylella fastidiosa* and *Mycobacterium tuberculosis*. Several draft models were reconstructed using the BiGG Integration Tool and different templates (*all*, *selected* and *random*). The variability of the models was assessed using the reactions and metabolic functions associated with the model's genes. This analysis showed that, even though the models shared a significant portion of reactions and metabolic functions, models from different organisms are still differentiated. Moreover, there also seems to be variability among the templates used to generate the draft models to a lower extent. This study concluded that the BiGG Integration Tool provides a fast and reliable alternative for draft reconstruction for bacteria.

Keywords: Genome-Scale Metabolic Models, *merlin*, BiGG Integration Tool, BiGG models.

1 Introduction

The reconstruction of comprehensive Genome-Scale Metabolic Models (GEMs) is nowadays a common approach in systems biology. The reconstruction of GEMs relies on using genomic data of a given organism to assemble a genome-wide metabolic network, which can predict the metabolic behaviour in different conditions [1, 2], using simulation methods like Flux Balance Analysis (FBA) [3]. Furthermore, these models are used for *in silico* strain optimisation and drug target discovery [4].

A wide variety of models are available in several online databases. Even though most reconstructed models correspond to bacterial organisms, models for more complex organisms, such as plants and mammals, have become more relevant lately [5].

M. Rocha et al. (Eds.): PACBB 2021, LNNS 325, pp. 136-144, 2022.
https://doi.org/10.1007/978-3-030-86258-9_14

The BiGG Models is a centralised online database of high-quality, manually curated GEMs collected from available literature [6]. Since 2010, BiGG has compiled accessible information on the models' reactions, metabolites and genes. Currently, this knowledge base contains 108 metabolic models from a wide variety of organisms, ranging from bacteria, such as *Escherichia coli*, to more complex organisms like *Homo sapiens*. In addition, BiGG attempts to connect the information it contains with external databases and with the standardisation of reactions and metabolites identifiers across GEMs to allow direct comparison between models [6].

The first step of the model reconstruction, in a bottom-up approach, is to create a draft metabolic network using the organism's annotated genome and biochemical databases. However, this draft network corresponds to an incomplete set of reactions that includes gaps, dead-end metabolites and blocked reactions, requiring further curation to obtain the final model [1].

An alternative approach is to use existing GEMs as templates to create the initial draft. In this approach, reactions are added to a draft model when homologous genes in the template models are available. CarveMe implements such a top-down approach using all reactions and metabolites from BiGG to build a universal model, which will then be carved into a final simulation-ready gapless model [7].

Hence, this study aims to assess draft GEMs generated using BiGG models as a template for three distinct organisms: *Streptococcus thermophilus*, *Xylella fastidiosa* and *Mycobacterium tuberculosis*. For this, an inhouse developed tool available in *merlin*, named BiGG Integration Tool (BIT) [8, 9], was used. Furthermore, we assessed the variability of the generated draft models' reactions and metabolic functions for different reconstruction approaches and compared them with the models generated using CarveMe.

2 Results and Discussion

We created 21 draft models using three distinct approaches and analysed the models through a multivariate analysis. In detail, we reconstructed seven draft models for each bacteria: *M. tuberculosis*, *S. thermophilus* and *X. fastidiosa*. BIT allows creating draft reconstructions automatically using three templates. For the first template (*all*), BIT uses all information available in BiGG models, whereas for the second template (*selected*), the user selected a set of models from BiGG models, in this case three models were used. Finally, for the last template (*random*), BIT will randomly select a set of three models from the database. Likewise, we used CarveMe [7] to obtain a draft model for each bacterium. The models were then analysed regarding the variability of reactions. The metabolic functions of the draft models were compared through the Clusters of Orthologous Genes (COGs) database.

2.1 Genomes' Comparative Functional Analysis

Besides two bacteria, *S. thermophilus* and *X. fastidiosa,* unavailable in BiGG, the recogniser [10] tool was used to collect COG identifiers for all species present in BiGG. As shown in Figure 1, we analysed the principal components contributing to the variability of the COG-annotated metabolic functions. The analysis results suggest that the BiGG database's organisms are grouped by phylum. Moreover, there is a clear separation between eukaryotes and prokaryotes. These results corroborate the database authors latest publication [6], in which models from eukaryotes and prokaryotes were segregated using PCA.

