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

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 mechanisms governing such assemblages, it is fundamental to identify the taxa present (*who*) and the processes 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 such state-of-the-art methods to study microbial assemblages in extreme environments. To this end, the Tristomo marsh in Karpathos island (Greece), was chosen as a study case.

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 as well as increase computing resources. Workflows along with computing infrastructures ease this need to a great extent; in this nontion, a Pipeline for environmental DNA Metabarcoding Analysis (PEMA) was developed (Chapter **??**). However, 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 derived Operational Taxonomic Unitss (OTUs) get no taxonomic hits. The presence of pseudogenes but also of non-eukaryotic taxa among the amplicon data, with the simultaneous absence of the latter from the most commonly-used reference databases justify this phenomenon to a great extent. To identify such cases the Dark mAtteR iNvestigator (DARN) software was developed; DARN makes use of a COI-oriented tree of life to provide further insight to such known unknown sequences (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 returns 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, especially as the dimension of a metabolic model increases. To address such challenges, a Python library called dingo was developed using a Multiphase Monte Carlo Sampling algorithm (Chapter 1).

Finally, sediment and microbial mat samples as well as microbial aggregates from a hypersaline marsh in Tristomo bay (Karpathos, Greece) were analyzed. 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) were measured at the sampling sites. Key functions supporting life in such environments were identified and metagenome-assembled genomes (MAGs) of novel species found were built (Chapter **??**).

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. Such a framework is described for the study of microbial interactions at real-world communities.

Περίληψη

Οι μικροβιακές κοινότητες αποτελούν ακρογωνιαίο λίθο για τους περισσότερους τύπους οικοσυστημάτων. Για να διευκρινιστούν οι μηχανισμοί που καθορίζουν τέτοιες κοινότητες είναι καθοριστικής σημασίας η αναγνώριση των τάξων που τις απαρτίζουν (ποιος) καθώς και των διεργασιών που πραγματοποιούνται (τι) στους διάφορους τύπους περιβαλλόντων (που). Χάρη σε μια σειρά τεχνολογικών επιτευγμάτων, ιδιαίτερα μεγάλες ποσότητες πληροφορίας/δεδομένων από όλα τα επίπεδα οργάνωσης της ζωής έχουν σωρευτεί τις τελευταίες δεκαετίες. Σε αυτό το πλαίσιο, οι μελέτες μικροβιακής οικολογίας είναι άρρηκτα συνδεδεμένες και βασίζονται σε βιοπληροφορικές μεθόδους και αναλύσεις. Ωστόσο, έχει προκύψει ένας σημαντικός αριθμός προκλήσεων τόσο από την βιολογική σκοπιά όσο και από αυτήν την επιστήμης υπολογιστών. Μεταξύ αυτών, καθοριστικό ερώτημα αποτελεί το τι μπορούμε να κάνουμε με όλα αυτά τα επιμέρους κομμάτια πληροφορίας·. Το παράδειγμα της Βιολογίας Συστημάτων απαντά σε αυτό το ερώτημα περνώντας από πιο αναγωγικές σε πιο ολιστικές προσεγγίσεις προσπαθώντας να ερμηνεύσει το πως προκύπτουν και συνδέονται οι ιδιότητες ενός συστήματος.

Στόχος αυτής της διδακτορικής διατριβής ήταν να ενισχύσει την ανάλυση δεδομένων από μικροβιώματα αναπτύσσοντας λογισμικά εργαλεία που να απαντούν σε τρέχουσσες υπολογιστικές προκλήσεις για την μελέτη μικροβιακών κοινοτήτων. Επιπλέον, να μελετήσει μικροβιακές κοινότητες σε ακραία περιβάλλοντα εφαρμόζοντας σύγχρονες μεθόδους για την αναγνώριση τάξων και διεργασιών. Για την επίτευξη αυτού του στόχου, το έλος Τριστόμου στο νησί της Καρπάθου, επιλέχθηκε ως περιοχή μελέτης.

Το περιβαλλοντικό DNA και η μέθοδος της μετακωδικοποίησης έχουν χρησιμοποιηθεί σημαντικά για την εκτίμηση της βιοποικιλότητας (ποιος) και τη δομή των κοινοτήτων. Σημαντικός αριθμός αλληλουχικών δεδομένων που στοχεύουν σε ορισμενα γονίδια δείκτες και που εξαρτόνται από τις ταξινομικές ομάδες στόχους, είναι διαθέσιμα χάρη στις τεχνικές αλληλούχισης υψηλής απόδοσης HTS. Ωστόσο, η βιοπληροφορική ανάλυση τέτοιων δεδομένων απαιτούν μεγάλο αριθμό βημάτων και παραμέτρων καθώς και σημαντικούς υπολογιστικούς πόρους. Οι ροές εργασιών σε συνδυασμό με υπολογιστικές υποδομές μπορούν να απαντήσουν σε αυτές τις απαιτήσεις σε σημαντικό βαθμό. Σε αυτό το πλαίσιο αναπτύχθηκε η ροή εργασίας PEMA με στόχο την ανάλυση δεδομένων μετακωδικοποίησης από περιβαλλοντικό DNA. Κεφάλαιο **??**. Ωστόσο, η μέθοδος μετακωδικοποίησης χαρακτηρίζεται από σειρά περιορισμών. Η υπομονάδα I της κυτοχρωμικής οξειδάσης c (COI), αποτελεί έναν δείκτη που χρησιμοποιείται ευρέως, ειδικά στην περίπτωση ευκαρυωτικών τάξων - στόχων. Είναι γνωστό πως σε μελέτες όπου ο δείκτης αυτός χρησιμοποιείται, ένας μεγάλος αριθμός των λειτουργικών ταξινομικών μονάδων (OTUs) που προκύπτουν, δεν καταφέρνουν να ταυτοποιηθούν. Η παρουσία τοσο

ψευδογονιδίων όσο όμως και μη-ευκαρυωτικών τάξων ανάμεσα σε τέτοια αλληλουχικά δεδομένα, με την ταυτόχρονη απουσία των τελευταίων από τις βάσεις αναφοράς, εξηγεί την μη ταυτοποίησή τους σε σημαντικό βαθμό. Για την αναγνώριση τέτοιων περιπτώσεων, αναπτύχθηκε το υπολογιστικό εργαλείο DARN το οποίο αξιοποιεί ένα φυλογενετικό δέντρο που καλύπτει και τις 3 επικράτειες του δέντρου της ζωής, βασισμένο σε αλληλουχίες του δείκτη Κεφάλαιο COI, Κεφάλαιο **??**.

Μέθοδοι γονιδίων δεικτών και μεταγονιδιωματικής καθώς και το σύνολο των μεθόδων αλληλούχισης υψηλή απόδοσης, έχουν οδηγήσει στην σώρευση σημαντικά μεγάλου αριθμού δεδομένων και μεταδεδομένων καταγράφοντας τάξα και διεργασίες σε σειρά τύπους περιβαλλόντων. Για να επιτρέψουν την βέλτιστη προσβασιμότητα και αξιοποίηση αυτής της πληροφορίας, έχουν δημιουργηθεί σειρά βάσεων δεδομένων, οντολογιών αλλά και προτύπων - κανόνων για να ακολουθεί η κοινότητα για την καταχώρηση τους. Αξιοποιώντας μεθόδους ενσωμάτωσης/ολοκλήρωσης δεδομένων **data integration** για την εύρεση των διάφορων κομματιών πληροφορίας και την συσχέτισή τους, καθώς και τεχνικών εξόρυξης κειμένου **text mining** για την ανάκτηση γνώσης από το σύνολος της δημόσια διαθέσιμης βιβλιογραφίας αναπτύχθηκε η βάση-γνώσης PREGO, Κεφάλαιο **??**. η οποία επιστρέφει χιλιάδες σχέσεις μεταξύ τάξων, περιβαλλόντων και διεργασιών.

Καθοριστικό ερώτημα ωστόσο σε οτι αφορά τις μικροβιακές κοινότητες, αποτελεί το ¨πώς' τα διάφορα μικροβιακά τάξα εξασφαλίζουν την θέση τους ως μέλη της κοινότητας. Μεταβολικές αλληλεπιδράσεις μεταξύ των διάφορων τάξων παίζουν καθοριστικό ρόλο για την συγκρότηση τέτοιων κοινοτήτων. Μεταβολικά δίκτυα στην κλίμακα του γονιδιώματος (GEMs) επιτρέπουν την αναγνώριση τέτοιων αλληλεπιδράσεων. Η τυχαία δειγματοληψία στον χώρο που ορίζεται από τις πιθανές τιμές που μπορεί να πάρουν οι ροές των αντιδράσεων (**flux sampling**) επιτρέπει την αναπαράσταση των τιμών που μπορεί να λάβουν αυτές οι ροές κάτω από συγκεκριμένες συνθήκες. Ωστόσο η μέθοδος **flux sampling** είναι ιδιαίτερα απαιτητική από υπολογιστική σκοπιά, ιδιαίτερα όσο η διάσταση του μεταβολικού μοντέλου αυξάνει. Για τον σκοπό αυτό αναπτύχθηκε η βιβλιοθήκη **dingo** η οποία κάνει χρήση ενός πολυφασικού αλγορίθμου **Monte Carlo**, Κεφάλαιο 1.

Τέλος, αναλύθηκαν δείγματα ιζήματος από το έλος Τριστόμου Καρπάθου, καθώς επίσης δείγματα από μικροβιακούς τάπητες **mat** και από μικροβιακά συσσωματώματα (**aggregates**). Για τον σκοπό αυτό, χρησιμοποιήθηκε τόσο η μέθοδος μετακωδικοποίησης με γονίδιο-δείκτη το 16S όσο και η μέθοδος μεταγονιδιωματικής **shotgun**. Επίσης μετρήθηκαν περιβαλλοντικές παράμετροι (όπως αλατότητα, συγκέντρωση οξυγόνου). Βασικές λειτουργίες που υποστηρίζουν τη ζωή σε τέτοιες συνθήκες εντοπίστηκαν ενώ ακόμη γονιδιώματα τάξων που εντοπίζονται για πρώτη φορά ανασκευαάστηκαν από τις αλληλουχίες του μεταγονιδιώματος (MAGs), Κεφάλαιο **??**.

Όπως συμβαίνει και στις μικροβιακές κοινότητες, οι βιοπληροφορικές μέθοδοι σπάνια στέκουν απομονωμένες, αντίθετα τείνουν να συγκροτούν κι αυτές τις δικές τους ¨κοινότητες¨. Οι μέθοδοι που αναπτύχθηκαν στα πλαίσια αυτής της διατριβής επιδιώκουν να συγκροτήσουν ένα πλαίσιο μελέτης από το επίπεδο της κοινότητας σε αυτό του είδους και από εκεί, πίσω πάλι στην κοινότητα. Ένα τέτοιο πλαίσιο αναλύεται για την μελέτη μικροβιακών αλληλεπιδράσεων.

List of Figures and Tables

List of Figures

List of Tables

List of Abbreviations and Symbols

Abbreviations

COI	Cytochrome c oxidase subunit I
ITS	Internal Transcribed Spacer
NGS	Next Generation Sequencing
eDNA	environmental DeoxyriboNucleic Acid
OTU	Operational Taxonomic Unit
ASV	Amplicon Sequence Variant
HPC	High Performance Computing
MCMC	Markov Chain Monte Carlo
MMCS	Multiphase Monte Carlo Sampling
PREGO	PRocess Environment OrGanism
PEMA	Pipeline for Environmental DNA Metabarcoding Analysis
DARN	Dark mAtteR iNvestigator
PSRF	Potential Scale Reduction Factor
ESS	Effective Sample Size

Symbols

$\mathbb{R}$	Set of real numbers
$\mathcal{O}$	Algorithm complexity
$\tilde{\mathcal{O}}$	Algorithm complexity ignoring polylogarithmic factors
$\mathbb{E}$	Expected Value operator
$\mathcal{U}$	utility function....
ℓ	a ray
P	a polytope
τ	a trajectory
∂P	boundary of the P polytope
v	a flux vector
v_i	flux value of the reaction i
W	walk length
ρ	number of reflections

Chapter 1

Software development to establish metabolic flux sampling approaches at the community level

1.1 A New MCMC Algorithm for Sampling the Flux Space of Metabolic Networks

Citation: Chalkis, A., Fisikopoulos, V., Tsigaridas E. and Zafeiropoulos H. Geometric Algorithms for Sampling the Flux Space of Metabolic Networks. 37th International Symposium on Computational Geometry (SoCG 2021)
DOI: 10.4230/LIPIcs.SoCG.2021.21 [1]

1.1.1 Abstract

Systems Biology is a fundamental field and paradigm that introduces a new era in Biology. The crux of its functionality and usefulness relies on metabolic networks that model the reactions occurring inside an organism and provide the means to understand the underlying mechanisms that govern biological systems. Even more, metabolic networks have a broader impact that ranges from resolution of ecosystems to personalized medicine.

The analysis of metabolic networks is a computational geometry oriented field as one of the main operations they depend on is sampling uniformly points from polytopes; the latter provides a representation of the steady states of the metabolic networks. However, the polytopes that result from biological data are of very high dimension (to the order of thousands) and in most, if not all, the cases are considerably skinny. Therefore, to perform uniform random sampling efficiently in this setting, we need a novel algorithmic and computational framework specially tailored for the properties of metabolic networks.

[1] Authors' names are in alphabetical order. Here a modified version of the published version is presented in terms of relevance, coherence and formatting. The dingo Python library, a wrapper of the C++ code of the MMCS algorithm, is available at https://github.com/geomscale/dingo and a relative publication is under preparation. Proofs for the lemmas mentioned and parameter tuning can be found in the original publication.

We present a complete software framework to handle sampling in metabolic networks. Its backbone is a Multiphase Monte Carlo Sampling (MMCS) algorithm that unifies rounding and sampling in one pass, obtaining both upon termination. It exploits an improved variant of the Billiard Walk that enjoys faster arithmetic complexity per step. We demonstrate the efficiency of our approach by performing extensive experiments on various metabolic networks. Notably, sampling on the most complicated human metabolic network accessible today, Recon3D, corresponding to a polytope of dimension 5 335, took less than 30 hours. To our knowledge, that is out of reach for existing software.

1.1.2 Introduction

The field of Systems Biology

Systems Biology establishes a scientific approach and a paradigm. As a research approach, it is the qualitative and quantitative study of the systemic properties of a biological entity along with their ever evolving interactions [Klipp et al., 2016, Kohl et al., 2010]. By combining experimental studies with mathematical modeling it analyzes the function and the behavior of biological systems. In this setting, we model the interactions between the components of a system to shed light on the system's *raison d'être* and to decipher its underlying mechanisms in terms of evolution, development, and physiology [Ideker et al., 2001].

