67750: Advanced Practical Machine Learning 2018/9

Exercise 4 — Biological Background and Dataset

Dr. Tommy Kaplan *TA: Daniel Gissin*

The most simplistic model for genes, RNA and proteins is the following - DNA is the information for the making of the cell and it's encoded in units termed genes. Every cell uses only a portion of this information at each point in time, according to environmental signals (and the noise they come with). These signals can be abiotic conditions: salinity, humidity, temperature, etc., or biochemical: molecules secreted by other cells. How does the cell uses the information? It copies the DNA to messenger RNA (mRNA) in various copy numbers. These mRNA molecules are exported out of the nucleus, and then translated by ribosomes to proteins. The proteins are the actual macro-molecules that perform cellular functions, e.g. metabolism, structure, cell-cell communication, etc. So to sum this up:

$$\underset{"information"}{DNA} \overset{transcription}{\rightarrow} mRNA \overset{translation}{\rightarrow} \underset{"function"}{Protein}$$

It turns out that a lot of the decisions are mediated by a class of proteins (transcription factors) that respond to signals and bind the DNA where they find specific "words". For example, imagine a protein that under heat conditions changes its structure and is then able to bind ATTGTTA sequences in the gnome. These proteins act like bookmarks that the mRNA copying machines recognize, and thus, transcription of new messages is directed to relevant genes.

But these proteins are themselves a result of transcribed (and translated) genes, so how does this work? You can think of it as a very large and complex dynamic system whose elements control the state of other elements (and possibly their own state as well). The "microarray" dataset was used to infer quite a bit about these very complex regulatory mechanisms.

The following sections are descriptions and tips regarding the provided dataset.

1 Gene Expression Profiles from Yeast - "microarray"

The underlying thought of the people who generated the data is simple, and is very "reverse engineering" based: Imagine the cell as a very complex machine you wish to understand, but unfortunately have no knowledge of the inside or a "User's Manual". Now suppose you are able to hack the machine and take snapshots of its internal state (e.g. memory, variables, etc). In addition, you can change the machines input.

A naïve approach to understanding the machine can be - let's feed the machine with many different inputs, examine the internal state at each run, and try to understand the circuits that control the machine. For example, if we always see a certain collection of variables having correlated values, we might think they are under a common control.

In this exercise, our machine is the cell, and the variables are the expression levels of the different genes (i.e. mRNA copy numbers). Every row in the matrix corresponds to a specific gene, and every column corresponds to a condition.

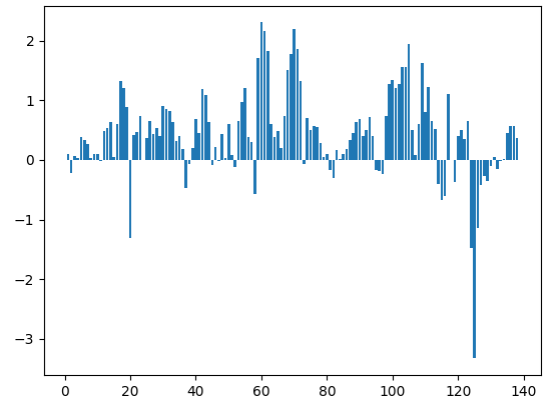

Figure 1: Bar Plot of a Single Gene

In the provided "microarray_exploration" you can see an exploration of the data set, starting with a single gene and consition and moving up all the way to a full view of the data.

1.1 Single Entry

We first read a single entry from the data.

our data corresponds to the expression levels of 6,701 genes in 138 different environmental conditions. So the (i, j) entry in the data is a gene with codename "HIS5" (you can look it up online), and the condition in which the measurement was made. The consition here was a measurement of the gene expression after 12 hours of being exposed to Nitrogen Depletion. The value was 0.391, log2 of the change in expression level of the gene between the beginning of the experiment and the measurement (after 12 hours). In other words - If the expression level at time 0 was x, then the expression level after a single day was $x \cdot 2^{0.391}$, i.e. - it went up by a factor of $\sim$1.3. Note that this may mean that the gene went up from 6 copies to 9, or from $3 \cdot 10^4$ copies to $9 \cdot 10^4$, and we wouldn't be able to tell the difference from this data.

