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

developing computational approaches to better understand microbial assemblages

Haris Zafeiropoulos
PhD candidate

1. Microbial ecology from a computational point-of-view

2. PEMA: a metabarcoding pipeline

3. PREGO: a knowledge-base for organisms - environments - processes associations

4. dingo: a Python library for metabolic flux sampling

5. Take home messages & future work

Microbial ecology & biogeochemical cycles

a corner-stone for life on earth

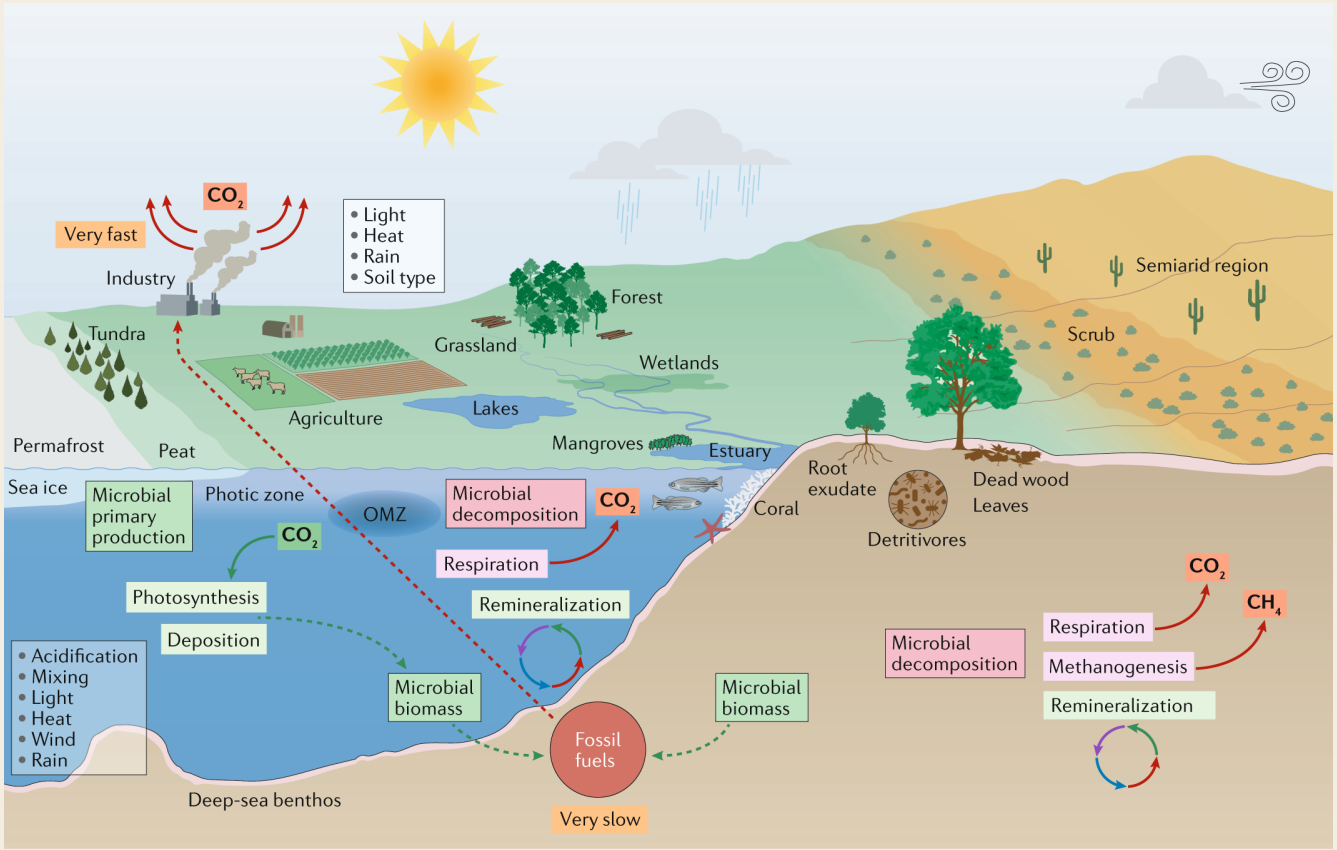


Figure from: Cavicchioli et al. Nature Reviews Microbiology 17.9 (2019): 569-586.

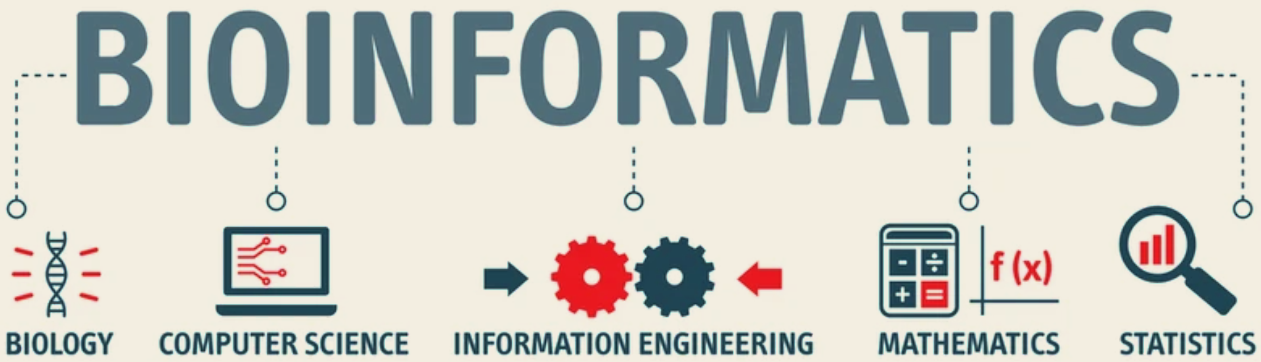
Questions to address

for a deeper understanding of microbial assemblages

Community structure ***who***	Functional potential ***what***	Microbial interactions ***why / how***
eveyrone is everywhere	*zero-sum game*	*the entagnled bank*

We are living in a computational era

both a challenge & an opportunity



1. Microbial ecology from a computational point-of-view

2. PEMA: a metabarcoding pipeline

3. PREGO: a knowledge-base for organisms - environments - processes associations

4. `dingo`: a Python library for metabolic flux sampling

5. Take home messages & future work

PEMA

a pipeline for eDNA metabarcoding analysis

https://github.com/hariszaf/pema

pema.hcmr.gr

eDNA metabarcoding

for biodiversity assessment

Marker genes

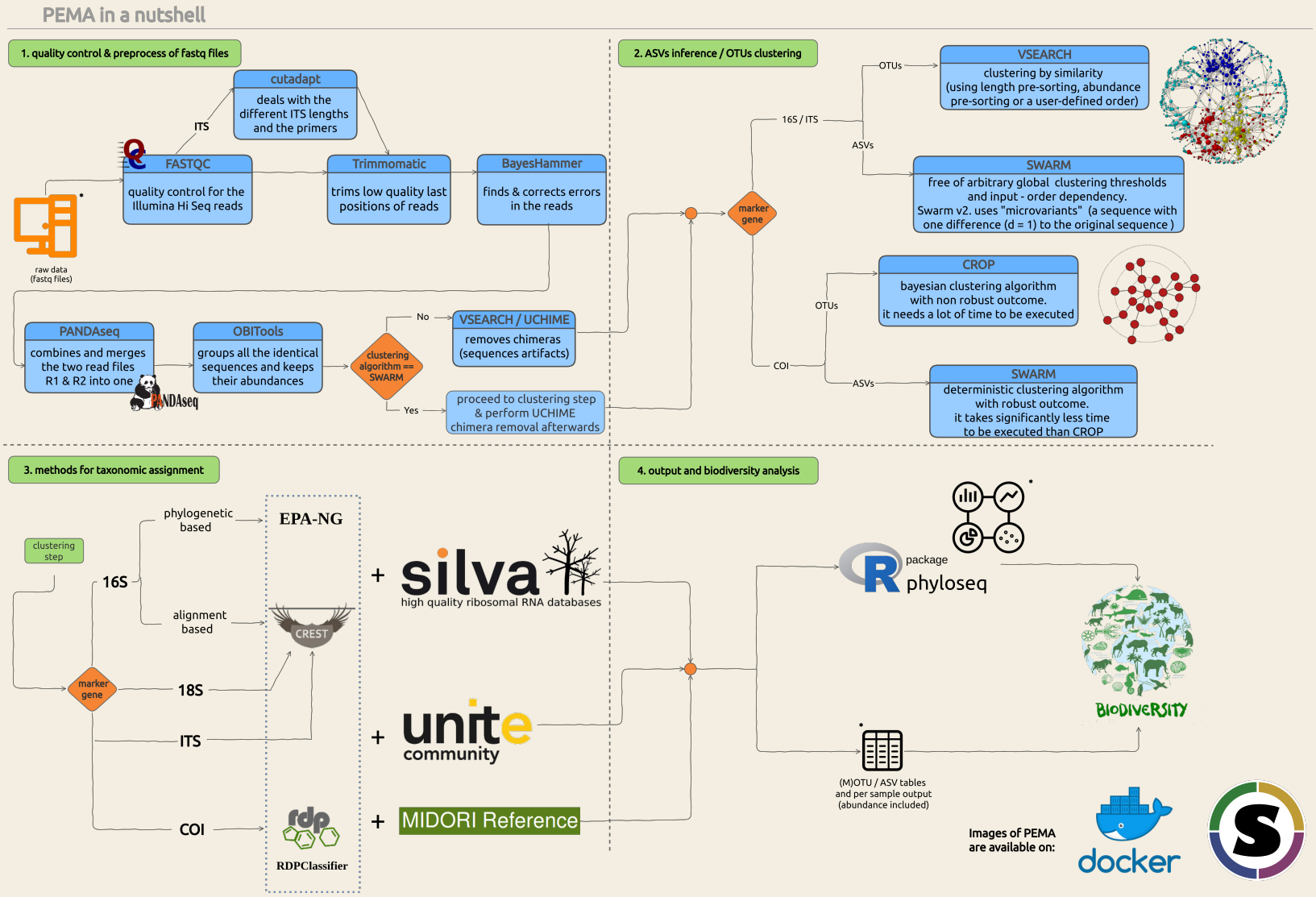
1. **16S rRNA:** Bacteria, Archaea
2. **12S rRNA:** Vertebrates
3. **18S rRNA:** Small eukaryotes, Metazoa
4. **ITS:** Fungi
5. **COI:** Eukaryotes
6. ***rbcl*:** Plants
7. ***dsrb*:** Bacteria, Archaea
8. ...

