Diet Optimization for Dysbiotic Microbiome

Amirhesam Abedsoltan, Deevanshu Goyal, Mahsa Nafisi

CSE 282A
Bioinformatics II: Intro to Bioinformatics Algorithms

Final Project
Winter 2023

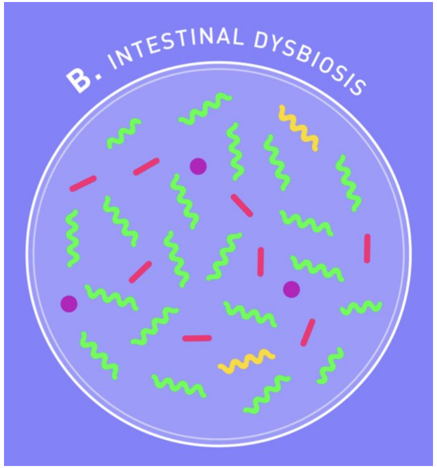
Problem Statement: Biological Context

- Dysbiosis: **Imbalance** in the gut microbe community leading to decreased diversity of microbial population

- Characterised by ‘**over-representation**’ of certain microbial communities and ‘**under-representation**’ of other microbial communities

- Related to **several syndromes** such as Inflammatory Bowel Syndrome, Metabolism-Associated Liver Disease, and Gut-brain axis (CNS syndromes)

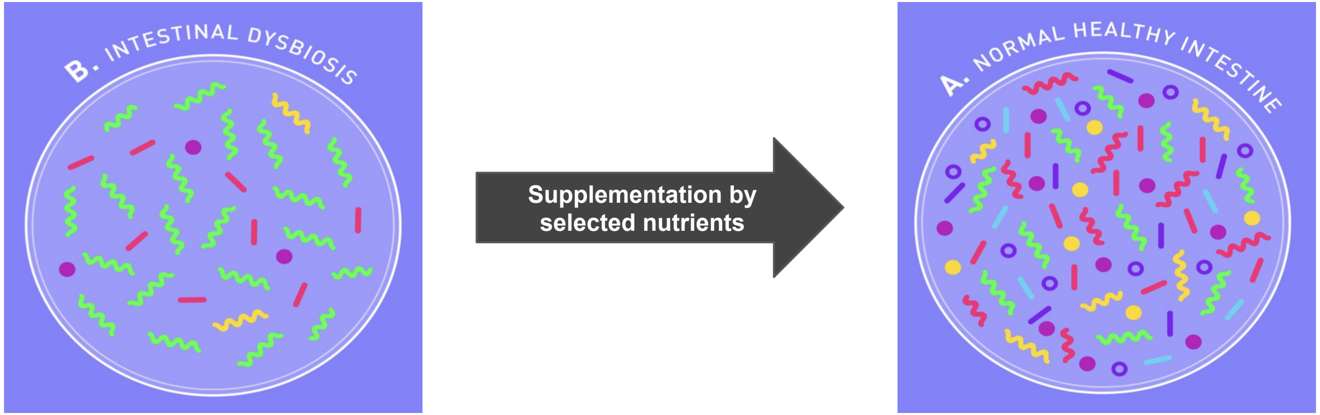
A. NORMAL HEALTHY INTESTINE



Impact of supplemental diet

- Diet can change the number and activity of different microbial communities by providing or depleting the necessary nutrients

- This suggests that a **personalized diet** could re-establish a healthy and balanced microbiome from a diseased state.



Supplementation by selected nutrients

Mathematical formulation

Given,

1. *Test Microbial Sample* (*TMS*) : taxonomic profile with the relative abundances (*in %*) of top 400 Taxons or Amplicon Sequence Variants (ASVs) for a dysbiotic HGM

2. *Reference Sample Collection* (*RSC*) : taxonomic profile with the relative abundances (*in %*) of top 400 taxons (or ASVs) for 1000s of healthy HGM

3. *Nutrient Impact Matrix* (*NIM*) : nutrient-taxon interaction profile between 79* nutrients and 400 taxons (or ASVs), where we interpret each interaction ($\in (0,1]$) as the likelihood of observing a growth in the population of a particular taxon given the use of the corresponding nutrient

