Deciphering microbial complexity using genomic data

Aaron N. Brooks, PhD

Our planet is teeming with microbes. A single gram of soil (less than a teaspoon) can contain hundreds of millions single-celled organisms, representing tens of thousands of different species. Microbes are everywhere. They colonize our skin, flourish in our guts, dominate the oceans, and even grow on nuclear waste. Collectively, their biomass likely exceeds that of both plants and animals. It should be no surprise, then, that microbes have a profound impact on us and our planet. Individually and collectively these little creatures sculpt ecosystems and influence our health in fundamental ways, many of which we have yet to understand fully.

The purpose of my dissertation was to build computer-based models to better understand how microorganisms function in different environments at a molecular level. To make the problem more tractable, I focused specifically on the process of transcription, whereby the levels of particular genes are increased or decreased. I decided to focus on transcription since it is well known that this is one of the first things to happen when cells sense a change in their environment.

Transcription is orchestrated by intricate molecular control networks that have been shaped and rewired gradually over evolutionary timescales to pair microbes with their habitats. Deciphering the control structures and subsequent dynamics of these so-called *gene regulatory networks* has been a longstanding challenge in biology. My dissertation contributed to the field by providing a computational framework for reverse-engineering highly-accurate gene regulatory networks directly from widely-available, high-throughput measurements of all transcripts in a cell. I did this by combining state-of-the-art ensemble learning algorithms with graph-theoretic methods. For my dissertation, I constructed models for two very different kinds of microorganisms (one a bacteria and the other an archaea). The models predicted a number of surprising programmed molecular events that I subsequently confirmed directly with experiments, including the environment-dependent internal segmentation of transcripts that encode several proteins called operons. Overall, my work suggests that microbial genomes are controlled in ways that are far more complex than people had previously imagined; yet it also demonstrates that we can predict when, where, and (to some extent) why these unexpected events occur if we use sophisticated computer-based tools to recast our intuition.

In popular culture and even scientific circles single-celled microrganisms are considered to be relatively *"simple"*. In a sense, the stereotype is fair: their genomes are much smaller than ours; they encode far fewer genes; and, as their name implies, they consist of only a single-cell rather than an complicated assembly of different cell types. Yet, despite this relative simplicity, it is difficult to interpret much less predict how microbes will respond to new conditions at a molecular level. More fundamentally, we are still trying to decipher how the various traits we can observe are encoded in their genomes. If we did know their genetic basis, we could modify microbial genomes to alter their behavior. This would be a disruptive innovation for biotechnology and industry. My work suggests that there are few easy, universal "rules" for controlling expression of a microorganism's genome. Nevertheless, using the tools I established, we can build computer models that help tame some of this complexity, guiding our focus towards specific hypothesis and insights that would have been difficult or impossible to conceive in their absence.

In addition to conducting the primary research, I also built several online resources to help the community explore the models and produce models of their own. A comprehensive web resource (egrin2.systemsbiology.net) and a suite of computational methods (github.com/baliga-lab/egrin2-tools) have been made freely available to promote extension of these models.