Deciphering the Temporal Changes in Microbiome Communities

Author: Lusine Adunts

BS in Data Science

American University of Armenia

Supervisor: Lilit Nersisyan

Armenian Bioinformatics Institute

Abstract—**Living organisms harbor diverse bacterial communities, collectively termed as microbiome, residing in areas such as the gut and skin. Alterations in microbiome composition in human body are linked to numerous disorders, from inflammatory bowel disease to colorectal cancer. The analysis of microbiome datasets poorly deals with the dynamic nature of the communities. Approaches that take the microbial interactions into accounts are needed. This paper aims to integrate graph theory and single cell transcriptomics methodologies into microbiome data analysis which will untangle the complex web of interactions within microbial communities and between these communities and their hosts.**

Index Terms—**microbial communities, subgraph analysis, microbiome velocity, pseudotime analysis, mathematical modeling**

I. INTRODUCTION

A. Microbiome communities: complex networks of microorganisms

Hundreds of thousands of bacteria normally inhabit living organisms, particularly the human body - the gut, skin, and reproductive organs [1]. Bacterial communities provide numerous advantages to the host, by performing various physiological functions such as reinforcing gut health, molding the intestinal lining, extracting energy, defending against harmful pathogens, and controlling the host's immune system [2]. Dynamic changes in the composition of bacterial communities, or the microbiome have been shown to be linked to a diverse array of disorders, including obesity, diabetes, cardiovascular disorders, cancer, hypertension, and IBDs [3]. Moreover, besides microbiome-host interactions a wide range of interactions take place among the bacteria. The competition and cooperation between them play a crucial role in community composition formation and functioning [4]. Within these communities, bacteria compete with their neighbors for nutrients, including carbon, nitrogen, metals, phosphate, as well as for space where these nutrients are abundant. Microbial communities also engage in cooperation to aid in resource digestion, combat antibiotics, or manage other environmental stresses.

Methodologies that assess the changes of microbiome composition affected by environmental perturbations mainly focus on the bacteria separately, ignoring the interactions between them [5]. However, numerous diseases are associated with the network of bacteria [6]. Therefore, the independent analysis of the bacterial abundance may not be enough.

B. Approaches to study dynamics of microbiome communities

Numerous statistical and computational methods are used for analyzing microbiome data. The downstream analysis of these datasets may have different components mainly consisting of the exploration of the change of a single feature in different conditions [7]. Differential abundance analysis is done for identification of the taxa that show significant difference between two or more conditions [8]. Alpha diversity metrics such as Shannon and Simpson indices summarize the structure of a community [9]. Identification of keystone species is often employed to provide insights into the potential drivers of change within the system. Time series clustering done at either the participant level or the feature level identifies naturally occurring clusters focusing on recognizing patterns that evolve over time.

Besides single features, the interactions between features are of great importance. The methods detect only indirect effects which are propagated through the network of interactions. In the presence of an external effect on one feature, in later timepoints an effect will be visible on the features interacting with it. This effect may falsely be assumed to be associated with the disease if holistic approaches are not used to differentiate direct and indirect effects. To assess the interactions co-occurrence networks are created [10]. The networks are produced by applying an association metric or correlation coefficient in a pair-wise manner. However, the methods do not take into account neither the influence of external perturbations and the longitudinal nature of some datasets, nor the dynamics of the communities. There is a need for further exploration of microbial datasets, and techniques inspired from graph theory and single cell data analysis can be helpful with right choices of meaningful equivalents of the concepts.

C. Lessons from graph theory

Graph theory offers multiple methods for the analysis of its components. An essential problem is to find the core subgraph in a graph responsible for most of the weight in it. The Maximum Weight Connected Subgraph Problem (MWCS) is the problem of finding a connected subgraph with maximum total weight in a node-weighted graph [11]. It is an NP-hard problem and its solution is usually approximated by transforming the given instance to the Prize-Collecting Steiner Tree Problem or by using an integer programming formulation.

(a) Number of identified genes in single cell.

(b) Number of identified species in microbiome.

(c) Velocity values in single cell.

(d) Velocity or log change values in microbiome.

(e) Change values in microbiome (relative).
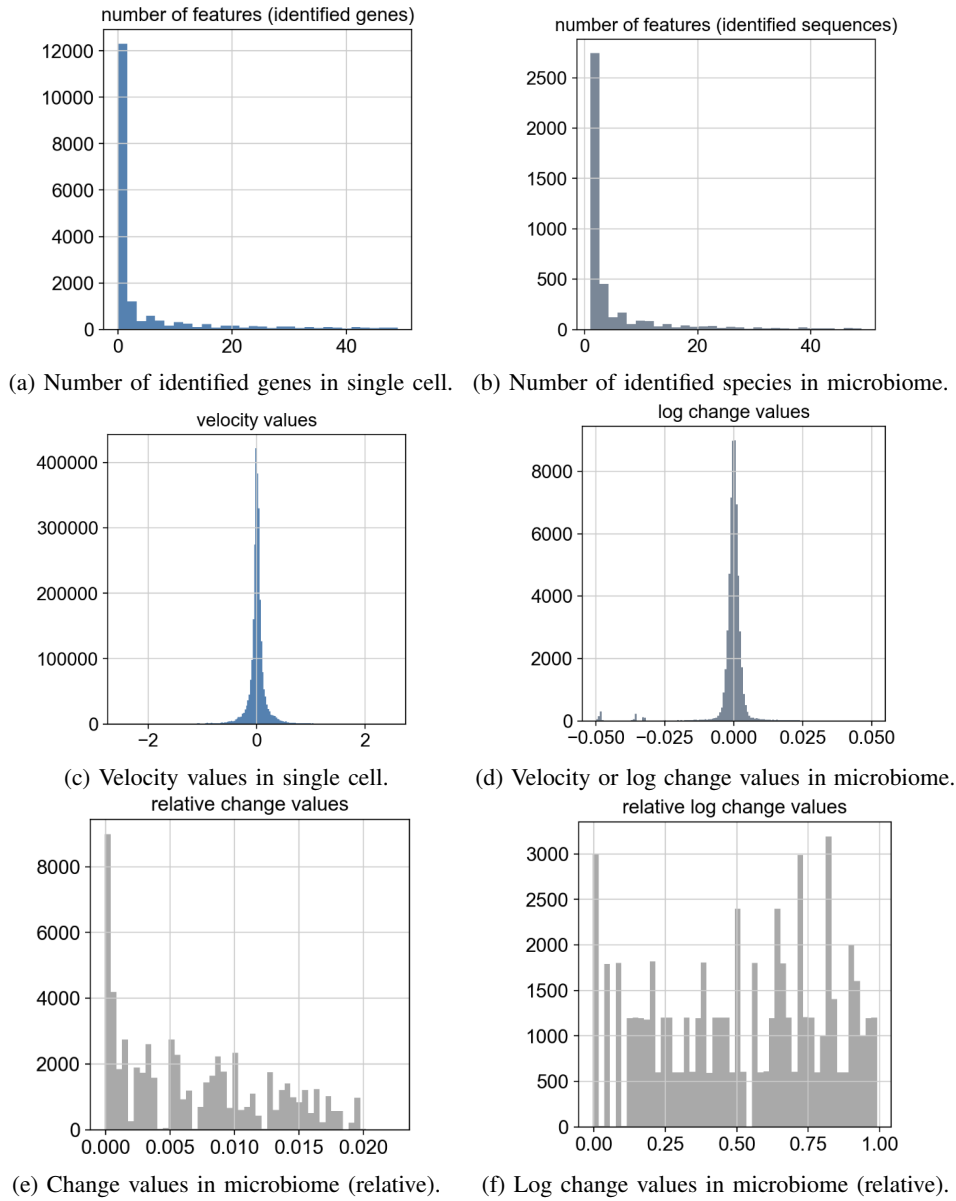
(f) Log change values in microbiome (relative).