Methodology

- Sampling
- Extraction
- Bioinformatics
- Biodiversity analysis

PEMA features

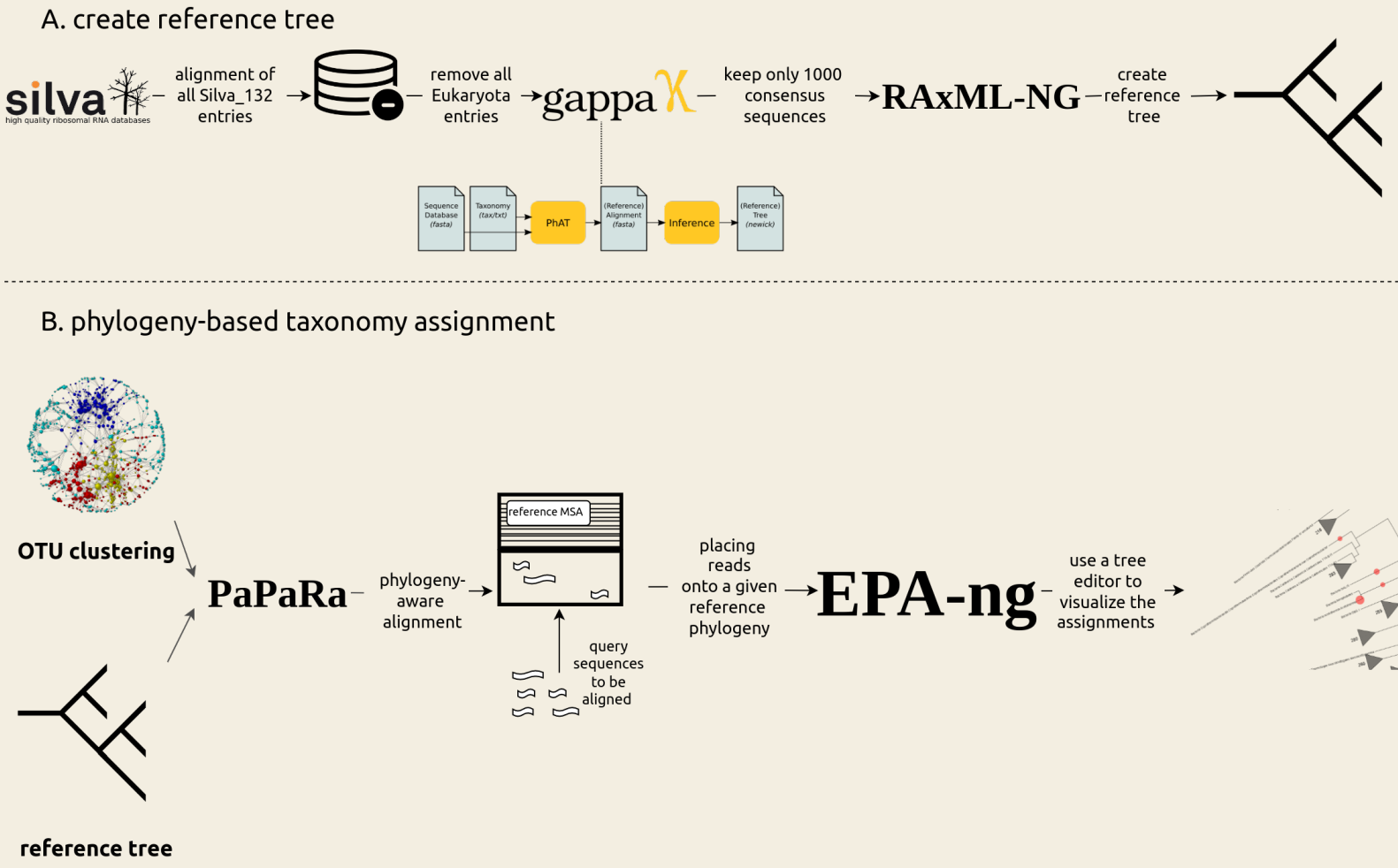
one step at a time!



* Illustrations with asterisk are from the Noun Project

PEMA features

one step at a time!



PEMA coding insights

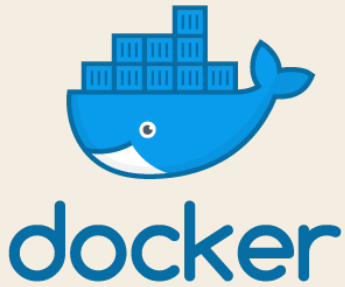
Being a geek just for a bit !

```
for(int i : range(1,
    in  := "in_$i.tx
    sys date > $in

    out := "out_$i.t
    task( out <- in
        sys echo Tas
    }
}
```

`BigDataScript`

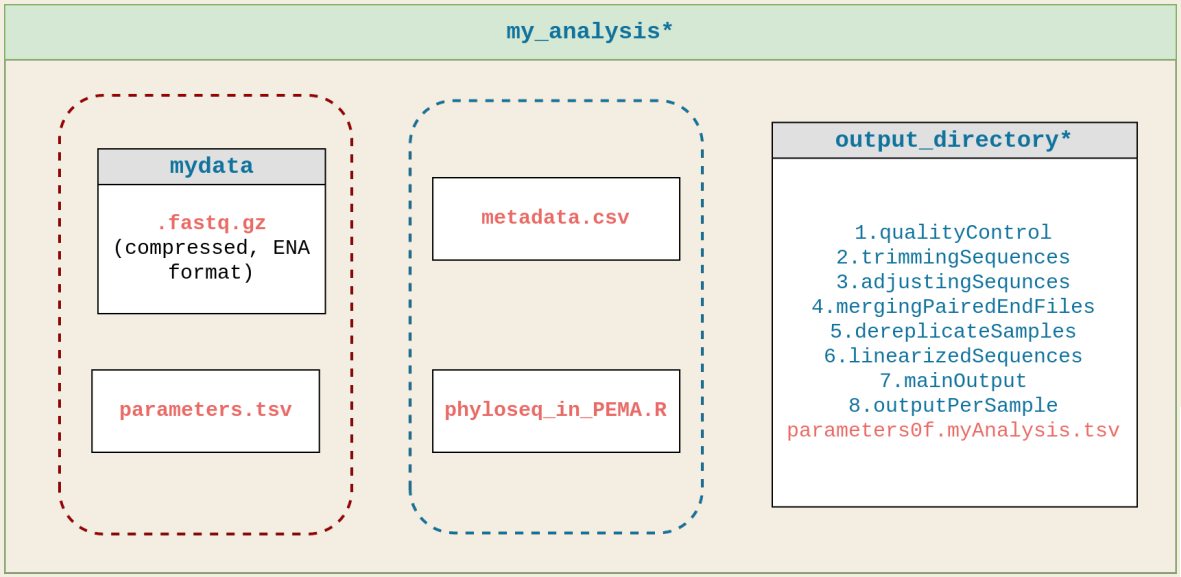
programming language



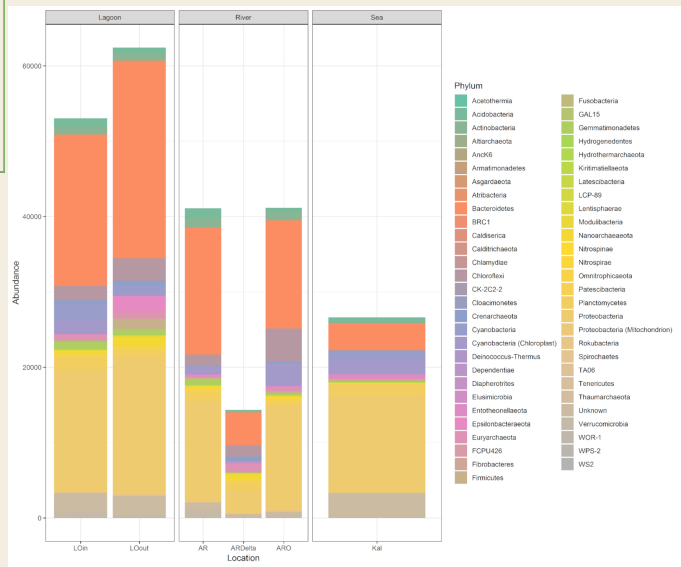
Containerization

Mount your I/O

give & take



	Sample 1	Sample 2	Sample 3	Sample 4
Taxon 1	1	0	1	2
Taxon 2	0	1	0	2
Taxon 3	1	1	0	4

