Google Summer of Code 2021
CODING PROJECT PROPOSAL
From DNA sequences to metabolic interactions: building a pipeline to extract key metabolic processes

Name: Haris Zafeiropoulos
Affiliation: Department of Biology, University of Crete
Program: PhD candidate, Second participation in GSoC
Mentors: Elias Tsigaridas, Apostolos Chalkis, Zafeirakis Zafeirakopoulos
email: haris.zafr@gmail.com
GitHub: https://github.com/hariszaf
Address: Valestra 99, Heraklion, Crete, Greece, 71202
Phone: +30 694 909 3089

April 13, 2021

Contents

1 Synopsis

Twenty years after Covert W.Covert, Bernhard O. Palsson and their colleagues published their review on metabolic modelling of microbial strains [1], the value of this method has been well established.

From the beginning, metabolic modeling has been interwoven with constraing-based methods [2]. The value of randomized sampling in the framework of metabolic modeling has been proved itself over the years [3], [4].

High Throughput Sequencing technologies have allowed process in the genetic information (DNA) in a cost-efficient and easy way, especially for microbial species as their genome is only but a few hundreds of genes long.

Once the complete genome of a species is obtained, the complete reconstruction of the metabolic network of the species is enabled, called genome-scale metabolic models (GEMs) [5].

Such models for all the species present in a microbial community, allows the study of metabolic interactions, thus an insight for the actual microbial intaractions [6].

Aim of this project is to integrate the produced data and knowledge of these twenty (and more) years and make use of the randomized flux sampling method to evaluate the metabolic interactions retrieved. To this end, thousands of publicly available reference microbial genomes will be selected, and their automatic metabolic network reconstructions will be implemented. Based on these models, cross-feeding interactions algorithms will be performed for groups of species to extract key metabolic processes. New functions, implementing the recently developed Multiphase Monte Carlo Sampling (MMCS) approach [7] in the framework of the `dingo` library, will make use of the randomized flux sampling concept to evaluate the processes retrieved.

2 The Project

2.1 Background

Microbial communities populate most environments on earth; from the seafloor to the human gut, they literay live everywhere [8]. The play a critical role in shaping the environment as we know it. By driving biogeochemical cycles, bacteria, along with geochemical (abiotic) transformations (atmospheric, tectonic and geothermal), shape Earth's climate [9]. At the same time, *the human body is inhabited by millions of tiny living organisms* having a fundamental part in keeping us healty[10]. However, up to nowm scientists are able to cultivate approximately 1% of known Bacteria [11]. Since the growth of various bacteria depends on their interactions with others [12], inferring microbial interactions would strongly support cultivating taxa for the first time, allowing the production of secondary metabolites and their biotechnological applications. At the same time, it would be an essential tool to further expand our understanding regarding the underlying mechanisms governing a range of phenomena, from ecosystem functioning to human health disorders. This is why researchers from all these scientific fields have been studied the community structure of the microbial communities of their interest, focusing both on the taxa present and the metabolic processes (referred as *functions*).

Metabolism is a network of the metabolic pathways that occur in an organism; thus, a metabolic network is the representation of all these pathways [2]. Using the stoichiometry of each reaction, which is always the same in the various species, we convert the metabolic network of an organism into a mathematical model [2]. High Throughput Sequencing has made access to the complete genomic information of an organism rather easy. However, building the complete metbolic network of a species is not that trivial yet [13]. However, over the last few years, automatic reconstruction approaches for building genome-scale metabolic models [14] of relatively high quality have been developed.

The study of metabolic dependencies to infer interspecies microbial interactions has been estrablished the last years (Fig. 2.1), for more see [15].

