Metabolite Overproduction through Engineering and Optimization of Microbiome Composition Dynamics

Stephen Lincoln
Department of Chemical and Biomolecular Engineering
University of Connecticut, Storrs, CT, USA
191 Auditorium Road
Storrs, CT 06268
Stephen.Lincoln@uconn.edu

Jacquelynn Benjamino
Department of Molecular And Cell Biology
University of Connecticut, Storrs, CT, USA
91 North Eagleville Road
Storrs, CT 06268
Jacquelynn.Benjamino@uconn.edu

Joerg Graf
Department of Molecular And Cell Biology
University of Connecticut, Storrs, CT, USA
91 North Eagleville Road
Storrs, CT 06268
Joerg.Graf@uconn.edu

Ranjan Srivastava
Department of Chemical and Biomolecular Engineering
University of Connecticut, Storrs, CT, USA
191 Auditorium Road
Storrs, CT 06268
Srivasta@engr.uconn.edu

ABSTRACT

Although many advances have been made in genome sequencing for analyzing the composition of microbiomes, very few studies have attempted to learn and model their dynamics. Furthermore, no studies have attempted to exploit the dynamics of compositional changes of a microbiome for overproducing a metabolite of interest. This task proves to be computationally difficult at best and intractable at worst. The challenge is due to the complex, interdependent, and large number of highly non-linear interactions among members of a microbiome, as well as environmental factors, e.g. substrate. Here, we present a computationally tractable strategy using machine learning methods and stochastic optimization to characterize and potentially engineer a microbiome. In this work, an artificial neural network (ANN) is utilized to learn how six different lignocellulose food sources affect the temporal composition of the hindgut microbiome of *Reticulitermes flavipes*, the eastern subterranean termite. The learned dynamics from the ANN are optimized using either a genetic algorithm or artificial immune system approach. Specifically, the optimization objective is the maximization of the *Rhodospirillales*, an acetate producing order of bacteria, which will intrinsically maximize acetate production from the microbiome. The genetic algorithm and artificial immune system are compared for robustness and speed.

Keywords

Biology and chemistry; Artificial immune systems; Genetic algorithms; Artificial neural networks; Microbiome

1. INTRODUCTION

The emergent field of microbiome research has become increasingly important with the surge in genome sequencing technology. For example, one of the main goals of the Human Microbiome Project (HMP) is to analyze genetic sequences of microbiomes from various parts of the human body, such as the gut and skin, and determine if relationships between microbiome composition and diseases exist [6]. Consequently, metagenomics studies which analyze large amount of sequencing data have been able to link changes in human microbiome composition to maladies, such as obesity [5], inflammatory bowel disease [2], and type 2 diabetes [3]. Although metagenomic advances have helped give insight to the microbiome, there have been only a few studies attempting to model and learn the dynamics of the microbiome due to the difficulty in identifying the numerous interrelationships among community members, the computational difficulty of modeling those highly nonlinear, and other external influences such as substrate, temperature, pH, micronutrient concentrations, etc.

In this work, we propose an algorithm which utilizes an artificial neural network to learn the dynamics of the microbiome present in the hindgut of *R. flavipes*. The learned dynamics are then used in conjunction with an artificial immune system and a genetic algorithm to engineer the microbiome and determine a substrate regimen to maximize the relative abundance of the order *Rhodospirillales*. *Rhodospirillales* was chosen due to the order's ability to create acetic acid as a product.

GECCO'16 Companion, July 20-24, 2016, Denver, CO, USA
ACM 978-1-4503-4323-7/16/07.
http://dx.doi.org/10.1145/2908961.2908999

2. METHODS

2.1 Termite Hindgut Sequencing

R. flavipes termites were separated into colonies based on which single substrate diet they were to receive. These substrates included spruce, cardboard, oak, maple, mulch, birch, and one colony was starved for a total of seven colonies. Termites from each colony were sampled on the day of arrival (day 0) and on days 1, 2, 3, 7, 14, 21, 28, 35, 42, 49 after arrival. Before arrival and on day 0 the termites were fed mulch. Hindgut samples were sequenced with an Illumina MiSeq using 16S rRNA sequencing to determine the relative abundance of operational taxonomic units (OTUs) at a given time. Each OTU which was identified was originally drilled down to the species level. Before being fed to the neural network, the OTUs were grouped by taxonomic order in order to reduce noise in the algorithm.

2.2 Artificial Neural Network

A deep backpropagation artificial neural network (ANN) was created using Fast Artificial Neural Network (FANN) [4] with Python bindings. The number of input nodes was set to the number of taxonomic orders plus the number of substrates present/lacking in the given colony for a total of 70 inputs nodes. The number of output nodes was set to 64, or the number of taxonomic orders present. The ANN utilized two hidden layers with 67 and 60 nodes, respectively. The ANN was trained by feeding the relative abundance of each taxonomic order and substrates present/lacking at time period t as inputs and the relative abundance of each taxonomic order for time period t+1 as

the target for each time point in each colony. The network was trained until the error was below 10^{-5}.

