The fundamental theorem of molecular biology is thus:

DNA, the molecular memory that stores the information of life itself, is read and transcribed into a series of comparatively short orders, which are transported around the cell in the form a molecule called messenger RNA (mRNA). Messenger RNA encodes the information necessary to build protein. When an mRNA is delivered within the cell, the mRNA is translated from the four letter genetic code into the 20 letter amino acid code. Amino acids assemble to form proteins, which are the basic building blocks of the molecular machinery of all cells.

Some proteins in the cell control the rate at which messenger RNAs are transcribed from the DNA. These proteins are called transcription factors. Transcription factors are one tool cells use to respond to their environment. As an example, recall that all of us began as a single cell – the union of a sperm and egg – and that every one of the cells in our body derives from the division of that single cell. At every division, the DNA of that original single cell is copied exactly into the newly formed cells.

Why, then, are our liver cells so much different from our nerve cells? This is due entirely to what parts of the DNA are expressed, or transcribed into RNA and then translated into amino acids, to form the proteins that carry out the various functions of the cell. Furthermore, this is a dynamic process – while a nerve cell does not turn into a liver cell spontaneously, as far as we know, nerve cells do respond to their environments in other ways. That response sometimes requires that different parts of the DNA are transcribed at different rates.

Transcription factors are one important aspect of the complex system that allows a cell to change. Understanding how and why cells change, and possibly even controlling that change, is a long term goal of biology, both as an exercise in understanding the fundamental nature of life, and also in applications such as medicine.

There are a number of methods that biologists use to identify proteins as transcription factors. Once a protein is identified as a transcription factor, we can identify where on the DNA that transcription factor is able to attach. However, even when we can determine that the transcription factor is, in fact, located on a specific region of DNA, we have no direct method of measuring the transcription factor’s affect on the rate of transcription.

Due to the size of the genome in even a very simple organism, and our inability to directly measure the rate of transcription, establishing the relationships between transcription factors and the genes that they regulate is currently an open question in biology.

Michael Brent and his lab at Washington University develop computational tools that take advantage of a measure of the amount of RNA in a cell, a technique called RNA sequencing (RNA-seq), to infer transcription factor networks. A transcription factor network is simply a collection of relationships between identified transcription factors, and the genes they are likely to regulate.

RNA-seq allows biologists to count the number of copies of a given messenger RNA in a cell at a certain time. Used in concert with genetic manipulation, this tool can be used as a proxy measure for the rate of transcription.

The Brent lab’s computational method is called NetProphet. For a given organism, as an example, yeast, the input to NetProphet is a list of transcription factors and a large matrix with dimensions genes (rows) by samples (columns). The matrix entries are the RNA-seq counts of messenger RNA of a given gene in a given sample. NetProphet also needs a design matrix which describes the genetic manipulation performed on each of the samples. A common and useful genetic manipulation is, for a set of samples, to genetically delete a single transcription factor. The theory is that the genes which are most affected by this deletion are the most likely targets of the deleted transcription factor.

The output of NetProphet is a graph in the discrete math sense, where the nodes are defined as genes (remember, the transcription factors themselves derive from genes, too) and the edges are regulatory relationships between transcription factors and the genes they regulate.

This is, however, a very large network. Uncompressed, the data that is provided on the Brent lab website for a given organism (currently, yeast and the fruit fly) are ~300GB. To access information in the network, one needs some amount of computer fluency. While these skills are becoming far more common in the biology research community, they are not ubiquitous. The purpose of this project is to provide an interactive portal through which a biologist can explore and extract information from the NetProphet transcription factor network.