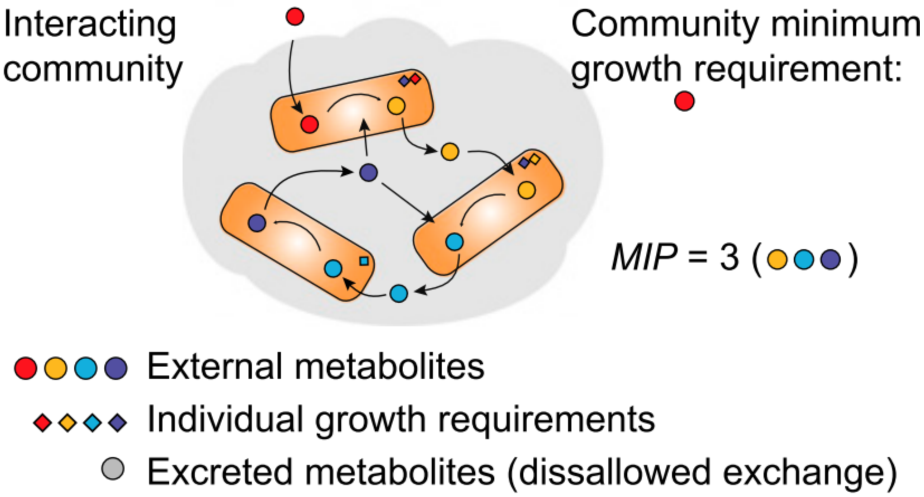
Interacting community
Community minimum growth requirement:
MIP = 3 (●●●)
●●●● External metabolites
◆◆◆◆ Individual growth requirements
● Excreted metabolites (dissallowed exchange)

Figure 1: Representation of a microbial community of 3 interacting species. The Metabolic Interaction Potential metric (MIP) quatifies the propensity of a community to exhange metabolites. The figure is part of Fig. 3. at [15]

Constrained-based modelling approaches, such as as Flux Balance Aanalysis and flux randomized sampling, have enabled the investigation of the properties of the possible steady states of a metabolic network, end up with essential insights on metabolism at every level [16], [17]. Microbial communities harbor metabolically interdependent cooperative groups. As a matter of fact, such groups play a key role for species co-occurrence [15].

The aim of this project is to bring together data (reference microbial genomes), automatic genome scale metabolic networks reconstruction tools, cross-feeding interactions algorithms and flux randomized sampling. This way, we will enable the evaluatation of the effect of each of the predicted metabolic interactions, to the various species of the community.

2.2 Methodology

This project will make use of third party software tools and will also develop some new functions in the framework of the `dingo` project.

Global repositories including thousands of reference microbial genomes such as the Genome Taxonomy Database (GTDB) will be exploited [18], [19].

Automatic genome scale metabolic networks reconstruction tools, such as AuReMe [20] and CarveMe [14], have been recently developed. Such tools will be implemented to get the corresponding GEMs for the genomes retrieved and/or for the genomed provided by the user.

The SMETANA package, is a set of algorithms looking for possible cross-feeding interactions between the species of the microbial community under study [15]. Such algorithms will be implemented in the GEMs built from the original genomes gathered or/and provided by the user, to return key metabolites.

Flux randomized sampling using the MMCS algorithm of the `dingo` Python library will be implemented to evaluate the effect of the metabolites returned from the previous steps. To this end, new functions

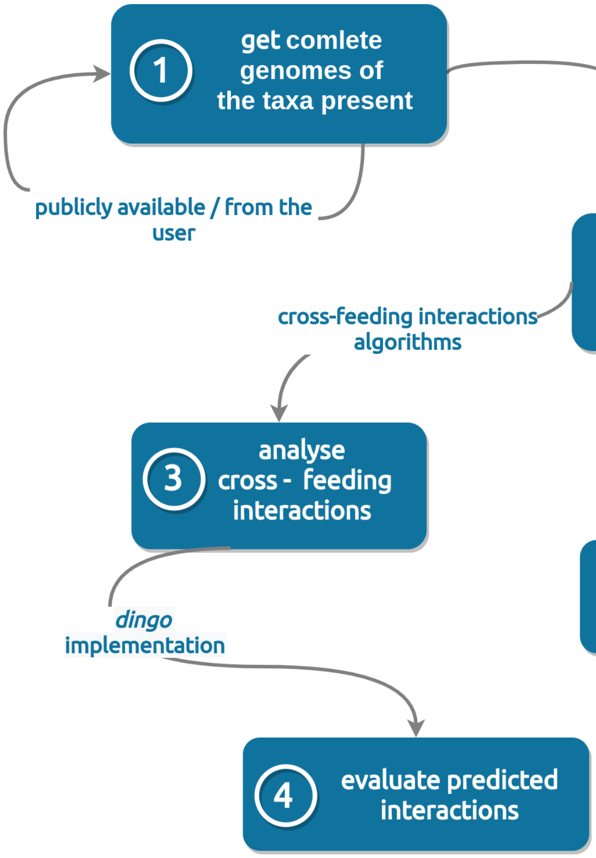
1
get comlete genomes of the taxa present
publicly available / from the user
automatic GEM reconstruction tools
2
build / retrive correspondng metabolic networks
cross-feeding interactions algorithms
3
analyse cross - feeding interactions
dingo implementation
5
infer microbial associations
dingo implementation
4
evaluate predicted interactions

Figure 2: Project's workflow. This proposal consists of 4 major steps combining third party tools implementation and development of new code in the framework of the `dingo` Python library.

will be developed in the framework of the `dingo` library to check the response of the biomass function [21] or the function of maximization of ATP per unit flux [22] in alterations regarding the corresponding metabolites. Furthermore, functions will be developed to perform sampling in the flux spaces of all the paired associations between the various species, to strengthen these associations, depending on whether they respond robustly on exchange/competition for metabolites.

