Dr. Trevor M. Cickovski
Assistant Teaching Professor of Computer Science
Florida International University

Philosophy of Research

Completed: July 2020

This research philosophy is built on **reality as an infinitely evolving picture, composed of a puzzle consisting of multiple interdependent pieces that fit together to complete this picture**. Yet a deeper look at these pieces reveals there is knowledge beneath and above the surface, or "worlds-within-worlds" (*cf.* Kevin McCormick). An important *world* that exists within not just humans but many living organisms and environmental niches, is the **microbiome**. And in this "worlds-within-worlds" model, **behavior at the microscale impacts behavior at the macroscale**. Likewise we as humans perform at the level of the collective function of our parts, and that includes our microbiome. Progressing to larger scales society will perform at the level of its collective members, which include humans, other organisms, and the environment. Therefore microbiomes are naturally ubiquitous, and very powerful.

A microbiome represents the collection of microbes that inhabit some environmental niche, with some of these niches existing inside other organisms ("hosts"). Microbes include bacteria, single-celled archaea, fungi, protists, and viruses. The "human gut microbiome", for example, is the collection of microbes inhabiting a human gut (every human has a distinct gut microbiome). The human itself has many other microbiomes, including the skin, oral, lung, *etc.* Most of these are essential for our survival. While mapping DNA was an enormous accomplishment, these implications lead to the correct conclusion that organisms are not just composed of their own cells but others. In fact recent counts estimate the human body as composed of at least 50% microbial cells (1).

The role of the microbiome in collective health is illustrated by the diagram at the end of this philosophy. The central dogma in biology (heavily summarized) states that based on unique cellular DNA of an organism, a specific set of *metabolites* (in other words, products that are essential for an organism's metabolism, or sustenance), will be produced (M1- in the figure). These interact in a variety of different ways, making new products that in turn interact, and so forth. Microbial cells have different DNA from the host, with each strain having unique DNA. This genetic material produces other metabolites, which interact with those produced by other microbes and the host, resulting in a complex interaction web that determines functionality of the organism. Some metabolites are also produced and consumed by microbes as well as the host, serving as nutrients or toxins. Therefore the microbiome and host compose an *ecosystem* of organisms, with *relationships* (i.e. cooperation or competition) heavily dependent on this interaction web.

Returning to the base, microbes and their corresponding metabolites are simply other pieces of this dynamic picture. They can interact in ways that advance ('positive') or restrict ('negative'). One way or another, a properly functioning microbiome is an essential component for optimal performance of the host. We are only cracking the iceberg as far as untangling the complex interaction web that forms the host-microbiome backbone. To progress efficiently, several contributions will be necessary:

1. ***Application to mental illness***. As if this picture were not complex enough, the recent discovery of the **gut-brain axis** (2) implies a dual "communication channel" involving gut and brain metabolites, meaning behavior in the gut impacts behavior in the brain, and vice-versa. Further, these metabolites include *neuromodulators* such as dopamine, serotonin and noradrenaline – where dysregulation is the key component of mental illness. For example, an estimated 90% of serotonin is actually synthesized in the gut, not the brain (3). Numerous studies have discovered microbes capable of synthesizing neuromodulators. Multiple correlations have been discovered between GI dysfunction and mental illnesses such as depression, anxiety, ADHD and autism. Studies have illustrated that mental illness implies a unique gut microbiome profile compared to controls. Probiotics have already shown positive effects when administered to autism patients (4). This establishes quite strong evidence of a connection

between the gut microbiome and mental illness, implying that the specific collection of microbes in the gut and their metabolites participate in downstream chemical reactions that directly or indirectly impact brain behavior. Recently proposed renaming of this communication channel to the gut-brain-microbiome axis (5) indicate this to be a strongly accepted idea in the research community.

Mental illness is an area that realistically we are only beginning to understand. The twentieth century witnessed an incredible shift in illness type frequency due to medical and antibiotic advancements, from a decline in infectious disease to a rise in mental illness. According to the National Institute of Mental Health, roughly one-fifth of Americans are suffering from some form of mental illness, illustrating the importance of understanding this field. As with any field, understanding requires knowledge of the players, which strong evidence suggests includes the gut microbiome and the underlying metabolic pathways between its members and the host (6).