Fig. 1: Translatability check from single cell transcriptomics to microbiome.

Markov chains are mathematical models that represent systems undergoing transitions from one state to another with certain probabilities [12]. These states can be viewed as nodes in a graph, and the probabilities of transitioning from one state to another are represented as directed edges between these nodes. Markov chains are helpful for construction of transitions from one state of the biological system to another.

Random walks are stochastic processes where a path is formed by taking successive steps in random directions [13]. In the context of graph theory, a random walk can be visualized on a graph where each vertex represents a state and edges represent the possible steps from one vertex to another. Each move from one vertex to a connected vertex is chosen at random, often with equal probability, making the graph a useful tool for analyzing the properties and behaviors of random

walks, such as convergence and cover times. Numerically, the convergence can show the probability of one biological state landing in another one in the future.

D. Lessons from single-cell transcriptomics analysis

Human body is a highly complex system composed of approximately 3.72×10^{13} cells of various types [14]. Although almost all cells contain the same set of genetic material, the transcriptomic information differs and reflects the heterogeneity of the cells by identifying the unique activity of a specific set of genes. Profiling the gene expression activity in cells of an organism is considered as one of the most authentic approaches to examine cell identity, state, function and response. Single-cell RNA sequencing enables the investigation of millions of cells in a single study to distinguish each cell at

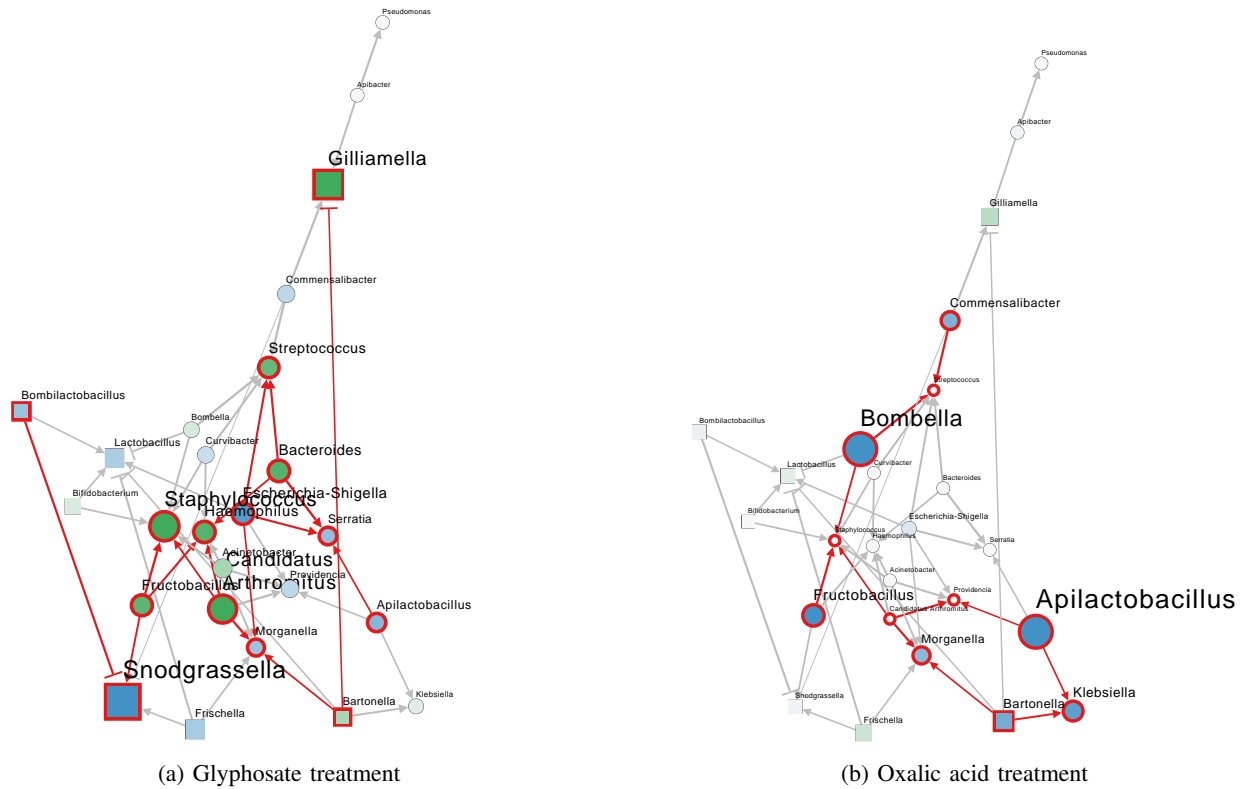(a) Glyphosate treatment (b) Oxalic acid treatment

Fig. 2: Subgraph analysis on honeybee gut microbiome datasets.

the transcriptome level [15]. They give matrices where rows are genes and columns are cells with gene expression values, showing how active a gene in a cell is.

