ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ
UNIVERSITY OF CRETE

Microbial communities through the lens of high throughput sequencing, data integration and metabolic networks analysis

Haris Zafeiropoulos

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Science (PhD) in Biology

Promotors:
Prof. Emmanouil Ladoukakis
Dr Evangelos Pafilis
Dr Christoforos Nikolaou

Academic year 2021 – 2022

Members of the examination committee
&
reading committee

Prof. Emmanouil Ladoukakis
Univeristy of Crete
Biology Department

Dr. Evangelos Pafilis
Hellenic Centre for Marine Research
Institute of Marine Biology, Biotechnology and Aquaculture

Dr. Christoforos Nikolaou
Biomedical Sciences Research Center "Alexander Fleming"
Institute of Bioinnovation

Suggested members to contact with to join the 7-member committee

Prof. Konstadia (Dina) Lika
Univeristy of Crete
Biology Department

Prof. Panagiotis Sarris
University of Crete
Department of Biology

Dr. Jens Carlsson
University College Dublin
School of Biology and Environmental Science/Earth Institute

Prof. Karoline Faust
KU Leuven
Department of Microbiology and Immunology, Rega Institute

Contents

Abstract

Microbial communities are a cornerstone for most ecosystem types. To elucidate the mecahinsms governing such assemblages, it is fundamental to identify the taxa present (*who*) and the processeses that occur (*what*) in the various environments (*where*). Thanks to a series of technological breakthroughs vast amounts of information/data from all the various levels of the biological organization have been accumulated over the last decades. In this context, microbial ecology studies are now relying on bioinformatics methods and analyses. Therefore, a great number of challenges both from the biologist- and the computer scientist point-of-view have arisen; one among the most emerging ones being: *"what shall we do with all these pieces of information?"*. The paradigm of Systems Biology addresses this challenge by moving from reductionism to more holistic approaches attempting to interpret how the properties of a system emerge.

Aim of this PhD was to enhance microbiome data analyses by developing software addressing on-going computational challenges on the study of microbial communities. On top of that, to exploit state-of-the-art methods to identify taxa, functions and microbial interactions in assemblages of various aquatic environmnets. To this end, a number of publically available datasets were used while a swamp from the Karpathos island (Greece), was chosen as a study case for the described framework.

Environmental DNA and metabarcoding have been widely used to estimate the biodiversity (the *who*) and the structure of communities. Vast amount of sequencing data targeting certain marker genes depending the taxonomic group of interest become available thanks to High Throughput Sequencing technologies. However, the bioinformatics analysis of such data require multiple steps and parameter settings. Software workflow-oriented tools along with computing infustructures ease this need to a great extent and PEMA was developed to this end (Chapter **??**). Moreover, eDNA metabarcoding has limitations too. Cytochrome c oxidase subunit I (COI) marker gene is a commonly used marker gene, especially in studies targeting eukaryotic taxa. It is well known that in COI studies a great number of the OTUs/ASVs returned get no taxonomic hits. The presence of non-eukaryotic taxa with their simultaneous absence from the most commonly-used reference databases justify this phenomenon to a great extent. DARN makes use of a COI-oriented tree of life to provide further insight to such known unknown sequences (Chapter **??**).

Shotgun metagenomics provide further information regarding the processes that occur in a community (the *what*). Sediment and microbial mat samples as well as microbial aggregates from a hypersaline swamp in Tristomo bay (Karpathos, Greece) were analysed. Both amplicon (16S rRNA) and shotgun sequencing data were used to characterize the microbial structure of the communities and environmental parameters (e.g. salinity, oxygen concentration, granulometric composition) were measured at the sampling sites. Key functions supporting life in such environments were identified and metagenome-assembled genomes (MAGs) of major species found were built (Chapter **??**).

