Dear Mr. Hector Nazario,

To fully understand the function of a gene, one must first learn how it affects other genes. The basic question asked when inferring gene regulatory networks is: If there is a change in one gene, how does it affect all of the other genes? Will that change activate a chain reaction that spreads throughout the entire organism? Or will it simply be alone in its change, accomplishing its task without any interaction from other genes? To answer these questions, we can quantify how active or inactive a gene is at a given time point with by its “expression value”. We can then draw edges between genes by asking “If gene A has a high expression value at time 1, what happens to gene B at time 2?” By asking this for each pair of genes, we can create a network.

However, this is a more difficult problem than it seems on its surface. There are two major issues: the data are noisy and the data are expensive and difficult to collect. To deal with these issues, a wide array of different mathematical and statistical techniques has been employed to attempt to separate the signal from the noise. Some algorithms seem to work better for some datasets than others. For example, Dynamical Bayesian Network approaches seem to work extremely well when there are a lot of experimental conditions and few genes. Other algorithms work better for datasets with many genes and few conditions. There are also ways to use different types of data as prior knowledge to give algorithms a good starting point.

The primary goal of this work is to provide an easy-to-follow and comprehensive approach to the inference of gene regulatory networks. It will bring together the wide array of techniques and schools of thought currently used in this rapidly developing field. Though we focus on genetic networks, most of the ideas we outline apply to any networks in which entity values are measured and we want to determine causal relationships among entities.

The approach is to first describe the current state-of-the-art when it comes to algorithms. Running examples will be worked through for each algorithm, giving the reader an understanding of how the numbers are actually being manipulated by the algorithm. We provide necessary background information to facilitate the understanding of each algorithm. Then, we will describe how to combine these algorithms to obtain a better answer than only using one alone. Several algorithms will combine to form a pipeline and the user may choose to form a variety of different pipelines. These different pipelines will be presented in the context of varying types and sizes of datasets. The main idea behind this pipelining approach is that there is likely no one-size-fits-all algorithm that will work with every dataset, so teaching the reader how to design his or her own algorithms by mixing and matching will contribute to the state of the art. Code and datasets will be provided in online supplemental materials.

We are submitting this proposal simultaneously to you and to Morgan Claypool. Dennis has published with both and likes you both. Our main goal is to contribute to the state of the art, so we want as wide a readership as possible.

Sincerely,

Jesse Lingeman

Prof. Dennis Shasha