** number of nutrients considered under this study*

Objective of the study

We wish to achieve,

1. The definition of a ‘normal’ population of any microbial taxon (or ASV) in a healthy human gut

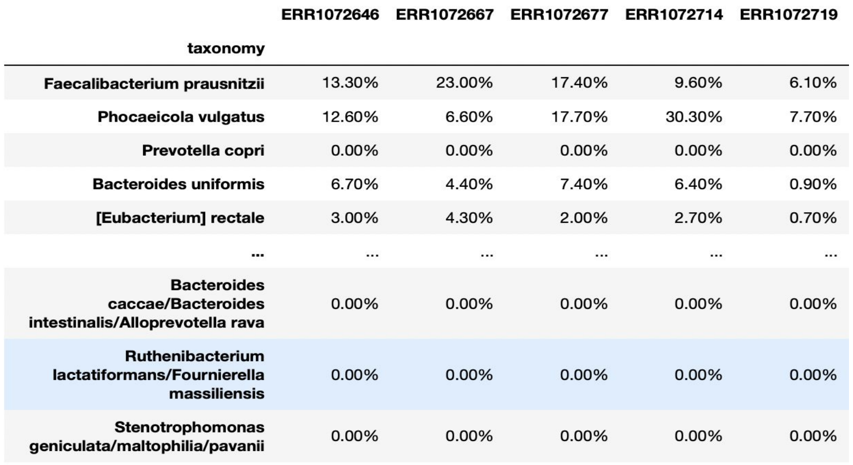
    a. How do we know what relative abundance % of any particular ASV translates to it being present in the necessary ‘amount’ to ensure a healthy gut microbial population?

2. The optimal selection of nutrients that would collectively:

    a. Promote the growth of maximum number of ASVs deemed as ‘under-represented’ in the TMS
    b. Minimize the impact on ASVs deemed as ‘over-represented’ in the TMS

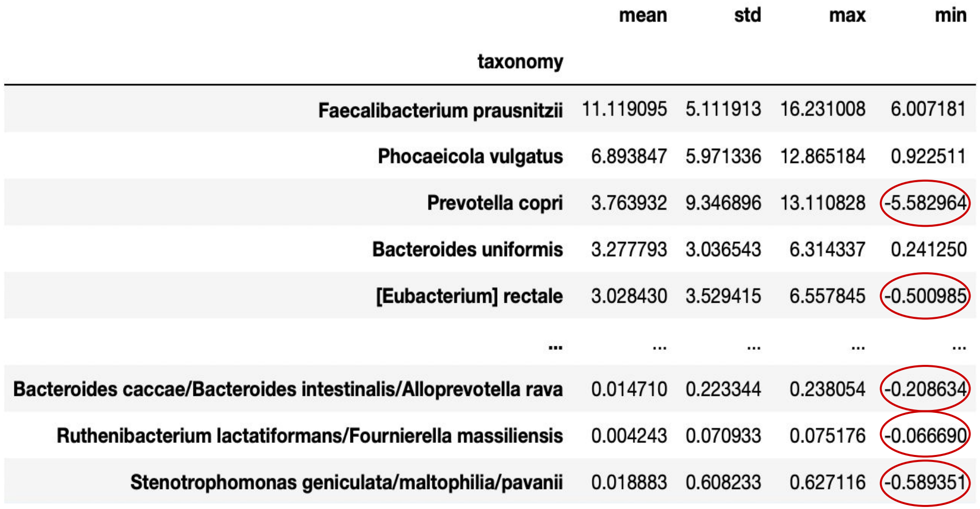
Defining the ‘normal’ population

*Healthy HGM samples (RSC)**



taxonomy	ERR1072646	ERR1072667	ERR1072677	ERR1072714	ERR1072719
Faecalibacterium prausnitzii	13.30%	23.00%	17.40%	9.60%	6.10%
Phocaeicola vulgatus	12.60%	6.60%	17.70%	30.30%	7.70%
Prevotella copri	0.00%	0.00%	0.00%	0.00%	0.00%
Bacteroides uniformis	6.70%	4.40%	7.40%	6.40%	0.90%
[Eubacterium] rectale	3.00%	4.30%	2.00%	2.70%	0.70%
...	...	...	...	...	...
Bacteroides caccae/Bacteroides intestinalis/Alloprevotella rava	0.00%	0.00%	0.00%	0.00%	0.00%
Ruthenibacterium lactatiformans/Fournierella massiliensis	0.00%	0.00%	0.00%	0.00%	0.00%
Stenotrophomonas geniculata/maltophilia/pavanii	0.00%	0.00%	0.00%	0.00%	0.00%