To analyze single cell datasets numerous bioinformatics techniques have been created [16]. Dimensionality reduction techniques help to reduce the high dimensional gene expression data and to visualize the cells on a two dimensional space. Later, the cells are arranged according to gene expression similarities and trajectories are built. Pseudotime analysis is a time-series examination of the data and can be used to infer the trajectory of the cells at the single-cell level which is expected to discover cell types and cryptic states [17]. Pseudotime analysis methods computationally order the cells along a trajectory topology and give a broader idea about the underlying biological process. Monocle is one of broadly used tools for pseudotime analysis [18]. It learns an explicit principal graph to describe the gene expression data and rebuilds single-cell trajectories by embedding a reversed graph to improve the robustness and accuracy of predicted trajectories. Another algorithm is the diffusion pseudotime (DPT) which uses random-walk-based distances in diffusion map space [19].

Gene expression values assist in RNA velocity estimation which indicates where each cell or cell cluster will move on the derived two dimensional space and predicts the future state of a cell [20]. The tool CellRank integrates this directionality information layer with the gene expression values in a weighted manner, and derives a transition probability matrix

that shows the probabilities of transition from one cell state to another [21]. Using Markov Chain modeling it calculates fate probabilities of cells towards terminal states of the cellular trajectory.

E. Microbiome communities and single cells: feature space comparison

Microbial communities, although different from single cells in nature, behave similarly from the data distribution perspective. Both single cell RNA-seq (scRNA-seq) and metagenomics datasets are described by sparsity of feature matrices that arises from insufficient sampling of RNA in single cells or DNA in microbiome communities, particularly leading to missing values for features having a low abundance. The sparsity, however, is more pronounced in single cell datasets than in microbiome datasets. Therefore, data imputation applied in scRNA-seq datasets may not be as relevant as in the case of metagenomics datasets. Although both types of datasets are compositional in nature they have a key difference. While in single cell datasets the total number of expression values is more or less the same from one cell to the next, it is not the case in microbiome. Taking into account the differences, the translation from one dataset to another is possible.

By choosing the analogous concepts, the tools used in single cell transcriptomics can be adapted for microbial communities. Each sample or microbial community is equivalent to a cell with the taxonomic abundance derived from 16S amplicon or WGS datasets taken as an analogue of the gene expression

value. The transitions between these entities represent community composition shifts along a real or pseudotime axis. In microbiome, the velocity value equivalent is the growth rate either derived from relative read coverage comparing sites closer to the origin of replication to the terminus or predicted from networks of interactions between species. An alternative for the growth rate is the peak to through ratio (PTR) which is a metric calculated based on DNA sequencing coverage differences between genomic regions closer or further from the replication origin [22]. The PTR approach serves as an empirical, model-independent measure of individual species behavior neglecting the growth-promoting or inhibiting interactions among different species.

II. AIMS AND OBJECTIVES

The aim of this work is to transform our approach to analyzing microbiome datasets. Utilizing cross-sectional datasets, we aim to map the landscape of pseudotemporal dynamics and identify potential attractor states within microbiome communities. The goal is to arrange the microbiome communities derived from the cross-sectional studies along a pseudotemporal axis, reflecting the gradual shifts that potentially link one individual's microbiome to another. We aim to identify the microbiome community shifts using microbiome velocities and to predict the future state of a community. Through network analysis we aim to identify the subgroups of species that are most changed from external perturbation such as a treatment. After the validation on simulated datasets the methodologies will be used to infer important knowledge from real microbiome datasets.

III. METHODS

A. Co-occurrence network construction

The relationships of bacteria sharing the same environment have been characterized by generating co-occurrence networks. These networks are a proxy for interactions between the bacteria as interaction causes co-occurrence. The aim is to combine the co-occurrence information of the bacteria from different samples and to construct a uniform interaction network which later can be used to derive for each sample subsets of bacteria that has the biggest influence in it.

To construct the uniform co-occurrence network of a study, we calculated the Pearson correlation coefficients between all pairs of bacteria and derived p-adjusted values with Bonferroni multiple testing correction for each sample in the study separately. We filtered out the pairs that had correlation coefficient lower than 0.6 or p-adjusted value larger than 0.1 as these pairs of bacteria have either no or little interaction. As a result, a co-occurrence network was constructed where nodes are bacteria and edges show interaction between them. To construct the underlying network for further analysis, we merged the separate networks with overlapping edges having mean correlation coefficient of the separate networks. The resulting underlying network shows overall interactions between bacteria independent of separate samples.

B. Maximum weight connected subgraph analysis

In order to find the subset of bacteria that has the biggest influence or change, we did subgraph analysis for each sample separately. We analyzed the dynamics using the co-occurrence or interaction network constructed based on all samples and the abundance values of the sample. Then, to see the difference between untreated and treated samples log fold change values were used. It is calculates as $log\frac{A_t}{A_u}$ where A_t is the abundance of the species in the treated group and A_u is the abundance of the species in the untreated group. To convert the continuous values into discrete weights for the nodes, we set the weight for the species with absolute logFC value higher than 2 as 3, for those with absolute logFC value in the range [1, 2) as 1 and the others as -2. If the bacteria was not present in the sample the weight was set as 0.

To infer the subset of bacteria with the greatest overall abundance, the maximum weight connected subgraph (MWCS) was extracted. Formally, MWCS is: given an undirected graph $G = (V, E)$ and weights $w : V \rightarrow R$, find a subset of nodes V^* such that the induced graph $G^* = (V^*, E^*)$, where $E^* = \{\{u, v\} \in E | u, v \in V^*\}$, is connected and the overall weight $\sum_{v \in V} w(v)$ is maximal. The problem is NP-hard and exact solution requires exponential number of constraints. Its approximation methods mainly convert it into a Prize-Collecting Steiner Tree Problem or use linear integer programming (ILP) formulations. Another way to come to an acceptable solution with heuristics is to use a Relax-and-Cut scheme, i.e. Lagrangian relaxation combined with constraint generation [23].