2.3 Benefits to the Community

The proposed project will benefit the most the `dingo` Python library as an open source code project, as it will provide a thorough use case in a big-data scale and will also develop further functions for the analysis of microbial communities.

Furthermore, it will be essential for biologists studying microbial interactions in all the different biological fields, from ecosystem functioning to precision medicine.

3 Deliverables

3.1 Overview

The tasks of this project will be split in the two main periods of the project; i.e. the Community Bonding period between the 17th of May to the 7th of June (3 weeks) and the Coding period between the 7th of June and the 16th of August (10 weeks). At the end of the coding period, all the implementations described in the Deliverables described above, will have been completed.

The `dingo` library provides a great basis for the implementation of this project. Therefore, a limited amount of time is needed to set up all the necessary coding environments. As I have been involved with the GeomScale group and the `dingo` library for over a year, it will not need time to get used to the group's working framework. In addition, the proposed subject has been discussed and thorough study of the scientific issues that might occur, will be addressed before the bonding period starts.

That said, the coding tasks could start before the bonding period is over. Furthermore, a blog will be created to report the project's progress periodically and a thorough documentation for the `dingo` Python library will be delivered as well as a repository with the models built.

The `dingo` library is under the LGPL-3 license. However, as the workflow will make use of third-party tools separate licenses will apply.

3.2 Bonding period (May 17 - June 7)

Main milestone of the bonding period will be to communicate with biologists to address and confirm all the scientific-oriented issues that might occur. A group of molecular ecologists have already knowledge of the project and I will be in contact with them even before bonding period starts.

The first two weeks will require a strong effort from me in designing, the `dingo` functions as well as the setup of the repository for the models that will be about to build and the download/storage of their corresponding genomes. The mentors' guidance will be essential during this time.

In the third week of the bonding period, it is my intention to be ready to start coding. I do not count it as a coding week, however I believe I will be ready.

Furthermore, I will build a `git` branch from the `dingo` GitHub repository and a blog to report my progress periodically.

3.3 Coding period (June 7 - August 16)

This 10 weeks period will be split in the following time windows, each of which corresponds to a certain milestone:

- **Weeks: #1 and #2**

In the first 2 weeks, the repository where the genomes and the models will be stored will be built, Fig. 2.2 box 1. In addition, the more than 190,000 genomes from the GTDB will be downloaded. A deamon for getting new or updated genomes will be also set. A first documentation of the repository and how to access it will be delivered.

- **Week: #3**

In this week, the two software tools mentioned in the 3 section for automatic GEM reconstruction, will be thoroughly tested and it will be decided which will be used, Fig. 2.2 box 2. GEMs then will be built and stored. Format-oriented issues will be also addressed. The models built during this week will be available via the blog of the project as well.

- **Weeks: #4 to #7**

During this period, the SMETANA algorithms will be implmemented for all the models built by that time. Thourough examination of the output will be performed Fig. 2.2 box 3. Once a small number of successful runs of the SMETANA algorithms are completed, I will start developing the `dingo` functions for the evaluation of the metabolities returned, Fig. 2.2 box 4. As this will be the most complicated task, both from the coding and the biology point of view, extensive communication both with the mentors of the project and biologists will be held to its fine completion. A short report with the process of the project during this period will be delivered in the project's blog.

- **Weeks: #8 and #9**

In those two weeks, `dingo` functions for inferring microbial interactions in pairs of GEMs will be developed, Fig. 2.2 box 5. Just like the previous milestone, this part will be developed side-by-side with biologists ensuring that the new functions developed, provide valid outcome. As this part is rather scientific, if further developments is needed, I will discuss this with the mentors to continue working on that even after the project is complete. By the end of this period, I will have a report delivered in the project's blog.

- **Week: #10**

During this last week, tests for the complete pipeline will be held. Thourough documentation for the pipeline will be provided. I will make a Pull Request, so my branch to merge with the `main` branch of the `dingo` library. I will finally submit the final evaluation of the GSoC project.