The similarity of metabolic functions among the three bacteria *M. tuberculosis*, *S. thermophilus* and *X. fastidiosa* was further analysed using the metabolic COG identifiers. According to Figure 2, *M. tuberculosis* had the highest number of COG identifiers (1080), of which 544 were unique. On the other hand, *X. fastidiosa* and *S. thermophilus* shared most of their COG identifiers with *M. tuberculosis*, having only 155 and 126 unique COG identifiers, respectively. All organisms in this study share 226 COG metabolic identifiers. Hence, there seems to be a clear distinction of the functional annotation among the organisms selected for this study.

BIT and CarveMe were then used to generate draft models for *S. thermophilus*, *X. fastidiosa* and *M. tuberculosis* (Supplementary Material 1), representing different microorganisms, namely a lactic acid bacterium (gram-positive), a plant-pathogen (gram-negative) and a well-studied bacteria, which already has a GEM available on BiGG, respectively. Next, we assessed the variability of reactions and genes' metabolic functions included in these draft reconstructions.

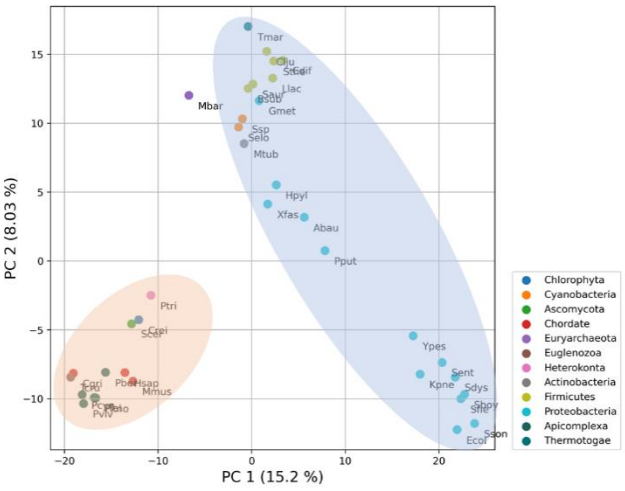

Fig. 1. PCA plot comparing metabolic COG identifiers obtained for the organisms present in BiGG as well as *S. thermophilus* and *X. fastidiosa*. Principal Components (PC) 1 and 2 are depicted with the percentage of explained variance. PCA scores have been plotted and coloured according to the organism's phylum, and ellipses represent the clusters obtained for *Eukarya* (orange) and *Bacteria* (blue), with the point outside both belonging to Archaea. Each dot is annotated with an organism-specific identifier, using the first letter of the genus and the first three letters of the species second name.

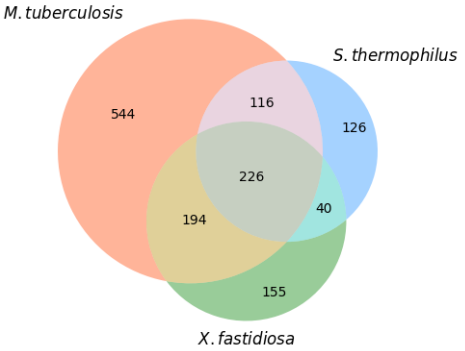

Fig. 2. Venn diagram of the COG identifiers obtained with recogniser for *M. tuberculosis*, *S. thermophilus* and *X. fastidiosa*. Numbers indicate total number of unique COG identifiers.

2.2 Models' Analysis

The draft models were generated from BiGG using BIT with the mentioned templates (*all*, *selected* and *random*). The content of the models generated by each template was compared in a Venn diagram, representing the number of unique and shared reactions in the different models of a given organism (Figure 3). This analysis allowed us to assess the influence of the template on the content of the models. The models obtained using the *all*-template have a larger number of reactions, of which over 55% are missing in the other templates' models. Nevertheless, *all*-template models include most reactions of the other templates, though to a lesser extent for *M. tuberculosis*. One possible explanation is that BIGG includes models for *M. tuberculosis*, which have been used to create the draft model of this organism in the *all*-template. Concerning the remaining templates, homology searches may return other matches based on the selected models. Hence, this analysis suggests that the BIT's template will influence the number of reactions included in a draft model.