Initially, Systems Biology emerged as a need. New technologies in Biology accumulate vast amounts of information/data from different levels of the biological organization, i.e., genome, transcriptome, proteome, metabolome [Quinn et al., 2016]. This leads to the emerging question *"what shall we do with all these pieces of information"?* The answer, if we consider Systems Biology as a paradigm, is to move away from reductionism, still the main conceptual approach in biological research, and adopt holistic approaches for interpreting how a system's properties emerge [Noble, 2008]. The following diagram provides a first, rough, mathematical formalization of this approach.

$$\textit{components} \rightarrow \textit{networks} \rightarrow \textit{in silico models} \rightarrow \textit{phenotype}$$ [Palsson, 2015].

Systems Biology expands in all the different levels of living entities, from the molecular, to the organismal and ecological level. The notion that penetrates all levels horizontally is *metabolism*; the process that modifies molecules and maintains the living state of a cell or an organism through a set of chemical reactions [Schramski et al., 2015]. The reactions begin with a particular molecule which they convert into some other molecule(s), while they are catalyzed by enzymes in a key-lock relationship. We call the quantitative relationships between the components of a reaction *stoichiometry*. Linked reactions, where the product of the first acts as the substrate for the next, build up metabolic pathways. Each pathway is responsible for a certain function. We can link together the aggregation of all the pathways that take place in an organism (and their corresponding reactions) and represent them mathematically using the reactions' stoichiometry. Therefore, at the species level, metabolism is a network of its metabolic pathways and we call these representations *metabolic networks*.

From metabolism to computational geometry

The complete reconstruction of the metabolic network of an organism is a challenging, time consuming, and computationally intensive task; especially for species of high level of complexity such as *Homo sapiens*. Even though sequencing the complete genome of a species is becoming a trivial task providing us with quality insight, manual curation is still mandatory and large groups of researchers need to spend a great amount of time to build such models [Thiele and Palsson, 2010]. However, over the last few years, automatic reconstruction approaches for building genome-scale metabolic models [Machado et al., 2018] of relatively high quality have been developed. Either way, we can now obtain the metabolic network of a bacterial species (single cell species) of a tissue and even the complete metabolic network of a mammal. Biologists are also moving towards obtaining such networks for all the species present in a microbial community. This will allow us to further investigate the dynamics, the functional profile, and the inter-species reactions that occur. Using the stoichiometry of each reaction, which is always the same in the various species, we convert the metabolic network of an organism to a mathematical model. Thus, the metabolic network becomes an *in silico* model of the knowledge it represents.

In metabolic networks analysis mass and energy are considered to be conserved [Palsson, 2009]. As many homeostatic states, that is steady internal conditions [Shishvan et al., 2018], are close to steady states (where the production rate of each metabolite equals its consumption rate [Cakmak et al., 2012]) we commonly use the latter in metabolic networks analysis.

Stoichiometric coefficients are the number of molecules a biochemical reaction consumes and produces. The coefficients of all the reactions in a network, with m metabolites and n reactions ($m < n$), form the *stoichiometric matrix* $S \in \mathbb{R}^{m \times n}$ [Palsson, 2015]. The nullspace of S corresponds to the steady states of the network:

$$S \cdot x = 0, \tag{1.1}$$

where $x \in \mathbb{R}^n$ is the *flux vector* that contains the fluxes of each chemical reaction of the network. Flux is the rate of turnover of molecules through a metabolic pathway.

All physical variables are finite, therefore the flux (and the concentration) is bounded [Palsson, 2015]; that is for each coordinate x_i of the x, there are $2n$ constants $x_{ub,i}$ and $x_{lb,i}$ such that $x_{lb,i} \leq x_i \leq x_{ub,i}$, for $i \in [n]$. We derive the constraints from explicit experimental information. In cases where there is no such information, reactions are left unconstrained by setting arbitrary large values to their corresponding bounds according to their reversibility properties; i.e., if a reaction is reversible then its flux might be negative as well [Lularevic et al., 2019]. The constraints define a n-dimensional box containing both the steady and the dynamic states of the system. If we intersect that box with the nullspace of S, then we define a polytope that encodes all the possible steady states and their flux distributions [Palsson, 2015]. We call it the steady-state *flux space*. Figure 1.1 illustrates the complete workflow from building a metabolic network to the computation of a flux distribution.

Using the polytopal representation, a commonly used method for the analysis of a metabolic network is Flux Balance Analysis (FBA) [Orth et al., 2010]. FBA identifies a single

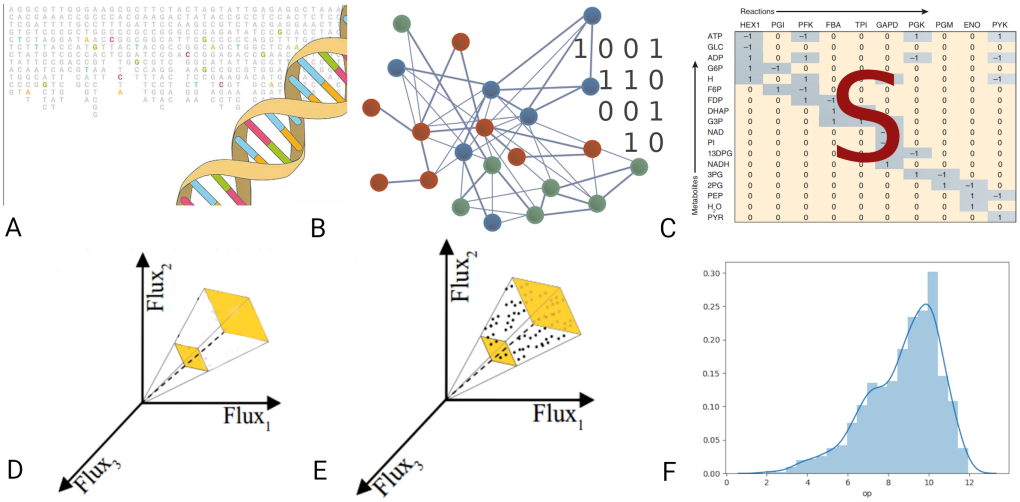

FIGURE 1.1: From DNA sequences to distributions of metabolic fluxes. (A) The genes of an organism provide us with the enzymes that it can potentially produce. Enzymes are like a blueprint for the reactions they can catalyze. (B) Using the enzymes we identify the reactions in the organism. (C) We construct the stoichiometric matrix of the metabolic model. (D) We consider the flux space under different conditions (e.g., steady states); they correspond to polytopes containing flux vectors addressing these conditions. (E) We sample from polytopes that are typically skinny and of high dimension. (F) The distribution of the flux of a reaction provides great insights to biologists.

optimal flux distribution by optimizing a linear objective function over a polytope [Orth et al., 2010]. Unfortunately, this is a *biased* method because it depends on the selection of the objective function. To study the global features of a metabolic network we need *unbiased methods*. To obtain an accurate picture of the whole solution space we exploit sampling techniques [Schellenberger and Palsson, 2009]. If collect a sufficient number of points uniformly distributed in the interior of the polytope, then the biologists can study the properties of certain components of the whole network and deduce significant biological insights [Palsson, 2015]. Therefore, efficient sampling tools are of great importance.

Make a comment on the bad ones from the fig. according to the IEEE review

Metabolic networks through the lens of random sampling

Efficient uniform random sampling on polytopes resulting from metabolic networks is a very challenging task both from the theoretical (algorithmic) and the engineering (implementation) point of view. First, the dimension of the polytopes is of the order of certain thousands. This requires, for example, advanced engineering techniques to cope with memory requirements and to perform linear algebra operations with large matrices; e.g., in Recon3D [Brunk et al., 2018] we compute the null space of a 8 399 × 13 543 matrix. Second, the polytopes are rather skinny (Section 1.1.5); this makes it harder for sampling algorithms to move in the interior of polytopes and calls for novel practical techniques to

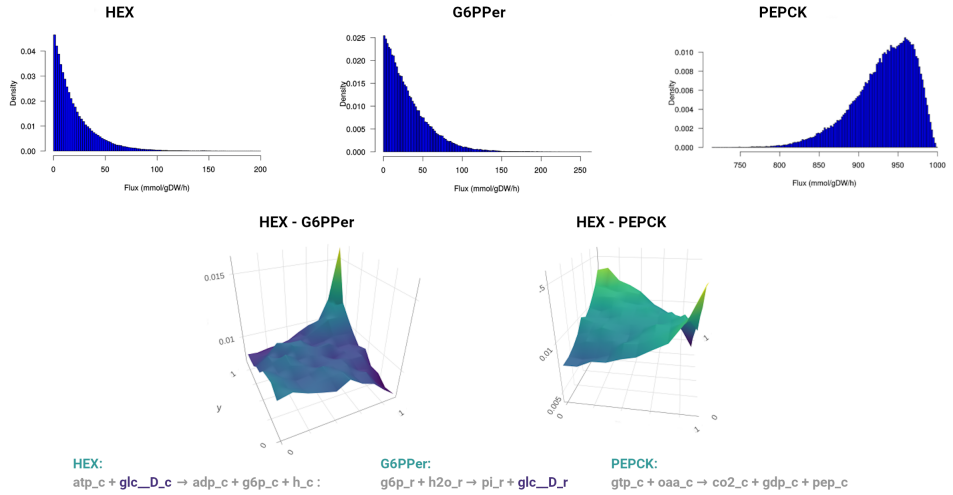

FIGURE 1.2: Flux distributions in the most recent human metabolic network Recon3D [Brunk et al., 2018]. We estimate the flux distributions of the reactions catalyzed by the enzymes Hexokinase (D-Glucose:ATP) (HEX), Glucose-6-Phosphate Phosphatase, Edoplasmic Reticular (G6PPer) and Phosphoenolpyruvate carboxykinase (GTP) (PEPCK). As we sample steady states, the production rate of $glc\_\_D\_c$ should be equal to its consumption rate. Thus, in the corresponding copula, we see a positive dependency between HEX, i.e., the reaction that consumes $glc\_\_D\_c$ and G6PPer, that produces it. Furthermore, the PEPCK reaction operates when there is no $glc\_\_D\_c$ available and does not operate when the latter is present. Thus, in their copula we observe a negative dependency between HEX and PEPCK. A copula is a bivariate probability distribution for which the marginal probability distribution of each variable is uniform. It implies a positive dependency when the mass of the distribution concentrates along the up-diagonal (HEX - G6PPer) and a negative dependency when the mass is concentrated along the down-diagonal (HEX - PEPCK). The bottom line contains the reactions and their stoichiometry.

sample.

There is extended on-going research concerning advanced algorithms and implementations for sampling metabolic networks over the last decades. Markov Chain Monte Carlo algorithms such as Hit-and-Run (HR) [Smith, 1984] have been widely used to address the challenges of sampling. Two variants of HR are the non-Markovian Artificial Centering Hit-and-Run (ACHR) [Kaufman and Smith, 1998] that has been widely used in sampling metabolic models, e.g., [Saa and Nielsen, 2016], and Coordinate Hit-and-Run with Rounding (CHRR) [Haraldsdóttir et al., 2017]. The latter is part of the cobra toolbox [Heirendt et al., 2019], the most commonly used software package for the analysis of metabolic networks. CHRR enables sampling from complex metabolic networks corresponding to the highest dimensional polytopes so far. There are also stochastic formulations where the inclusion of experimental noise in the model makes it more compatible with the stochastic nature of biological networks [MacGillivray et al., 2017]. The recent study in [Fallahi et al., 2020] offers an overview as well as an experimental comparison of the currently available samplers.

These implementations played a crucial role in actually performing in practice uniform sampling from the flux space. However, they are currently limited to handle polytopes of dimension say ≤ 2500 [Fallahi et al., 2020, Haraldsdóttir et al., 2017]. This is also the order of magnitude of the most complicated, so far, metabolic network model built, Recon3D [Brunk et al., 2018]. By including 13543 metabolic reactions and involving 4140 unique metabolites, Recon3D provides a representation of the 17% of the functionally of annotated human genes. To our knowledge, there is no method that can efficiently handle sampling from the flux space of Recon3D.

Apparently, the dimension of the polytopes will keep rising and not only for the ones corresponding to human metabolic networks. Metabolism governs systems biology at all its levels, including the one of the community. Thus, we are not only interested in sampling a sole metabolic network, even if it has the challenges of the human. Sampling in polytopes associated to network of networks are the next big thing in metabolic networks analysis and in Systems Biology [Bernstein et al., 2019, Perez-Garcia et al., 2016].

Regarding the sampling process, from the theoretical point of view, we are interested in the convergence time, or *mixing time*, of the Markov Chain, or geometric *random walk*, to the target distribution. Given a d-dimensional polytope P, the mixing time of several geometric random walks (e.g., HR or Ball Walk) grows quadratically with respect to the sandwiching ratio R/r of the polytope [Lovász et al., 1997, Lovász and Vempala, 2006]. Here r and R are the radii of the smallest and largest ball with center the origin that contains, and is contained, in P, respectively; i.e., $rB_d \subseteq P \subseteq RB_d$, where B_d is the unit ball. It is crucial to reduce R/r, i.e., to put P in well a rounded position where $R/r = \widetilde{\mathcal{O}}(\sqrt{d})$; the $\widetilde{\mathcal{O}}(\cdot)$ notation means that we are ignoring polylogarithmic factors. A powerful approach to obtain well roundness is to put P in *near isotropic position*. In general, $K \subset \mathbb{R}^d$ is in isotropic position if the uniform distribution over K is in isotropic position, that is $\mathbb{E}_{X\sim K}[X] = 0$ and $\mathbb{E}_{X\sim K}[X^T X] = I_d$, where I_d is the $d \times d$ identity matrix. Thus, to put a polytope P into isotropic position one has to generate a set of uniform points in its interior and apply to P the transformation that maps the point-set to isotropic position; then iterate this procedure until P is in c-isotropic position [Cousins and Vempala, 2016, Lovász and Vempala, 2006], for a constant c. In [Adamczak et al., 2010] they prove that $\mathcal{O}(d)$ points suffice to achieve 2-isotropic position. Alternatively in [Haraldsdóttir et al., 2017] they compute the maximum volume ellipsoid in P, they map it to the unit ball, and then apply to P the same transformation. They experimentally show that a few iterations suffice to put P in John's position [John, 2014]. Moreover, there are a few algorithmic contributions that combine sampling with distribution isotropization steps, e.g., the multi-point walk [Bertsimas and Vempala, 2004] and the annealing schedule [Kalai and Vempala, 2006].

An important parameter of a random walk is the walk length, i.e., the number of the intermediate points that a random walk visits before producing a single sample point. The longer the walk length of a random walk is, the smaller the distance of the current distribution to the stationary (target) distribution becomes. For the majority of random walks there are bounds on the walk length to bound the mixing time with respect to a statistical distance. For example, HR generates a sample from a distribution with total variation distance less than ϵ from the target distribution after $\widetilde{\mathcal{O}}(d^3)$ [Lovász and Vempala, 2006] steps, in a well rounded convex body and for log-concave distributions.