For the subgraph extraction from the node-weighted underlying graph we used mwscr [24]. The algorithm formulates the problem as an ILP, where decision variable is either 0 or 1. In the relaxation step the problem is converted to a relaxed LP problem and is solved. Then, the algorithm finds the inequalities violating the solution of the problem and updates the Lagrangian multipliers for the next iteration. The process repeats until the solution is sufficiently good or the maximum number of iterations set in the beginning is reached. As a result, the subgraphs are inferred for each of the studies.

The data used for underlying graph construction and subgraph analysis is cross-sectional, i.e. observations of many different individuals (subjects, objects) at a given time, each observation belonging to a different individual. The subgraph analysis data contains honeybee gut microbiome samples, taken from Texas and Denmark studies. Texas study contains 84 untreated samples and 59 samples treated with Glyphosate (Herbicide). Denmark study has 182 untreated samples and 39 treated with Oxalic acid (Pesticide). For the analysis, the mean-based merged samples were used both for untreated and untreated groups.

To visualize the resulting networks we used Cytoscape 3.9.0 with Spring Embedded layout. The square nodes stand for core species. The red bordered nodes and edges show the identified subgraph. Green nodes stand for positive logFC, and blue for negative with shade standing for the value. The text size is set
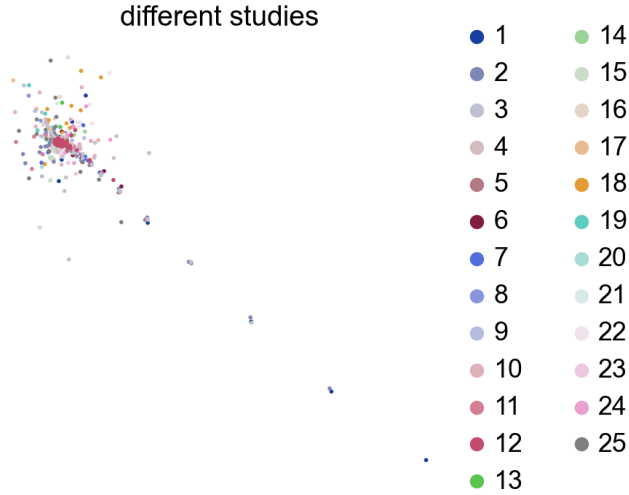
Fig. 3: State space analysis of different studies.

based on the absolute logFC. The edge size is set based on species correlation and the shape shows if the correlation is positive or not, i.e. the specie activates or inhibits activity of the other.

C. Data simulation

To have a ground truth for method validation, we did data simulation with miaSim 1.5.5 [25]. Given growth rates of species, an interaction matrix of the interactions between the species, metacommunity probabilities, and initial abundances of the species in the community miaSim simulates the abundances in the next time points. The simulation is done based on Lotka-Volterra dynamics model:

$$\frac{d}{dt}N_i(t) = N_i(t)\left(g_i + \sum_{j=1}^{D} A_{ij}N_j(t)\right) \tag{1}$$

where N_i is the (relative) abundance of species i, g_i is the growth rate of species i, and A_{ij} is the interaction between species i and species j.

As miaSim generates with fixed growth rates, we changed the tool to integrate different growth rates for each of the timepoints if needed. This was done for external perturbation simulation. We adjusted the growth rates through time to simulate antibiotic effect and resistance to it.

For the analysis the same growth rates, interaction matrix and metacommunity probabilities were used. The growth rates for the species were uniformly generated in [0, 1] range. The interaction matrix was generated with the default parameters. The metacommunity probabilities were randomly generated from Dirichlet distribution.

D. Pseudotime analysis

The temporal order of the bacterial communities are encoded in their abundance profiles. Diffusion pseudotime (DPT) efficiently estimates this ordering. It measures the progression

through branching lineages using a random-walk-based distance in diffusion map space and allows for branching and pseudotime analysis on large-scale datasets. The pseudotime is computed in three steps. First, a transition matrix T is constructed that approximates the dynamic transitions through stages of the process. Second, the distance metric is defined which shows the pseudotime of the point from the predefined root point. Third, branching points are identified by comparing two random walks over cells, one starting at the root point and one starting from the point most distant from the root, measuring the pseudotime with respect to both separately. The two sequences of pseudo times are anticorrelated until the two walks merge in a new branch, where they become correlated.

E. Growth rate and PTR

In the presence of the growth rates, abundances and interaction matrix, the Lotka-Volterra model allows us to estimate the derivative of population size $\frac{dN}{dt}$. Having the growth rates, the interaction matrix and the abundances one can calculate the $\frac{dN}{dt}$ change from 1, i.e. the velocity value. Another analogue for velocity value can be $\frac{dlogN}{dt}$ log change, derived from an equivalent equation

$$\frac{d}{dt}logN_i(t) = g_i + \sum_{j=1}^{D} A_{ij}N_j(t) \tag{2}$$

However, in the case of real datasets another metric should be used as the growth rates and interaction matrices are not available. The PTR value is a measure of individual species growing behavior [26, 27, 28, 29]. CoPTR estimates PTR by estimating the ratio of DNA sequencing coverage at the replication origin and the coverage at the terminus [30]. Although it ignores the interactions between the species, the PTR measure effectively estimates the growth rates of the species and can be used as species velocity value.