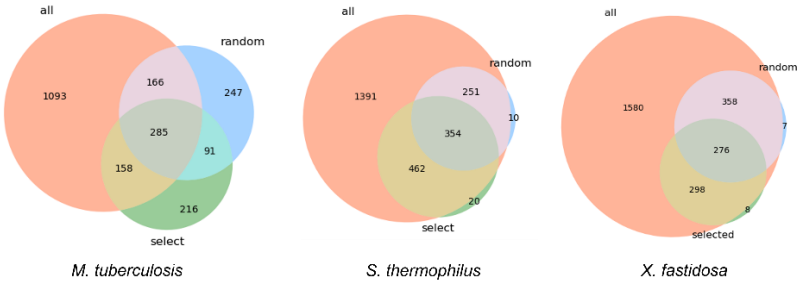

Fig. 3. Venn diagrams for reactions by organism. Several draft models have been generated from BiGG using the *merlin*'s BIT and different templates: *all, random* and *selected*. Venn Diagrams illustrate the number of reactions shared between the different models for each organism.

The draft reconstructions derived from the *selected*-template were analysed together with CarveMe's models. The number of reactions shared among the draft models of the three bacteria is presented in the Venn diagram in Figure 4, whereas the diagrams for the remaining BIT's templates are presented in Supplementary Material 2. With the *selected*-template, the three resulting models shared 180 reactions among them. Moreover, the draft model for *S. thermophilus* contained more unique reactions (556 reactions), while *X. fastidiosa* shared 351 reactions, mostly with *M. tuberculosis*, which represents 60% of its total reactions. Nevertheless, this result is different from the metabolic annotation, as *M. tuberculosis* had more unique COG identifiers than any other bacteria (Figure 2).

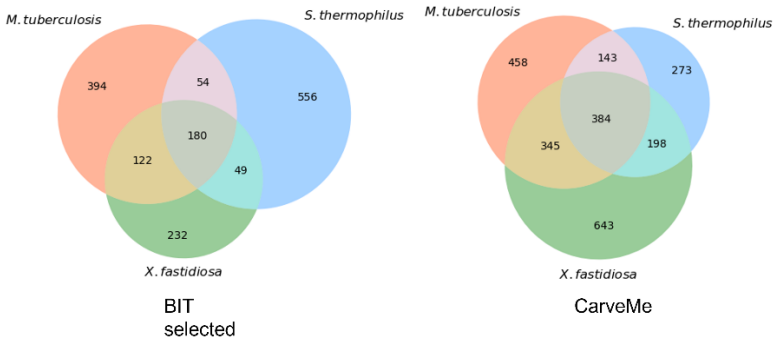

Fig. 4. Venn diagrams for reactions of draft models created with BIT's *selected*-template and CarveMe, showing the number of reactions shared between the models of different organisms.

Regarding the draft models created with CarveMe, 384 reactions are shared among the three bacteria. However, in contrast with BIT's *selected*-template results, CarveMe's draft model for *X. fastidiosa* has the highest number of unique reactions (443 reactions), whereas the *S. thermophilus* model shares 725 reactions with the other models, 527 with *M. tuberculosis*, and 582 with *X. fastidiosa*.

Figure 5 displays the comparison of the draft models created with both tools. Here, we analysed the number of reactions shared among draft models of the same organism but created with BIT's *selected*-template and CarveMe. Almost half of the reactions in BIT's *selected*-template models are not present in CarveMe's models. Thus, although both tools use BiGG to generate draft reconstructions, the obtained models are significantly different. However, models created with CarveMe include more reactions than BIT's *selected*-template models, as the former tool generates a simulation-ready gapless model [7]. Hence, CarveMe's models will also include artifacts, like sink and demand reactions, that are not included in the drafts generated with BIT, which can explain some of the variability. On the other hand, BIT's models still require curation and gap filling to obtain a simulation-ready model.