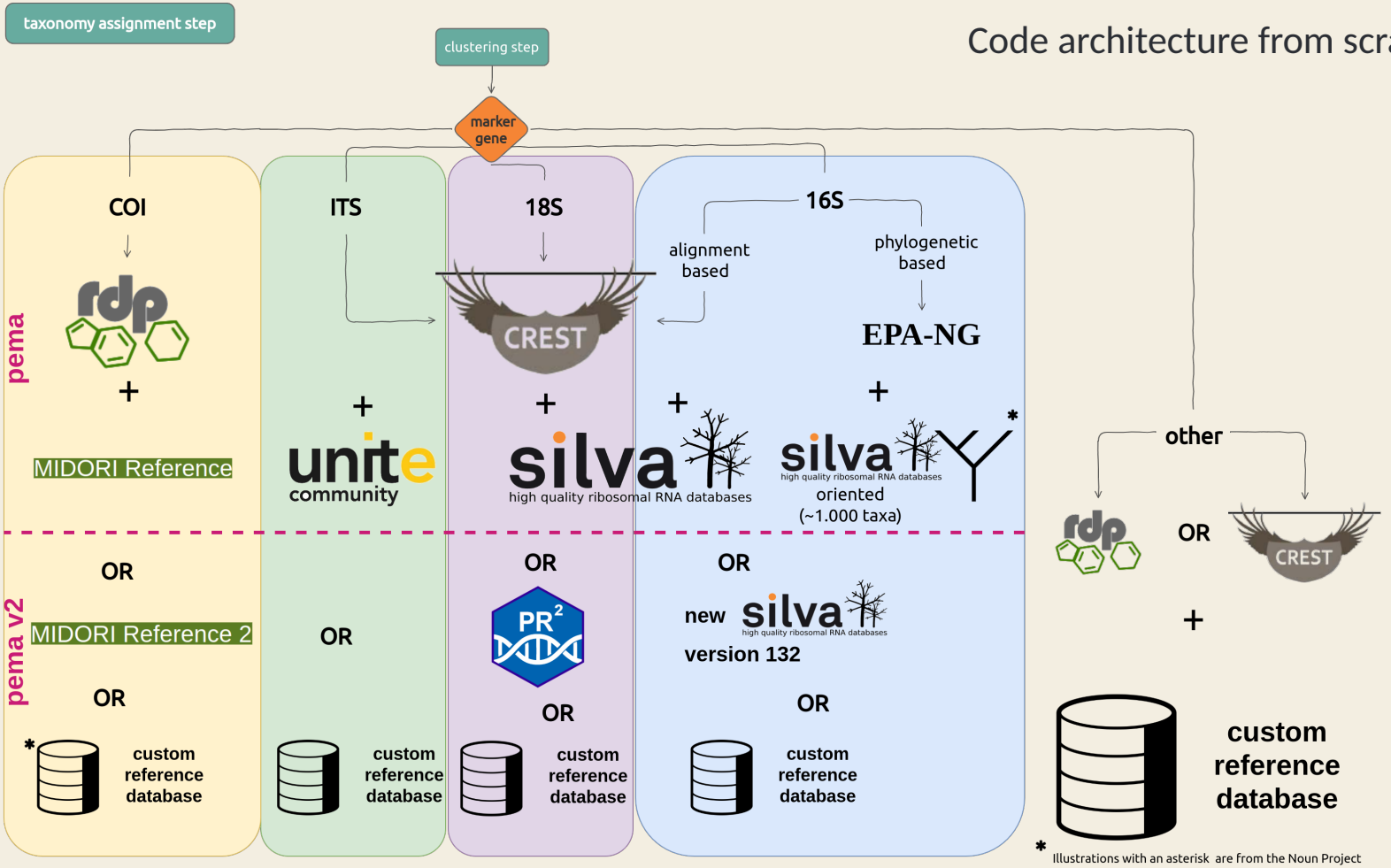


text file
directory
***user can edit the name**
(the rest **need** or will have the exact names shown)

- - - - - mandatory input files
- - - - - optional input files

PEMA v.2

addressing some of the challenges



Code architecture from scratch!

Latest PEMA version

addressing the challenges of the community

ASSEMBLE
ASSOCIATION OF EUROPEAN MARINE BIOLOGICAL LABORATORIES EXPANDED

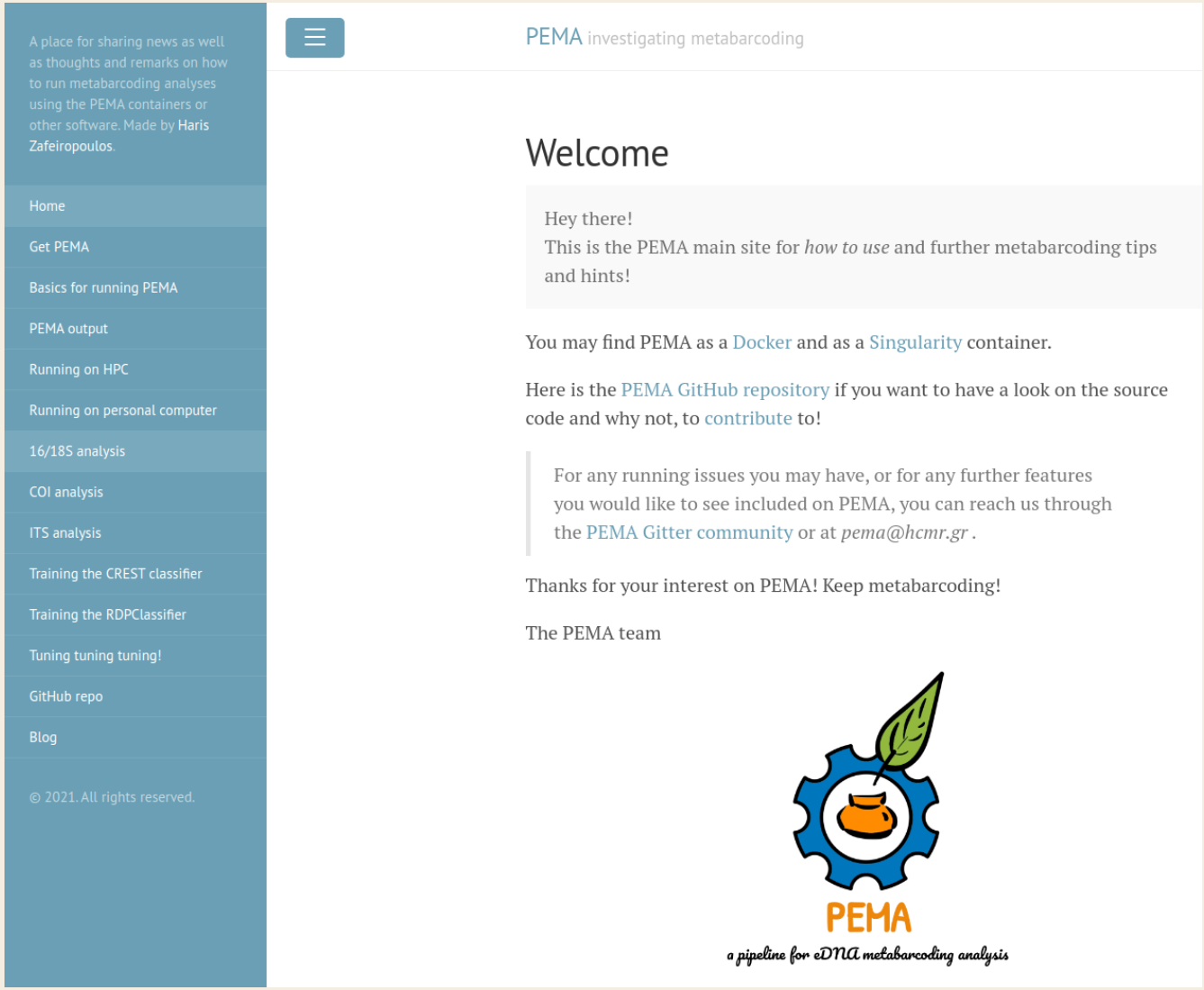
MBON
Marine Biodiversity Observation Network

`pema:v.2.1.4` includes:

1. analysis of 12S rRNA data now supported (12S Vertebrate Classifier v2.0.0-ref database)
2. PR2 as an alternative reference database for the case of 18S rRNA
3. the `ncbi-taxonomist` tool was added to return the NCBI Taxonomy Id of the taxonomies found