F. Simulation state space analysis

In the presence of the microbiome velocities, Cellrank integrates the directionality information with the abundance values and derives the transition probability matrix [21]. Using Markov-Chains it infers macrostates, defined groups of states that are likely to transition between each other but to states of other groups. Based on the transition probabilities the macrostates are classified as initial, intermediate and terminal. With random walks and the inferred transition probability matrix Cellrank identifies the fate probabilities towards the terminal states, i.e. the probability that the entity will land in particular terminal state in the future. By computing Pearson's correlation between the activity and fate probabilities it infers driver features. The driver features can later be examined either based on taxonomic information or by BLAST (Basic Local Alignment Search Tool) which searches the query sequence against a database of other DNA known sequences to find regions of similarity.
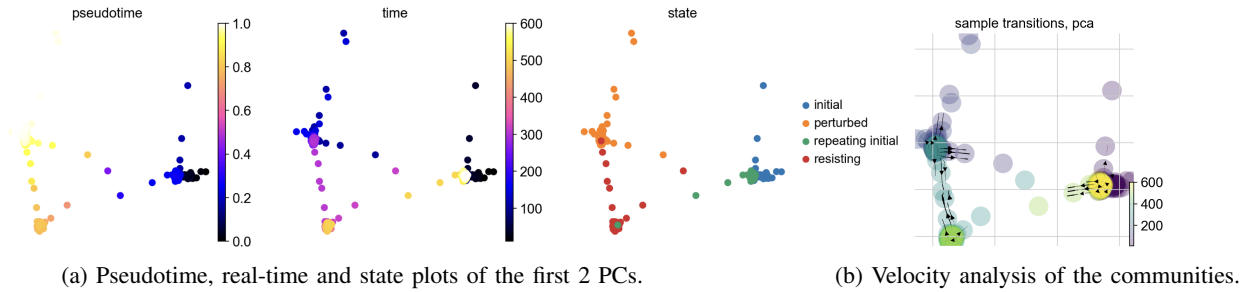
(a) Pseudotime, real-time and state plots of the first 2 PCs.

(b) Velocity analysis of the communities.

Fig. 4: Simulation 1 results. Each point represents one sample.

(a) Pseudotime, real-time and state plots of the first 2 PCs.
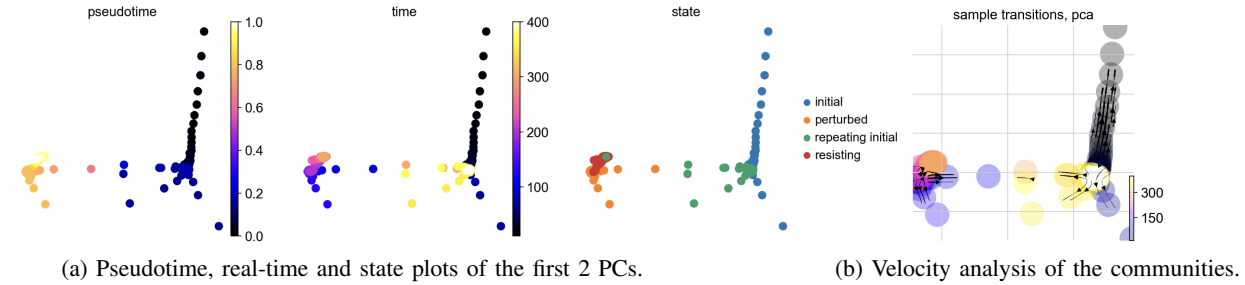
(b) Velocity analysis of the communities.

Fig. 5: Simulation 2 results. Each point represents one sample.

G. Translatability

To ensure the translatability between RNA transcriptomics and microbial data we checked the data distribution. For comparison we used the Cellrank pancreas single cell RNA-seq dataset. Not filtered distribution of the number of features, i.e. identified genes and sequences, is negative binomial for both cases (Fig. 1a,b). The RNA velocity and log change values have normal distribution (Fig. 1c,d). We used the absolute abundances in the simulations, as the distributions of log change and change values for relative abundance are not normally distributed and cannot safely be taken as velocity equivalent (Fig. 1e,f). However, in case of real datasets this should be taken into account as usually we deal with relative abundances in the real life scenarios.

IV. RESULTS AND DISCUSSION

A. Finding subnetworks affected by a perturbation

To understand how a treatment and other external perturbations affect the microbiome, it is useful to infer a network of species that is treatment-specific or is most affected by the external perturbation. The species in the network may be a part of important pathways and biological processes. This information can give insights about perturbation or treatment effect and community behavioral changes after the perturbation happened.

To identify this kind of networks from bee gut microbiome datasets we did subgraph analysis. The aim was to identify subnetworks that were most affected by treatment with either Glyphosate or Oxalic acid. Based on the visualization, Glyhposate and Snodgrassella are highly affected by Glyphosate treatment (Fig. 2a). The change in Snodgrasella activity negatively affects numerous other species activity, such as

Staphylacoccus and Streptocaccus. Interestingly, in case of Oxalic acid most of the species in the identified subgraph are negatively affected by the treatment (Fig. 2b). Further analysis can be done do explore the behavior of specific bacteria and the role of these groups of species.

B. Microbial state space analysis

As the aim of the work was to model the dynamics of microbiome networks over time we needed to set a ground truth to further apply perturbations and see the effects. We started off with a random set of species with different biological properties. These properties are the growth rates, which show how fast the species divide, interactions that tell how the abundance of one species affects other species in the community and metacommunity probabilities, that show the likelihood that species from the metacommunity can enter the community. We use the generalized Lotka-Voltera predator-pray model that models the interactions using differential equations for change.

To see what the state space looks like, we did simulations for 100 species with the same growth rates, interaction matrix and metacommunity probabilities, but different initial abundances. We did one simulation with the initially set parameters, and used the generated abundances of the first 20 time points as initial states for 20 different simulations (Fig. 3). Although having slightly different initial state, the microbial communities grow in the same direction, i.e. they have propensity to move toward attractor states. These attractor states are differently reachable depending on the starting state. This means that small changes do not affect the microbiome composition drastically, but strong perturbations might take the system out of a local state space and push it to reach

(a) Pseudotime ordering of the communities.

(b) Trajectory inference based on similarity.

(c) Correlation of sequences in different states.