Amplicon and shotgun metagenomics approaches along with the rest of the omics technologies, have led to vast amount of data and metadata, recording the *who*, the *what* and the *where*. To enable optimal accessibility and usage of this information, a great number of databases, ontologies as well as community-standards have been developed. By exploiting data integration techniques to bring such bits of information together, as well as text mining methods to retrieve knowledge "hidden" among the billions of text lines in already published literature, the PREGO knowledge-base generates thousands of *what* - *where* - *who* potential associations (Section **??**).

The driving question though is *how* the different microbial taxa ascertain their endurance as part of a community. Metabolic interactions among the various taxa play a decisive role for the composition of such assemblages. Genome-scale metabolic networks (GEMs) enable the inference of such interactions. Random sampling on the flux space of such metabolic models, provides a representation of the flux values a model can get under various conditions. However, flux sampling is challenging from a computational point of view. To address such challenges, a Python library called dingo was developed using a Multiphase Monte Carlo Sampling algorithm (Chapter **??**). GEMs were built using the MAGs retrieved from the Tristomo swamp and metabolic interactions between them and their environment were investigated.

Similar to microbial communities, bioinformatics methods tend to build assemblages while "living" on your own is quite rare. The methods developed during this PhD project combined with state-of-the-art methods anticipate to build a framework that enables moving from the community to the species level and then back again to the one of the community.

Chapter 1

Short CV

Education

- **Doctor of Philosophy** (2018 – 2022), University of Crete, Biology Department
  Thesis: Microbial communities through the lens of high throughput sequencing, data integration and metabolic networks analysis
  Thesis conducted at IMBBC - HCMR

- **M.Sc. in Bioinformatics** (2016 – 2018), University of Crete, School of Medicine
  Thesis: eDNA metabarcoding for biodiversity assessment: Algorithm design and bioinformatics analysis pipeline implementation
  Thesis conducted at IMBBC - HCMR

- **B.Sc. in Biology** (2011 – 2016), National and Kapodistrian University of Athens, department of Biology
  Thesis: Morphology, morphometry and anatomy of species of the genus *Pseudamnicola* in Greece

Research projects - working Experience

- **A workflow for marine Genomic Observatories data analysis** (2021 - ongoing)
  Role: scientific responsiblse & developer
  This EOSC-Life funded project aims at developing a workflow for the analysis of EMBRC's Genomic Observatories (GOs) data, allowing researchers to deal better with this increasing amount of the data and make them more easily interpretable.

- **PREGO: Process, environment, organism (PREGO)** (2019 - 2021)
  Role: PhD candidate
  PREGO is a systems-biology approach to elucidate ecosystem function at the microbial dimension.

- **ELIXIR-GR** (2019 - 2021)
  Role: technical support

ELIXIR-GR is the Greek National Node of the ESFRI European RI ELIXIR, a distributed e-Infrastructure aiming at the construction of a sustainable European infrastructure for biological information.

- **RECONNECT** (2018 - 2020)
  Role: technical support
  RECONNECT is an Interreg V-B "Balkan-Mediterranean 2014-2020" project. It aims to develop strategies for sustainable management of Marine Protected Areas (MPAs) and Natura 2000 sites.

Publications

- **PREGO: A Literature and Data-Mining Resource to Associate Microorganisms, Biological Processes, and Environment Types.**
  Zafeiropoulos, H., Paragkamian S.[2], Stelios Ninidakis, Georgios A. Pavlopoulos, Lars Juhl Jensen, and Evangelos Pafilis. *Microorganisms* 10, no. 2 (2022): 293., DOI: 10.3390/microorganisms10020293

- **The Dark mAtteR iNvestigator (DARN) tool: getting to know the known unknowns in COI amplicon data**
  Zafeiropoulos H., Gargan L., Hintikka S., Pavloudi C., & Carlsson J. *Metabarcoding and Metagenomics*, 5, p.e69657, 2021, DOI: 10.3897/mbmg.5.69657

- **0s & 1s in marine molecular research: a regional HPC perspective.**
  Zafeiropoulos H., Gioti A., Ninidakis S., Potirakis A., ..., & Pafilis E. *GigaScience*, 9(3), p.giab053, 2021 DOI: 10.1093/gigascience/giab053