‘Normal’ population statistics



taxonomy	mean	std	max	min
Faecalibacterium prausnitzii	11.119095	5.111913	16.231008	6.007181
Phocaeicola vulgatus	6.893847	5.971336	12.865184	0.922511
Prevotella copri	3.763932	9.346896	13.110828	-5.582964
Bacteroides uniformis	3.277793	3.036543	6.314337	0.241250
[Eubacterium] rectale	3.028430	3.529415	6.557845	-0.500985
...	...	...	...	...
Bacteroides caccae/Bacteroides intestinalis/Alloprevotella rava	0.014710	0.223344	0.238054	-0.208634
Ruthenibacterium lactatiformans/Fournierella massiliensis	0.004243	0.070933	0.075176	-0.066690
Stenotrophomonas geniculata/maltophilia/pavanii	0.018883	0.608233	0.627116	-0.589351

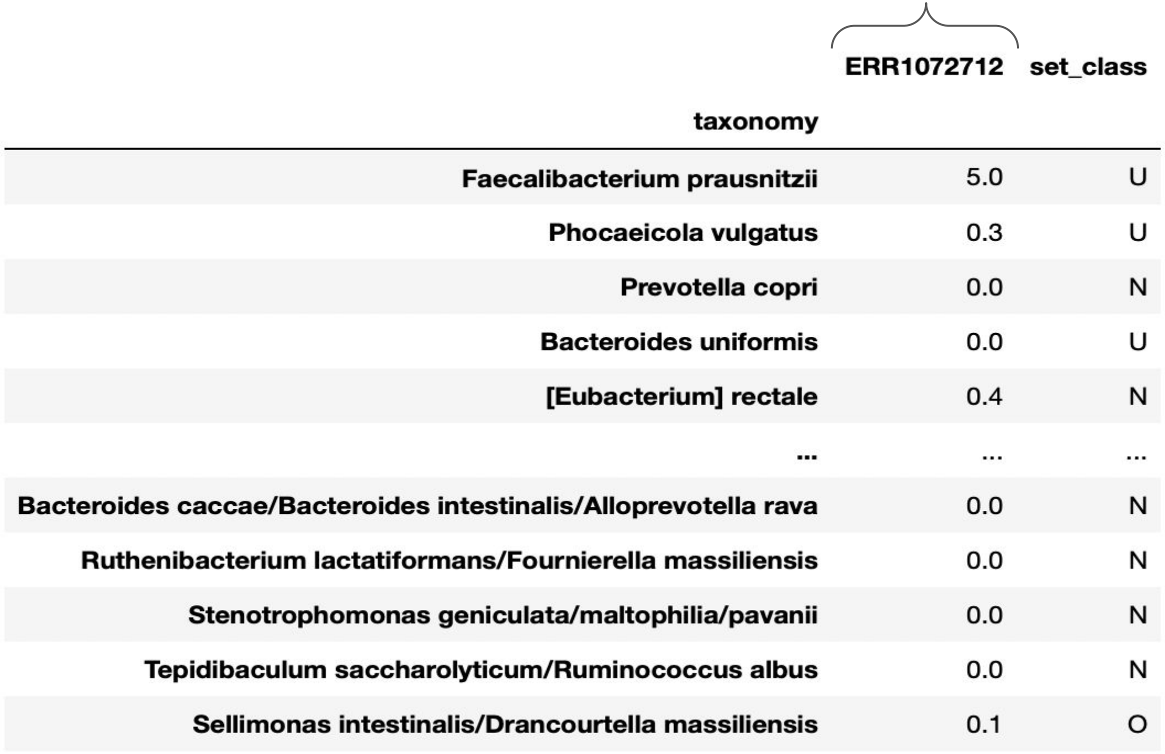
Effectively 0

Here, *max = mean + std.* and *min = mean - std.*

** screenshot shows a limited number of samples, actual healthy HGM samples in 1000s; similarly, 400 taxons considered in total as opposed to what is shown*

Dividing taxons to under/over-represented or normal sets

*Example dysbiotic HGM sample**



taxonomy	ERR1072712	set_class
Faecalibacterium prausnitzii	5.0	U
Phocaeicola vulgatus	0.3	U
Prevotella copri	0.0	N
Bacteroides uniformis	0.0	U
[Eubacterium] rectale	0.4	N
...	...	...
Bacteroides caccae/Bacteroides intestinalis/Alloprevotella rava	0.0	N
Ruthenibacterium lactatiformans/Fournierella massiliensis	0.0	N
Stenotrophomonas geniculata/maltophilia/pavanii	0.0	N
Tepidibaculum saccharolyticum/Ruminococcus albus	0.0	N
Sellimonas intestinalis/Drancourtella massiliensis	0.1	O