Much of this evidence has come from *differential analysis*, intended to determine whether or not there is a difference between the gut microbiome of mental illness patients compared to controls. It is now time to delve deeper and analyze the gut microbiome and host as a *collective* system. Algorithms have been developed to build *Microbial Social Networks* (7) based on abundance correlations of microbes within gut samples. In other words, microbes that tend to co-occur or co-avoid can be estimated to have cooperative or competitive ecological relationships. *Clustering* on such networks can determine social and rival microbial communities. *Centrality* algorithms (8,9) can run on these signed and weighted social networks to determine important entities within the ecosystem. Since the microbiome is not static, but dynamic, *time-series analysis* (10) will help form a fully accurate view of this picture. Not all ecological relationships are two-way; there is commensalism (positive one-way), amensalism (negative one-way), and even parasitism (positive one-way and negative the other). *Causal analysis* (11) will illuminate this area, and has already been used to predict colonization order in the oral microbiome. *Multi-omics* analysis, which incorporates metagenomics (analysis of microbial and host genomes) and metabolomics (analysis of underlying metabolic pathways), helps complete the picture by integrating the underlying web of interactions (12,13). The more complete this picture becomes, the more accurately mental illness can be both understood and diagnosed, as particularly the latter is currently heavily symptomatic in nature.

2. ***Implications for more natural treatments and individualized medicine***. The "antibiotic debate" has been raging for years, and occupying a polar stance on either side of this argument holds dangerous implications. Antibiotics cannot be avoided at the risk of allowing pathogenic bacteria to survive and thrive. At the same time, the reality of the microbiome as an ecosystem makes the traditional purpose of the antibiotic, to target and kill one specific microbe, a risky investment. Killing off one microbe that has supportive and oppositional relationships with other entities in the microbiome implies a change to the overall ecosystem (similar to the implications involved with wiping out bunnies in a forest). Antibiotics are not the only treatment subject to these implications, but also medications and even diet – which introduce chemicals into the underlying web of interactions thereby influencing microbiome and host functionality, must be as well. Indeed we may even be seeing this manifest itself right now in the form of potential side effects, which are often long lists suggesting effects can be different depending on the individual. This offers further support for involvement of the microbiome in this equation, which is also unique to every individual.

The field of pre- and pro-biotics is gaining momentum and can sometimes be an alternative to more targeted treatments like antibiotics and medications. The latter may continue to have a place. Our current system of checks-and-balances for administration is the monitoring by doctors for potential side effects who may stop treatment if observed. Representing an individual's microbiome and metabolic network *in silico* and modeling the effect of a targeted treatment on the overall ecosystem holds the advantage that

these checks and balances take place before (as opposed to after) administration, reducing patient risk and doctor time. These models could also develop customized pre- and pro-biotics for individuals, with optimal percentages of microbial abundance depending on their unique ecosystem, determined through *in silico* parameter sweeps and observations.

3. ***Free and open exchange of ideas***. Advancing microbiome research requires looking at the players, facilitating ways to maximize communication, and minimizing re-inventing the wheel. Microbiome research is interdisciplinary and its players include biologists, chemists, physicists, mathematicians, bioinformaticians and computer scientists. Microbiome analyses can generally be broken into two categories: (a) *in vitro*/*in vivo* lab experiments and (b) *in silico* downstream analyses. Not all of (a) must take place before (b), for example (a) can sometimes validate (b). Some of (a) will take place before (b), since microbial samples must first be placed in an electronic format for *in silico* analysis. This often involves sequencing a sample using techniques such as PCR, producing a set of DNA sequences that can then be subsequently analyzed by computational tools.

In silico is of course the most natural role for a computer scientist in this dominion, and downstream analysis is *algorithmic* in nature. Sequences are often *clustered* by some similarity metric then looked up in a *database* to determine a most likely microbe match (called a "taxon"). With resulting estimates of member taxa and counts, *diversity analysis* can be performed on these collective populations. There are multiple algorithms for this (alpha- or beta-, Chao, Simpson, Shannon indices). Metabolic interactions, as well as ecological relationships between microbes and the host, are conveniently represented as networks. This implies that *network analysis* is applicable. Important members of the microbial community can be ascertained using *biomarker analysis* on a sample or *centrality* analysis on a network. *Differential analysis* can also be performed on sample sets (i.e., healthy and diseased) to determine the degree to which their microbiome is distinct, and there a multitude of algorithms for this (from simple Euclidean distance to feature detection). Networks can be *visualized* in a variety of ways.