1.2 Single Gene

Next, we explore a single gene instead of a single entry by looking at a bar plot of all the changes in expression of the gene, given the different conditions the cell was exposed to (seen in figure 1).

Note that the gene sometimes goes up (relative to what? depends on the experiment...), and sometimes goes down, but there's a pattern - usually a few consecutive conditions will resemble each other.

1.3 Correlation of Two Conditions

How can we see if this phenomenon is just a "fluke" and appears only in this gene, or has a more global nature? Well, we can select all the genes, and look at their levels in two conditions that seem suspicious to us, say condition 27 and condition 29. This can be seen in figure 2, and these 2

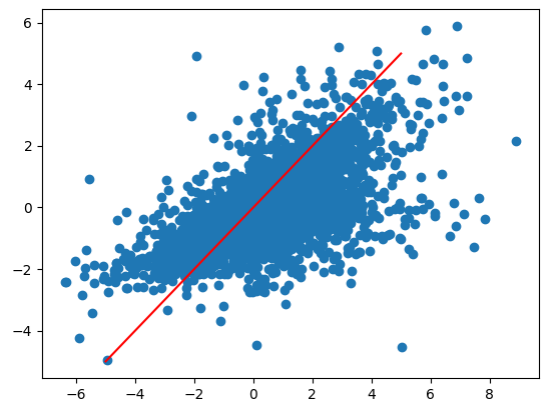

Figure 2: Correlatio between condition 27 and 29

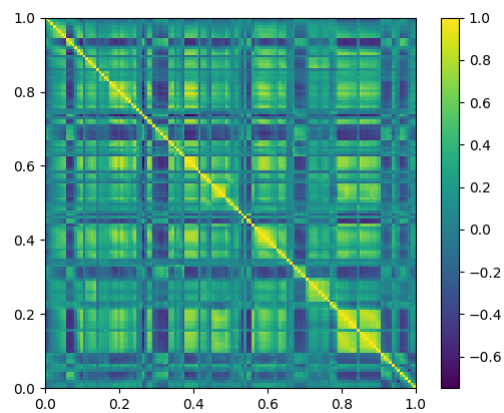

Figure 3: Correlation Matrix of all Conditions

experiments seem to have a correlation (genes that go up in 27, also go up in 29, and those that go down 27, also go down in 29), but how can we get an overview over the entire dataset?

1.4 Correlation of All Conditions

Figure 3 shows us the bigger picture, and indeed, along the diagonal of the matrix we see these high correlation "boxes", that correspond to the batches we saw in the bar plot. There is a lot of interesting information in this plot, but we'll stop at this point (you are welcome to explore it...).

So why do these batch effects appear in the data? simple - a single experiment usually looks at a specific condition in various time points (you can examine the condition names). Thus, in a specific condition (e.g. hot temperature) the yeast responds by expressing certain genes and shutting down other genes, but this response by the cells takes time, and is maintained through time, so we see it through the various time points.

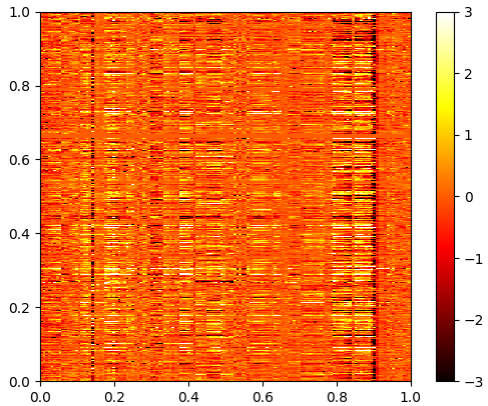

Figure 4: The Entire Data Set

1.5 The Entire Data Set

Finally, now that we know how to read a single entry in the data, and we've examined a gene, we can look at the entire data matrix (seen in figure 4). And it looks aweful!

So how can make sense of this data? clustering is one option... But besides allowing us to look at the data, clustering also makes sense in the context of our question. What does it mean if two genes are similar in this matrix? It means that they respond in a similar manner in a variety of conditions - so a cluster of genes is a good candidate for genes that are regulated together (or co-regulated).