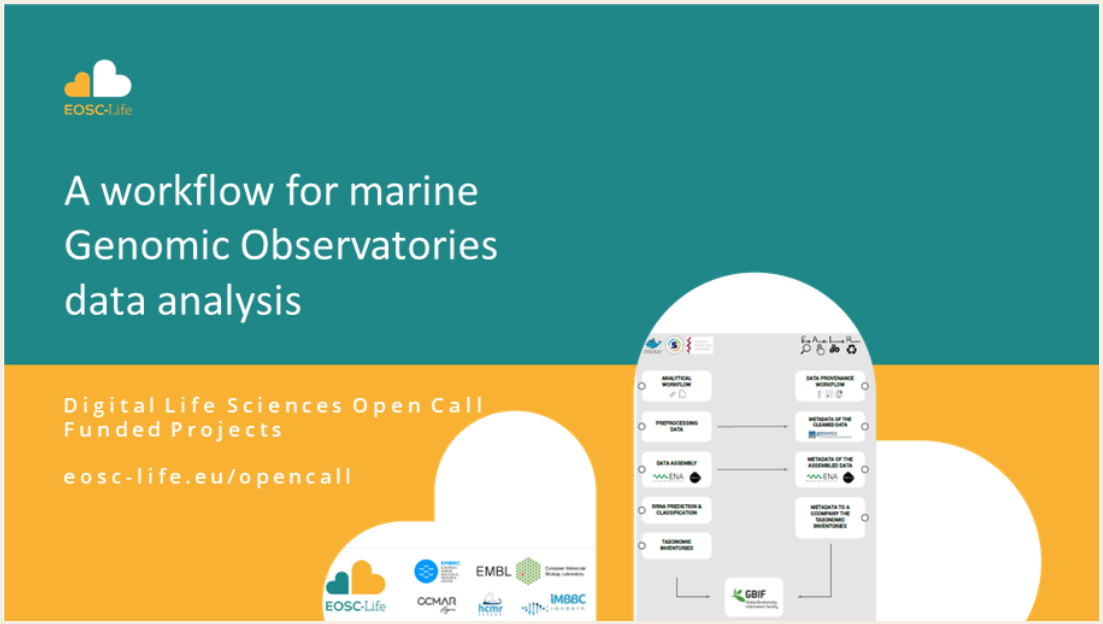
How to and further documentation

at pema.hcmr.gr



Metagenome go deeper than amplicon studies

yet come with vast computational challenges



Work in progress

To be available in the months to come.

You may find more about this project through its GitHub repository

1. Microbial ecology from a computational point-of-view

2. PEMA: a metabarcoding pipeline

3. PREGO: a knowledge-base for organisms - environments - processes associations

4. dingo: a Python library for metabolic flux sampling

5. Take home messages & future work

related repositories under

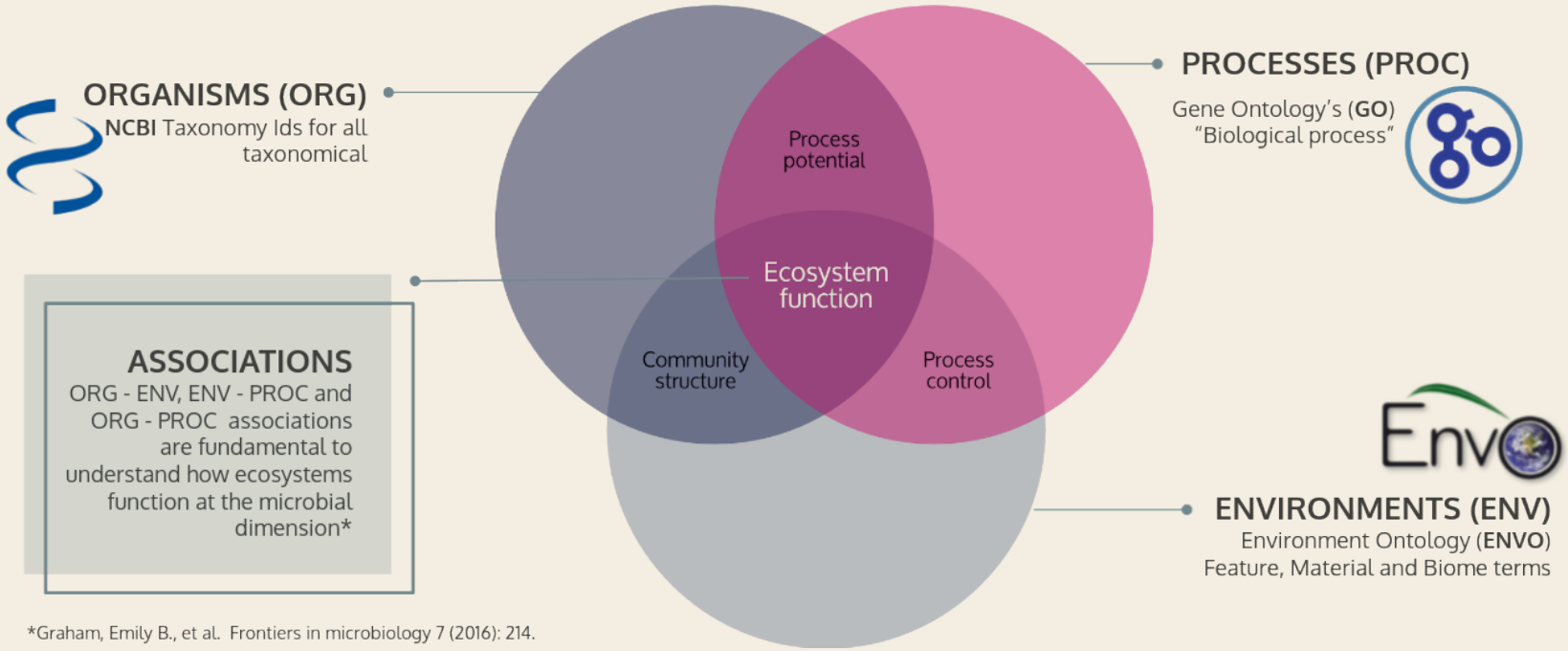
https://github.com/orgs/lab42open-team

web-app under

http://prego.hcmr.gr/

PREGO as *processes - environments - organisms*

and how to link them



*Graham, Emily B., et al. Frontiers in microbiology 7 (2016): 214.

Metadata example

from various metagenome repositories

Sample metadata [-]

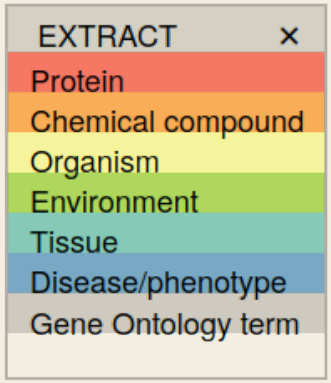
Collection date:	11/1/11
Elevation:	200
Environment (biome):	soil
Environment (feature):	nosZ
Environment (material):	soil DNA
Environmental package:	MIGS/MIMS/MIMARKS.soil
Geographic location (depth):	15-20cm
Instrument model:	454 GS FLX Titanium
Investigation type:	miens-survey
NCBI sample classification:	410658
Project name:	EcoFINDERS

Project Information	
Cultured	No
Ecosystem	Environmental
Ecosystem Category	Aquatic
Ecosystem Subtype	Oceanic
Ecosystem Type	Marine

MG-RAST ID	name	biome	feature	material	sample	library	location	country	coordinates	download
mgm4702467.3	06032015b_S2_L001_R2_001	Large lake biome	lake	water	mgs485560	mgl485562	Cincinnati	USA	39.11, -84.5	metadata submitted
mgm4702469.3	06052015a_S3_L001_R2_001	Large lake biome	lake	water	mgs485566	mgl485568	Cincinnati	USA	39.11, -84.5	metadata submitted
mgm4702471.3	06032015a_S1_L001_R2_001	Large lake biome	lake	water	mgs485554	mgl485556	Cincinnati	USA	39.11, -84.5	metadata submitted

Named Entity Recognition

tagging the literature



Identification of potentially important pathways missing from the model

From the metagenomic bins, we were able to identify two metabolic processes that were not previously included in the model. A number of MAGs (bin.59, bin.15, bin.73) clustered to the KEGG genomes of freshwater sulfur-oxidizing autotrophs capable of denitrification, *Sulfuritalea hydrogenivorans* [41], and *Sulfuricella denitrificans* [42]. These MAGs contained the diagnostic genes for carbon fixation (*rbcLS*), sulfur cycling (*dsrAB*), and denitrification (*nosZ*). One MAG (bin.59) also clustered with iron oxidizing autotroph *Sideroxydans lithotrophicus* ES-1. Bin.59 is the most relatively abundant bin from 17 to 21 m depth. Thus, if this MAG is associated with iron oxidation, it also contains sulfur-cycling genes that add to metabolic flexibility, which was previously observed [40]. The model did not include sulfide oxidation with nitrate, so it is unclear from the current model predictions where this process is expected to occur within the water column to compare to the MAG distributions.

Example text from Arora-Williams et al. Microbiome 6.1 (2018): 1-16.

PREGO methodology

co-occurrence again!