Considering the relative abundance (RA_a) of any particular taxon (or AVS), it is considered;

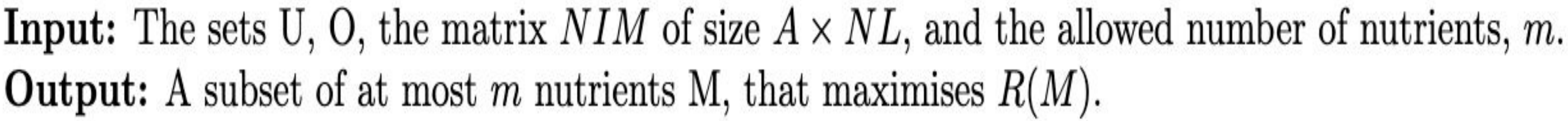
‘U’ or ‘Under-represented’ if
$RA_a < \mathit{min}$

‘O’ or Over-represented’ if
$RA_a > \mathit{max}$

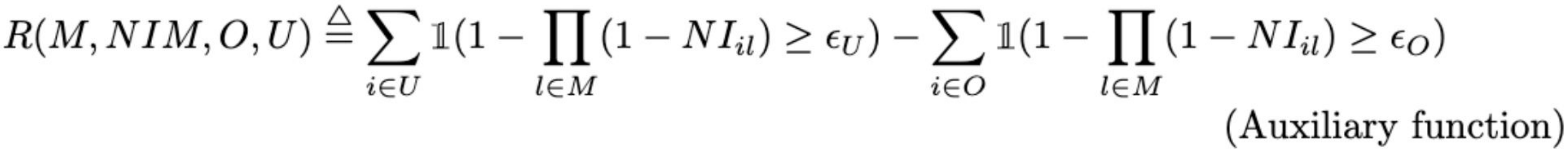
‘N’ or ‘Normal’ if
$\mathit{min} < RA_a < \mathit{max}$

** Further results for this particular example sample as well*

Algorithmic formulation

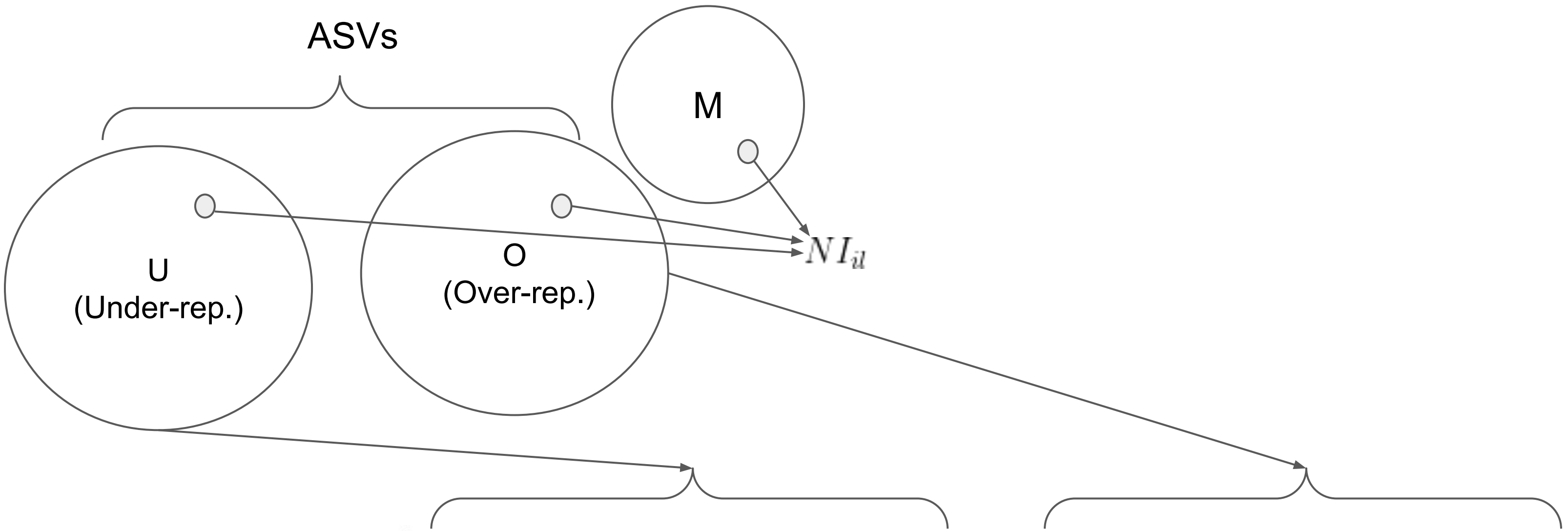


Input: The sets U, O, the matrix NIM of size $A \times NL$, and the allowed number of nutrients, m.
Output: A subset of at most m nutrients M, that maximises $R(M)$.