Similarly, CDHR mixes after a polynomial, in the diameter and the dimension, number of steps [Laddha and Vempala, 2020, Narayanan and Srivastava, 2020] for the case of uniform distribution. However, extended practical results have shown that both CDHR and HR converges after $\mathcal{O}(d^2)$ steps [Chalkis et al., 2020, Cousins and Vempala, 2016, Haraldsdóttir et al., 2017]. The leading algorithms for uniform polytope sampling are the Riemannian Hamiltonian Monte Carlo sampler [Lee and Vempala, 2018] and the Vaidya walk [Chen et al., 2018], with mixing times $\widetilde{\mathcal{O}}(md^{2/3})$ and $\widetilde{\mathcal{O}}(m^{1/2}d^{3/2})$ steps, respectively. However, it is not clear if these random walks can outperform CDHR in practice, because of their high cost per step and numerical instability.

Billiard Walk (BW) [Gryazina and Polyak, 2014] is a random walk that employs linear trajectories in a convex body with boundary reflections; alas with an unknown mixing time. The closest guarantees for its mixing time are those of HR and stochastic billiards [Dieker and Vempala, 2015]. Interestingly, [Gryazina and Polyak, 2014] shows that, experimentally, BW converges faster than HR for a proper tuning of its parameters. The same conclusion follows from the computation of the volume of zonotopes [Chalkis et al., 2020]. It is not known how the sandwiching ratio of P affects the mixing time of BW. Since BW employs reflections on the boundary, we can consider it as a special case of Reflective Hamiltonian Monte Carlo [Chevallier et al., 2018].

For almost all random walks the theoretical bounds on their mixing times are pessimistic and unrealistic for computations. Hence, if we terminate the random walk earlier, we generate samples that are usually highly correlated. There are several *MCMC Convergence Diagnostics* [Roy, 2020] to check if the quality of a sample can provide an accurate approximation of the target distribution. For a dependent sample, a powerful diagnostic is the *Effective Sample Size* (ESS). It is the number of effectively independent draws from the target distribution that the Markov chain is equivalent to. For autocorrelated samples, ESS bounds the uncertainty in estimates [Geyer, 1992] and provides information about the quality of the sample. There are several statistical tests to evaluate the quality of a generated sample, e.g., potential scale reduction factor (PSRF) [Gelman and Rubin, 1992], maximum mean discrepancy (MMD) [Gretton et al., 2012], and the uniform tests [Cousins, 2017]. Interestingly, the copula representation we employ in Figure 1.2 to capture the dependence between two fluxes of reactions was also used successfully in a geometric framework to detect financial crises capturing the dependence between portfolio return and volatility [Calès et al., 2018].

1.1.3 Contribution

We introduce a Multi-phase Monte Carlo Sampling (MMCS) algorithm (Section 1.1.4 and Algorithm 2) to sample from a polytope P. In particular, we split the sampling procedure in phases where, starting from P, each phase uses the sample to round the polytope. This improves the efficiency of the random walk in the next phase, see Figure 1.3. For sampling, we propose an improved variant of Billiard Walk (BW) (Section 1.1.4 that enjoys faster arithmetic complexity per step. We also handle efficiently the potential arithmetic inaccuracies near to the boundary, see [Chevallier et al., 2018]. We accompany the MMCS algorithm with a powerful MCMC diagnostic, namely the estimation of Effective Sample Size (ESS), to identify a satisfactory convergence to the uniform distribution. However,

our method is flexible and we can use any random walk and combination of MCMC diagnostics to decide convergence.

The open-source implementation of our algorithms[2] provides a complete software framework to handle efficiently sampling in metabolic networks. We demonstrate the efficiency of our tools by performing experiments on almost all the metabolic networks that are publicly available and by comparing with the state-of-the-art software packages as cobra (Section 1.1.6). Our implementation is faster than cobra for low dimensional models, with a speed-up that ranges from 10 to 100 times; this gap on running times increases for bigger models (Table 1.1). The quality of the sample our software produces is measured with two widely used diagnostics, i.e., ESS and potential scale reduction factor (PSRF) [Gelman and Rubin, 1992]. The highlight of our method is the ability to sample from the most complicated human metabolic network that is accessible today, namely Recon3D. In Figure 1.2 we estimate marginal univariate and bivariate flux distributions in Recon3D which validate:

- the quality of the sample by confirming a mutually exclusive pair of biochemical pathways, and that
- our method indeed generates steady states

In particular, our software can sample $1.44 \cdot 10^5$ points from a 5335-dimensional polytope in a day using modest hardware. This set of points suffices for the majority of systems biology analytics. To our understanding this task is out of reach for existing software. Last, MMCS algorithm is quite general sampling scheme and so it has the potential to address other hard computational problems like multivariate integration and volume estimation of polytopes.

1.1.4 Methods & Implementation

Efficient Billiard walk

The geometric random walk of our choice to sample from a polytope is based on Billiard Walk (BW) [Gryazina and Polyak, 2014], which we modify to reduce the per-step cost.

For a polytope $P = \{x \in \mathbb{R}^d \mid Ax \leq b\}$, where $A \in \mathbb{R}^{k \times d}$ and $b \in \mathbb{R}^k$, BW starts from a given point $p_0 \in P$, selects uniformly at random a direction, say v_0, and it moves along the direction of v_0 for length L; it reflects on the boundary if necessary. This results a new point p_1 inside P. We repeat the procedure from p_1. Asymptotically it converges to the uniform distribution over P. The length is $L = -\tau \ln \eta$, where η is a uniform number in $(0,1)$, that is $\eta \sim \mathcal{U}(0,1)$, and τ is a predefined constant. It is useful to set a bound, say ρ, on the number of reflections to avoid computationally hard cases where the trajectory may stuck in corners. In [Gryazina and Polyak, 2014] they set $\tau \approx diam(P)$ and $\rho = 10d$. Our choices for τ and ρ depend on a burn-in step that we detail in Section 1.1.5.

At each step of BW we compute the intersection point of a ray, say $\ell := \{p + tv,\ t \in \mathbb{R}_+\}$, with the boundary of P, ∂P, and the normal vector of the tangent plane at the intersection point. The inner vector of the facet that the intersection point belongs to is a row of A. To

[2] https://github.com/GeomScale/volume_approximation/tree/socg21

compute the point $\partial P \cap \ell$ where the first reflection of a BW step takes place, we solve the following m linear equations

$$a_j^T(p_0 + t_j v_0) = b_j \Rightarrow t_j = (b_j - a_j^T p_0)/a_j^T v_0,\ j \in [k], \tag{1.2}$$

and keep the smallest positive t_j; a_j is the j-th row of the matrix A. We solve each equation in $\mathcal{O}(d)$ operations and so the overall complexity is $\mathcal{O}(dk)$. A straightforward approach for BW would consider that each reflection costs $\mathcal{O}(kd)$ and thus the per step cost is $\mathcal{O}(\rho kd)$. However, our improved version performs more efficiently both *point* and *direction updates* by storing computations from the previous iteration combined with a preprocessing step. The preprocessing step involves the normal vectors of the facets, that takes $m^2 d$ operations, and the amortized per-step complexity of BW becomes $\mathcal{O}((\rho + d)k)$.

Lemma 1. *The amortized per step complexity of BW is $\mathcal{O}((\rho + d)k)$ after a preprocessing step that takes $\mathcal{O}(k^2 d)$ operations, where ρ is the maximum number of reflections per step.*

The use of floating point arithmetic could result to points outside P due to rounding errors when computing boundary points. To avoid this, when we compute the roots in Equation (1.2) we exclude the facet that the ray hit in the previous reflection.

Algorithm 1 Billiard Walk(P, p, ρ, τ, W)

Require: polytope P; point $p \in P$; upper bound on the number of reflections ρ;
parameter τ to adjust the length of the trajectory; walk length W.
Ensure: a point in P (uniformly distributed in P).
for $j = 1, \ldots, W$ **do**
$L \leftarrow -\tau \ln \eta$; $\eta \sim \mathcal{U}(0,1)$ {*length of the trajectory*} $i \leftarrow 0$ {*current number of reflections*} $p_0 \leftarrow p$ {*initial point of the step*} pick a uniform vector u_0 from the unit sphere {*initial direction*}
while $i \leq \rho$ **do**
$\ell \leftarrow \{p_i + t u_i, 0 \leq t \leq L\}$ {*this is a segment*}
if $\partial P \cap \ell = O$ **then**
$p_{i+1} \leftarrow p_i + L u_i$ **break**
end if
$p_{i+1} \leftarrow \partial P \cap \ell$; {*point update*}
the inner vector, s, of the tangent plane at p,
s.t. $||s|| = 1$, $L \leftarrow L - |P \cap \ell|$, $u_{i+1} \leftarrow u_i - 2(u_i^T s)s$ {*direction update*}
$i \leftarrow i + 1$
end while
if $i = \rho$ **then**
$p \leftarrow p_0$
else
$p \leftarrow p_i$
end if
end for
return p

At each step of Billiard Walk, we compute the intersection point of a ray, say $\ell := \{p + tu,\ t \in \mathbb{R}_+\}$, with the boundary of P, ∂P, and the normal vector of the tangent plane of P at the intersection point. The inner vector of the facet that the intersection point belongs to is a row of A. To compute the point $\partial P \cap \ell$ where the first reflection of a Billiard Walk step takes place we need to compute the intersection of ℓ with all the hyperplanes that define the facets of P. This corresponds to solve (independently) the following m linear equations

$$a_j^T(p_0 + t_j u_0) = b_j \Rightarrow t_j = (b_j - a_j^T p_0)/a_j^T u_0,\ j \in [k], \tag{1.3}$$

and keep the smallest positive t_j; a_j is the j-th row of the matrix A. We solve each equation in $\mathcal{O}(d)$ operations and so the overall complexity is $\mathcal{O}(dk)$, where k is the number of rows of A and thus an upper bound on the number of facets of P. A straightforward approach for Billiard Walk would consider that each reflection costs $\mathcal{O}(kd)$ and thus the per step cost is $\mathcal{O}(\rho kd)$. However, our improved version performs more efficiently both *point* and *direction updates* in pseudo-code by storing some computations from the previous iteration combined with a preprocessing step. The preprocessing step involves the normal vectors of the facets and takes $k^2 d$ operations. So the amortized per-step complexity of Billiard Walk becomes $\mathcal{O}((\rho + d)k)$. The pseudo-code appear in Algorithm 1.

Multiphase Monte Carlo Sampling algorithm

To sample steady states in the flux space of a metabolic network, with m metabolites and n reactions, we introduce a Multiphase Monte Carlo Sampling (MMCS) algorithm; it is multiphase because it consists of a sequence of sampling phases.

Let $S \in \mathbb{R}^{m\times n}$ be the stoichiometric matrix and $x_{lb},\ x_{ub} \in \mathbb{R}^n$ bounds on the fluxes. The flux space is the bounded convex polytope

$$\text{FS} := \{x \in \mathbb{R}^n \mid Sx = 0,\ x_{lb} \le x \le x_{ub}\} \subset \mathbb{R}^n. \tag{1.4}$$

The dimension, d, of FS is smaller than the dimension of the ambient space; that is $d \le n$. To work with a full dimensional polytope we restrict the box induced by the inequalities $x_{lb} \le x \le x_{ub}$ to the null space of S. Let the H-representation of the box be $\left\{x \in \mathbb{R}^n \,\middle|\, \begin{pmatrix} I_n \\ -I_n \end{pmatrix} x \le \begin{pmatrix} x_{ub} \\ x_{lb} \end{pmatrix}\right\}$, where I_n is the $n \times n$ identity matrix, and let $N \in \mathbb{R}^{n\times d}$ be the matrix of the null space of S, that is $SN = 0_{m\times d}$. Then $P = \{x \in \mathbb{R}^d \mid Ax \le b\}$, where $A = \begin{pmatrix} I_n N \\ -I_n N \end{pmatrix}$ and $b = \begin{pmatrix} x_{ub} \\ x_{lb} \end{pmatrix} N$, is a full dimensional polytope (in $\mathbb{R}^d$). After we sample (uniformly) points from P, we transform them to uniformly distributed points (that is steady states) in FS by applying the linear map induced by N.

MMCS generates, in a sequence of sampling phases, a set of points, that is almost equivalent to n independent uniformly distributed points in P, where n is given. At each phase, it employs Billiard Walk (Section 1.1.4) to sample approximate uniformly distributed points, rounding to speedup sampling, and uses the Effective Sample Size (ESS) diagnostic to decide termination. The pseudo-code of the algorithm appears in

Algorithm 2.

Overview.

Initially we set $P_0 = P$. At each phase $i \geq 0$ we sample at most λ points from P_i. We generate them in chunks; we also call them *chain* of sampling points. Each chain contains at most l points (for simplicity consider $l = \mathcal{O}(1)$). To generate the points in each chain we employ BW, starting from a point inside P_i; the starting point is different for each chain. We repeat this procedure until the total number of samples in P_i reaches the maximum number λ; we need $\frac{\lambda}{l}$ chains. To compute a starting point for a chain, we pick a point uniformly at random in the Chebychev ball of P_i and we perform $\mathcal{O}(\sqrt{d})$ burn-in BW steps to obtain a warm start.

After we have generated λ sample points we perform a rounding step on P_i to obtain the polytope of the next phase, P_{i+1}. We compute a linear transformation, T_i, that puts the sample into isotropic position and then $P_{i+1} = T_i(P_i)$. The efficiency of BW improves from one phase to the next one because the sandwiching ratio decreases and so the average number of reflections decreases and thus the convergence to the uniform distribution accelerates (Section 1.1.6). That is we obtain faster a sample of better quality. Finally, the (product of the) inverse transformations maps the samples to $P_0 = P$. Figure 1.3 depicts the procedure.

Termination.

There are no bounds on the mixing time of BW [Gryazina and Polyak, 2014], hence for termination we rely on ESS. MMCS terminates when the minimum ESS among all the univariate marginals is larger than a requested value. We chose the marginal distributions (of each flux) because they are essential for systems biologists, see [Bordel et al., 2010] for a typical example. In particular, after we generate a chain, the algorithm updates the ESS of each univariate marginal to take into account all the points that we have sampled in P_i, including the newly generated chain. We keep the minimum, say n_i, among all marginal ESS values. If $\sum_{j=0}^{i} n_j$ becomes larger than n before the total number of samples in P_i reaches the upper bound λ, then MMCS terminates. Otherwise, we proceed to the next phase. In summary, MMCS terminates when the sum of the minimum marginal ESS values of each phase reaches n.

Rounding step.

This step is motivated by the theoretical result in [Adamczak et al., 2010] and the rounding algorithms [Lovász and Vempala, 2006, Cousins and Vempala, 2016]. We apply the linear transformation T_i to P_i so that the sandwiching ratio of P_{i+1} is smaller than that of P_i. To find the suitable T_i we compute the SVD decomposition of the matrix that contains the sample row-wise [Artstein-Avidan et al., 2020].

Updating the Effective Sample Size.