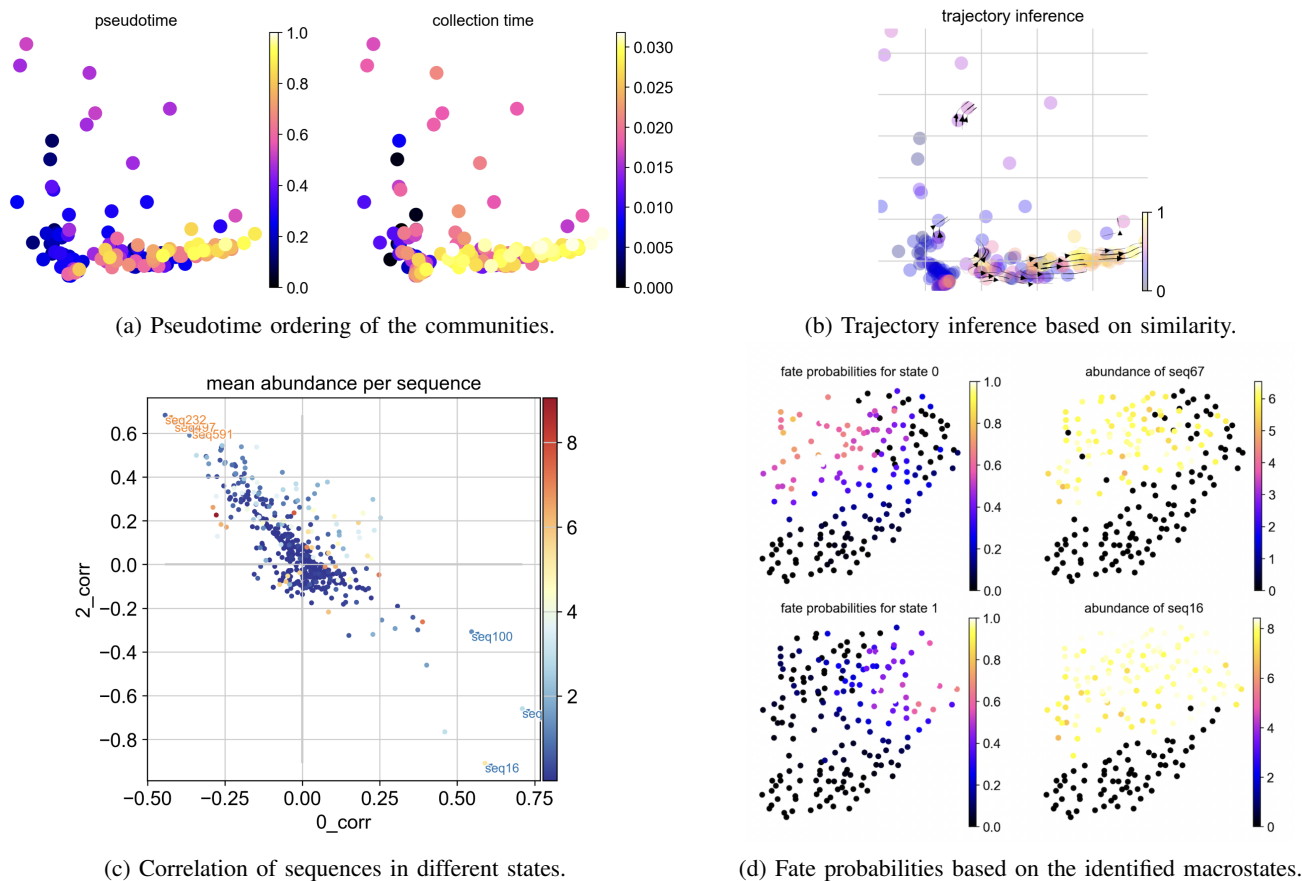(d) Fate probabilities based on the identified macrostates.

Fig. 6: Analysis on a real dataset.

another attractor state which might be an irreversible change. The question is: how strong should a perturbation be to take the system out of the local attractor state. We will attempt to answer this question with real-world data.

C. Applying pseudotime approach and velocity analysis in single-cell transcriptomics to simulated microbial networks

We were interested in adopting the diffusion pseudotime from single-cell RNA sequencing analysis that orders the cell states according to their similarity, and assigns a starting point and branching points to the trajectory that describes gradual changes leading from one cell state to another along a pseudotime axis. We took the microbiome states derived from the simulation experiments described above, and applied the algorithm to order the states according to their similarity. We were interested in identifying whether the psuedotime analysis would order the states not only according to simulation time, but also according to the state compositions after different types of perturbations.

For the pseudotime analysis many simulations were done two of which are presented (Fig. 4, Fig. 5). For both of the simulations the initially generated interaction matrix and metacommunity probabilities were used. In the first simulation abundances for 100 species for 600 timepoints were generated with the initially generated growth rates. We next modeled a

perturbation that can be caused by introducing an antibiotic. The antibiotics are usually affecting the growth rate of a subset of bacterial species present in the community. Other effects are indirect, as the species that change in abundance propagate this change through interactions with their neighbors, thus, leading to overall changes in the composition. Therefore, to simulate antibiotic effect, after timepoint 100 the growth rates of 7 species that had high growth rates were reduced by 0.1. To further recapitulate a real-world scenario, where some bacteria develop mutations over time that make them resistant to the antibiotic, we gradually increased the growth rate of 1 of these 7 species after the time point 300 and kept it steady until the time point 500. For the last 100 time points the growth rates were set the same as in the beginning of the simulation, as if the antibiotic is removed from the environment (and no longer consumed).

In the simulation, as the initial and final phases biologically are the same state, the pseudotime ordering correctly identified both as early states (Fig. 4a). Also, it put resisting and perturbation states late in pseudotime with perturbation states followed by resisting. We thus show that the pseudotime approach can be adapted to microbiome datasets. However, we can also notice that even though the pseudotime orders the initial and last phase states close to each other, it cannot detect the trajectory of change starting from initial states and ending

at the last states (Fig. 4b). Therefore, later we also introduce the concept of RNA velocity that helps to order states not only based on similarity but also on directionality of change from one state to another.

The integration of log change values as species velocity did not significantly affect the trajectory. Although the communities correctly move from the initial states towards the perturbation and resisting states, the arrows point in contradictory directions instead of a continuous movement from first state to the last.

In the second simulation, we used the same interaction matrix and metacommunity probabilities. We generated the growth rates similar to the first simulation with perturbation and resisting states lasting 100 timepoints each instead of 200. Similarly, the pseudotime ordering captured the initial and last states as early in time followed by perturbed and resisting states (Fig. 5a). The velocity arrows do not perform as expected by again pointing to contradicting directions (Fig. 5b). Further work needs to be done to adjust the methodology in order to have arrows pointing in the right direction.

D. Applying pseudotime analysis to human gut dataset

To check the validated pseudotime analysis on real dataset we performed the analysis on longitudinal human gut data. The dataset contained metagenomic data of fecal samples taken from 89 hosts in different time points over 18 months. For the analysis we used the data from one of the donors, having overall 205 samples.