$$R(M, NIM, O, U) \triangleq \sum_{i\in U} \mathbb{1}\Big(1 - \prod_{l\in M}(1 - NI_{il}) \geq \epsilon_U\Big) - \sum_{i\in O} \mathbb{1}\Big(1 - \prod_{l\in M}(1 - NI_{il}) \geq \epsilon_O\Big) \tag{Auxiliary function}$$

Reward Function for nutrients’ selection



$$R(M, NIM, O, U) \triangleq \sum_{i \in U} \mathbb{1}\Big(1 - \prod_{l \in M}(1 - NI_{il}) \geq \epsilon_U\Big) - \sum_{i \in O} \mathbb{1}\Big(1 - \prod_{l \in M}(1 - NI_{il}) \geq \epsilon_O\Big)$$

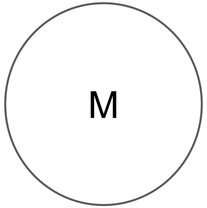
(Auxiliary function)

(Number of under-represented taxons benefitting from *M)*

(Number of over-represented taxons benefitting from *M)*

Reward Function for nutrients’ selection

If ϵ_O is small -> we want to avoid feeding over-represented
If ϵ_U is small -> we want to feed as much as possible under-represented

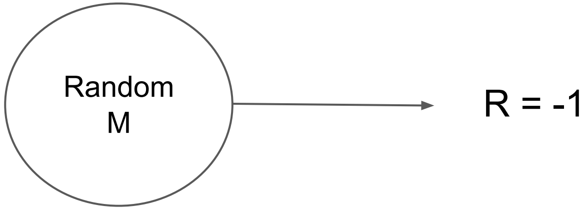


$$R(M, NIM, O, U) \triangleq \sum_{i \in U} \mathbb{1}(1 - \prod_{l \in M}(1 - NI_{il}) \geq \epsilon_U) - \sum_{i \in O} \mathbb{1}(1 - \prod_{l \in M}(1 - NI_{il}) \geq \epsilon_O)$$

(Auxiliary function)

Defining baseline - Naive Randomized Selection

- We always need a baseline
- Classification task 10 classes -> random guess 10% accuracy
- Task is complicated and what is random guess?



Randomized Search Algorithm

```
Do this 5000 times:
    Start with random set
    Do this for 1000 times:
        Every time we either eliminate/add (with probability of 0.2) & replace (with probability of 0.8)
        We keep the change if it improves the score
```

Best score -1 -> 5

Main caveat is slow

Randomized Divide & Conquer

Note that directly applying divide and conquer does not work!

Do this 5000 times:

> Divide the nutrients set randomly to M set.
>
> From each set choose the nutrient with highest reward function value.

Best score -1 -> 5

Very fast

Summary of results

Approach	Best score	Caveat/benefit
Naive Randomized Algorithm	-1	Score low/fast
Randomized Search	5	Score high/slow
Divide and Conquer	5	Score high/fast

Future steps

1. Weighted reward function

$$R(M, NIM, O, U) \triangleq \mu \cdot \sum_{i \in U} \mathbb{1}(1 - \prod_{l \in M} (1 - NI_{il}) \geq \epsilon_U) - (1 - \mu) \cdot \sum_{i \in O} \mathbb{1}(1 - \prod_{l \in M} (1 - NI_{il}) \geq \epsilon_O) \quad \text{(Auxiliary function)}$$

2. Directly optimizing the probability ratio instead of hard thresholding

$$R(M, NIM, O, U) \triangleq \frac{\prod_{i \in U} (1 - \prod_{l \in M} (1 - NI_{il}))}{\prod_{i \in O} (1 - \prod_{l \in M} (1 - NI_{il}))} \qquad \text{(Auxiliary function)}$$

3. Clustering of RSC to define multiple ‘normal’ population definitions and incorporating subjective nature of a ‘healthy’ diversity

Thank you for your attention!