The effective sample size of a sample of points generated by a process with autocorrelations ρ_t at lag t is function (actually an infinite series) in the ρ_t's; its exact value is unknown. Following [Geyer, 1992], we efficiently compute ESS employing a finite sum

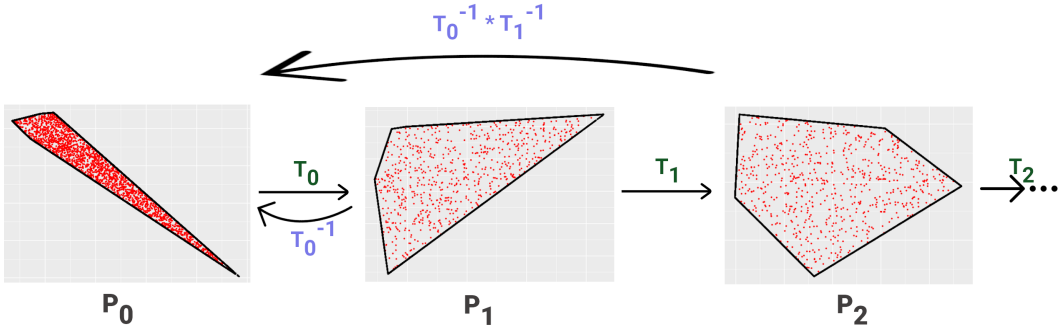

FIGURE 1.3: An illustration of our Multiphase Monte Carlo Sampling algorithm. The method is given an integer n and starts at phase $i = 0$ sampling from P_0. In each phase it samples a maximum number of points λ. If the sum of Effective Sample Size in each phase becomes larger than n before the total number of samples in P_i reaches λ then the algorithm terminates. Otherwise, we proceed to a new phase. We map back to P_0 all the generated samples of each phase.

of monotone estimators $\hat{\rho}_t$ of the autocorrelation at lag t, by exploiting Fast Fourier Transform. Furthermore, given M chains of samples, the autocorrelation estimator $\hat{\rho}_t$ is given by, $\hat{\rho}_t = 1 - \frac{C - \frac{1}{M}\sum_{i=1}^{M} \hat{\rho}_{t,i}}{B}$, where C and B are the within-sample variance estimate and the multi-chain variance estimate given in [Gelman and Rubin, 1992] and $\hat{\rho}_{t,i}$ is an estimator of the autocorrelation of the i-th chain at lag t. To update the ESS, for every new chain of points the algorithm generates, we compute $\hat{\rho}_{t,i}$. Then, using Welford's algorithm we update the average of the estimators of autocorrelation at lag t, as well as the between-chain variance and the within-sample variance estimators given in [Gelman and Rubin, 1992]. Finally, we update the ESS using these estimators.

To update the ESS, for every new chain of points the algorithm generates, we compute the estimator of its autocorrelation. Then, using Welford's algorithm we update the average of the estimators of autocorrelation at lag t, as well as the between-chain variance and the within-sample variance estimators [Gelman and Rubin, 1992]. Finally, we update the ESS using these estimators.

Lemma 2 (Complexity of MMCS per phase)**.** *Let $P = \{x \in \mathbb{R}^d \mid Ax \leq b\}$, where $A \in \mathbb{R}^{k \times d}$ and $b \in \mathbb{R}^k$, be a full dimensional polytope in $\mathbb{R}^d$. To sample n points (approximately) uniformly distributed in P, MMCS (Algorithm 2) performs $\mathcal{O}(W(\rho + d)k\lambda + \lambda^2 d + d^3)$ arithmetic operations per phase, where W is the walk length of Billiard Walk, ρ is an upper bound on the number of reflections, and λ and upper bound on the points generated at each phase.*

In Section 1.1.5 we discuss how to tune the parameters of MMCS to make it more efficient in practice. We also comment on the (practical) complexity of each phase, based on the tuning.

Algorithm 2 Multiphase Monte Carlo Sampling$(P, n, l, \lambda, \rho, \tau, W)$

Require: A full dimensional polytope $P \in \mathbb{R}^d$;
requested effectiveness $n \in \mathbb{N}$ (number of sampled points);
l length of each chain;
λ upper bound of the number of generated points in each phase λ;
upper bound on the number of reflections ρ;
parameter τ to adjust the length of the trajectory; walk length W.

Ensure: a set n of approximate uniformly distributed points $S \in P$

Set $P_0 \leftarrow P$, $sum\_ess \leftarrow 0$, $S \leftarrow \emptyset$, $i \leftarrow 0$, $T_0 = I_d$
while $sum\_ess < n$ **do**
  $sum\_point\_phase \leftarrow 0$, $U \leftarrow \emptyset$
  while $sum\_point\_phase < \lambda$; **do**
    Set $Q \leftarrow \emptyset$; Generate a starting point $q_0 \in P_i$;
    for $j = 1, \ldots, l$ **do**
      $q_j \leftarrow$Billiard_Walk$(P_i, q_{j-1}, \rho, \tau, W)$, Store the point q_j to the set Q
    end for
    $S \leftarrow S \cup T_i^{-1}(Q)$, $U \leftarrow U \cup Q$, $sum\_point\_phase \leftarrow sum\_point\_phase + l$ Update ESS n_i of this phase
    if $sum\_ess + n_i \geq n$ **then**
      break
    end if
  end while
  $sum\_ess \leftarrow sum\_ess + n_i$, Compute T such that $T(U)$ is in isotropic position, $P_{i+1} \leftarrow T(P_i)$, $T_{i+1} \leftarrow T_i \circ T$, $i \leftarrow i + 1$
end while
return S

1.1.5 Results

Implementation and Experiments

This section presents the implementation of our approach and the tuning of various parameters. We present experiments in an extended set of BiGG models [King et al., 2016], including the most complex metabolic networks, the human Recon2D [Swainston et al., 2016] and Recon3D [Brunk et al., 2018]. We end up to sample from polytopes of thousands of dimensions and show that our method can estimate precisely the flux distributions. We analyze various aspects of our method such as the run-time, the efficiency, and the quality of the output.

We compare against the state-of-the-art software for the analysis of metabolic networks, which is the Matlab toolbox of cobra [Heirendt et al., 2019]. Our implementation for low dimensional networks is two orders of magnitude faster than cobra. As the dimension grows, this gap on the run-time increases. The workflow of cobra for sampling first

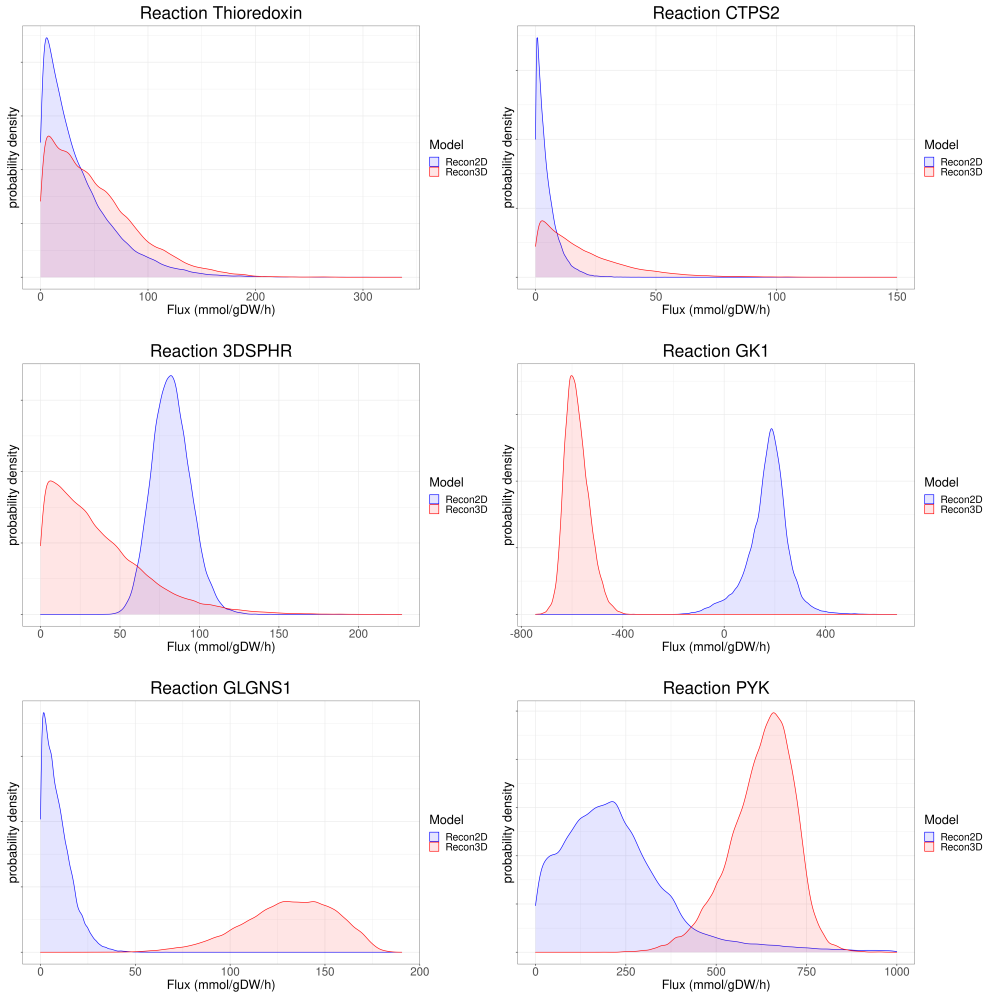

FIGURE 1.4: Estimation, using our tools, of the marginal distribution of 6 reaction fluxes in two constraint-based model of Homo Sapiens metabolism, namely Recon2D [Swainston et al., 2016] (blue color) and Recon3D [Brunk et al., 2018] (red color). In the case of GK1 we observe how the flux distribution of a reaction may change once the direction of the reaction changes.

performs a rounding step and then samples using Coordinate Directions Hit-and-Run (CDHR).

In [Jadebeck et al., 2020] they provide a C++ implementation of the sampling method that cobra uses and they show that their implementation is approximately 6 times faster than cobra. Nevertheless, we choose to compare against cobra, since it additionally provides efficient preprocessing methods that are crucial for the experiments, and give an implicit comparison with [Jadebeck et al., 2020].

say something about the immense differences

The fast mixing of billiard walk allow us to use all the generated samples to approximate each flux distribution and so we compute a better flux distribution estimation. To estimate each marginal flux distribution, using the samples, we exploit Gaussian kernel

				MMCS		cobra	
model	m	n	d	Time (sec)	N	Time (sec)	N
e_coli_core	72	95	24	6.50e-01	3.40e+03 (8)	7.20e+01	4.61e+06
iLJ478	570	652	59	9.00e+00	5.40e+03 (5)	4.54e+02	2.79e+07
iSB619	655	743	83	1.70e+01	8.20e+03 (5)	9.56e+02	5.51e+07
iHN637	698	785	88	2.00e+01	6.80e+03 (4)	1.03e+03	6.19e+07
iJN678	795	863	91	2.50e+01	8.10e+03 (4)	1.17e+03	6.62e+07
iNF517	650	754	92	1.70e+01	6.20e+03 (4)	1.33e+03	6.77e+07
iJN746	907	1054	116	5.70e+01	8.70e+03 (3)	2.22e+03	1.07e+08
iAB_RBC_283	342	469	130	5.20e+01	1.07e+04 (5)	7.85e+03	4.05e+08
iJR904	761	1075	227	2.98e+02	1.62e+04 (4)	8.81e+03	4.12e+08
iAT_PLT_636	738	1008	289	3.25e+02	1.04e+04 (2)	1.73e+04	6.68e+08
iSDY_1059	1888	2539	509	2.813e+03	2.31e+04 (3)	6.66e+04	2.07e+09
iAF1260	1668	2382	516	6.84e+03	5.33e+04 (6)	7.04e+04	2.13e+09
iEC1344_C	1934	2726	578	4.86e+03	3.95e+04 (4)	9.42e+04	2.67e+09
iJO1366	1805	2583	582	6.02e+03	5.14e+04 (5)	9.99e+04	2.71e+09
iBWG_1329	1949	2741	609	3.06e+03	4.22e+04 (4)	1.05e+05	2.97e+09
iML1515	1877	2712	633	4.65e+03	5.65e+04 (5)	1.15e+05	3.21e+09
Recon1	2766	3741	931	8.09e+03	1.94e+04 (2)	3.20e+05	6.93e+09
Recon2D	5063	7440	2430	2.48e+04	5.44e+04 (2)	~ 140 days	$1.57e{+}11$
Recon3D	8399	13543	5335	1.03e+05	1.44e+05 (2)	–	–

TABLE 1.1: Several, 17, metabolic networks from [King et al., 2016]; also Recon2D and Recon3D from [Noronha et al., 2019]. The semantics of the tables are as follows: (m) the number of Metabolites, (n) the number of Reactions, (d) the dimension of the polytope; (N) is the total number of sampled points × walk length; for MMCS we stop when the sum of the minimum value of ESS among all the univariate marginals in each phase is 1 000 (we report the number of phases in parenthesis); for cobra we set the walk length to $8d^2$ and $1.57e{+}08$ for Recon2D stop when all marginals have PSRF < 1.1; the run-time of cobra for Recon2D is an estimation of the sequential time and we report it to have a rough comparison with our implementation.

density estimation. This is a non-parametric way to estimate the probability density function of a random variable. For more details we refer to [Jones et al., 1996].

We provide a complete open-source software framework to handle big metabolic networks. The framework loads a metabolic model in some standard file formats (e.g., mat and json files) and performs an analysis of the model, e.g., it estimates the marginal distributions of a given reaction flux. All the results are reproducible using our publicly available code

The core of our implementation is in C++ to optimize performance while the user interface is implemented in R. The package employs [Guennebaud et al., 2010] for linear algebra, number generation, [Chalkis and Fisikopoulos, 2020], an open-source package for high dimensional sampling and volume approximation.

All experiments were performed on a PC with Intel Core i7-6700 3.40GHz × 8 CPU and 32GB RAM. In the sequel, MMCS refers to our implementation.

1.1.6 Experiments

We test and evaluate our software on 17 models from the BIGG database [King et al., 2016] as well as Recon2D and Recon3D from [Noronha et al., 2019]. In particular, we sample from models that correspond to polytopes of dimension less than 100; the simplest model in this setting is the well known bacteria *Escherichia Coli*. We also sample from models that correspond to polytopes of dimension a few thousands; this is the case for Recon2D and Recon3D. We do not employ parallelism for any implementation, thus we report only sequential running times.

We assess the quality of our results by employing both the Effective Sample Size (ESS) and the potential scale reduction factor (PSRF) [Gelman and Rubin, 1992]. In particular, we compute the PSRF for each univariate marginal of the sample that MMCS outputs. Following [Gelman and Rubin, 1992], a convergence is satisfying according to PSRF when all the marginals have PSRF smaller than 1.1.

In Table 1.1, we report the results of MMCS and cobra. For cobra, we report only the run-time of the sampling phase (we do not add to it the preprocessing time). We run MMCS until we get a value of ESS equal to 1 000; i.e. we stop when the sum over all phases of the minimum values of ESS among all the marginals is larger than 1 000. All the marginals of the MMCS samples reported in Table 1.1 have PSRF < 1.1. This is a strong statistical evidence on the quality of the generated sample.