Comment: The above time schedule is a worst-case schedule. However, if the coding exceeds my present expectations I will discuss with mentors new additions and implementations.

There are not schedule conflicts during the summer.

---

4 Related work

I have been a member of the GeomScale project over the last year. Based on an idea from the GSoC of 2020 we have been working on sampling the flux space of metabolic networks. I contributed with wrapping the C++ code of `VolEsti` to build the Python interface. Furthermore, I implemented the MMCS method developed by the GeomScale group on metabolic networks of high dimensions [7]. This work was accepted in the proceedings of the 37th Symposium on Computational Geometry.

5 Tests

This proposal is not among the ideas listed in the table of proposed coding projects of the GeomScale group. As the scope of my proposal is close to the Inferring microbial interactions project, and after contacting the mentors, I implemented the tests of the latter project in the framework of my proposal.

You can find my answers on this link and the correspondig code on its correspondig GitHub repository.

6 Biographical information

6.1 Education

- PhD candidate at University of Crete (2018 - ongoing). Dissertation on: "Merging NGS data, knowledge aggregation and data integration techniques, along with ecological network analysis (ENA): an attempt to decipher microbial community ecology and ecosystem functioning by taking advantage of the hypothesis-generating method".
- MSc in Bioinformatics at the University of Crete (2016 - 2018). Thesis: "eDNA metabarcoding for biodiversity assessment: Algorithm design and bioinformatics analysis pipeline implementation"
- BSc in Biology at the National and Kapodistrian University of Athens (2010 - 2016). Thesis: "Morphology, morphometry and anatomy of species of the genus Pseudamnicola in Greece"

6.2 Publications

- Zafeiropoulos, Haris, Anastasia Gioti et al. 0s and 1s in marine molecular research: a regional HPC perspective (**under review in GigaScience journal**)
- Zafeiropoulos, Haris, Anastasia Gioti et al. (2021, April 5). The IMBBC HPC facility: history, configuration, usage statistics and related activities (Version 1.0.0). Zenodo. DOI: 10.5281/zenodo.4665308
- Polymenakou, Paraskevi N., et al. "The Santorini Volcanic Complex as a Valuable Source of Enzymes for Bioenergy." Energies 14.5 (2021): 1414. DOI: 10.3390/en14051414
- Chalkis, Apostolos, et al. "Geometric algorithms for sampling the flux space of metabolic networks." arXiv preprint arXiv:2012.05503 (2020) https://hal.inria.fr/hal-03047049v2
- Zafeiropoulos, Haris, et al. "PEMA: a flexible Pipeline for Environmental DNA Metabarcoding Analysis of the 16S/18S ribosomal RNA, ITS, and COI marker genes." GigaScience 9.3 (2020): giaa022.. DOI:10.1093/gigascience/giaa022

6.3 Programming

- Author of the `PEMA` workflow
- Very experienced in programming with scripting programming languages (Python, BigDataScript, R).
- Basic knowledge of C++
- Very experienced in container-based technologies (Docker, Singularity)
- Experienced in web

6.4 Personal motivation

Bioinformatics play a great part in almost every aspect of modern biology. Metabolic interactions [23] is now the next big thing in systems biology. That is why after graduating Biology school (BSc), I focused on computer science and Bioinformatics (MSc). I am now quite experienced with a series of scripting programming languages (Python, BigDataScript, R) and with container-based technologies (Docker, Singularity).

My research interests focus on ecology and ecosystem functioning at the microbial dimension. As Systems Biology approaches can benefit the most this field, I have spent the last 2 years working on knowledge aggregation and data integration techniques as well as networks analysis are employed. Metabolic interactions [23] is now the next big thing in systems biology. Over the last year, I have been working with the GeomScale group on the `VolEstipy` project. Developing algorithms for sampling on the flux space of metabolic networks have helped me in getting a more holistic point of view of the potentials of sampling in the study of metabolic networks.

This GSoC project would allow me to continue working with the GeomScale group, whilst building the proposed application would benefit me the most, especially as it will allow me to extend the scopus of my PhD dissertation.

References

[1] M. W. Covert, C. H. Schilling, I. Famili, *et al.*, "Metabolic modeling of microbial strains in silico," *Trends in biochemical sciences*, vol. 26, no. 3, pp. 179–186, 2001.

[2] B. Palsson, *Systems biology.* Cambridge university press, 2015.