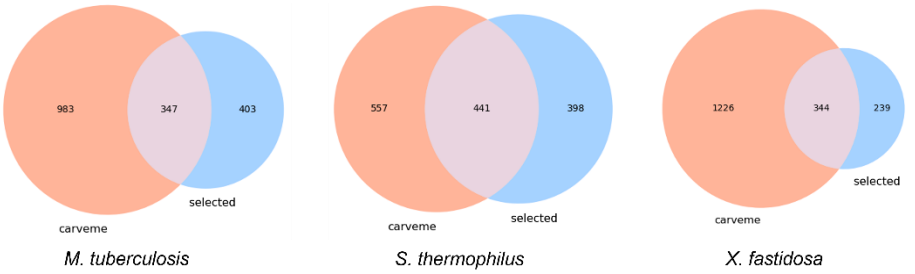

Fig. 5. Venn diagram for reactions by organism's model. The number of reactions shared between the models created with the *merlin*'s BIT using the *selected*-template and those created with CarveMe was assessed for each organism.

Finally, the reaction space of all draft models was analysed by PCA, and the score plots for the first three principal components are shown in figure 6. These components explained 32.7% of the variability in the reaction's space. Principal Component (PC) 1 separates the data into three groups. The group with the lowest score contains four random models from *M. tuberculosis,* while the group with the highest score covers all *S. thermophilus'* models. The other group comprises all models of *X. fastidiosa* and the remaining for *M. tuberculosis*. PC2 separates this last group by organism. PC3 does not clearly separate models by organism though it converges CarveMe's models. According to the reactions' PCA, models of the same organism seem more identical to each other rather than to models of a different organism. Nonetheless, the template used also contributes to the variability in the models' reaction space, but to a lesser extent.

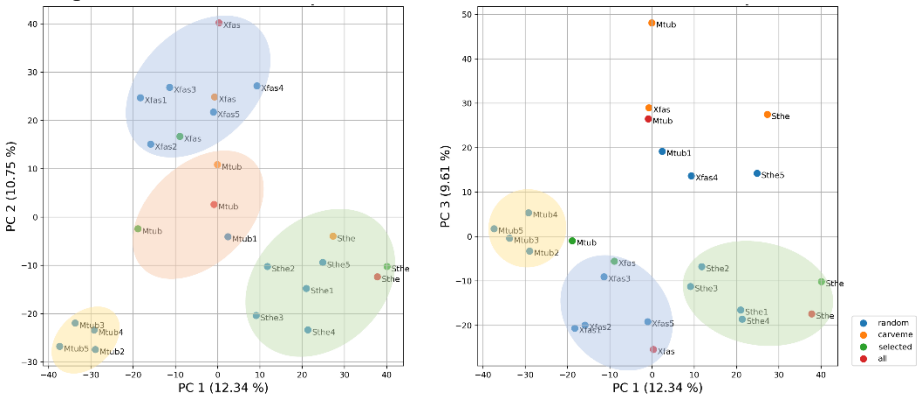

Fig. 6. PCA of the draft models' reaction space. Several draft models have been generated from the BiGG Models database using BIT (varying the template) and CarveMe package (using the universal BiGG model). Principal Components (PC) 1, 2, and 3 are depicted with the percentage of explained variance. PCA scores have been plotted and coloured according to the set of template models. Each dot is annotated with a model-specific identifier, using the first letter of the genus and the first three letters of the species second name. Since five random-template models have been created using a different set of template models, these models are also numbered accordingly. Ellipses surrounding a given set of models are merely presented for illustration purposes and do not represent real k-means clusters.

Genes used in the draft models were retrieved and cross-referenced with the COG annotation of the genomes to assess the metabolic functions included in the draft models. The metabolic annotation of BIT's *selected*-template and CarveMe models was assessed using a Venn diagram (Figure 7). A substantial portion of COG functions is shared among all models created using BIT's *selected*-template. Likewise, models created with CarveMe also reveal the same pool of common metabolic functions. However, a similar number of COG functions is unique to each draft model created with BIT's *selected*-template and CarveMe tools.

In contrast, the metabolic COG annotation performed on the three organisms indicates a smaller portion of common metabolic functions and higher percentages of unique metabolic COGs for each organism, suggesting that the representation of the metabolism in the draft models is still incomplete. Interestingly, the large number of unique reactions among the draft models created with both BIT's *selected*-template and CarveMe tools does not support the metabolic annotation.