The histograms in Figure 1.4 illustrate an approximation for the flux distribution of 6 reaction fluxes in Recon2D and Recon3D, respectively. Notice the difference in estimated densities due to the stoichiometric matrix update from Recon2D to Recon3D. The marginal flux distribution of reaction Thioredoxin in Recon2D was estimated also in [Haraldsdóttir et al., 2017] and used as an evidence for the quality of the sample. In Figure 1.2, we employ the copula representation to capture the dependency between two fluxes of reactions and confirm a mutually exclusive pair of biochemical pathways.

Comparing runtime performance, MMCS is one or two orders of magnitude faster than cobra and this gap becomes much larger for higher dimensional models such as Recon2D and Recon3D. Considering the experiments reported in [Jadebeck et al., 2020], they report the run-time of CDHR for each model until it generates a sample with PSFR 1.2; for Recon3D they report $\sim$ 1 day. Interestingly, for Recon3D, MMCS achieves PSRF 1.2 after $\sim$ 1 hour while reach PSRF 1.1 after $\sim$ 1 day.

For some models –we report them in Table 1.2– we introduce a further improvement to obtain a better convergence. If there is a marginal in the generated sample from MMCS that has a PSRF larger than 1.1, then we do not take into account the k first phases, starting with $k = 1$ until we get both ESS equal to 1 000 and all the PSRF values smaller than 1.1 for all the marginals. By "we do not take into account" we mean that we neither store the generated sample –for the first k phases– nor we sum up its ESS to the overall ESS considered for termination by MMCS. Note that for these models it is not practical to repeat MMCS runs for different k until we get the required PSRF value. We can obtain the final results –reported in Tables 1.1– in one pass. We simply drop a phase when the ESS reaches the requested value but the PSRF is not smaller than 1.1 for all the marginals. In Table 1.2, we separately report the MMCS runs for different k just for performance analysis reasons.

model	k	Time (sec)	PSRF < 1.1	M	N
iAF1260	0	6955	41%	6	56100
	1	6943	56%	6	54100
	2	6890	76%	6	55200
	3	6867	95%	6	53200
	4	6840	100%	6	53300
iBWG_1329	0	3067	50%	4	42100
	1	3189	97%	5	48800
	2	4652	100%	5	56500
iEC1344	0	4845	77%	4	41100
	1	4721	96%	4	42500
	2	4682	100%	4	39500
iJO1366	0	3708	66%	5	51500
	1	6022	100%	5	51400

TABLE 1.2: During our experiments we do not take into account the sample of the k first phases, thus we do not also count the value of the Effective Sample Size (ESS) in these phases, before we start storing the generated sample and sum up the ESS of each phase. In all cases MMCS stops when the sum of ESS reaches 1000. For each case we report the total run-time, the percentage of the marginals that have PSRF smaller than 1.1, the total number of phases (M) needed (including the k first phases), and the total number of Billiard Walk steps (N), including those performed in the k first phases.

Interestingly, the total number of Billiard Walk steps –and consequently the run-time– does not increase as k increases in Table 1.2. This means that the performance of our method improves for these models when we do not take into account the k first phases of MMCS. This happens because the performance of Billiard Walk improves as the polytope becomes more rounded from phase to phase.

In Table 1.3, we analyze the performance of Billiard Walk for the model iAF1260. We sample $20d$ points per phase with walk length equal to 1 and we report the average number of reflections, the ESS, the run-time, and the ratio $\sigma_{\max}/\sigma_{\min}$ per phase. The latter is the ratio of the maximum over the minimum singular value of the point-set. The larger this ratio is the more skinny the polytope of the corresponding phase is. As the method progresses from the first to the last phase, the average number of reflections and the run-time decrease and the ESS increases. This means that as the polytope becomes more rounded from phase to phase, the Billiard Walk step becomes faster and the generated sample has better quality. This explains why the total run-time does not increase when we do not take into account the first k phases: the initial phases are slow and they contribute poorly to the quality of the final sample; the last phases are fast and contribute with more accurate samples.

1.2 Conclusions and future work

We propose a novel method for sampling that can sample from a convex polytope in a few thousands of dimensions within a day on modest hardware. This way, we are able, for the first time, to perform accurate sampling from the latest human metabolic network, Recon3D.

Sampling from iAF1260				
Phase	Avg. #reflections	ESS	$\frac{\sigma_{\max}}{\sigma_{\min}}$	Time (sec)
1st	7819	67	43459	2271
2nd	4909	68	922	1631
3rd	3863	77	582	1278
4th	3198	71	360	1080
5th	1300	592	29	454
6th	1187	4821	3.5	417
7th	1181	4567	2.8	415

TABLE 1.3: We sample $20d = 10320$ points per phase with Billiard Walk and walk length equal to 1, where $d = 516$ is the dimension of the corresponding polytope. For each phase we report the average number of reflections per BIlliard Walk step, the minimum value of Effective Sample Size among all the univariate marginals, the ratio between the maximum and the minimum singular value of the SVD decomposition of the generated sample, and the run-time.

Regarding future work, parallelism could lead to a speedup in the run-time of our method as the algorithm is rather straightforward to parallelize. An additional improvement would be to exploit the sparsity of the stoichiometric matrix S and sample directly from the low dimensional polytope in $\mathbb{R}^n$ without projecting to a lower dimensional space.

Moreover, our method could be extended to any log-concave distribution restricted to the flux space and combined with bayesian metabolic flux analysis, to sample from multivariate, possibly multi-modal target distribution [Heinonen et al., 2019] addressing multiple challenges of the method from the biological point of view (e.g., unrealistic assumptions, uncertainty etc.). Last but not least, flux sampling in metabolic models built out from multiple metabolic networks, e.g., representing a microbial community, could also lead to important biological insights.

Chapter 2

Conclusions

2.1 Bioinformatics approaches enhance microbial diversity assesment based on HTS data

Main goal of this PhD project was to address on-going challenges related to the bioinformatics analysis of HTS-oriented studies as well as to provide ways for the optimal exploitation of such data and of the current knowledge that is linked to them.

The 16S rRNA gene has been used for decades as the golden standard for the study of microbial communities. It has been shown that the full-length 16S sequence combined with appropriate treatment of the intragenomic copy variants has the potential to provide taxonomic resolution of bacterial communities even at the strain level [Johnson et al., 2019]. However, when the region is chosen carefully and a thorough alignment procedure is applied, even short short reads may return phylogenetic information comparable with the one fron full-length 16S rRNA reads [Jeraldo et al., 2011]. This was also shown in Chapter **??** as the 16S rRNA amplicon analysis was in line with the taxonomy assignment of the shotgun reads.

Even if amplicon studies have proven themselves essential for the assessment of microbial diversity, the bioinformatics analaysis in such studies, usually comes with several issues; with the lack of parameter tuning being among the most crucial ones. As shown in Chapter **??** where mock communities were used to validate the PEMA results, it is parameter tuning that determines the precision and recall scores in such analyses. Sequencing mock communities along with the rest of the samples allows the tuning of the bioinformatics analysis based on a known assemblage and thus, it enables parameter tuning based on the idiosyncracy of each particular experiment/study [Bokulich et al., 2020].

When studying a microbial community, non-prokaryotic species need to be considered too. In that case, 16S rRNA is not the best marker to use; instead, several markers have been used for different taxonomic groups. Thus, several studies aiming at the biodiversity assessment of environmental samples, make use of several markers and apparently, workflows supporting their analysis are vital. As shown in Chapter **??**, the PEMA approach attempts to address this challenge by supporting the analysis of several markers but also by supporting the semi-automatic analysis of any marker since training of the classifiers

invoked with any local database is possible.

Moreover, it is also commonly known that pseudogenes as well as nuclear mitochondrial pseudogenes (numts) can lead to several biases in such studies [Song et al., 2008]. To address this challenge multiple computational efforts have been implemented [Porter and Hajibabaei, 2021] This issue also applies for the case of Bacteria and Archaea and the 16S rRNA gene [Pei et al., 2010] even if it has been shown that bacterial pseudogenes have a great chance of being removed almost directly after their formation; so fast that to be governed by a strictly neutral model of stochastic loss [Kuo and Ochman, 2010]. As shown in Chapter **??**, a great part of the OTUs/ASVs retrieved from COI amplicon data may actually come from bacterial and/or archaeal taxa. Such approaches need to be merged in amplicon studies as an extra quality control stop but also to enable further investigation of the unassigned OTUs/ASVs. In Chapter **??** is also shown the need for reference databases to also include non-target sequences so they can distinguish actual hits.

However, there is still a major question regarding the microbial diversity assessment; how could HTS methods be used to recognise novel taxa and their metabolic potential? As shown in Chapter **??**, the reconstruction of MAGs from shotgun metagenomics data may play a great role in the description of unknown and currently uncultivated taxa. Such studies and their corresponding MAGs have enriched our knowledge on the tree of life to a great extent over the last few years, uncovering several prokaryotic phyla, leading to radical challenges on their taxonomy and the taxonomy scheme [Parks et al., 2022]. Long-read sequencing technologies such as Nanopore and PacBio, have improved their accuracy to a great extent, offering high-quality, cutting-edge alternatives for testing hypotheses about microbiome structure and functioning as well as assembly of eukaryote genomes from complex environmental DNA samples [Tedersoo et al., 2021].

2.2 Containerization technologies and e-infrastructures provide the means for computational capacity and reproducibility

As shown in Chapter **??** the computing resources required for the analysis of several microbiome studies may range from those covered by a personal computer to overwhelming the capacity of Tier-2 HPC facilities; this also applies for any biological topic using HTS data. On top of that, as also shown in Chapter **??**, the maintance of bioinformatics-oriented HPC facilities comes with a great number of challenges.

By encapsulating a software along with its dependencies in an isolated and easy to reinstall environment (container) containerization addresses several of them at once; first, the distribution and the installation becomes now a straight-forward task, requiring only for a containerization technology present on the facility, second, versioning of the various software used is not an issue anymore as a container may either "save" a version from being obsolete in case it is strongly dependent on that, either keep track of the latest version of the encapsulated software, morevover, several versions of the same software may be part of different containers without any conficts. In addition, the creation and management of standardized workflows/pipelines is facilitated to a great extent. Workflow

tools such as Common Workflow Language (CWL) [1], Snakemake [2] and Nextflow [3] have been proven of high value in building such pipelines as they support the connection of multiple independent software. MGnify [Mitchell et al., 2020] is a great example of this case.

However, more often than not, such workflows require major computing resources to analyse real-world microbiome data sets. To this end, HPC facilities and cloud solutions are required. Therefore, efforts such as those discussed in Section **??** for containerised tools such as the PEMA workflow [Zafeiropoulos et al., 2020] to be integrated in e-infrustructures, are rather significant. This way, reproducability is secured and analyses that cannot be performed in a personal computer is accessible to researches that have no access to local servers or HPCs.

2.3 High quality metadata enable efficient exploitation of sequecning data in a meta-analysis level

It is common knowledge that both shotgun and long-read metagenomics provide high quality data enabling the study of real-world microbial communities even at the strain level [Meyer et al., 2022]. However, these data are not fully exploited unless they come with thorough and standardized metadata; indicatively it has been shown that more than the 20% of the metagenomes published between 2016 and 2019 were not even accessible [Eckert et al., 2020]. To address such challenges, community-driven initiatives can be of great importance [Yilmaz et al., 2011, Vangay et al., 2021]. Such initiatives could be of great help for more specific topics as well, e.g. protocols suitting the Ocean Best Practices System [Samuel et al., 2021].

FAIR principles have set a new era on how data are stored and ditributed. Data and metadata provenance goes even further by allowing not only the reuse of the data, but also keeping track of where the data came from, potential edits, as well as of the analyses that are linked to those, both regarding their outcome and the analysis *per se*. Ensuring FAIR (meta)data could be considered as the first step of data integration [Freire et al., 2008, Deelman et al., 2010]. Nevertheless, (meta)data provenance and data integration share several challenges [Cheney et al., 2009].

As shown in Section **??**, the higher the quality of the accompanying metadata, the higher the confidence for meta-analysis approaches may get. Furthermore, metadata accompanying the analyses, i.e. the workflows and their implementation, can provide further insight on the effect of the various softwares and databases used. In general, community efforts to set best-practice for formal metadata description and their use for packaging research data with their accompanying metada such as the one of RO-Crate may have a great impact. Focusing on the microbiome community, further challenges need to be addressed to this end, with taxonomy and nomenclature being among the most crucial ones. The GTDB [Parks et al., 2022] that provides *a phylogenetically consistent and rank normalized genome-based taxonomy for prokaryotic genomes sourced from the*

[1] https://github.com/common-workflow-language/common-workflow-language
[2] https://snakemake.github.io
[3] https://www.nextflow.io

NCBI Assembly database combined with efforts such as those of Pallen et al. to ensure valid names for novel taxa that will keep being discovered over the next years, could play a great role on addressing this issue. However, moving all the so-far knowledge in a specific format is far from easy. At the same time, as discussed in **??**, bringing together data of different types can return great insight by either generate hypotheses or further supporting hypotheses such as the one of Pavloudi et al. on the potential use of various electron acceptors from the different strains present in different environmental types (see **??**). This becomes clearer considering that up to now only a small fraction of the microbial diversity has been described and named (24,000 species [Parte et al., 2020]) with almost 10,000 of them to have done so, over the last 6 years (see LPSN statistics [4]).

Moreover, linking the various ontologies related to a certain domain without loosing the benefits of each of those, is essential for the community. That means that efforts for mapping entities of an ontology or a database to those of others sharing a common field, even if there is not a one-to-one relationship, will increase considerably their impact. Efforts such the one of the Rhea reaction knowledgebase [Bansal et al., 2022]

By addessing these challenges and by ensuring high quality metadata, data integration techniques will be improved considerably. Combined with machine learning techniques, such methods can contribute the most in exploiting the full potential of the HTS data produced and the so-far knowledge for the utmost biological insights that can be drawn [Noor et al., 2019].

2.4 Markov Chain Monte Carlo approaches enable flux sampling at the microbial community level

Metabolic modelling and genome-scale metabolic models in particular, provide a great framework to study the genotype - phenotype relationship [Lewis et al., 2012]. Thus, it enables the investigation on how a species respond under changing environmental conditions too [Herrmann et al., 2019]. Flux sampling on microbial GEMs has been proved the most valuable for the identification of specific reactions that are transcriptionally regulated [Bordel et al., 2010] or required under certain conditions for the species to survive [Herrmann et al., 2019]. But also in the study of the variant by-produts produced by the different strains of a species [Scott et al., 2021].