Note the above analysis has a sequential, not spatial, flavor. Clustering must be done before the database lookup, which must be done before diversity, biomarker or network analysis. Network analysis is a prerequisite to centrality or visualization. Even limiting ourselves to these types of analysis, by no means a complete set, still creates a four-stage pipeline summarized as: (1) cluster, (2) lookup, (3) population analysis and (4) visualization. New algorithms are continuously being developed for these four stages, by independent teams of researchers.

Witness then the resulting challenge. Each independent team contains interdisciplinary researchers, and each has a preferred programming language for algorithm development and testing. One team builds a new algorithm for a particular stage (i.e. stage 3), but must test it inside a pipeline alongside the other stages (in this case (1), (2) and (4)). There are tools available for these latter stages built by other research teams, using languages and file formats unfamiliar to the developer. They may have external dependencies or require a large CPU or memory footprint, exceeding the computational resources of their lab. The amount of time spent integrating their new stage 3 with multiple tools developed by independent research teams now becomes extremely significant, possibly (maybe even likely) exceeding the amount required to recreate the entire pipeline from scratch in their language of choice. Indeed particularly the early years of microbiome research resulted in an explosion of software pipelines (Mothur (14), Qiime (15), Galaxy (16), MEGAN (17) – in fact a 2017 count by BioStar (18) tallied nearly fifty). Many of these shared pipeline stages, differing in only a few places, pointing to a great deal of reinventing the wheel. Each of these tools are well-constructed, and many emerged as top-quality analysis tools. The deficiency lies not in the construction, but the lack of communication channels between algorithm developers. The field will progress faster when reinventing the wheel is minimized.

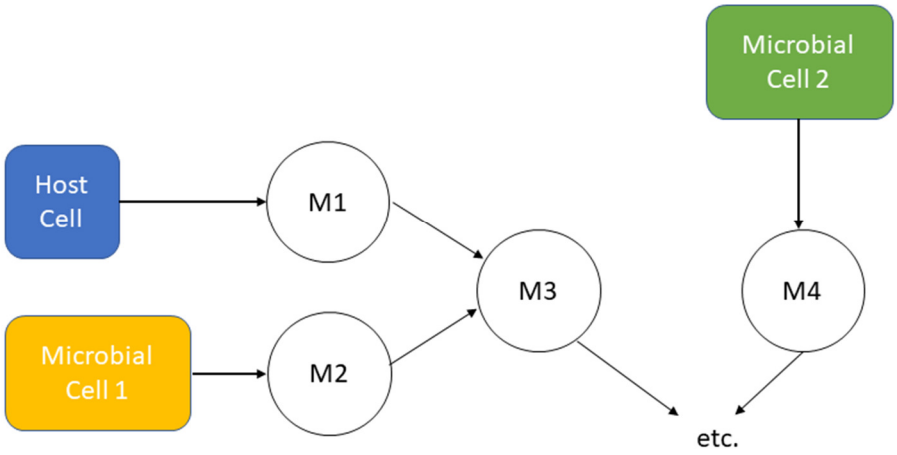
A pipeline can be viewed as a series of stages, where at each stage a different algorithm can be "plugged in" to accomplish the task (for the clustering stage, this can be "KMeans" (19), "PAM" (20), "AffinityPropagation" (21), *etc*., or a new algorithm). This is illustrated in the attached supplementary figure, with known potential algorithms for the mentioned example four-stage pipeline. The task then becomes to create an environment where a researcher can develop a plugin in their programming language of choice, and test aside other plugins through a uniform user interface that is independent of source language or file formats.

This environment is available through the toolkit Plugin-Based Microbiome Analysis (PluMA, (22)), a lightweight package freely available on the web at (http://biorg.cs.fiu.edu/pluma/). Dynamically loading plugins at runtime creates a backend that is lightweight (under 1 MB) in size, keeping the package manageable for users with limited computational resources (23). Several compiled (C++, CUDA) and scripted (Python, Perl, and R) languages are acceptable for plugin construction, allowing flexibility for algorithm developers to represent their ideas in their language of choice. Through a uniform user interface between plugins that consists of just an input and output file, a user can assemble pipelines using a wide array of plugins without knowing anything about their underlying implementation. By maintaining centralized online plugin and pipeline pools with a standardized method for testing and monitoring to ensure reliability, developers have easy access to plugins they need to assemble test pipelines. In 2019, the PluMA plugin pool contained 70 plugins. This count has risen to more than 225 upon the new release of PluMA 1.1 in 2020, an encouraging result that this "natural" growth in microbiome research is currently taking place.