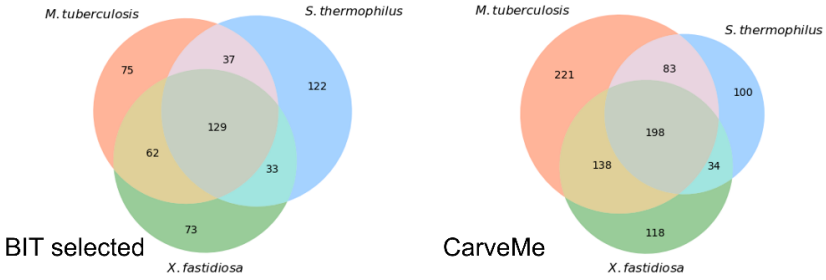

Fig. 7. Venn's diagram for the metabolic COG annotation of the draft models generated using BIT's *selected*-template and CarveMe.

The collections of COG identifiers obtained for each model were now represented in a scatter plot using PCA scores (Figure 8). This PCA suggested that neither the tool nor the template used to generate draft models significantly impact the model's metabolic characterisation. The metabolic COG annotation obtained for each model seems not to change significantly with the template or method. According to PC 1 (Figure 7), the functional characterisations of the *S. thermophilus* models can be differentiated from *X. fastidiosa* and *M. tuberculosis* models. Likewise, *X. fastidiosa* and *M. tuberculosis* models obtained different PCA scores according to PC 2. These results show that all three methodologies result in similar sets of metabolic genes, distinct from other organisms' sets of metabolic genes.

Although the metabolic characterisation of the draft models' genes allows differentiating models by organism rather than by template or method, the analysis of both reaction spaces still suggests that models seem to share a significant portion of reactions (Figure 6).

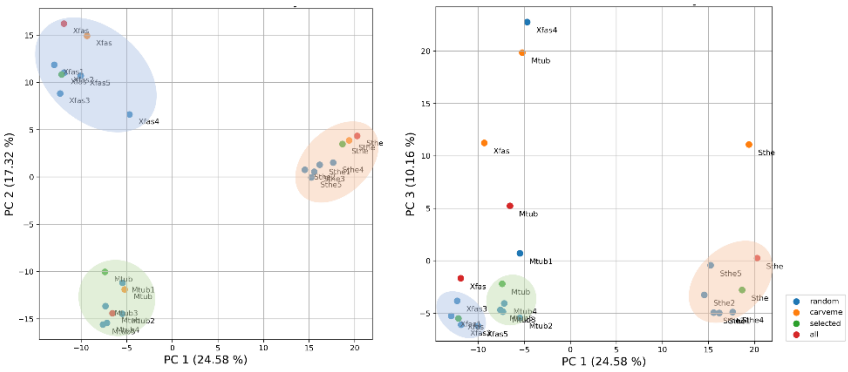

Fig. 8. PCA of the draft models' metabolic COG annotation. Several draft models have been generated from the BiGG Models database using BIT (varying the template) and CarveMe package (using the universal BiGG model). Principal Components (PC) 1, 2, and 3 are depicted with the percentage of explained variance. PCA scores have been plotted and coloured according to the set of template models. Each dot is annotated with a model-specific identifier, using the first letter of the genus and the first three letters of the species second name. Since five random-template models have been created using a different set of template models, these models are numbered accordingly. Ellipses surrounding a given set of models are merely presented for illustration purposes and do not represent real k-means clusters.

3 Conclusion

This study concludes that BIT can reconstruct differentiated draft models from BiGG, regarding reactions and metabolic functions of the models' genes. This means that BiGG can be used as a source of templates in a bottom-up approach, as it appears to generate distinct models for different species. Nevertheless, because of the distribution of organisms analysed in this work, this can only be stated for simple bacterial organisms. Therefore, further analysis is required to assess the applicability of this method in more complex organisms.

Moreover, this tool presents an easy and fast alternative to reconstruct GEMs. However, it must be considered that the template used can also affect the resulting drafts. Thus, it must be carefully selected for higher-quality results. Further curation and gap-filling will still be required to obtain the final simulation-ready model.