As shown in Chapter 1 the higher the dimension of the polytope derived from a metabolic model, the more challenging the sampling on its flux space gets. The dimension of a polytope derived from any single microbial GEM is at least one order of magnitude lower than the one from species such as *H. sapiens*. However, in real-world microbial communities it is rare for a species to be on its own. Based on Perez-Garcia et al. [2016] there are various approaches for modelling a community as a whole, each coming with several pros and cons. As the *lumped network* approach neglects microbial diversity dynamics and the dynamics of their corresponding processes, assuming a *super-organism* where all species share the same reactions and exploit their environment in the same way, it cannot be used for the study of microbial interactions. To this end, models that

[4] https://lpsn.dsmz.de/statistics/figure/10

integrate a GEM for every species present, taking into account the relative abundance of each species, and also support the exchange of compounds between each species and the environment are required [Diener et al., 2020]. Dynamic versions of such models also allow the changes in the biomass concentrations of each individual species to be taken into account [Zhuang et al., 2011]. Approaches like COMETS [Dukovski et al., 2021] and MICOM have been essential towards this direction. However, flux sampling has not been merged yet to large-scale approaches mostly because of the computational challenges that arise as the dimension of the polytope that derives from a model increases. Therefore, approaches such as the MMCS algorithm desribed in Chapter 1 may benefit the community to this end.

2.5 Future work: more holistic approaches are essential to uncover the underlying mechanisms governing microbial communities

Metagenomics and the rest of the 'omics technologies have turned the page on microbial ecology studies. However, to address questions such as the seasonality effects on the community structure and its functioning (see Section **??**) Systems Biology approaches are required. As discussed by Bajic and Sanchez [2020] *"a combination of quantitative high-throughput experiments and predictive metabolic models can help us map the genotype - phenotype map of microbial metabolic strategies.* Further technologies, such as Raman micro-spectroscopy [Jing et al., 2018], can also be of great use to this end. The prediction of such strategies based on genomic information according to Bajic and Sanchez will also provide great insight on the evolvability of metabolic decisions and will shed light on how these decisions affect microbial coexistence in the communities. The software developed in the framework of the PhD can benefit such approaches, either by improving HTS data analysis, or by enhancing their exploitation or with novel predictive algorithms. Apparently, for each of these fields, there are still several challenges that need to be addressed.

Approaches such as PREGO (see Chapter **??**) will not reveal their true potential unless the community embrace a series of standards and metadata protocols. Thus, it is essential both to develop such checklists but also to convincr and train the community using them. As already discussed, machine learning methods (see Conclusion 2.3) as well as network theory and visualizations may benefit such approaches notably.

Moreover, to infer microbe - microbe associations, co-occurrence networks can be of help and the incorporation of phenotypical data such as pH values, optimal temperatures etc. to such networks, may benefit the inference of such associations to a great extent. Resources such as FAPROTAX [Louca, Parfrey, and Doebeli, 2016], PhenDB [Feldbauer et al., 2015] and BugBase [Ward et al., 2017] provide great input for such a task.

Moreover, by coupling data integration with metabolic modelling approaches more

MAKE IT WHOLE

As already discussed, metabolic models at the community level. to infer microbial interactions but also to study the fitness of the community.

Eco-evolutionary dynamics of complex social strategies in microbial communities [Harrington and Sanchez, 2014]

Acknowledgements

This dissertation has recieved funding from the Hellenic Foundation for Research and Innovation (HFRI) and the General Secretariat for Research and Technology (GSRT), under grant aggrement No 241 (PREGO project).

Appendices

Appendix A

Computational Geometry

A.1 Moving from the concentration to the flux vector

The dynamic mass balance on a chemical compound is the difference between the sum of the fluxes of all the reactions that form it and the sum of all the reactions that degrade it.

In general, the following ordinary differential equation expresses such a mass balance:

$$\frac{d\omega_i}{dt} = \sum_k s_{ik} v_k = \langle s_i \cdot v \rangle, \tag{A.1}$$

where ω_i is the total mass of the i-th metabolite, s_{ik} is the stoichiometric coefficient for this metabolite in the k-th reaction, and v_k is the flux of the k-th reaction. By considering all the differential equations expressing the dynamic mass balance of all the compounds present in a metabolic network, we have

$$\frac{d\omega}{dt} = Sv, \tag{A.2}$$

where S is the stoichiometric matrix having as rows the vectors s_i. In this setting, S is the map of the linear transformation that sends the flux vector to a vector of time derivatives of the concentration vector Palsson [2015].

A.2 Definitions & concepts

A *hyperplane* is a set of the form

$$H = \{x \in \mathbb{R}^n : p^T \cdot x = t\} \tag{A.3}$$

and defines two closed *halfspaces*. Such a halfspaces would be denoted as

$$\begin{aligned} H_- &= \{x \in \mathbb{R}^n : p^T \cdot x \leq t\} \\ H_+ &= \{x \in \mathbb{R}^n : p^T \cdot x \geq t\} \end{aligned} \tag{A.4}$$

The intersection of a finite number of halfspaces builds a *polyhedron*. A system of inequalities arises:

$$a_i^t \cdot x \leq b_i, i \in \{1, .., m\} \tag{A.5}$$

where m is the number of halfspaces. Thus, a polyhedron can be denoted as:

$$P = \{x \in \mathbb{R}^n : A \cdot x \leq b\} \tag{A.6}$$

where A is a $m * n$ matrix with m being the number of halfspaces and n the dimension of the space. Finally, b is a vector of the right side of the inequalities (b_i).

A *bounded* polyhedron meaning, $\exists M > 0$ such that $\|x\| \leq M$ for all $x \in P$, is called a ***polytope***.

Some of the inequalities in A though can be geometrically redundant, meaning that if these are removed P remains the same. The *dimension* of a polytope P is equal to $n - r(P)$ where $r(P)$ is the maximum number of linearly independent defining hyperplanes containing P.

We call *defining hyperplanes* the **total** hyperplanes defined from the system, meaning those coming from the $A \cdot x \leq 0$ plus those coming from any constraints. For example, maybe we would have $x \geq 0$ then these hyperplanes would be also considered as defining hyperplanes.

We consider P as a ***fully dimensional*** polytope if and only if $dim(P) = n$. In other words, a d-polytope is full-dimensional in d-space. Each (nonredundant) inequality corresponds to a facet of the polytope

In case that our system has only inequalities, then the polytope derived is always full dimensional. However, in case that extra cosntraints as equalities are included, then the polytope derived could be full-dimensional or not. If the space defined by the equalities intersects the one defined by the inequalities, then the polytope is not full-dimensional.

A *face* is a set of points $F \subseteq P$ that belongs to the intersection of a nonempty set of defining hyperplanes To show that a valid inequality is a face we just need to find a point in the intersection of the hyperplane it defines and our polytope. To show that a face is a ***facet***, i.e. a face of dimension $n-1$, we need to show that it belongs to exactly one defining hyperplane. If it belongs to more, then it is no longer a facet.

Facets are necessary and sufficient for the complete description of a polytope in terms of valid inequalities.

If P is full-dimensional then it has a unique minimal description:

$$P = \{x \in \mathbb{R}^n : a_i^T \cdot x \leq b_i, i = \{1, ..m\}\} \tag{A.7}$$

where each of the m inequalities is unique to within a positive multiple.

Points $x_1, x_2, ..., x_k \in \mathbb{R}^n$ are *affinely independent* if the $k-1$ directions: $x_i - x_1, i \in \{2, k\}$ are linearly independent. The maximum number of affinely independent points in P is denoted as $i(P)$. Now the dimension of P can be defined as: $dim(P) = i(P) - 1$

To show that P is full-dimensional we just need to show that it has exactly $n+1$ affinely independent points.

A matrix is said to have full rank if its rank equals the largest possible for a matrix of the same dimensions, which is the lesser of the number of rows and columns.

Markov Chain Monte Carlo

Definition 1. *A Markov chain or Markov process is a stochastic model describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event*

Markov Chain Monte Carlo (MCMC) methods are algorithms that sample from a probabilistic distribution.

Appendix B

PREGO

B.1 Mappings

PREGO produces entity identifiers either by Named Entity Recognition (NER) with the EXTRACT tagger or by mapping retrieved identifiers to the selected ones. PREGO adopted NCBI taxonomy identifiers for taxa, Environmental Ontology for environments and Gene Ontology as a structure knowledge scheme for Processes (GObp) and Molecular Functions (GOmfs). The latter was for reasons that are two-fold, first Gene Ontology has a Creative Commons Attribution 4.0 License and second there are many resources that have mapped their identifiers to Gene Ontology. MG-RAST metagenomes and JGI/IMG isolates annotations come with KEGG orthology (KO) terms; Struo-oriented genome annotations, on the other hand, have Uniprot50 ids. The mapping from KO to GOmf and Uniprot50 to GOmf is implemented via UniProtKB mapping files of their FTP server (see idmapping.dat and idmapping_selected.tab files). By using the 3-column mapping file, the initial annotations were mapped to GOmf. As a complement, a list of metabolism-oriented KEGG ORTHOLOGY (KO) terms has been built (see *prego_mappings* in the Availability of Supporting Source Codes section). Finally, as STRUO annotations refer to GTDB genomes, publicly available mappings (accessed on 24 December 2021) were used to link the genomes used with their corresponding NCBI Taxonomy entries.

B.2 Daemons

An important component PREGO approach (Figure A1) is the regular updates which keep PREGO in line with the literature and microbiology data advances. The updates are implemented with custom scripts called daemons that are executed regularly spanning from once a month up to six-month cycles. This variation occurs because of the API requirements of each web resource as well as the computational intensity of the association extraction from the retrieved data.

Each Daemon is attached to a resource because its data retrieval methods (API, FTP) and following steps, shown in Figure A1, require special handling and multiple scripts (see *prego_daemons* in the Availability of Supporting Source Codes section).

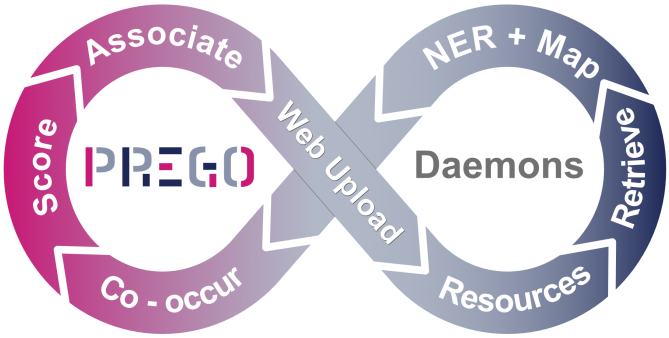

FIGURE B.1: Software daemons perform all steps of the PREGO methodology in a continuous manner similar to the Continuous Development and Continuous Integration method.

B.3 Scoring

Scoring in PREGO is used to answer the questions:

- Which associations are more thrustworthy?
- Which associations are more relevant to the user's query?

Relevant, informative, and probable associations are presented to the user through the three channels that were discussed previously. Each channel has its own scoring scheme for the associations it contains and all of them are fit in the interval $(0,5]$ to maintain consistency. The values of the score are visually shown as stars. The Genome Annotation and Isolates channel has fixed values of scores depending on the resource because Genome Annotation is straightforward, and the microbe id is known a priori. On the other hand, Environmental Samples channel data are based on samples, which contain metagenomes and OTU tables. Thus, it has two levels of organization, microbes with metadata, and sample identifiers. Each association of two entities is scored based on the number of samples they co-occur. A Literature channel scoring scheme is based on the co-mention of a pair of entities in each document, paragraph, and sentence. The differences in the nature of data require different scoring schemes in these channels. The contingency table (Table B.1) of two random variables, X and Y are the starting point for the calculation of scores. The term $X = 1$ might be a specific NCBI id and $Y = 1$ a ENVO term. The $c_{1,1}$ is the number of instances that two terms of $X = 1$ and $Y = 1$ are co-occurring, i.e., the joint frequency. The marginals are the $c_{1,.}$ and $c.,1$ for x and y, respectively, which are the backgrounds for each entity type. Different handling of these frequencies leads to different measures. There is not a perfect scoring scheme, just the one that works best on a particular instance. Consequently, scoring attributes require testing different measures and their parameters.

		Y = y		
		Yes	No	Total
X = x	Yes	$c_{x,y}$	$c_{x,0}$	$c_{x,.}$
	No	$c_{0,y}$	$c_{0,0}$	$c_{0,.}$
	Total	$c_{.,y}$	$c_{.,0}$	$c_{.,.}$

TABLE B.1: Contingency table of co-occurrences between entities $X = x$ and $Y = y$. This is the basic structure for all scoring schemes. $c_{x,y}$ is the count of the co-occurrence of these entities. $c_{x,.}$ is the count of the x with all the entities of Y type (e.g., Molecular function). Conversely, $c_{.,y}$ is the count of y with all the entities of X type (e.g., taxonomy

Literature Channel

Scoring in the Literature channel is implemented as in STRING 9.1 [Franceschini et al., 2012] and COMPARTMENTS [Binder et al., 2014], where the text mining method uses a three-step scoring scheme. First, for each co-mention/co-occurrence between entities (e.g., Methanosarcina mazei with Sulfur carrier activity), a weighted count is calculated because of the complexity of the text.

$$c_{x,y} = \sum_{k=1}^{n} w_d \delta_{dk}(x,y) + w_p \delta_{p,k}(x,y) + w_s \delta_{sk}(x,y) \tag{B.1}$$

Different weights are used for each part of the document (k) for which both entities have been co-mentioned, $w_d = 1$ for the weight for the whole document level, $w_p = 2$ for the weight of the paragraph level, and $w_s = 0.2$ for the same sentence weight. Additionally, the delta functions are one (Equation B.1) in cases the co-mention exists, zero otherwise. Thus, the weighted count becomes higher as the entities are mentioned in the same paragraph and even higher when in the same sentence. Subsequently, the co-occurrence score is calculated as follows:

$$score_{x,y} = c_{x,y}^{a} \left(\frac{c_{x,y} c_{.,.}}{c_{x,.} c_{.,y}}\right)^{1-a} \tag{B.2}$$

where $a = 0.6$ is a weighting factor, and the $c_{x,.}$, $c_{.,.}$, $c_{.,y}$ are the weighted counts as shown in Table B.1 estimated using the same Equation B.2. This value of the weighting factor has been chosen because it has been optimized and benchmarked in various applications of text mining [34,70,71]. The value of Equation B.2 is sensitive to the increasing size of the number of documents (MEDLINE PubMed—PMC OA). Therefore, to obtain a more robust measure, the value of the score is transformed to z-score. This transformation is elaborated in detail in the COMPARTMENTS resource [Binder et al., 2014]. Finally, the confidence score is the z-score divided by two. Cases in which the scores exceed the (0,4] interval are capped to a maximum of 4 to reflect the uncertainty of the text mining pipeline.

Environmental Samples Channel

Data from environmental samples are OTU tables and metagenomes. Thus, for each entity x, the number of samples is calculated as the background and a number of samples of the associated entity (metadata background) c.,y (see Table A1). Each association between entities x, y has a number of samples, cx,y that they co-occur. Note that each resource is independent and the scoring scheme is applied to its entities. This means that the same association can appear in multiple resources with different scores. The score is calculated with the following formula:

$$score_{x,y} = 2.0 * \sqrt{\frac{c_{x,y}}{c_{.,y}}}^{a} \tag{B.3}$$

This score is asymmetric because the denominator is the marginal of the associated entity. Thus, the score decreases as the marginal of y is increasing, i.e., the number of samples that y is found. On the other hand, it promotes associations in which the number of samples of the association are similar to the marginal of y. The exponents on the numerator and denominator equal to 0.5 and to 0.1, respectively, in order to reduce the rapid increase of score. Lastly, the value of the score is capped in the range $(0, 4]$.