The pseudotime ordering matches with the collection time with directionality from the early timepoints to the late ones (Fig. 6ab). We identified three macrostates, defined as subsets of states that are more likely to transition between each other than with states from other macrostates (Fig. 6cd). In addition, we also identified driver features (bacterial sequences) that correlated most with the transition between these macrostates. We identified sequences belonging to species of the Genus Bacteroides as the main drivers between these states. Interestingly, the original paper that has also performed high resolution sequencing of individual bacterial strains over time, has identified Bacteroides ovatus (a species from the same genus) that shows differential abundances of three different strains between three periods in time (Fig. 7) [31]. This means that this species may have an important role in driving longitudinal dynamics of the human gut microbiome.

This also demonstrates one of the advantages of adapting the methods from single-cell RNA sequencing data to not only cluster the microbiome states, but also identify species that drive the transitions between the states and thus understand the underlying biological mechanisms.

V. CONCLUSION

Microbial communities play an important role in organisms as the change in their composition are associated with numerous disorders. The computational methodologies for the analysis of microbiome datasets do not usually use a network approach for studying the dynamic changes in the community.
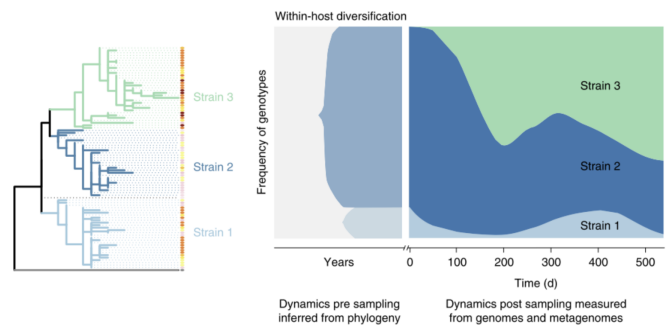

Fig. 7: Bacteroides ovatus abundance analysis [31].

Inspired by single cell transcriptomics and graph theory m we suggest subgraph extraction, pseudotime ordering and velocity analysis for microbial datasets. We show their validity on both simulated and real datasets and suggest a few ways the methods can be useful for further exploration.

REFERENCES

[1] Jack A Gilbert et al. "Current understanding of the human microbiome". In: *Nature Medicine* 24.4 (Apr. 2018), pp. 392–400. ISSN: 1546-170X. DOI: 10.1038/nm.4517. URL: http://dx.doi.org/10.1038/nm.4517.

[2] Catherine A. Lozupone et al. "Diversity, stability and resilience of the human gut microbiota". In: *Nature* 489.7415 (Sept. 2012), pp. 220–230. ISSN: 1476-4687. DOI: 10.1038/nature11550. URL: http://dx.doi.org/10.1038/nature11550.

[3] Muhammad Afzaal et al. "Human gut microbiota in health and disease: Unveiling the relationship". In: *Frontiers in Microbiology* 13 (Sept. 2022). ISSN: 1664-302X. DOI: 10.3389/fmicb.2022.999001. URL: http://dx.doi.org/10.3389/fmicb.2022.999001.

[4] Shuang Wang et al. "Microbial collaborations and conflicts: unraveling interactions in the gut ecosystem". In: *Gut Microbes* 16.1 (Dec. 2023). ISSN: 1949-0984. DOI: 10.1080/19490976.2023.2296603. URL: http://dx.doi.org/10.1080/19490976.2023.2296603.

[5] Cerano C. Da Silva, Michele A. Monteil, and Elaine Monica Davis. "Overweight and Obesity in Children Are Associated with an Abundance of Firmicutes and Reduction of Bifidobacterium in Their Gastrointestinal Microbiota". In: *Childhood Obesity* 16.3 (Apr. 2020), pp. 204–210. ISSN: 2153-2176. DOI: 10.1089/chi.2019.0280. URL: http://dx.doi.org/10.1089/chi.2019.0280.

[6] Andrew B. Onderdonk, Mary L. Delaney, and Raina N. Fichorova. "The Human Microbiome during Bacterial Vaginosis". In: *Clinical Microbiology Reviews* 29.2 (Apr. 2016), pp. 223–238. ISSN: 1098-6618. DOI: 10.1128/cmr.00075-15. URL: http://dx.doi.org/10.1128/CMR.00075-15.

[7] Ruiqi Lyu et al. "Methodological Considerations in Longitudinal Analyses of Microbiome Data: A Comprehensive Review". In: *Genes* 15.1 (Dec. 2023), p. 51. ISSN: 2073-4425. DOI: 10.3390/genes15010051. URL: http://dx.doi.org/10.3390/genes15010051.

[8] Jose Lugo-Martinez et al. "Dynamic interaction network inference from longitudinal microbiome data". In: *Microbiome* 7.1 (Apr. 2019). ISSN: 2049-2618. DOI: 10.1186/s40168-019-0660-3. URL: http://dx.doi.org/10.1186/s40168-019-0660-3.

[9] Amy D. Willis. "Rarefaction, Alpha Diversity, and Statistics". In: *Frontiers in Microbiology* 10 (Oct. 2019). ISSN: 1664-302X. DOI: 10.3389/fmicb.2019.02407. URL: http://dx.doi.org/10.3389/fmicb.2019.02407.

[10] David Berry and Stefanie Widder. "Deciphering microbial interactions and detecting keystone species with co-occurrence networks". In: *Frontiers in Microbiology* 5 (May 2014). ISSN: 1664-302X. DOI: 10.3389/fmicb.2014.00219. URL: http://dx.doi.org/10.3389/fmicb.2014.00219.

[11] Eduardo Álvarez-Miranda, Ivana Ljubić, and Petra Mutzel. "The Maximum Weight Connected Subgraph Problem". In: *Facets of Combinatorial Optimization*. Springer Berlin Heidelberg, 2013, pp. 245–270. ISBN: 9783642381898. DOI: 10.1007/978-3-642-38189-8_11. URL: http://dx.doi.org/10.1007/978-3-642-38189-8_11.

[12] David Freedman. *Markov chains*. Springer, 2012.