- **Geometric Algorithms for Sampling the Flux Space of Metabolic Networks**
  Chalkis, A., Fisikopoulos, V., Tsigaridas, E. & **Zafeiropoulos, H.** *37th International Symposium on Computational Geometry (SoCG 2021)*, 21:1–21:16, 189, 2021 DOI: 10.4230/LIPIcs.SoCG.2021.21

- **The Santorini Volcanic Complex as a Valuable Source of Enzymes for Bioenergy**
  Polymenakou, P.N., Nomikou, P., **Zafeiropoulos, H.**, Mandalakis, M., Anastasiou, T.I., Kilias, S., Kyrpides, N.C., Kotoulas, G. & Magoulas,A. *Energies*, 14(5), p.1414, 2021 DOI: 10.3390/en14051414

- **PEMA: a flexible Pipeline for Environmental DNA Metabarcoding Analysis of the 16S/18S ribosomal RNA, ITS, and COI marker genes**
  Zafeiropoulos, H., Viet, H.Q., Vasileiadou, K., Potirakis, A., Arvanitidis, C., Topalis, P., Pavloudi, C. & Pafilis, E. *GigaScience*, 9(3), p.giaa022, 2020 DOI: 10.1093/gigascience/giaa022

[2] ZH and PS contibuted equally in this study

In preparation

- dingo: a Python library for metabolic networks analysis
- Deciphering the functional potential of a hypersaline swamp microbial mat community

Awards

- **European Molecular Biology Organization Short-Term Fellowship** (2022)
  Project title: Exploiting data integration, text-mining and computational geometry to enhance microbial interactions inference from co-occurrence networks
  https://hariszaf.github.io/microbetag/
- **Mikrobiokosmos travel grant in memorium of Prof. Kostas Drainas** (2021)
- **Google Summer of Code** (2021)
  Project title: From DNA sequences to metabolic interactions: building a pipeline to extract key metabolic processes
  Report, GSoC archive
- **Federation of European Microbiological Societies Meeting Attendance Grant** (2020)
  for joining the *Metagenomics, Metatranscript- omics and multi 'omics for microbial community studies* Physalia course
- **Short Term Scientific Mission (STSM) - DNAqua-net COST action** (2019)
  Project title: A comparison of bioinformatic pipelines and sampling techniques to enable benchmarking of DNA metabarcoding
  Report
- **Best Poster Award @ Hellenic Bioinformatics conference** (2018)
  for *PEMA: a Pipeline for Environmental DNA Metabarcoding Analysis*

Selected presentations

- **Bioinformatics Open Source Conference - BOSC2021** (2021)
  dingo: A python library for metabolic networks sampling & analysis, video poster - video
- **1st DNAQUA International Conference** (2021)
  PEMA v2: addressing metabarcoding bioinformatics analysis challenges, oral talk - video
- **Federation of European Microbiological Societies - FEMS2020** (2020)
  "Mining literature and -omics (meta)data to associate microorganisms, biological processes and environment types" - video poster

- **PyData Global PyData2020**
  “Geometric and statistical methods in systems biology: the case of metabolic networks”, oral talk - video

- **8th International Barcode of Life Conference** - 2019
  "P.E.M.A.: a pipeline for environmental DNA metabarcoding analysis" (flashtalk)

Participation in proposal writing

- "Climate Change Metagenomic Record Index (CCMRI)" project: submitted at the 2nd Call for H.F.R.I Research Projects to Support Faculty Members & Researchers (June 2020). Approved for funding

- "A workflow for marine Genomic Observatories data analysis" project: submitted at the second Training Open Call of EOSC-Life (November 2020). Approved for funding

Contact

You may find more about me and my work on my personal website.
You can also find me on GitHub, Twitter and ResearchGate.
Email: haris-zaf@hcmr.gr, haris.zafr@gmail.com