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 12, 2021

Contents

1 Synopsis

2 The Project

Microbial communities populate most environments on earth; from the seafloor to the human gut, they literay live everywhere [7]. 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 [5]. At the same time, *the human body is inhabited by millions of tiny living organisms* having a fundamental part in keeping us healty[4]. However, up to nowm scientists are able to cultivate approximately 1% of known Bacteria [8]. Since the growth of various bacteria depends on their interactions with others [9], 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 [6]. 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 [6]. 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 [**thiele2010protocol**]. However, over the last few years, automatic reconstruction approaches for building genome-scale metabolic models [**machado2018fast**] of relatively high quality have been developed.

The study of metabolic dependencies to infer interspecies microbial interactions has been estrablished the last years [11]. Constrained-based modelling approaches, suchas 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 [10,11].

The aim of this project is to

I propose the building of a pipeline to

and thus climate [1] climate [5, 6]

Over the last 20 years,

Their composition is thought to be largely shaped by interspecies competition for the available resources, but cooperative interactions, such as metabolite exchanges, have also been implicated in community assembly.

The prevalence of metabolic interactions in microbial communities, however, has remained largely unknown. Here, we systematically survey, by using a genome-scale metabolic modeling approach, the extent of resource competition and metabolic exchanges in over 800 communities. We find that, despite marked resource competition at the level of whole assemblies, microbial communities harbor metabolically interdependent groups that recur across diverse habitats. By enumerating flux-balanced metabolic exchanges in these co-occurring subcommunities we also predict the likely exchanged metabolites, such as amino acids and sugars, that can promote group survival under nutritionally challenging conditions. Our results highlight metabolic dependencies as a major driver of species co-occurrence and hint at cooperative groups as recurring modules of microbial community architecture.

Although metabolic interactions have long been implicated in the assembly of microbial communities, their general prevalence has remained largely unknown

Our results highlight metabolic dependencies as a major driver of species co-occurrence [11].

2.1 Methodology

Microbial interactions play a fundamental role in deciphering the underlying mechanisms that govern ecosystem functioning. Metabolic interactions allow to determine such interactions. There are several ways of identifying metabolic interactions. You can look at 1. pathway complementarity and overlap, but you could also look at 2. seed nodes, as in NetCooperate and NetCmpt (Borenstein and Freilich labs) and also at 3. metabolite exchanges predicted by metabolic modeling (which you plan to do as well).

Metabolic pathway complementarity between the species (nodes) of each association (edge) of a network will support the identification of cooperative potential.

In addition, uniform sampling in the flux spaces of all paired associations will strengthen the actual associations, demonstrating robustness on exchange/competition for metabolites.

Automatic genome scale metabolic networks reconstruction tools, such as AuReMe (18) and CarveMe (19), have been recently developed. Once such a metabolic network exists, metabolic flux sampling can be used to study its global features, allowing the study of properties of certain components of the whole network, and deduce significant biological insights (20,21). microbetag will search for potential matches between both taxa in each of the network edges and the 30.000 genomes available in GTDB. For those pairs of taxa that a match is found, microbetag will implement genome-scale metabolic reconstruction tools to build their metabolic networks.

Aim of this project is to get the flux distributions of the processes related to the pathways both taxa of an association are involved with.

Then, microbetag will check whether the taxa of the association under study exchange or compete for metabolites in a robust way.

I suggest to implement a pipelenine

1. get/provide a pair of .fasta files (genomes) 2. run `AuReMe` or `CarveMe` to get metabolic networks 3.

2.2 Benefits to the Community

Benefit to the `volesti` as an open source library

Benefit to the biologist

Build synthetic communities, cultivate the uncultivated.

3 Deliverables

The tasks of this project will be split in the two main periods of the project; i.e. the Community Bonding period between April 27 and May 18 (3 weeks) and the Coding period between May 18 and August (10 weeks). At the end of the coding period, all the implementations described in the Tasks chapter will have been completed.

`VolEsti` 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. A number of extra rounding algorithms will be added on the `VolEsti` package in the next few weeks; thus, working with members

of the `VolEsti` group would allow us for a simultaneous development of the source code and the interface, without unexpected delays. 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 `VolEsti` Python interface will be delivered. The `VolEsti` package is under the LGPL-3 license, while the cobrapy package is under the GPL and cobra-toolbox under the GNU General Public License version 3.0; all open-source. Thus, I intend to give the `VolEsti` Python interface the LGPL-3 license too.

3.1 Bonding period (May 17 - June 7)

During the bonding period everything that I need to know regarding the implementation of T1 will be fully covered. Thus, the source code of VolEsti will be thoroughly studied and a step-by-step guide for building the Python interface will be delivered by the end of the second week of the bonding period. In its last week, everything that might be needed to be able to start coding by the end of the bonding period, will be taken care of. Thus, I will build a branch of the current VolEsti project at my GitHub repository [18] and a blog to report my progress periodically will be also created in my web page*. The first two weeks will require a strong effort from me in comprehending the architecture of the VolEsti source code. To address this, 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.