4 Materials and Methods

4.1 Genomes' Comparative Functional Analysis

The functional comparison was performed for each organism with a corresponding BiGG model and for *S. thermophilus* and *X. fastidiosa*. The COG database is a popular resource for functional characterisation [11] and was used as the reference for functional annotation with recogniser [10], as described in Supplementary Material 3.

4.2 Draft models

BIT was used to reconstruct draft models from BiGG. A detailed description of how the tool works is presented in supplementary material 4. Seven drafts were reconstructed for each organism, using different templates: *all*, *selected* and *random*. In addition, three drafts were reconstructed using CarveMe for comparison. A detailed description of the methods used to reconstruct the drafts is described in Supplementary Material 1.

4.3 Multivariate analysis

The reactions and metabolic functions of the 24 draft models were compared using Venn diagrams and PCA plots. The methodology used for this analysis is presented in Supplementary Material 2.

5 Supplementary Materials

All Supplementary Material files mentioned in the manuscript are available at https://nextcloud.bio.di.uminho.pt/s/GZC2577Nz7K4AqP.
All the scripts used for this work are available at https://github.com/BioSystemsUM/bit-analysis.

Acknowledgements

This study was supported by the Portuguese Foundation for Science and Technology (FCT) under the scope of the strategic funding of UIDB/04469/2020 unit. A. Oliveira (DFA/BD/10205/2020), E. Cunha (DFA/BD/8076/2020), F. Cruz (SFRH /BD/139198/2018), J. Sequeira (SFRH/BD/147271/2019), and M. Sampaio (SFRH/BD/144643/2019) hold a doctoral fellowship provided by the FCT. Oscar Dias acknowledge FCT for the Assistant Research contract obtained under CEEC Individual 2018.

References

1. Thiele, I., Palsson, B.Ø.: A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat. Protoc. 5, 93–121 (2010).
2. Feist, A.M., Herrgård, M.J., Thiele, I., Reed, J.L., Palsson, B.: Reconstruction of biochemical networks in microorganisms. Nat. Rev. Microbiol. 7, 129–143 (2009).
3. Orth, J.D., Thiele, I., Palsson, B.Ø.: What is flux balance? Nat. Biotechnol. 28, 245–248 (2010).
4. O'Brien, E.J., Monk, J.M., Palsson, B.O.: Using genome-scale models to predict biological capabilities. Cell. 161, 971–987 (2015).
5. Zhang, C., Hua, Q.: Applications of genome-scale metabolic models in biotechnology and systems medicine. Front. Physiol. 6, 1–8 (2016).
6. Norsigian, C.J., Pusarla, N., McConn, J.L., Yurkovich, J.T., Dräger, A., Palsson, B.O., King, Z.: BiGG Models 2020: Multi-strain genome-scale models and expansion across the phylogenetic tree. Nucleic Acids Res. 48, D402–D406 (2020).
7. Machado, D., Andrejev, S., Tramontano, M., Patil, K.R.: Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. Nucleic Acids Res. 46, 7542–7553 (2018).
8. Dias, O., Rocha, M., Ferreira, E.C., Rocha, I.: Reconstructing genome-scale metabolic models with merlin. Nucleic Acids Res. 43, 3899–3910 (2015).
9. Capela, J., Lagoa, D., Rodrigues, R., Cunha, E., Cruz, F., Barbosa, A., Bastos, J., Lima, D., Ferreira, E.C., Rocha, M., Dias, O.: merlin v4.0: an updated platform for the reconstruction of high-quality genome-scale metabolic models. bioRxiv. (2021).
10. Sequeira, J.C., Rocha, M., Alves, M.M., Salvador, A.F.: UPIMAPI, reCOGnizer and KEGGCharter: three tools for functionalannotation. In: BOD 2021 - X Bioinformatics Open Days. Braga, Portugal. 57 (2021).
11. Galperin, M.Y., Kristensen, D.M., Makarova, K.S., Wolf, Y.I., Koonin, E. V.: Microbial genome analysis: The COG approach. Brief. Bioinform. 20, 1063–1070 (2019).