B.4 Bulk download

Users can also download programmatically all associations per channel through the links that are shown in Table B.2. The data are compressed to reduce the download size and md5sum files are provided as well for a sanity check of each download.

Channel	Link	md5sum	Size (in GB)
Literature	literature.tar.gz	literature.tar.gz.md5	5.4
Environmental Samples	environmental_samples.tar.gz	environmental_samples.tar.gz.md5	0.69
Annotated genomes and isolates	annotated_genomes_isolates.tar.gz	annotated_genomes_isolates.tar.gz.md5	0.26

TABLE B.2: Bulk download links and md5sum files.

Appendix C

Metagenome assembled genomes of novel prokaryotic taxa from a hypersaline marsh microbial mat

C.1 MAGs description

Using Barrnap [Seemann, 2014a] the 16S rRNA gene was extracted from the retrieved MAGs. For those where a 16S rRNA gene was found, Protologger (version 1.0) [Hitch et al., 2021] was used for a thorough description of their properties. On the Protologger framework, MAG's closest relatives based on 16S rRNA gene sequence similarity were retrieved using blastn (version 2.12.0+) and The All-Species Living Tree database [Ludwig et al., 2021]. Each MAG was placed on the GTDB phylogeny tree using the GTDB-Tk and average nucleotide identity (ANI) values were calculated to check whether the MAG is a representative of an already known species; no ANI value is reported for a genome pair if ANI value is much below 80%.

MAGs were then annotated with Prokka (Seemann 2014b) [Seemann, 2014b] and percentage of conserved proteins (POCP) values [Qin et al., 2014] was calculated between MAGs and the genomes that are close to it based on both the 16S rRNA and the genome-based assignment modules. This was the case only for genomes with validly published names according to the DSMZ nomenclature list. POCP analysis has been used to distinguish prokaryotic genera since a prokaryotic genus can be defined as a group of species with all pairwise POCP values higher than 50% [Qin et al., 2014]. The outcome of the Protologger tool for each archaeal MAG can be found on GitHub and for each bacterial too.

For MAG cases suggesting multiple novel entries in novel taxonomic groups higher than the species level, e.g. multiple novel genera within the same family, further POCP values were calculated between each of the MAGs and all of its closest relatives having a genome on GTDB using in-house scripts. For these cases, a phylogenetic tree using the MAG's alignment by the GTDB-Tk and the genomes' entries in the GTDB Multiple Sequence Alignment (MSA) was built. Both the phylogenetic trees and the POCP analyses for these cases are available here. In the "Etymology" section the names given to the new

taxa are described. For a thorough investigation of the so-far described characteristics of the reconstructed MAGs' higher taxonomic levels (e.g., genus, family etc.) the PREGO knowledge base [Zafeiropoulos et al., 2022] was exploited. All bioinformatics analyses were supported by the IMBBC High Performance Computing system [Zafeiropoulos et al., 2021].

Bibliography

R. Adamczak, A. Litvak, A. Pajor, and N. Tomczak-Jaegermann. Quantitative estimates of the convergence of the empirical covariance matrix in log-concave ensembles. *Journal of the American Mathematical Society*, 23(2):535–561, 2010. ISSN 0894-0347, 1088-6834. doi: 10.1090/S0894-0347-09-00650-X.

S. Artstein-Avidan, H. Kaplan, and M. Sharir. On radial isotropic position: Theory and algorithms, 2020.

D. Bajic and A. Sanchez. The ecology and evolution of microbial metabolic strategies. *Current opinion in biotechnology*, 62:123–128, 2020.

P. Bansal, A. Morgat, K. B. Axelsen, V. Muthukrishnan, E. Coudert, L. Aimo, N. Hyka-Nouspikel, E. Gasteiger, A. Kerhornou, T. B. Neto, et al. Rhea, the reaction knowledgebase in 2022. *Nucleic acids research*, 50(D1):D693–D700, 2022.

D. B. Bernstein, F. E. Dewhirst, and D. Segre. Metabolic network percolation quantifies biosynthetic capabilities across the human oral microbiome. *Elife*, 8:e39733, 2019.

D. Bertsimas and S. Vempala. Solving convex programs by random walks. *J. ACM*, 51 (4):540–556, July 2004. ISSN 0004-5411. doi: 10.1145/1008731.1008733. URL https://doi.org/10.1145/1008731.1008733.

J. X. Binder, S. Pletscher-Frankild, K. Tsafou, C. Stolte, S. I. O'Donoghue, R. Schneider, and L. J. Jensen. Compartments: unification and visualization of protein subcellular localization evidence. *Database*, 2014, 2014.

N. A. Bokulich, M. Ziemski, M. S. Robeson II, and B. D. Kaehler. Measuring the microbiome: Best practices for developing and benchmarking microbiomics methods. *Computational and Structural Biotechnology Journal*, 18:4048–4062, 2020.

S. Bordel, R. Agren, and J. Nielsen. Sampling the solution space in genome-scale metabolic networks reveals transcriptional regulation in key enzymes. *PLOS Computational Biology*, 6(7):1–13, 07 2010. doi: 10.1371/journal.pcbi.1000859. URL https://doi.org/10.1371/journal.pcbi.1000859.

E. Brunk, S. Sahoo, D. C. Zielinski, A. Altunkaya, A. Dräger, N. Mih, F. Gatto, A. Nilsson, G. A. P. Gonzalez, M. K. Aurich, et al. Recon3D enables a three-dimensional view of gene variation in human metabolism. *Nature biotechnology*, 36(3):272, 2018.

A. Cakmak, X. Qi, A. E. Cicek, I. Bederman, L. Henderson, M. Drumm, and G. Ozsoyoglu. A new metabolomics analysis technique: steady-state metabolic network dynamics analysis. *Journal of bioinformatics and computational biology*, 10(01):1240003, 2012.

L. Calès, A. Chalkis, I. Z. Emiris, and V. Fisikopoulos. Practical Volume Computation of Structured Convex Bodies, and an Application to Modeling Portfolio Dependencies and Financial Crises. In B. Speckmann and C. D. Tóth, editors, *34th International Symposium on Computational Geometry (SoCG 2018)*, volume 99 of *LIPIcs*, pages 19:1–19:15, Dagstuhl, Germany, 2018. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. doi: 10.4230/LIPIcs.SoCG.2018.19.

A. Chalkis and V. Fisikopoulos. volesti: Volume approximation and sampling for convex polytopes in R, 2020. https://github.com/GeomScale/volume_approximation.

A. Chalkis, I. Z. Emiris, and V. Fisikopoulos. Practical volume estimation of zonotopes by a new annealing schedule for cooling convex bodies. In A. M. Bigatti, J. Carette, J. H. Davenport, M. Joswig, and T. de Wolff, editors, *Mathematical Software – ICMS 2020*, pages 212–221, Cham, 2020. Springer International Publishing. ISBN 978-3-030-52200-1.

Y. Chen, R. Dwivedi, M. J. Wainwright, and B. Yu. Fast mcmc sampling algorithms on polytopes. *Journal of Machine Learning Research*, 19(55):1–86, 2018. URL http://jmlr.org/papers/v19/18-158.html.

J. Cheney, S. Chong, N. Foster, M. Seltzer, and S. Vansummeren. Provenance: a future history. In *Proceedings of the 24th ACM SIGPLAN conference companion on Object oriented programming systems languages and applications*, pages 957–964, 2009.

A. Chevallier, S. Pion, and F. Cazals. Hamiltonian Monte Carlo with boundary reflections, and application to polytope volume calculations. Research Report RR-9222, INRIA Sophia Antipolis, France, 2018. URL https://hal.archives-ouvertes.fr/hal-01919855.

B. Cousins. *Efficient high-dimensional sampling and integration*. PhD thesis, Georgia Institute of Technology, Georgia, U.S.A., 2017.

B. Cousins and S. Vempala. A practical volume algorithm. *Mathematical Programming Computation*, 8(2):133–160, 2016.

E. Deelman, B. Berriman, A. Chervenak, O. Corcho, P. Groth, and L. Moreau. Metadata and provenance management. *arXiv preprint arXiv:1005.2643*, 2010.

A. B. Dieker and S. S. Vempala. Stochastic billiards for sampling from the boundary of a convex set. *Mathematics of Operations Research*, 40(4):888–901, 2015. ISSN 0364765X, 15265471. URL http://www.jstor.org/stable/24540983.

C. Diener, S. M. Gibbons, and O. Resendis-Antonio. Micom: metagenome-scale modeling to infer metabolic interactions in the gut microbiota. *MSystems*, 5(1):e00606–19, 2020.

I. Dukovski, D. Bajić, J. M. Chacón, M. Quintin, J. C. Vila, S. Sulheim, A. R. Pacheco, D. B. Bernstein, W. J. Riehl, K. S. Korolev, et al. A metabolic modeling platform for the computation of microbial ecosystems in time and space (comets). *Nature protocols*, 16 (11):5030–5082, 2021.

E. M. Eckert, A. Di Cesare, D. Fontaneto, T. U. Berendonk, H. Bürgmann, E. Cytryn, D. Fatta-Kassinos, A. Franzetti, D. J. Larsson, C. M. Manaia, et al. Every fifth published metagenome is not available to science. *PLoS biology*, 18(4):e3000698, 2020.

S. Fallahi, H. J. Skaug, and G. Alendal. A comparison of Monte Carlo sampling methods for metabolic network models. *PLOS One*, 15(7):e0235393, 2020.

R. Feldbauer, F. Schulz, M. Horn, and T. Rattei. Prediction of microbial phenotypes based on comparative genomics. *BMC bioinformatics*, 16(14):1–8, 2015.

A. Franceschini, D. Szklarczyk, S. Frankild, M. Kuhn, M. Simonovic, A. Roth, J. Lin, P. Minguez, P. Bork, C. Von Mering, et al. String v9. 1: protein-protein interaction networks, with increased coverage and integration. *Nucleic acids research*, 41(D1): D808–D815, 2012.

J. Freire, D. Koop, E. Santos, and C. T. Silva. Provenance for computational tasks: A survey. *Computing in Science & Engineering*, 10(3):11–21, 2008.

A. Gelman and D. B. Rubin. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7(4):457–472, 1992. ISSN 0883-4237. URL https://www.jstor.org/stable/2246093. Publisher: Institute of Mathematical Statistics.

C. J. Geyer. Practical Markov chain Monte Carlo. *Statist. Sci.*, 7(4):473–483, 11 1992. doi: 10.1214/ss/1177011137. URL https://doi.org/10.1214/ss/1177011137.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL http://jmlr.org/papers/v13/gretton12a.html.

E. Gryazina and B. Polyak. Random sampling: Billiard walk algorithm. *European Journal of Operational Research*, 238(2):497 – 504, 2014. ISSN 0377-2217. doi: https://doi.org/10.1016/j.ejor.2014.03.041.

G. Guennebaud, B. Jacob, et al. *Eigen v3*, 2010. URL http://eigen.tuxfamily.org.

H. S. Haraldsdóttir, B. Cousins, I. Thiele, R. M. Fleming, and S. Vempala. CHRR: coordinate hit-and-run with rounding for uniform sampling of constraint-based models. *Bioinformatics*, 33(11):1741–1743, 01 2017. ISSN 1367-4803. doi: 10.1093/bioinformatics/btx052.

K. I. Harrington and A. Sanchez. Eco-evolutionary dynamics of complex social strategies in microbial communities. *Communicative & integrative biology*, 7(1):e28230, 2014.

M. Heinonen, M. Osmala, H. Mannerström, J. Wallenius, S. Kaski, J. Rousu, and H. Lähdesmäki. Bayesian metabolic flux analysis reveals intracellular flux couplings. *Bioinformatics*, 35(14):i548–i557, 2019.

L. Heirendt, S. Arreckx, T. Pfau, S. N. Mendoza, A. Richelle, A. Heinken, H. S. Haraldsdóttir, J. Wachowiak, S. M. Keating, V. Vlasov, et al. Creation and analysis of biochemical constraint-based models using the cobra toolbox v. 3.0. *Nature protocols*, 14(3):639–702, 2019.

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*, 5(1):1–8, 2019.

T. C. A. Hitch, T. Riedel, A. Oren, J. Overmann, T. D. Lawley, and T. Clavel. Automated analysis of genomic sequences facilitates high-throughput and comprehensive description of bacteria. *ISME Communications*, 1(1):1–16, May 2021. ISSN 2730-6151. doi: 10.1038/s43705-021-00017-z. URL https://www.nature.com/articles/s43705-021-00017-z. Number: 1 Publisher: Nature Publishing Group.

T. Ideker, T. Galitski, and L. Hood. A new approach to decoding life: systems biology. *Annual review of genomics and human genetics*, 2(1):343–372, 2001.

J. F. Jadebeck, A. Theorell, S. Leweke, and K. Noh. Hops: high-performance library for non uniform sampling of convex constrained models. *Bioinformatics*, 2020.

P. Jeraldo, N. Chia, and N. Goldenfeld. On the suitability of short reads of 16s rrna for phylogeny-based analyses in environmental surveys. *Environmental microbiology*, 13 (11):3000–3009, 2011.

X. Jing, H. Gou, Y. Gong, X. Su, L. Xu, Y. Ji, Y. Song, I. P. Thompson, J. Xu, and W. E. Huang. Raman-activated cell sorting and metagenomic sequencing revealing carbon-fixing bacteria in the ocean. *Environmental microbiology*, 20(6):2241–2255, 2018.

F. John. Extremum Problems with Inequalities as Subsidiary Conditions. In G. Giorgi and T. H. Kjeldsen, editors, *Traces and Emergence of Nonlinear Programming*, pages 197–215. Springer, Basel, 2014. ISBN 978-3-0348-0439-4. doi: 10.1007/978-3-0348-0439-4_9. URL https://doi.org/10.1007/978-3-0348-0439-4_9.

J. S. Johnson, D. J. Spakowicz, B.-Y. Hong, L. M. Petersen, P. Demkowicz, L. Chen, S. R. Leopold, B. M. Hanson, H. O. Agresta, M. Gerstein, E. Sodergren, and G. M. Weinstock. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nature Communications*, 10(1):5029, Nov. 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-13036-1. URL https://www.nature.com/articles/s41467-019-13036-1. Number: 1 Publisher: Nature Publishing Group.

M. C. Jones, J. S. Marron, and S. J. Sheather. A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, 91(433):401–407, 1996. ISSN 01621459. URL http://www.jstor.org/stable/2291420.

A. T. Kalai and S. Vempala. Simulated annealing for convex optimization. *Mathematics of Operations Research*, 31(2):253–266, 2006. ISSN 0364765X, 15265471. URL http://www.jstor.org/stable/25151723.

D. E. Kaufman and R. L. Smith. Direction choice for accelerated convergence in hit-and-run sampling. *Operations Research*, 46(1):84–95, 1998.