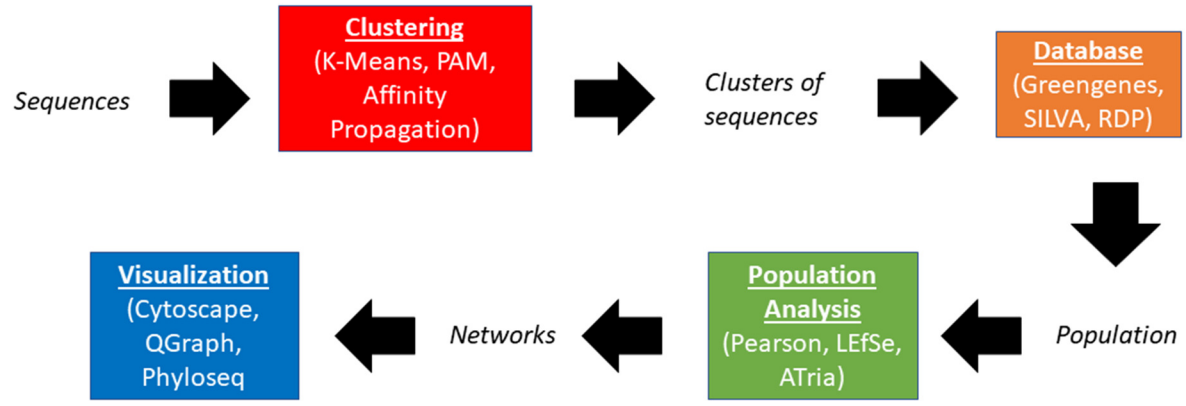
Independent of the underlying structure of a software package, building a solid user base will always demand usability. Now that the plugin pool is growing in size almost daily on average, 2020 begins the development of PluMA 2, with an entirely new shift of focus from internal flexibility to the external audience. The current "command-line" flavor of PluMA, which was natural given that many pipelines perform big data analysis and are ultimately run on server machines, carries limitations with respect to a user who simply wishes to build and test an algorithm on a small dataset. Microbiome research is interdisciplinary, and venturing outside of computer science will result in likely increases in more graphical systems (i.e. Windows) and visualization engines. Although because of its lightweight nature PluMA carries with it minimal dependencies, plugins can be developed by anyone and thosedependencies can be arbitrary, increasing the demand for self-installing libraries and executables or Dockerized containers. Finally, cloud-based versions of the package will be convenient for ensuring continued compatibility with multiple types of user devices. As its user-base grows, more continuous monitoring of the PluMA plugin and pipeline pools, as well as cloud-based computational resources, will be necessary. Through programs such as CISE Research Infrastructure (CRI), the National Science Foundation provides funding opportunities to develop and maintain such infrastructures.

Microbiome research is in its infancy. It is both enlightening and humbling to realize that we are more than simply our own cells, with internal ecosystems that must be maintained and kept healthy. There is much to be discovered with enormous implications, enriching this area of research with opportunity. It is important to be content to make small strides, as in the words of Hamilton, "legacy is about planting seeds you may never get to see grow." The future begins now!

SUPPLEMENTARY MATERIAL: WEB OF INTERACTIONS



SUPPLEMENTARY MATERIAL: SAMPLE PIPELINE



REFERENCES

1. Sender R, Fuchs S, Milo R. Revised Estimates for the Number of Human and Bacteria Cells in the Body. PLoS Biol. 2016 Aug 19;14(8):e1002533.
2. Berthoud H -r. Vagal and hormonal gut-brain communication: from satiation to satisfaction: Gut-brain communication. Neurogastroenterology & Motility. 2008 Apr 8;20:64–72.
3. Pettersson G. The neural control of the serotonin content in mammalian enterochromaffin cells. Acta Physiol Scand Suppl. 1979;470:1–30.
4. Shaaban SY, El Gendy YG, Mehanna NS, El-Senousy WM, El-Feki HSA, Saad K, et al. The role of probiotics in children with autism spectrum disorder: A prospective, open-label study. Nutritional Neuroscience. 2018 Oct 21;21(9):676–81.
5. Martin CR, Osadchiy V, Kalani A, Mayer EA. The Brain-Gut-Microbiome Axis. Cellular and Molecular Gastroenterology and Hepatology. 2018;6(2):133–48.