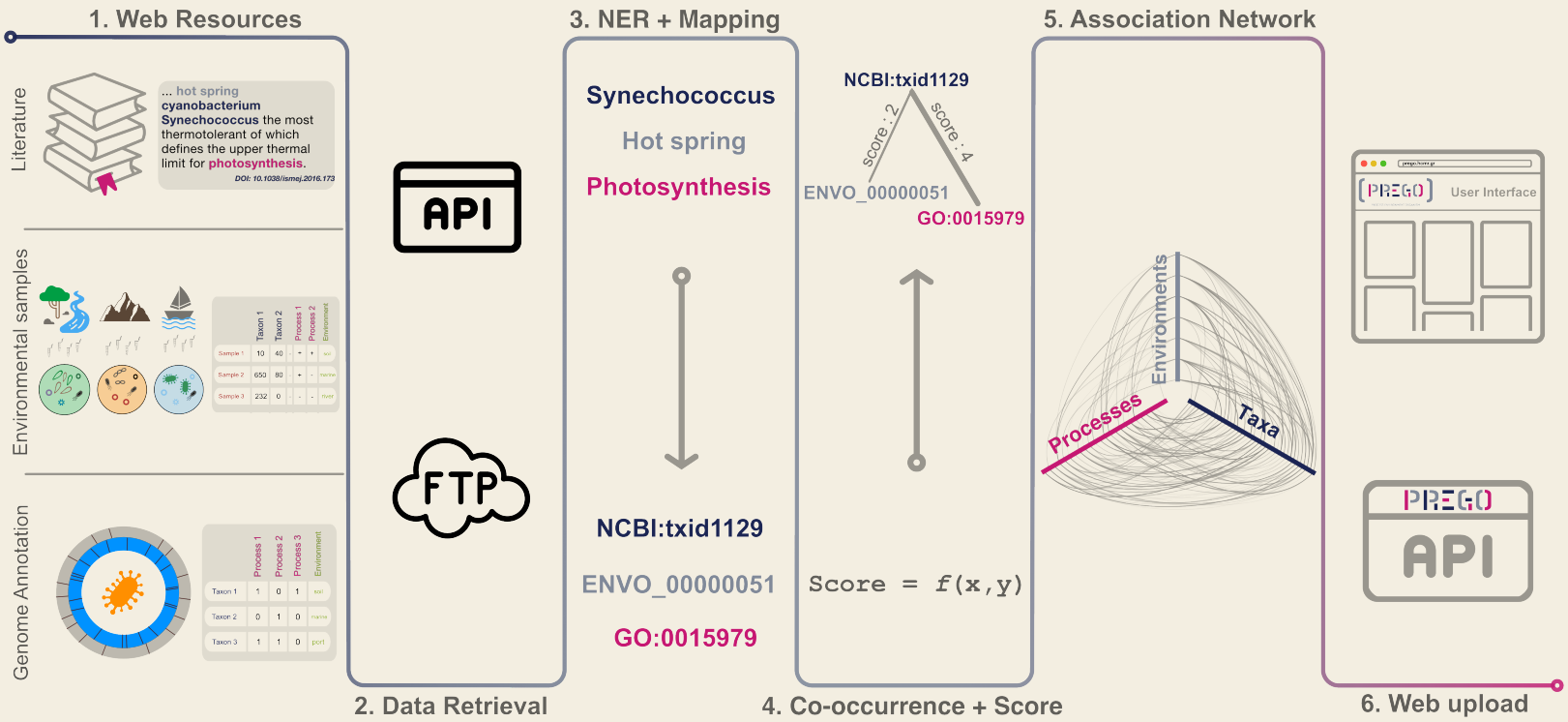
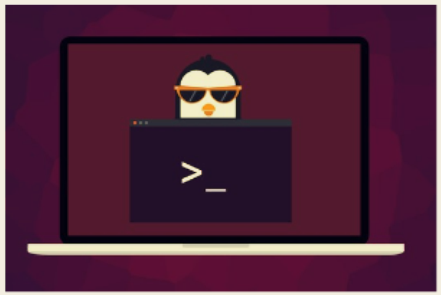


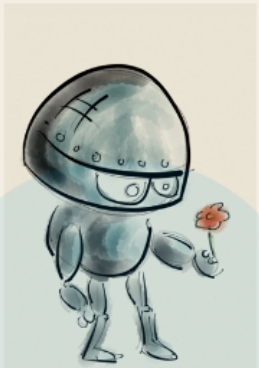
Figure from the PREGO publication that is now under review.

Building a knowledge-base

development and information technology operations

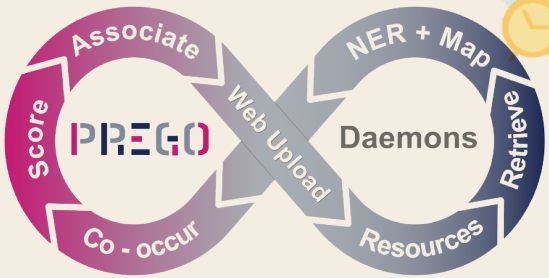


frequently updates;
`robots` allow to keep up with
literature and metagenomic data



building a
knowledge base (KB)

server dedicated
for building and
accessing the KB



What about taxa or functions of your interest?

or even environments!

Uranium bioremedation
by adding high
concentrations of acetate
What taxa support
"Response to acetate"
Have a look!
Geobacter is the first hit !

What about ***Posidonia*** ?
Literature suggests that Planctomycetes
and especially *Blastopirellula*
and *Rhodopirellula* are commonly
found in its microbiome.
Why so ?
You may have a look here

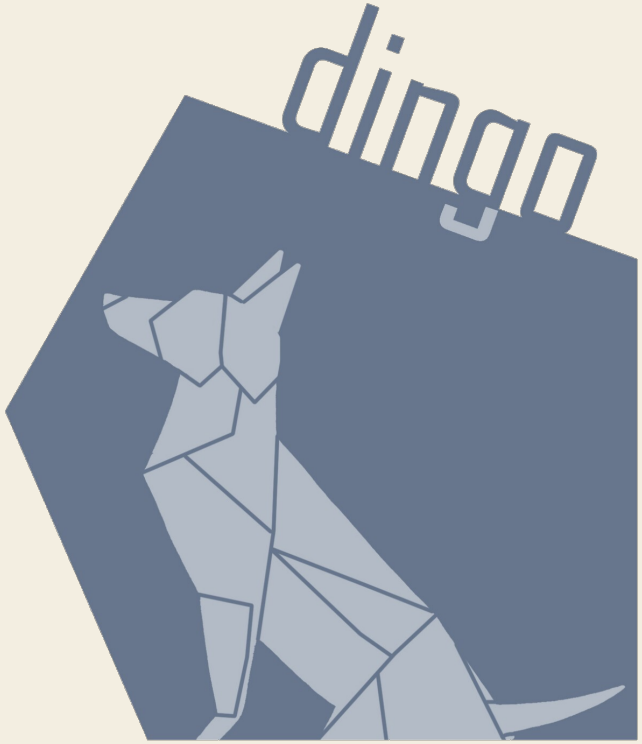
1. Microbial ecology from a computational point-of-view