Z. A. King, J. Lu, A. Drager, P. Miller, S. Federowicz, J. A. Lerman, A. Ebrahim, B. O. Palsson, and N. E. Lewis. Bigg models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic acids research*, 44(D1):D515–D522, 2016.

E. Klipp, W. Liebermeister, C. Wierling, and A. Kowald. *Systems biology: a textbook*. John Wiley & Sons, 2016.

P. Kohl, E. J. Crampin, T. Quinn, and D. Noble. Systems biology: an approach. *Clinical Pharmacology & Therapeutics*, 88(1):25–33, 2010.

C.-H. Kuo and H. Ochman. The extinction dynamics of bacterial pseudogenes. *PLoS genetics*, 6(8):e1001050, 2010.

A. Laddha and S. Vempala. Convergence of Gibbs Sampling: Coordinate Hit-and-Run Mixes Fast, 2020.

Y. T. Lee and S. S. Vempala. Convergence rate of riemannian hamiltonian monte carlo and faster polytope volume computation. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2018, page 1115–1121, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355599. doi: 10.1145/3188745.3188774. URL https://doi.org/10.1145/3188745.3188774.

N. E. Lewis, H. Nagarajan, and B. O. Palsson. Constraining the metabolic genotype–phenotype relationship using a phylogeny of in silico methods. *Nature Reviews Microbiology*, 10(4):291–305, 2012.

S. Louca, L. W. Parfrey, and M. Doebeli. Decoupling function and taxonomy in the global ocean microbiome. *Science*, 353(6305):1272–1277, 2016.

L. Lovász, R. Kannan, and M. Simonovits. Random walks and an $O^*(n^5)$ volume algorithm for convex bodies. *Random Structures and Algorithms*, 11:1–50, 1997.

L. Lovász and S. Vempala. Simulated annealing in convex bodies and an $O^*(n^4)$ volume algorithms. *J. Computer & System Sciences*, 72:392–417, 2006.

W. Ludwig, T. Viver, R. Westram, J. Francisco Gago, E. Bustos-Caparros, K. Knittel, R. Amann, and R. Rossello-Mora. Release LTP_12_2020, featuring a new ARB alignment and improved 16S rRNA tree for prokaryotic type strains. *Systematic and Applied Microbiology*, 44(4):126218, July 2021. ISSN 0723-2020. doi: 10.1016/j.syapm.2021.126218. URL https://www.sciencedirect.com/science/article/pii/S0723202021000412.

M. Lularevic, A. J. Racher, C. Jaques, and A. Kiparissides. Improving the accuracy of flux balance analysis through the implementation of carbon availability constraints for intracellular reactions. *Biotechnology and bioengineering*, 116(9):2339–2352, 2019.

M. MacGillivray, A. Ko, E. Gruber, M. Sawyer, E. Almaas, and A. Holder. Robust analysis of fluxes in genome-scale metabolic pathways. *Scientific Reports*, 7, 12 2017. doi: 10.1038/s41598-017-00170-3.

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*, 46(15):7542–7553, 2018.

F. Meyer, A. Fritz, Z.-L. Deng, D. Koslicki, T. R. Lesker, A. Gurevich, G. Robertson, M. Alser, D. Antipov, F. Beghini, et al. Critical assessment of metagenome interpretation: the second round of challenges. *Nature Methods*, pages 1–12, 2022.

A. L. Mitchell, A. Almeida, M. Beracochea, M. Boland, J. Burgin, G. Cochrane, M. R. Crusoe, V. Kale, S. C. Potter, L. J. Richardson, et al. Mgnify: the microbiome analysis resource in 2020. *Nucleic acids research*, 48(D1):D570–D578, 2020.

H. Narayanan and P. Srivastava. On the mixing time of coordinate hit-and-run, 2020.

D. Noble. *The music of life: biology beyond genes*. Oxford University Press, 2008.

E. Noor, S. Cherkaoui, and U. Sauer. Biological insights through omics data integration. *Current Opinion in Systems Biology*, 15:39–47, 2019.

A. Noronha, J. Modamio, Y. Jarosz, E. Guerard, N. Sompairac, G. Preciat, A. D. Daníelsdóttir, M. Krecke, D. Merten, H. S. Haraldsdóttir, et al. The virtual metabolic human database: integrating human and gut microbiome metabolism with nutrition and disease. *Nucleic acids research*, 47(D1):D614–D624, 2019.

J. D. Orth, I. Thiele, and B. Ø.. Palsson. What is flux balance analysis? *Nature biotechnology*, 28(3):245–248, 2010.

M. J. Pallen, A. Telatin, and A. Oren. The next million names for archaea and bacteria. *Trends in Microbiology*, 29(4):289–298, 2021.

B. Ø.. Palsson. Metabolic systems biology. *FEBS letters*, 583(24):3900–3904, 2009.

B. Ø.. Palsson. *Systems biology*. Cambridge university press, 2015.

D. H. Parks, M. Chuvochina, C. Rinke, A. J. Mussig, P.-A. Chaumeil, and P. Hugenholtz. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Research*, 50(D1):D785–D794, Jan. 2022. ISSN 0305-1048. doi: 10.1093/nar/gkab776. URL https://doi.org/10.1093/nar/gkab776.

A. C. Parte, J. S. Carbasse, J. P. Meier-Kolthoff, L. C. Reimer, and M. Göker. List of prokaryotic names with standing in nomenclature (lpsn) moves to the dsmz. *International journal of systematic and evolutionary microbiology*, 70(11):5607, 2020.

C. Pavloudi, A. Oulas, K. Vasileiadou, G. Kotoulas, M. De Troch, M. W. Friedrich, and C. Arvanitidis. Diversity and abundance of sulfate-reducing microorganisms in a mediterranean lagoonal complex (amvrakikos gulf, ionian sea) derived from dsrb gene. *Aquatic Microbial Ecology*, 79(3):209–219, 2017.

A. Y. Pei, W. E. Oberdorf, C. W. Nossa, A. Agarwal, P. Chokshi, E. A. Gerz, Z. Jin, P. Lee, L. Yang, M. Poles, et al. Diversity of 16s rrna genes within individual prokaryotic genomes. *Applied and environmental microbiology*, 76(12):3886–3897, 2010.

O. Perez-Garcia, G. Lear, and N. Singhal. Metabolic network modeling of microbial interactions in natural and engineered environmental systems. *Frontiers in microbiology*, 7: 673, 2016.

T. M. Porter and M. Hajibabaei. Profile hidden markov model sequence analysis can help remove putative pseudogenes from dna barcoding and metabarcoding datasets. *BMC bioinformatics*, 22(1):1–20, 2021.

Q.-L. Qin, B.-B. Xie, X.-Y. Zhang, X.-L. Chen, B.-C. Zhou, J. Zhou, A. Oren, and Y.-Z. Zhang. A proposed genus boundary for the prokaryotes based on genomic insights. *Journal of Bacteriology*, 196(12):2210–2215, June 2014. ISSN 1098-5530. doi: 10.1128/JB.01688-14.

R. A. Quinn, J. A. Navas-Molina, E. R. Hyde, S. J. Song, Y. Vázquez-Baeza, G. Humphrey, J. Gaffney, J. J. Minich, A. V. Melnik, J. Herschend, J. DeReus, A. Durant, R. J. Dutton, M. Khosroheidari, C. Green, R. da Silva, P. C. Dorrestein, and R. Knight. From sample to multi-omics conclusions in under 48 hours. msystems 1: e00038-16. *Crossref, Medline*, 2016.

V. Roy. Convergence Diagnostics for Markov Chain Monte Carlo. *Annual Review of Statistics and Its Application*, 7(1):387–412, 2020. doi: 10.1146/annurev-statistics-031219-041300.

P. A. Saa and L. K. Nielsen. ll-ACHRB: a scalable algorithm for sampling the feasible solution space of metabolic networks. *Bioinform.*, 32(15):2330–2337, 2016. doi: 10.1093/bioinformatics/btw132. URL https://doi.org/10.1093/bioinformatics/btw132.

R. M. Samuel, R. Meyer, P. L. Buttigieg, N. Davies, N. W. Jeffery, C. Meyer, C. Pavloudi, K. Johnson Pitz, M. Sweetlove, S. Theroux, et al. Towards a global public repository of community protocols to encourage best practices in biomolecular ocean observing and research. *Frontiers in Marine Science*, page 1488, 2021.

J. Schellenberger and B. Ø.. Palsson. Use of randomized sampling for analysis of metabolic networks. *Journal of biological chemistry*, 284(9):5457–5461, 2009.

J. R. Schramski, A. I. Dell, J. M. Grady, R. M. Sibly, and J. H. Brown. Metabolic theory predicts whole-ecosystem properties. *Proceedings of the National Academy of Sciences*, 112(8):2617–2622, 2015.

W. T. Scott, E. J. Smid, D. E. Block, and R. A. Notebaart. Metabolic flux sampling predicts strain-dependent differences related to aroma production among commercial wine yeasts. *Microbial cell factories*, 20(1):1–15, 2021.

T. Seemann. Barrnap: BAsic Rapid Ribosomal RNA Predictor, 2014a. URL https://github.com/tseemann/barrnap.

T. Seemann. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30(14): 2068–2069, July 2014b. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu153. URL https://doi.org/10.1093/bioinformatics/btu153.

S. S. Shishvan, A. Vigliotti, and V. S. Deshpande. The homeostatic ensemble for cells. *Biomechanics and Modeling in Mechanobiology*, 17(6):1631–1662, 2018.

R. L. Smith. Efficient Monte Carlo procedures for generating points uniformly distributed over bounded regions. *Operations Research*, 32(6):1296–1308, 1984. ISSN 0030364X, 15265463.

H. Song, J. E. Buhay, M. F. Whiting, and K. A. Crandall. Many species in one: Dna barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. *Proceedings of the national academy of sciences*, 105(36):13486–13491, 2008.

N. Swainston, K. Smallbone, H. Hefzi, P. D. Dobson, J. Brewer, M. Hanscho, D. C. Zielinski, K. S. Ang, N. J. Gardiner, J. M. Gutierrez, S. Kyriakopoulos, M. Lakshmanan, S. Li, J. K. Liu, V. S. Martínez, C. A. Orellana, L.-E. Quek, A. Thomas, J. Zanghellini, N. Borth, D.-Y. Lee, L. K. Nielsen, D. B. Kell, N. E. Lewis, and P. Mendes. Recon 2.2: from reconstruction to model of human metabolism. *Metabolomics*, 12(7):109, June 2016. ISSN 1573-3890. doi: 10.1007/s11306-016-1051-4. URL https://doi.org/10.1007/s11306-016-1051-4.

L. Tedersoo, M. Albertsen, S. Anslan, and B. Callahan. Perspectives and benefits of high-throughput long-read sequencing in microbial ecology. *Applied and Environmental Microbiology*, 87(17):e00626–21, 2021.

I. Thiele and B. Ø.. Palsson. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature protocols*, 5(1):93, 2010.

P. Vangay, J. Burgin, A. Johnston, K. L. Beck, D. C. Berrios, K. Blumberg, S. Canon, P. Chain, J.-M. Chandonia, D. Christianson, et al. Microbiome metadata standards: Report of the national microbiome data collaborative's workshop and follow-on activities. *Msystems*, 6(1):e01194–20, 2021.

T. Ward, J. Larson, J. Meulemans, B. Hillmann, J. Lynch, D. Sidiropoulos, J. R. Spear, G. Caporaso, R. Blekhman, R. Knight, R. Fink, and D. Knights. Bugbase predicts

organism-level microbiome phenotypes. *bioRxiv*, 2017. doi: 10.1101/133462. URL https://www.biorxiv.org/content/early/2017/05/02/133462.

P. Yilmaz, R. Kottmann, D. Field, R. Knight, J. R. Cole, L. Amaral-Zettler, J. A. Gilbert, I. Karsch-Mizrachi, A. Johnston, G. Cochrane, et al. Minimum information about a marker gene sequence (mimarks) and minimum information about any (x) sequence (mixs) specifications. *Nature biotechnology*, 29(5):415–420, 2011.

H. Zafeiropoulos, H. Q. Viet, K. Vasileiadou, A. Potirakis, C. Arvanitidis, P. Topalis, C. Pavloudi, and E. Pafilis. Pema: a flexible pipeline for environmental dna metabarcoding analysis of the 16s/18s ribosomal rna, its, and coi marker genes. *GigaScience*, 9(3): giaa022, 2020.

H. Zafeiropoulos, A. Gioti, S. Ninidakis, A. Potirakis, S. Paragkamian, N. Angelova, A. Antoniou, T. Danis, E. Kaitetzidou, P. Kasapidis, J. B. Kristoffersen, V. Papadogiannis, C. Pavloudi, Q. V. Ha, J. Lagnel, N. Pattakos, G. Perantinos, D. Sidirokastritis, P. Vavilis, G. Kotoulas, T. Manousaki, E. Sarropoulou, C. S. Tsigenopoulos, C. Arvanitidis, A. Magoulas, and E. Pafilis. 0s and 1s in marine molecular research: a regional HPC perspective. *GigaScience*, 10(8):giab053, Aug. 2021. ISSN 2047-217X. doi: 10.1093/gigascience/giab053. URL https://doi.org/10.1093/gigascience/giab053.

H. Zafeiropoulos, S. Paragkamian, S. Ninidakis, G. A. Pavlopoulos, L. J. Jensen, and E. Pafilis. PREGO: A Literature and Data-Mining Resource to Associate Microorganisms, Biological Processes, and Environment Types. *Microorganisms*, 10(2):293, Feb. 2022. ISSN 2076-2607. doi: 10.3390/microorganisms10020293. URL https://www.mdpi.com/2076-2607/10/2/293. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

K. Zhuang, M. Izallalen, P. Mouser, H. Richter, C. Risso, R. Mahadevan, and D. R. Lovley. Genome-scale dynamic modeling of the competition between rhodoferax and geobacter in anoxic subsurface environments. *The ISME journal*, 5(2):305–316, 2011.

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], Ninidakis, S., Pavlopoulos, G. A., Jensen, L. J., and Pafilis, E. *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.[4], 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

Under review

- **Deciphering the functional potential of a hypersaline swamp microbial mat community** Pavloudi, C., **Zafeiropoulos, H.**
  Under reviw in *FEMS Microbiology Ecology*

[2] ZH and PS contibuted equally
[4] ZH and GA contributed equally

- **Automating the curation process of historical literature on marine biodiversity using text mining: the DECO workflow** Paragkamian, S., Sarafidou, G., Mavraki, D., .., **Zafeiropoulos H.**, Arvanitidis, C., Pafilis, E., Gerovasileiou, V.
  Under review in *Frontiers in Marine Science*

In preparation

- Metagenome assembled genomes of novel prokaryotic taxa from a hypersaline marsh microbial mat
- Metabolic models of human gut microbiota: what did we learn and what are the next steps
- dingo: a Python library for metabolic networks analysis

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

Personal website: https://hariszaf.github.io/
GitHub account: https://github.com/hariszaf
Twitter account: @haris_zaf
Account in ResearchGate
e-mail: haris.zafr@gmail.com