[3] J. Schellenberger and B. Ø. Palsson, "Use of randomized sampling for analysis of metabolic networks," *Journal of biological chemistry*, vol. 284, no. 9, pp. 5457–5461, 2009.

[4] H. A. Herrmann, B. C. Dyson, L. Vass, G. N. Johnson, and J.-M. Schwartz, "Flux sampling is a powerful tool to study metabolism under changing environmental conditions," *NPJ systems biology and applications*, vol. 5, no. 1, pp. 1–8, 2019.

[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*, vol. 20, no. 1, pp. 1–18, 2019.

[6] O. Ponomarova and K. R. Patil, "Metabolic interactions in microbial communities: Untangling the gordian knot," *Current opinion in microbiology*, vol. 27, pp. 37–44, 2015.

[7] A. Chalkis, V. Fisikopoulos, E. Tsigaridas, and H. Zafeiropoulos, "Geometric algorithms for sampling the flux space of metabolic networks," *arXiv preprint arXiv:2012.05503*, 2020.

[8] S. P. Reise and N. G. Waller, "Item response theory and clinical measurement," *Annual review of clinical psychology*, vol. 5, pp. 27–48, 2009.

[9] P. G. Falkowski, T. Fenchel, and E. F. Delong, "The microbial engines that drive earth's biogeochemical cycles," *science*, vol. 320, no. 5879, pp. 1034–1039, 2008.

[10] G. J. Da Silva and S. Domingues, "We are never alone: Living with the human microbiota," *Front Young Minds*, vol. 5, p. 35, 2017.

[11] L. Tang, "Microbial interactions," *Nature methods*, vol. 16, no. 1, pp. 19–19, 2019.

[12] W. Wade, "Unculturable bacteria—the uncharacterized organisms that cause oral infections," *Journal of the Royal Society of Medicine*, vol. 95, no. 2, pp. 81–83, 2002.

[13] I. Thiele and B. Ø. Palsson, "A protocol for generating a high-quality genome-scale metabolic reconstruction," *Nature protocols*, vol. 5, no. 1, p. 93, 2010.

[14] D. Machado, S. Andrejev, M. Tramontano, and K. R. Patil, "Fast automated reconstruction of genome-scale metabolic models for microbial species and communities," *Nucleic acids research*, vol. 46, no. 15, pp. 7542–7553, 2018.

[15] A. Zelezniak, S. Andrejev, O. Ponomarova, D. R. Mende, P. Bork, and K. R. Patil, "Metabolic dependencies drive species co-occurrence in diverse microbial communities," *Proceedings of the National Academy of Sciences*, vol. 112, no. 20, pp. 6449–6454, 2015.

[16] E. E. Muller, K. Faust, S. Widder, M. Herold, S. M. Arbas, and P. Wilmes, "Using metabolic networks to resolve ecological properties of microbiomes," *Current Opinion in Systems Biology*, vol. 8, pp. 73–80, 2018.

[17] V. Vieira, P. Maia, M. Rocha, and I. Rocha, "Comparison of pathway analysis and constraint-based methods for cell factory design," *BMC bioinformatics*, vol. 20, no. 1, pp. 1–15, 2019.

[18] D. H. Parks, M. Chuvochina, D. W. Waite, *et al.*, "A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life," *Nature biotechnology*, vol. 36, no. 10, pp. 996–1004, 2018.

[19] D. H. Parks, M. Chuvochina, P.-A. Chaumeil, C. Rinke, A. J. Mussig, and P. Hugenholtz, "A complete domain-to-species taxonomy for bacteria and archaea," *Nature biotechnology*, vol. 38, no. 9, pp. 1079–1086, 2020.

[20] M. Aite, M. Chevallier, C. Frioux, *et al.*, "Traceability, reproducibility and wiki-exploration for "à-la-carte" reconstructions of genome-scale metabolic models," *PLoS computational biology*, vol. 14, no. 5, e1006146, 2018.

[21] A. M. Feist and B. O. Palsson, "The biomass objective function," *Current opinion in microbiology*, vol. 13, no. 3, pp. 344–349, 2010.

[22] R. Schuetz, L. Kuepfer, and U. Sauer, "Systematic evaluation of objective functions for predicting intracellular fluxes in escherichia coli," *Molecular systems biology*, vol. 3, no. 1, p. 119, 2007.

[23] J. Cai, T. Tan, and S. Joshua Chan, "Predicting nash equilibria for microbial metabolic interactions," *Bioinformatics*, 2020.