2. PEMA: a metabarcoding pipeline

3. PREGO: a knowledge-base for organisms - environments - processes associations

4. dingo: a Python library for metabolic flux sampling

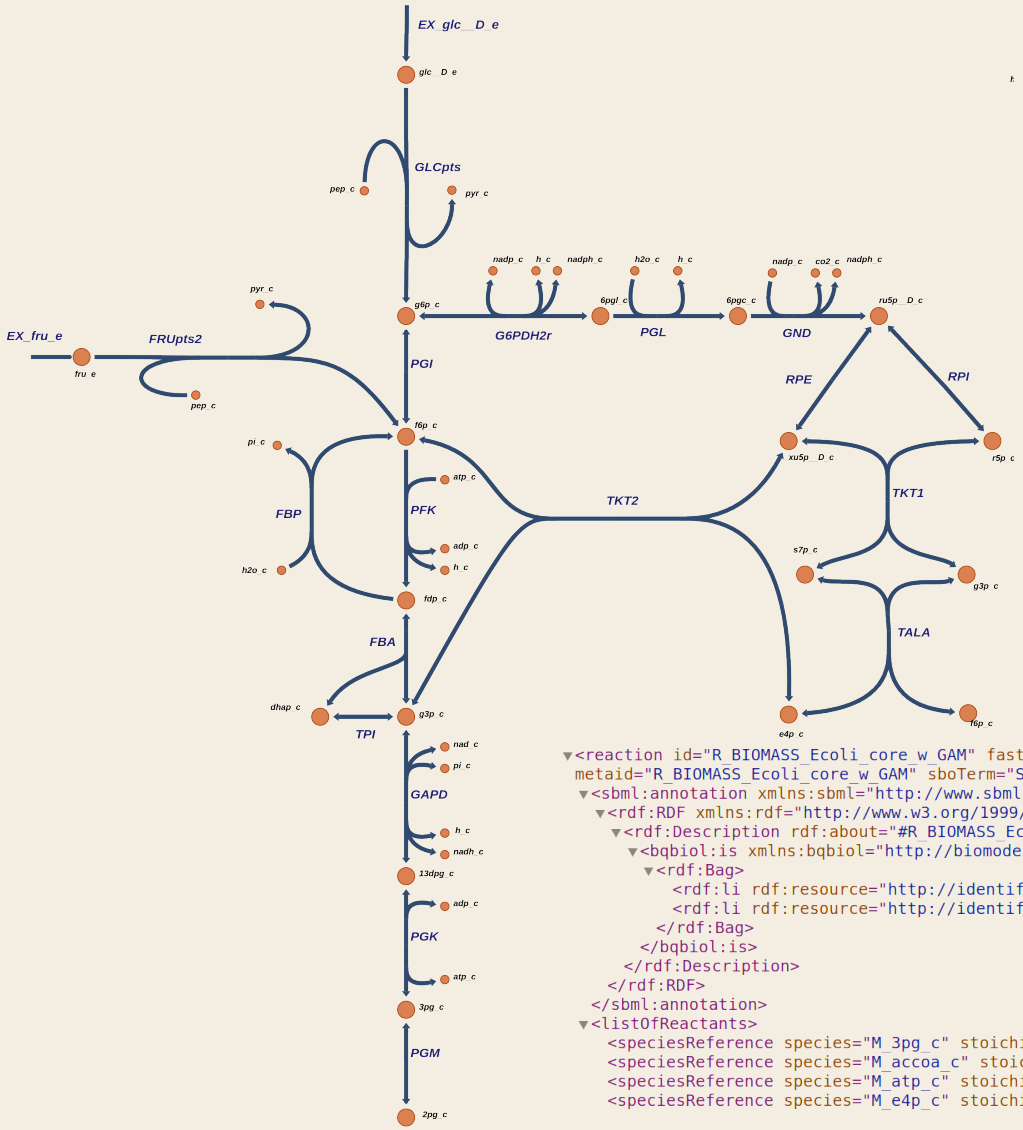
5. Take home messages & future work



https://github.com/GeomScale/dingo

Metabolic modelling

and the biomass function



Metabolic models allow us

to move from a metabolic map

to mathematical structures

the study of which may provide

fundamental biological insight

```
▼<reaction id="R_BIOMASS_Ecoli_core_w_GAM" fast="false" reversible="false" name="Biomass Objective Function with GAM"
 metaid="R_BIOMASS_Ecoli_core_w_GAM" sboTerm="SBO:0000629" fbc:upperFluxBound="cobra_default_ub" fbc:lowerFluxBound="cobra_0_bound">
 ▼<sbml:annotation xmlns:sbml="http://www.sbml.org/sbml/level3/version1/core">
  ▼<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
   ▼<rdf:Description rdf:about="#R_BIOMASS_Ecoli_core_w_GAM">
    ▼<bqbiol:is xmlns:bqbiol="http://biomodels.net/biology-qualifiers/">
     ▼<rdf:Bag>
        <rdf:li rdf:resource="http://identifiers.org/bigg.reaction/BIOMASS_Ecoli_core_w_GAM"/>
        <rdf:li rdf:resource="http://identifiers.org/metanetx.reaction/MNXR96280"/>
      </rdf:Bag>
     </bqbiol:is>
    </rdf:Description>
   </rdf:RDF>
  </sbml:annotation>
 ▼<listOfReactants>
    <speciesReference species="M_3pg_c" stoichiometry="1.496" constant="true"/>
    <speciesReference species="M_accoa_c" stoichiometry="3.7478" constant="true"/>
    <speciesReference species="M_atp_c" stoichiometry="59.81" constant="true"/>
    <speciesReference species="M_e4p_c" stoichiometry="0.361" constant="true"/>
```

Genome-scale metabolic reconstruction

approaches, pros and cons

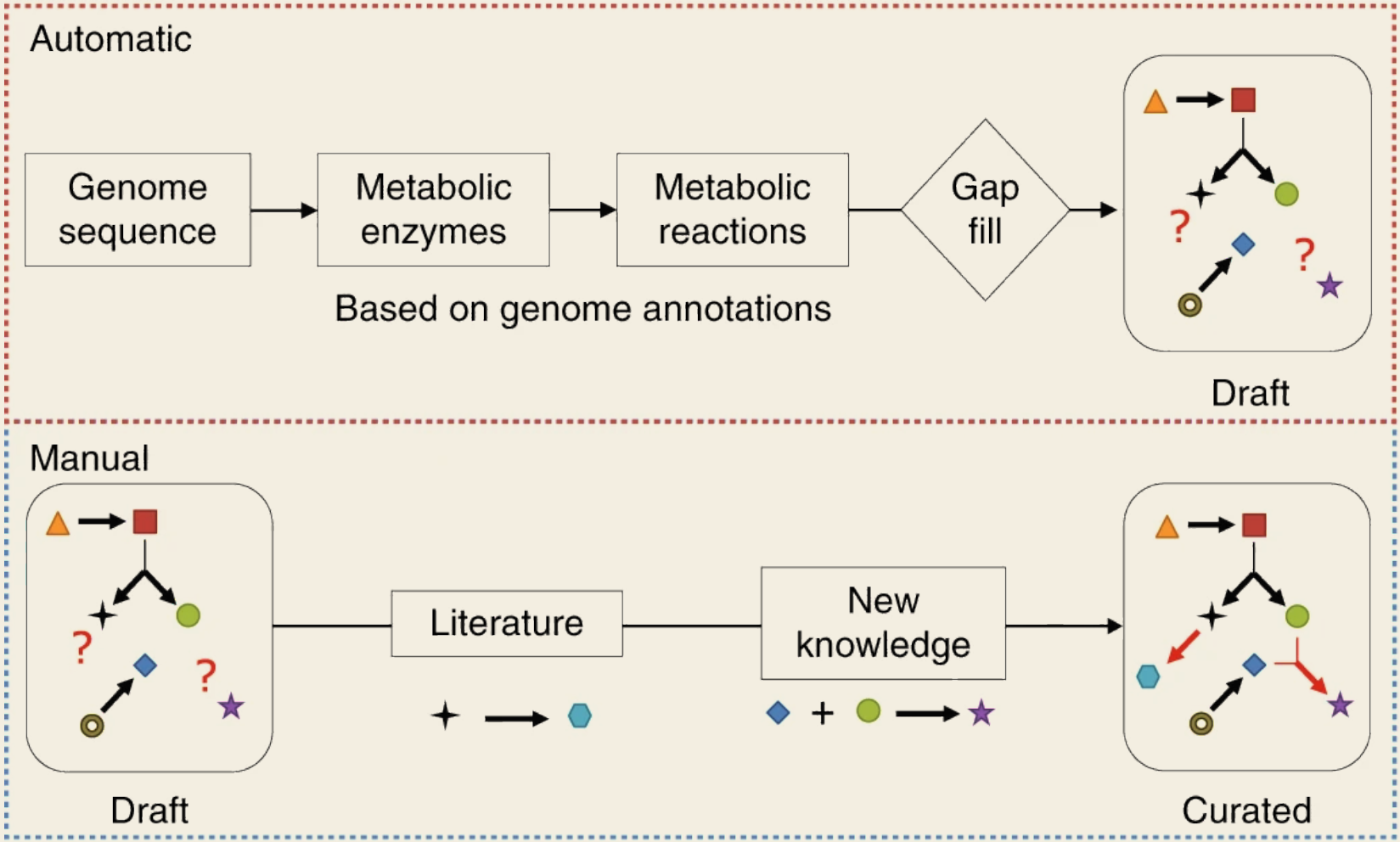
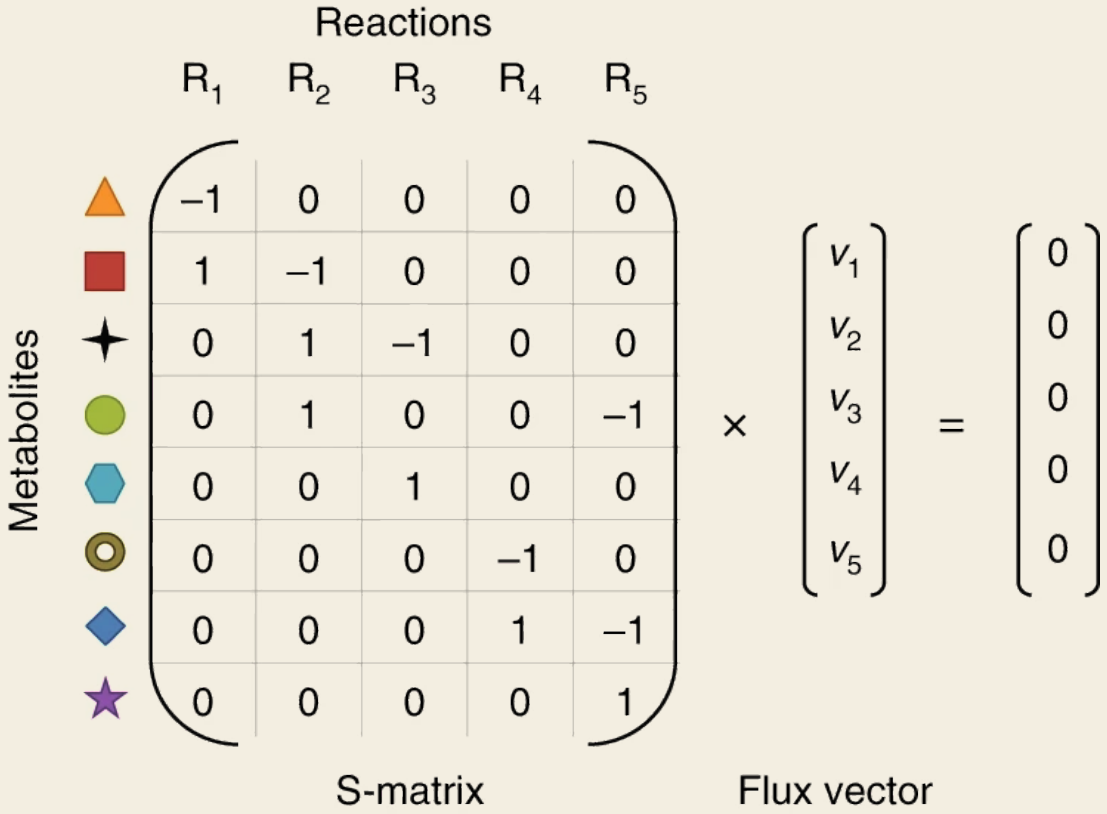


Figure from: Heirendt et al. Nature protocols 14.3 (2019): 639-702.