6. Mathee K, Cickovski T, Deoraj A, Stollstorff M, Narasimhan G. The gut microbiome and neuropsychiatric disorders: implications for attention deficit hyperactivity disorder (ADHD). Journal of Medical Microbiology. 2020 Jan 1;69(1):14–24.
7. Fernandez M, Riveros JD, Campos M, Mathee K, Narasimhan G. Microbial "social networks." BMC Genomics. 2015;16(Suppl 11):S6.
8. Cickovski T, Peake E, Aguiar-Pulido V, Narasimhan G. ATria: a novel centrality algorithm applied to biological networks. BMC Bioinformatics. 2017 Jun;18(S8):239.
9. Cickovski T, Aguiar-Pulido V, Narasimhan G. MATria: a unified centrality algorithm. BMC Bioinformatics. 2019 Jun;20(S11):278.
10. Lugo-Martinez J, Ruiz-Perez D, Narasimhan G, Bar-Joseph Z. Dynamic interaction network inference from longitudinal microbiome data. Microbiome. 2019 Dec;7(1):54.
11. Sazal MR, Ruiz-Perez D, Cickovski T, Narasimhan G. Inferring Relationships in Microbiomes from Signed Bayesian Networks. In: 2018 IEEE 8th International Conference on Computational Advances in Bio and Medical Sciences (ICCABS) [Internet]. Las Vegas, NV: IEEE; 2018 [cited 2020 Mar 27]. p. 1–1. Available from: https://ieeexplore.ieee.org/document/8542086/
12. Cickovski T, Manuel A, Mathee K, Campos M, Narasimhan G. Effects of Various Alpha-1 Antitrypsin Supplement Dosages on the Lung Microbiome and Metabolome. In: Măndoiu I, Murali TM, Narasimhan G, Rajasekaran S, Skums P, Zelikovsky A, editors. Computational Advances in Bio and Medical Sciences. Cham: Springer International Publishing; 2020. p. 90–101. (Lecture Notes in Computer Science; vol. 12029).
13. Aguiar-Pulido V, Huang W, Suarez-Ulloa V, Cickovski T, Mathee K, Narasimhan G. Metagenomics, Metatranscriptomics, and Metabolomics Approaches for Microbiome Analysis: Supplementary Issue: Bioinformatics Methods and Applications for Big Metagenomics Data. Evolutionary Bioinformatics. 2016 Jan;12s1:EBO.S36436.
14. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. AEM. 2009 Dec 1;75(23):7537–41.
15. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010 May;7(5):335–6.
16. Blankenberg D, Von Kuster G, Bouvier E, Baker D, Afgan E, Stoler N, et al. Dissemination of scientific software with Galaxy ToolShed. Genome Biol. 2014;15(2):403.
17. Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. Genome Research. 2007 Feb 6;17(3):377–86.
18. Parnell LD, Lindenbaum P, Shameer K, Dall'Olio GM, Swan DC, Jensen LJ, et al. BioStar: An Online Question & Answer Resource for the Bioinformatics Community. Bourne PE, editor. PLoS Comput Biol. 2011 Oct 27;7(10):e1002216.
19. Macqueen J. Some methods for classification and analysis of multivariate observations. In: Proc Fifth Berkeley Symp on Math Statist and Prob, Vol 1. University of California; 1967. p. 281–97.
20. Kaufman L, Rousseeuw P. Clustering by means of Medoids. In: Statistical Data Analysis Based on the L1-Norm and Related Methods. North-Holland: Y Dodge (ed); 1987. p. 405–16.
21. Frey BJ, Dueck D. Clustering by Passing Messages Between Data Points. Science. 2007 Feb 16;315(5814):972–6.
22. Cickovski T, Narasimhan G. Constructing lightweight and flexible pipelines using Plugin-Based Microbiome Analysis (PluMA). Berger B, editor. Bioinformatics. 2018 Sep 1;34(17):2881–8.
23. Cickovski T, Aguiar-Pulido V, Huang W, Mahmoud S, Narasimhan G. Lightweight microbiome analysis pipelines. In: International Work Conference on Bioinformatics and Biomedical Engineering. Granada, Spain; 2016.