2.3 Microbiome Engineering & Optimization

After the ANN is sufficiently trained, an artificial immune system (AIS) were compared to a genetic algorithm (GA) to maximize the order *Rhodospirillales*. In order to accomplish this goal, each algorithm generated a random population of 100 members. Since most colonies were sampled at ten time points and there are six substrates available to be fed to each colony, each member of the population was a random bit array of length 60. Each bit in each set of 6 bits in the array is representative of the presence or lack of a certain substrate given by a 1 or 0, respectively. Each set of six bits is representative of a sampling time period. To start, a test time point of the relative abundance of taxonomic orders was fed to the ANN with the first six bits of the array. The predicted relative abundance of *Rhodospirillales* was recorded, and the predicted relative abundances of each taxonomic order was then fed back into the ANN with the next 6 bits of the array. This is repeated for the length of the array. The final score of the member of the population is the composite trapezoidal rule for the relative abundance of *Rhodospirillales* at each time point. In other words, the area under the curve of predicted relative abundance of *Rhodospirillales* over time is the fitness score of the member of the population. After scoring all members, a new population was generated. For the AIS, the top 35% solutions were kept as memory solutions and the remaining population was generated based off of a simplified CLONALG [1] algorithm. For the GA, new solutions were generated by elitist selection and single point crossover. Each algorithm was run for 150 iterations. A range of mutation rates were tested for each algorithm. In addition, two versions of each algorithm were created: one version which imposed a limit of up to one substrate per time period and another version which had no substrate limit.

3. RESULTS

To test the ANN, seven time points were left out of the training set and used in the testing set. Each time point was fed into the ANN and the predicted relative abundances of each OTU was compared to the actual relative abundances of each OTU. The root mean squared error (RMSE) and Bray-Curtis similarity were determined for each test point and averaged across the set. The averaged RMSE and Bray-Curtis similarity were 0.0229 and 0.8576, respectively. The best solution for the AIS and GA were returned for each mutation rate, as presented in **Table 1** and **Table 2**.

Table 1. **TrapZ score of *Rhodospirillales* for the best solution using AIS with various mutation rates**

Mutation Rate	No Substrate Limit	Substrate Limit
5%	0.8393	0.6646
10%	0.9233	0.6646
15%	0.9225	0.6646
20%	0.9269	0.7131
25%	0.9242	0.7131
30%	0.9021	0.6646

Table 2. **TrapZ score of *Rhodospirillales* for the best solutions using GA with various mutation rates**

Mutation Rate	No Substrate Limit	Substrate Limit
2%	0.7596	0.6576
5%	0.7623	0.7342
7%	0.8099	0.6646
9%	0.8247	0.6646
12%	0.7814	0.6646
15%	0.6606	0.6646

Overall, the AIS performed better than the GA in terms of solution convergence and fitness. In terms of the predicted substrates present for each time period to maximize *Rhodospirillales*, both versions of the AIS and GA contained maple in the last seven time points for the majority of the top solutions. During the first three time points, the relative abundance of *Rhodospirillales* was the highest when multiple substrates were present.

4. CONCLUSION

This work focused on implementation of a deep, backpropagation artificial neural network in order to learn the dynamics of the hindgut microbiome of *R. flavipes*. The dynamics learned included the effect of substrate fed to the termite on the relative abundance of taxonomic orders in the microbiome. The goal of both the AIS and GA was to create a substrate regimen that would maximize the relative abundance of *Rhodospirillales*, an acetic acid producing order of bacteria. Two versions of both the AIS and GA were implemented; one version only allowed up to one substrate to be present at a time, while the other version had no limit on the number of substrates present. Overall, the AIS was able to converge on better solutions than the GA. Food regimens returned from the algorithms suggested that using maple as a substrate in the last seven time points, as well as multiple substrates in the first three time points are essential to maximize *Rhodospirillales* abundance in the microbiome. It is likely that this outcome occurred because the ANN was trained on samples taken from colonies only fed one substrate. To confirm these results, more experimental studies will need to be performed.

5. REFERENCES

[1] Brownlee, J. 2005. Clonal selection theory & CLONALG-The Clonal selection classification algorithm (CSCA). *Swinburne University of Technology*. (2005).

[2] Greenblum, S. et al. 2012. Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. *Proceedings of the National Academy of Sciences*. 109, 2 (Jan. 2012), 594–599.

[3] Larsen, N. et al. 2010. Gut Microbiota in Human Adults with Type 2 Diabetes Differs from Non-Diabetic Adults. *PLoS ONE*. 5, 2 (Feb. 2010), e9085.

[4] Nissen, S. 2003. Implementation of a fast artificial neural network library (fann). *Report, Department of Computer Science University of Copenhagen (DIKU)*. 31, (2003).

[5] Turnbaugh, P.J. et al. 2009. A core gut microbiome in obese and lean twins. *Nature*. 457, 7228 (Jan. 2009), 480–484.

[6] Turnbaugh, P.J. et al. 2007. The human microbiome project: exploring the microbial part of ourselves in a changing world. *Nature*. 449, 7164 (Oct. 2007), 804–810.