Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Application Note

OXFORD

Systems biology

MiSDEED: a synthetic multi-omics engine for microbiome power analysis and study design

Philippe Chlenski[1,*], Melody Hsu[1], and Itsik Pe'er[1,2,3]

[1] Department of Computer Science, Columbia University, New York, NY 10027, USA [2] Department of Systems Biology, Columbia University, New York, NY 10027, USA [3] Data Science Institute, Columbia University, New York, NY 10027, USA.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: MiSDEED is a command-line tool for generating synthetic longitudinal multi-omics data from simulated microbial environments. It generates relative-abundance timecourses under perturbations for an arbitrary number of samples and patients. All simulation parameters are exposed to the user to facilitate rapid power analysis and aid in study design. Users who want additional flexibility may also use MiSDEED as a Python package.

Availability and implementation: MiSDEED is written in Python and is freely available at https://github.com/pchlenski/misdeed.

Contact: pac@cs.columbia.edu

The behavior of the microbiome, greatly elucidated by improvements in genome sequencing and data analysis, is generating considerable research interest. For instance, the Human Microbiome Project (Turbaugh *et al*., 2007) endeavors to collect data on a mass scale to investigate the role of the microbiome in the context of human health and disease. Despite improvements in sequencing, sample collection itself still incurs significant overhead and many niches remain understudied. Furthermore, microbial relative abundance data, the most typical form of data collected in such studies, has a number of properties that make classical statistical analysis challenging: it is longitudinal, compositional, noisy, and stochastic (Antoine *et al*., 2019). The genetic power calculator (Purcell *et al*., 2003) streamlined research in statistical genetics by facilitating closed-form power analysis of hypothetical studies. Similarly, several tools help design microbiome studies: Web-GLV (Kuntal *et al*., 2019) enables researchers to visualize the dynamics of microbial systems using assumed ecological parameters, and Mattiello *et al*. (2016) provide a power calculator for case-control studies on microbial ecosystems near equilibrium. The generalized Lotka-Volterra (gLV) modeling assumptions underlying such tools are often used to generate synthetic data when designing inference methods for microbial relative abundance data (Joseph *et al*., 2020). Here we present MiSDEED: the *Mi*crobial *S*ynthetic *D*ata *E*ngine for *E*xperimental *D*esign, a flexible tool for generating synthetic longitudinal data from dynamic simulated ecosystems. These simulations can help investigators allocate resources in their study design, determine how to salvage underpowered studies, and design novel inference techniques with known ground truth.

MiSDEED's synthetic data generator (drawn in Figure 1) samples reads from probability distributions governed by gLV dynamics over a discrete set of time points T. Each generator has a set I of nodes which may represent different data types (e.g. metagenomics and metabolomics measurements of the same system) or two interacting ecosystems with the same data type. Each node $i \in I$ is initialized with a fixed dimensionality d_i, a vector of growth rates $\overrightarrow{g_i}$, and an initial abundance vector $\overrightarrow{x_{i,0}}$. A generator also has up to $|I|^2$ pairwise directed interactions between nodes. An interaction $\mathbf{A}$ between some nodes i and i' is a matrix of dimension $d_i \times d_{i'}$. Finally, the generator has a set J of interventions which may be applied to any node $i \in I$ such that each intervention j has a vector $\overrightarrow{u_j}$ of intervention magnitudes and another vector $\overrightarrow{b_j}$ of responses to the intervention. If intervention j is applied to node i, then $\overrightarrow{u_j}$ should have T dimensions and $\overrightarrow{b_j}$ should have d_i dimensions. To aid in parameter selection, MisDEED also contains convenience functions to initialize random but stable interaction matrices (Allesina and Tang, 2012) or to infer gLV parameters from known absolute abundance data (e.g. from a small pilot dataset) (Stein *et al*., 2013). Once generator parameters have been set, synthetic data can be produced in one of three ways: as a single timecourse, as multiple timecourses from varying initial conditions, or as multiple timecourses following a case-control split. In each case, the generator numerically solves the following equations with a biological