From a stoichiometric matrix

to a constraint-based model



In a **steady state**
the production rate
of each metabolite
equals its consumption rate

The **flux vector** is a vector with
the value of
each reaction flux
in a certain steady
state.

The steady state assumption
is ensured by
the **zero-vector**.

Figure from: Heirendt et al. Nature protocols 14.3 (2019): 639-702.

The region of steady states

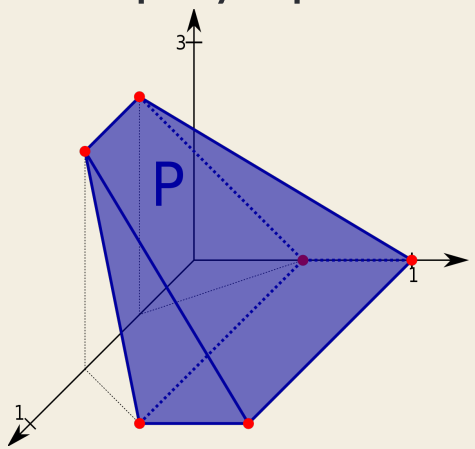
moving to full dimensional polytope

The *constraints* on the reactions fluxes.

$$Sv = 0,$$
$$v_{lb} \leq v \leq v_{ub} \quad (1) \quad v = Nx$$

$$S \in \mathbb{R}^{m \times n}, v \in \mathbb{R}^n$$

As a *full dimensional* polytope



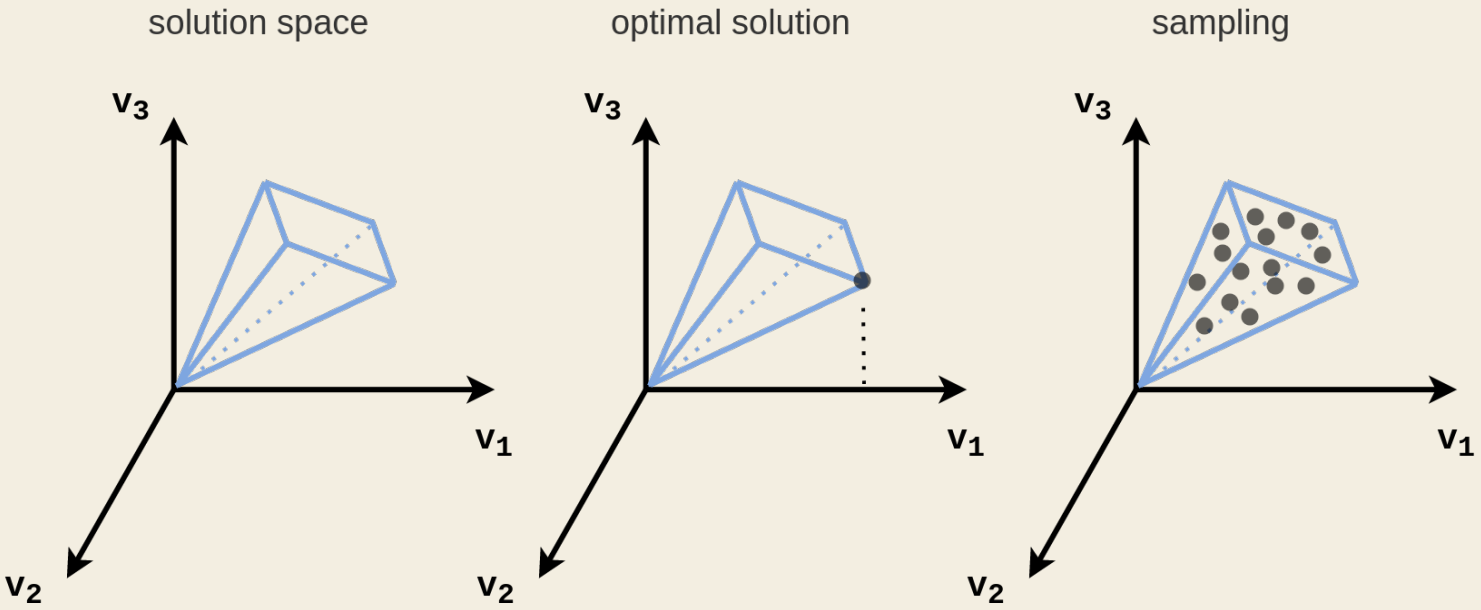
$$P := \{x \in \mathbb{R}^d | Ax \leq b\}$$

$N \in \mathbb{R}^{n \times d}$ denotes the matrix of the null space of S, i.e. $SN = 0_{m \times d}$.

By replacing v with Nx in Equation 1, we get the full dimensional polytope P, where $A = \begin{pmatrix} I_n N \\ -I_n N \end{pmatrix}$ and $b = \begin{pmatrix} v_{ub} \\ v_{lb} \end{pmatrix} N$, (in $\mathbb{R}^d$).

Flux sampling

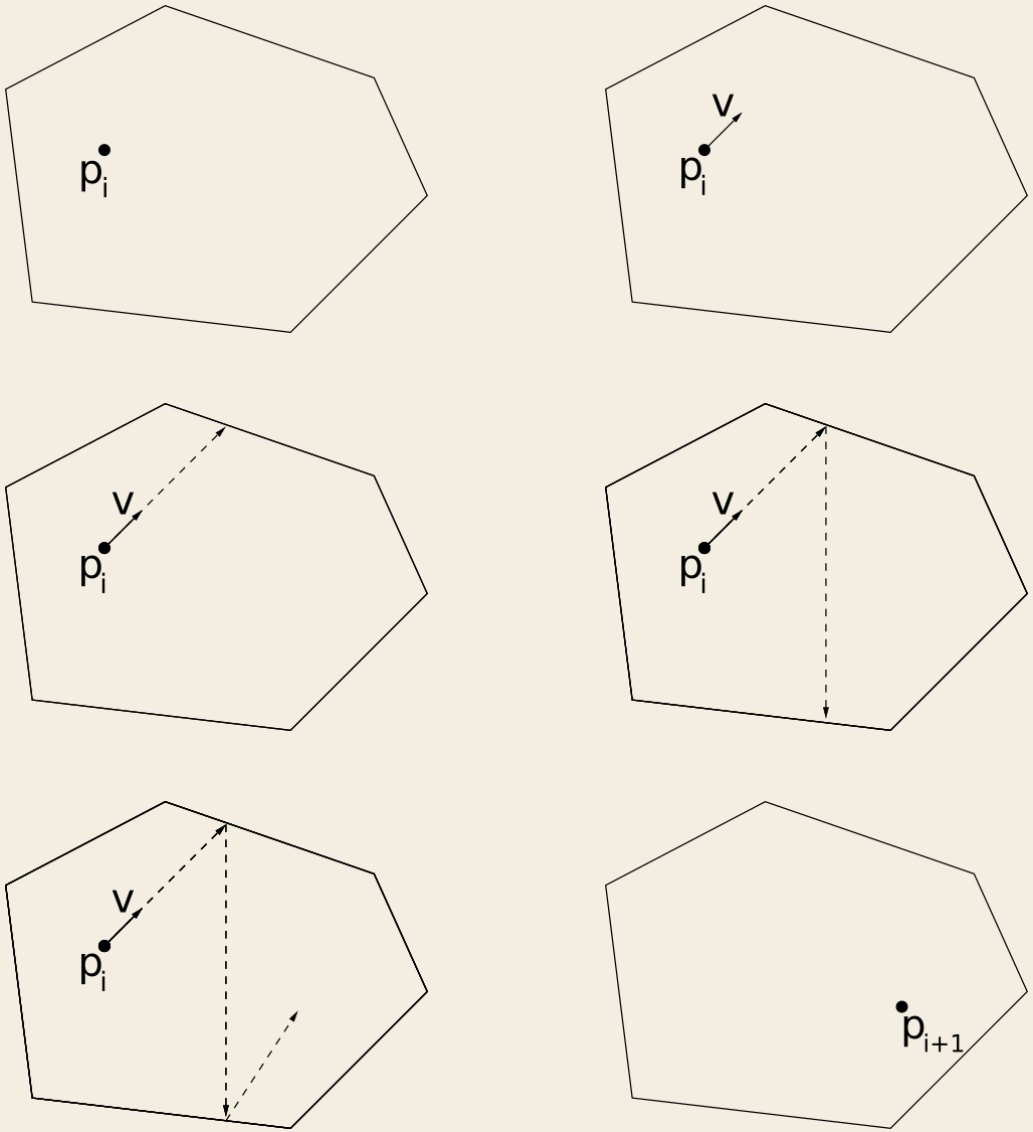
an alternative approach



- enables the analysis of GEMs without the need of an objective function
- determines the feasible solution spaces for fluxes in a network based on a set of conditions as well as the probability of obtaining a solution

Billiard walk

for random sampling



Generate the length of the trajectory $L \sim D$.