[13] Feng Xia et al. "Random Walks: A Review of Algorithms and Applications". In: *IEEE Transactions on Emerging Topics in Computational Intelligence* 4.2 (Apr. 2020), pp. 95–107. ISSN: 2471-285X. DOI: 10.1109/tetci.2019.2952908. URL: http://dx.doi.org/10.1109/TETCI.2019.2952908.

[14] Dragomirka Jovic et al. "Single-cell RNA sequencing technologies and applications: A brief overview". In: *Clinical and Translational Medicine* 12.3 (Mar. 2022). ISSN: 2001-1326. DOI: 10.1002/ctm2.694. URL: http://dx.doi.org/10.1002/ctm2.694.

[15] Lukas Heumos et al. "Best practices for single-cell analysis across modalities". In: *Nature Reviews Genetics* 24.8 (Mar. 2023), pp. 550–572. ISSN: 1471-0064. DOI: 10.1038/s41576-023-00586-w. URL: http://dx.doi.org/10.1038/s41576-023-00586-w.

[16] Malte D Luecken and Fabian J Theis. "Current best practices in single-cell RNA-seq analysis: a tutorial". In: *Molecular Systems Biology* 15.6 (June 2019). ISSN: 1744-4292. DOI: 10.15252/msb.20188746. URL: http://dx.doi.org/10.15252/msb.20188746.

[17] Koen Van den Berge et al. "Trajectory-based differential expression analysis for single-cell sequencing data". In: *Nature Communications* 11.1 (Mar. 2020). ISSN: 2041-1723. DOI: 10.1038/s41467-020-14766-3. URL: http://dx.doi.org/10.1038/s41467-020-14766-3.

[18] Cole Trapnell et al. "The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells". In: *Nature Biotechnology* 32.4 (Mar. 2014), pp. 381–386. ISSN: 1546-1696. DOI: 10.1038/nbt.2859. URL: http://dx.doi.org/10.1038/nbt.2859.

[19] Laleh Haghverdi et al. "Diffusion pseudotime robustly reconstructs lineage branching". In: *Nature Methods* 13.10 (Aug. 2016), pp. 845–848. ISSN: 1548-7105. DOI: 10.1038/nmeth.3971. URL: http://dx.doi.org/10.1038/nmeth.3971.

[20] Gioele La Manno et al. "RNA velocity of single cells". In: *Nature* 560.7719 (Aug. 2018), pp. 494–498. ISSN: 1476-4687. DOI: 10.1038/s41586-018-0414-6. URL: http://dx.doi.org/10.1038/s41586-018-0414-6.

[21] Marius Lange et al. "CellRank for directed single-cell fate mapping". In: *Nature Methods* 19.2 (Jan. 2022), pp. 159–170. ISSN: 1548-7105. DOI: 10.1038/s41592-021-01346-6. URL: http://dx.doi.org/10.1038/s41592-021-01346-6.

[22] H. Bremer and G. Churchward. "An examination of the Cooper-Helmstetter theory of DNA replication in bacteria and its underlying assumptions". In: *Journal of Theoretical Biology* 69.4 (Dec. 1977), pp. 645–654. ISSN: 0022-5193. DOI: 10.1016/0022-5193(77)90373-3. URL: http://dx.doi.org/10.1016/0022-5193(77)90373-3.

[23] Eduardo Álvarez-Miranda and Markus Sinnl. "A Relax-and-Cut framework for large-scale maximum weight connected subgraph problems". In: *Computers amp; Operations Research* 87 (Nov. 2017), pp. 63–82. ISSN: 0305-0548. DOI: 10.1016/j.cor.2017.05.015. URL: http://dx.doi.org/10.1016/j.cor.2017.05.015.

[24] Maximum weight connected subgraph (MWCS. *Alexander Loboda*. URL: https://github.com/ctlab/mwcsr.

[25] Yu Gao et al. "miaSim: an R/Bioconductor package to easily simulate microbial community dynamics". In: *Methods in Ecology and Evolution* 14.8 (May 2023), pp. 1967–1980. ISSN: 2041-210X. DOI: 10.1111/2041-210x.14129. URL: http://dx.doi.org/10.1111/2041-210X.14129.

[26] Tal Korem et al. "Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples". In: *Science* 349.6252 (Sept. 2015), pp. 1101–1106. ISSN: 1095-9203. DOI: 10.1126/science.aac4812. URL: http://dx.doi.org/10.1126/science.aac4812.

[27] Christopher T Brown et al. "Measurement of bacterial replication rates in microbial communities". In: *Nature Biotechnology* 34.12 (Nov. 2016), pp. 1256–1263. ISSN: 1546-1696. DOI: 10.1038/nbt.3704. URL: http://dx.doi.org/10.1038/nbt.3704.

[28] Akintunde Emiola and Julia Oh. "High throughput in situ metagenomic measurement of bacterial replication at ultra-low sequencing coverage". In: *Nature Communications* 9.1 (Nov. 2018). ISSN: 2041-1723. DOI: 10.1038/s41467-018-07240-8. URL: http://dx.doi.org/10.1038/s41467-018-07240-8.

[29] Yuan Gao and Hongzhe Li. "Quantifying and comparing bacterial growth dynamics in multiple metage-

nomic samples". In: *Nature Methods* 15.12 (Nov. 2018), pp. 1041–1044. ISSN: 1548-7105. DOI: 10 . 1038 / s41592-018-0182-0. URL: http://dx.doi.org/10.1038/ s41592-018-0182-0.

[30]   Tyler A. Joseph et al. "Accurate and robust inference of microbial growth dynamics from metagenomic sequencing reveals personalized growth rates". In: *Genome Research* 32.3 (Jan. 2022), pp. 558–568. ISSN: 1549-5469. DOI: 10.1101/gr.275533.121. URL: http://dx.doi. org/10.1101/gr.275533.121.

[31]   M. Poyet et al. "A library of human gut bacterial isolates paired with longitudinal multiomics data enables mechanistic microbiome research". In: *Nature Medicine* 25.9 (Sept. 2019), pp. 1442–1452. ISSN: 1546-170X. DOI: 10.1038/s41591-019-0559-3. URL: http://dx.doi.org/10. 1038/s41591-019-0559-3.