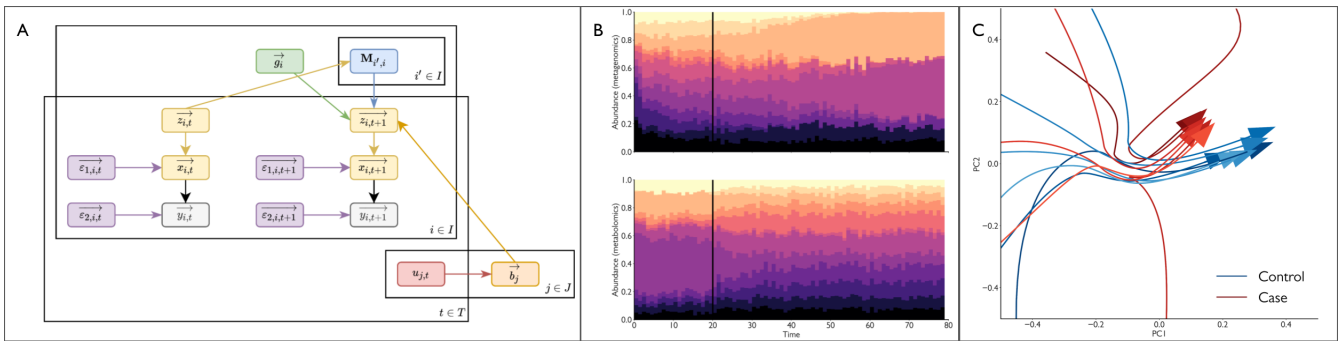

Fig. 1. (A) The graphical model underlying MiSDEED's synthetic data engine. The $\vec{y_t}$ vectors are sampled from a multinomial distribution parameterized by the total number of reads and the probability vector $\vec{x_t}$. (B) Simulated metagenomic (top) and metabolomic (bottom) relative-abundance timecourses with an intervention at $t = 20$ (black line). This intervention affects metabolite abundances directly and propagates into the metagenomics node gradually via metabolomics-metagenomics interactions. (C) 12 noiseless PCA-transformed case (red) and control (blue) metagenomic trajectories show how interventions induce convergence to distinct fixed points.

Category	Parameters
Generator	Number of time points, number of nodes, node names, node dimensions, time to first sample
Random interaction matrices	C (connectivity), d (negative self-interaction size) σ (multivariate normal variance), ρ (multivariate normal correlation)
Custom gLV parameters	Interaction matrices, growth rates, initial abundances, interventions and intervention responses
Synthetic data generation	Biological noise variance, number of reads, time step size, downsampling rate
Multiple samples	Number of individuals, probability of 0-valued initial abundances
Case-control	Case-control ratio, intervention node, intervention effect size

Table 1. Variable parameters in MiSDEED.

noise term $\vec{\varepsilon} \sim \mathcal{N}(\vec{0}, \sigma)$

$$\frac{d\vec{z_{i,t}}}{dt} = \vec{g_i} + \sum_{i' \in I} \mathbf{A}_{i,i'} \vec{z_{i',t}} + \sum_{j \in J} u_{j,t} \vec{b_j} + \vec{\varepsilon} \tag{1}$$

Each timecourse contains three derived matrices of synthetic data: Z (latent absolute abundances), X (latent relative abundances/probabilities), and Y (relative abundances sampled from Y).

MiSDEED is designed to be used as a standalone command line tool. The MiSDEED repository also contains the Python package underlying MiSDEED, a handful of utility scripts to support data visualization and learning gLV parameters, and a set of Jupyter notebooks showing common uses of the MiSDEED Python package. MisDEED can produce, save, and plot large amounts of synthetic data with varying initial conditions and model assumptions. To support power analysis, many variables can freely be changed by the user. These are listed in Table 1.

As an example use case, one may use a community matrix and growth rates learned from a pilot dataset and initialize 'metagenomics' and 'metabolomics' nodes such that the latter has no intrinsic growth rates, weak self-interactions, but strong interactions with the 'metagenomics' node according to some a priori assumptions. Perturbing metabolite abundances directly, the user may investigate how many patients must be enrolled in order to distinguish reliably between samples with and without this perturbation applied.

MiSDEED is a flexible framework for rapidly generating large amounts of realistic microbial trajectory data, thereby facilitating study design without assuming the system being studied is at equilibrium. Future development will focus on expanding code-free interfaces to MiSDEED; more flexible modeling assumptions for broader use cases, including non-uniform time points, individual variation in interaction matrices and growth rates, and population clusters; alternatives to gLV-based modeling for dynamics like mutualism; and investigation into the value of MiSDEED-generated data for transfer learning and algorithm development.

Acknowledgments

The work was supported by NIH/NCI Grant No. U54CA209997 Driving Biological Projects and Columbia University's 2020/2021 Data Science Institute Seed Grant.

References

Turnbaugh, P., *et al*. (2007) The Human Microbiome Project, *Nature* **449**, 804–810. doi: 10.1038/nature06244.

Antoine *et al*. (2019) A Generic Multivariate Framework for the Integration of Microbiome Longitudinal Studies With Other Data Types, *Frontiers in Genetics* **10**, 963. doi: 10.3389/fgene.2019.00963.

Purcell, S., Cherny, S., and Sham, P. (2003) Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* **19**, 149-50. doi: 10.1093/bioinformatics/19.1.149.

Kuntal, B., Gadgil, C., and Mande, S. (2019) Web-gLV: A Web Based Platform for Lotka-Volterra Based Modeling and Simulation of Microbial Populations, *Frontiers in Microbiology* **10**, 288. doi: 10.3389/fmicb.2019.00288.

Mattiello, F., *et al* (2016) A web application for sample size and power calculation in case-control microbiome studies. Bioinformatics **13**, 2038-40. doi: 10.1093/bioinformatics/btw099.

Joseph, T., Pasarkar, A., and Pe'er, I. (2020) Efficient and Accurate Inference of Mixed Microbial Population Trajectories from Longitudinal Count Data, *Cell Systems* **10.6**, 463-469. doi: 10.1016/j.cels.2020.05.006.

Allesina, S., and Tang, S. (2012) Stability criteria for complex ecosystems. *Nature* **483**, 205–208. doi: 10.1038/nature10832.

Stein, R., *et al*. (2013) Ecological Modeling from Time-Series Inference: Insight into Dynamics and Stability of Intestinal Microbiota. *PLOS Computational Biology* **9(12)**: e1003388. doi:10.1371/journal.pcbi.1003388