3.2 Coding period (June 7 - August 16)

This 13 weeks period will be split as shown in the following Gantt chart (Figure 3). More specifically:

• 1st - 4th week

In the first four weeks probably the most time-consuming task will be implemented. I will create the wrapper for the volesti package as described in T.1. Meanwhile, I will catalogue all the useful open-cobra visual components and I will include them in this first Python interface of volesti. Tests that the wrapper is operating well will be implemented. A first documentation of the Python interface will also be made.

• 5th - 7th week

In those four weeks, I am going to implement the cplex buildings; first for the case of volesti source code (C++ interface of the cplex package). Tests will be implemented as well.

• 8th - 9th week

And then, the 2 following weeks for the case of the Python wrapper built on T.1. A thorough report about the coding blocks regarding the "cplex" building will be delivered. 10th week By the time being, the GeomScale group will have added the extra sampling and rounding methods on the volesti source code. Thus, this week I will build those methods based on the cplex package as well.

• 11th - 12th week

In those two weeks, the final wrapper of the volesti Python interface will be implemented. This version will include the cplex-based methods of volesti both those that are currently included and those that are about to be added by the GeomScale group and the open-cobra visual components mentioned.

• 13th week

In the last week, the final volesti Python interface will be applied to a series of data as described in T.4. Likewise, the cobrapy library. Their performance will be reported. Finally, a complete documentation for the volesti Python interface will be delivered. This time schedule is a worst-case scenario. In case

that something is not going according to the plan, then the extra time that we attempt to earn from the bonding period, will come up as a backup plan.

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 [3]. 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. 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). PEMA [10], a pipeline for the analysis of metabarcoding data, was my first coding project; PEMA has now been published and selected from LifeWatch - ERIC a European Researh Infrustructure, to support . PREGO and DARN are ongoing bioinformatics projects in the framework of my PhD.

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.

The last year, I have been working with the GeomScale group on the `VolEstipy` project and sampling on the flux space of metabolic networks

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

Metabolic interactions [2] is now the next big thing in systems biology. In such a study, the changing phenotype of an ecosystem, as derived by the combination of phenotypes of all the ecosystem's individual entities, under different scenarios of constraint values could be predicted. The impact of such studies would be more than significant, especially on a time when nature suffers the most, mostly due to anthropogenic activities. Indicatively possible datasets such a study could be conducted on: "Impact of exogenous nitrogen on the cyanobacterial abundance and community in oil-contaminated sediment: A microcosm study" from Wang et al. [19] (peer-reviewed, no public data, email request) "Discovery of functional gene markers of bacteria for monitoring hydrocarbon pollution in the marine environment - a metatranscriptomics approach" from Knapik et al. [20] (not peer-reviewed, data not published yet) In their study [21], Gossart et al. describe the research framework of such studies we suggest to apply the aforementioned methods .

Coding has been a great part of my everyday routine through the recent years. The different needs of the different types of analysis in biology force you to deal with a great range of computing tasks, such as choosing the proper programming language for each issue, working with High Performance Computing environments, finding ways to make your code easy-to-use, easy-to-distribute and flexible at the same time. It would be both a profit and an enjoyment for me to be part of the Google Summer of Code and upgrade my so-far programming skills, particularly as `VolEsti` functions have been developed in C++, meaning that I will have to deal with new personal coding challenges.

References

[1] Kevin R Arrigo. "Marine microorganisms and global nutrient cycles". In: *Nature* 437.7057 (2005), pp. 349–355.

[2] Jingyi Cai, Tianwei Tan, and SH Joshua Chan. "Predicting Nash equilibria for microbial metabolic interactions". In: *Bioinformatics* (2020).

[3] Apostolos Chalkis et al. "Geometric algorithms for sampling the flux space of metabolic networks". In: *arXiv preprint arXiv:2012.05503* (2020).

[4] Gabriela Jorge Da Silva and Sara Domingues. "We are never alone: living with the human microbiota". In: *Front Young Minds* 5 (2017), p. 35.

[5] Paul G Falkowski, Tom Fenchel, and Edward F Delong. "The microbial engines that drive Earth's biogeochemical cycles". In: *science* 320.5879 (2008), pp. 1034–1039.

[6] Bernhard Palsson. *Systems biology*. Cambridge university press, 2015.

[7] Steven P Reise and Niels G Waller. "Item response theory and clinical measurement". In: *Annual review of clinical psychology* 5 (2009), pp. 27–48.

[8] Lei Tang. "Microbial interactions". In: *Nature methods* 16.1 (2019), pp. 19–19.

[9] William Wade. "Unculturable bacteria—the uncharacterized organisms that cause oral infections". In: *Journal of the Royal Society of Medicine* 95.2 (2002), pp. 81–83.

[10] Haris Zafeiropoulos et al. "PEMA: a flexible Pipeline for Environmental DNA Metabarcoding Analysis of the 16S/18S ribosomal RNA, ITS, and COI marker genes". In: *GigaScience* 9.3 (2020), giaa022.

[11] Aleksej Zelezniak et al. "Metabolic dependencies drive species co-occurrence in diverse microbial communities". In: *Proceedings of the National Academy of Sciences* 112.20 (2015), pp. 6449–6454.