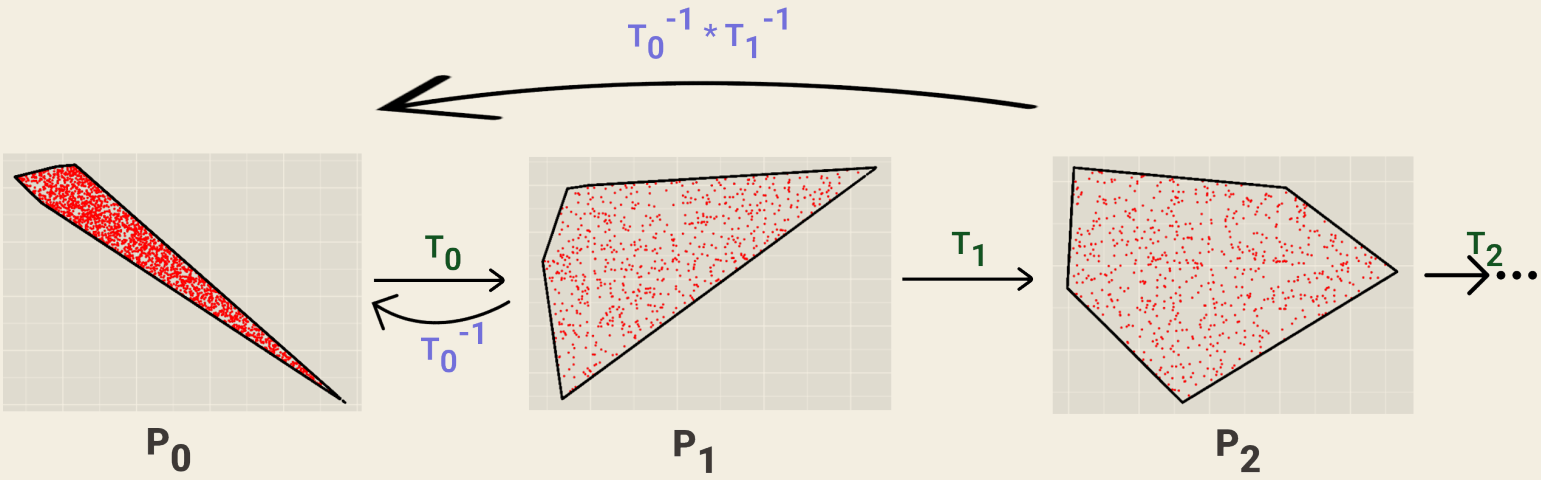
Pick a uniform direction v to define the trajectory.

The trajectory reflects on the boundary if necessary.

Return the end of the trajectory as $p_i + 1$.

Our Markov Chain Monte Carlo (MCMC) algorithm

for flux sampling

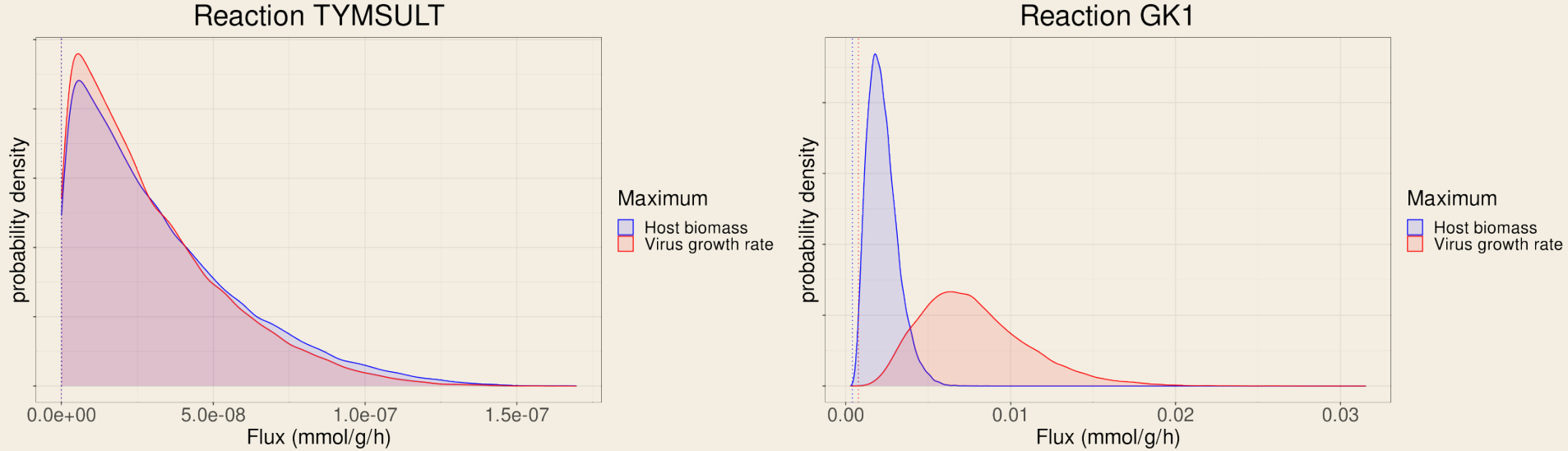


Steps of an MMCS phase

- **sampling step:** using a variant of the **Billiard walk**
- **rounding step:** calculate a linear transformation T_i that puts the sample into isotropic position and then apply it on P_i to obtain the polytope of the next phase
- check several statistic tests

Find possible targets against SARS-CoV-2

a flux sampling application

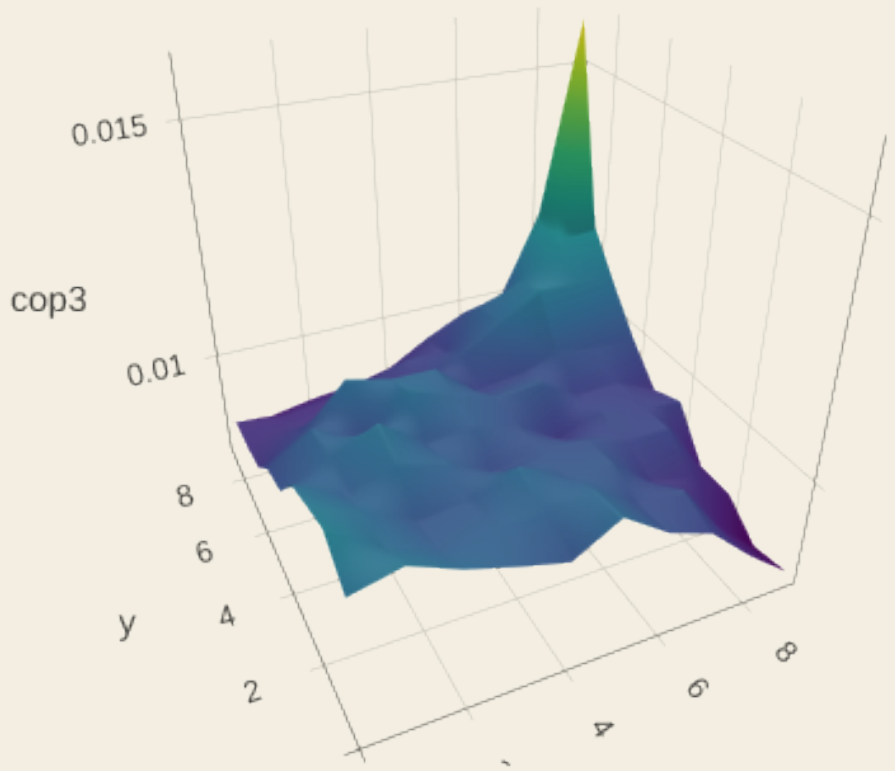


- Check if the flux distribution of a reaction changes.
- Find possible anti-viral targets and study further.

For more about this example case, you may check this blog-post.

What is the probability of a flux value of a reaction

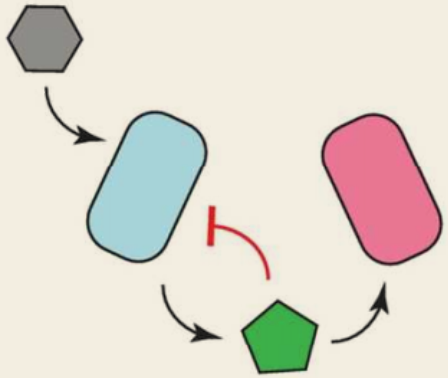
with respect to the flux value of another reaction



Further applications

of metabolic flux sampling

Scott, William T., et al. "Metabolic flux sampling predicts strain-dependent differences related to aroma production among commercial wine yeasts." Microbial cell factories 20.1 (2021): 1-15.



What about microbial interactions ?

1. Microbial ecology from a computational point-of-view

2. PEMA: a metabarcoding pipeline

3. PREGO: a knowledge-base for organisms - environments - processes associations

4. dingo: a Python library for metabolic flux sampling

5. Take home messages & future work

Bioinformatics allow us to address several challenges in microbial ecology

yet, it is more than a means to an end

- Bioinformatics approaches enhance microbial diversity assesment based on HTS data
- Containerization technologies and e-infrastructures provide the means for computational capacity and reproducibility
- High quality metadata enable efficient exploitation of sequecning data in a meta-analysis level
- Markov Chain Monte Carlo approaches enable flux sampling in high-dimensional polytopes

Perspective for future work

More holistic approaches are essential to uncover the underlying mechanisms governing microbial communities

Thank you for your attention

and your patience ;)

GitHub : https://github.com/hariszaf

email : haris-zaf@hcmr.gr

Twitter : haris_zaf

web-site : https://hariszaf.github.io/

Spacial thanks to:

Dr. Pavloudi C.

PhD Paragkamian S.

Mr. Ninidakis St.

Mr. Potirakis Ant.

Dr. Chalkis A.

Dr. Fisikopoulos V.

Prof. Tsigaridas E.

Dr. Pafilis E.

GeomScale