A New Genetic Logic for Vectorial Phenotypes

David Angeles-Albores

November 16, 2016

Contents

1 Introduction

The purpose of this manuscript is to clearly and formally set out the genetic logic that we developed for the genetic analysis of RNA-seq data. Briefly, our approach is derived from a set of simple set logic rules. The text below deals with straightforward genetic cases. We have not yet formalized our approach to account for measurement error, nor have we fully formalized our approach to deal with stimuli of variable strength. That said, because our set of operations is fundamentally linear in nature, we expect that an algebra could be constructed from this set of rules in the future. Said algebra could account for errors in measurement or variability in response to a perturbation.

2 Definition of a Genetic Logic and Some Simple Examples

Definition 2.1. $\{X\}$ is a set of genes, g, drawn from a genome G, where each specific gene has an associated measurement, s. For the purposes of this manuscript, such measurements are the regression coefficients of a generalized linear model, which can be loosely interpreted as the log-fold change relative to a perturbation.

$x_i \in \{X\}$ is a tuple (g, s), where $s \in \mathbb{R}$, and we refer to the set of genes in $\{X\}$ as g_X.

Because of its ability to query global gene expression, RNA-seq provides an unbiased way to measure $\{X\}$. However, different experiments will perturb genes in different ways. Therefore, it will be necessary to define transcriptomes relative to how they respond to a perturbation, not just relative to what gene controls them.

Definition 2.2. Given a gene X, if a perturbation increases the activity of X, then the transcriptomic response to the increased activity of X is defined to be $\left\{X^{\Uparrow}\right\}$. Likewise, if a perturbation decreases the activity of X, then the transcriptomic response to the decreased activity of X is defined to be $\left\{X^{\Downarrow}\right\}$.

Intuitively, it makes sense that knocking out a gene is the mathematical opposite of increasing the activity of a gene. Next, we will formalize the notions of knock-out and knock-up as they pertain to transcriptomic phenotypes.

Definition 2.3. If a perturbation knocks down X and a second perturbation increases X, then two transcriptomic phenotypes can be measured, $\{X \Uparrow\}, \{X \Downarrow\}$. We define a transcriptome to be the opposite of another if for every gene i, the change in expression is of the same magnitude but of the opposite direction:

$$x_i^{\Uparrow} = -x_i^{\Downarrow}, \text{for } i \in g_X, \tag{1}$$

is true.

More generally, we could write:

$$x_i^{\Uparrow} = f(x_i^{\Downarrow}). \tag{2}$$

If two transcriptomes are inverses of each other, we will write:

$$\left\{X^{\Uparrow}\right\} = f(\left\{X^{\Downarrow}\right\}). \tag{3}$$

Finally, two transcriptomes may partially overlap in the set of genes that they contain. Throughout this text, the overlap in gene names between two genes X and Y, denoted by $g_X \cap g_Y$, will be important, but I find the notation by gene name cumbersome. Therefore, we define the gene overlap of two transcriptomes to be:

$$g_X \cap g_Y \equiv \{X\} \cap \{Y\}. \tag{4}$$

Simple Genetic Circuits

Given these definitions, we can begin to write down genetic circuits. Before we can begin, we need one more thing. It is not clear at this point that all genes have transcriptomic profiles associated with them. In order to denote that a particular gene has an associated transcriptome, we will write:

$$B \multimap \{B\} \tag{5}$$

As an example, we can begin to apply our logic to genetic pathways. As a brief example (mainly to test out this new machinery), we can imagine the following example. Suppose 2 interacting genes exist, A, B. Only B has transcriptomic effects. If $A \longrightarrow B$ what is the transcriptional output of A^- or B^-?

$$\begin{aligned} \text{if } A \rightarrow B &\Rightarrow \{B\} \\ A^- &= \left\{B^{\Downarrow}\right\} \\ B^- &= \left\{B^{\Downarrow}\right\} \\ \therefore A^- &= B^- \end{aligned} \tag{6}$$

We could combine genetic logic with our transcriptomic notation to predict the phenotype of the knockout of A and B. Generally speaking, if the pathway in question led to a visible phenotype, then we would need to generate three mutants to observe an interaction: A^-, B^-, AB^-. In fact, before generating the double mutant, there would have been no way to predict that A and B were interacting partners. Using transcriptomic phenotypes, in theory it would be sufficient to generate only the single mutants. In reality, the added information enables us to treat the single mutants as a predictive data set of interaction. The double mutant would provide an independent test of the interaction. That said, even with all the mutants, we would not be able to identify the biochemical ordering.

As a matter of fact, it would also be possible to infer negative interactions:

$$\begin{aligned} \text{if } A \dashv B &\multimap \{B\} \\ A^- &= \left\{B^{\Uparrow}\right\} \\ B^- &= \left\{B^{\Downarrow}\right\} \\ \therefore A^- &= f(B^-) = -B^- \end{aligned} \tag{7}$$

In this case, the double mutant satisfies the following epistasis relationship:

$$A^-B^- = B^- \tag{8}$$

This epistasis relationship is sufficient to establish that A inhibits B, in that precise order.

3 Progressive Decorrelation Provides Information for Ordering Nodes in a Network

Let's reexamine the circuit where A activates B. In this case, suppose that (and this is a crucial assumption) A and B both have independent transcriptomes ($g_A \cap g_b = \varnothing$). Let's re-examine the pathway again:

$$
\begin{aligned}
&A \rightarrow B \\
&A^{-} = \left\{A^{\Downarrow}\right\}, \left\{B^{\Downarrow}\right\} \\
&B^{-} = \left\{B^{\Downarrow}\right\} \\
&\therefore B^{-} \subset A^{-}
\end{aligned}
\tag{9}
$$

In this admittedly fictitious example, it would be possible to identify, simply from the single mutants, the fact that A is upstream of B by virtue of their independent transcriptomic profiles. Simply put, A should have a larger effect than B as measured by the number of altered genes. This could only happen if A is upstream of B. Notice that a key assumption in this argument is that A is an absolute controller of B: No A, no B.

As linear pathways get longer, we should still be able to infer position on a chain based on set intersections. Namely, two adjacent nodes will always have a greater overlap than two nodes that are not adjacent. This is the basis for the contruction of figure X in the main text.

4 Layers of Regulation

In any genetic network, there are different layers of regulation. We could imagine a protein that regulates another via transcriptional regulation, or via protein-protein interactions. We could even imagine an enzyme that generates a chemical product that is a contrl signal. For this reason, we will need to define different ways to represent control.

Definition 4.1. Def: if gene a regulates b, $a \rightarrow b$, at the transcriptional level, we will hencforth use standard *C. elegans* nomenclature to write perturbations. Briefly, that nomenclature states that when referring to the DNA of a gene or RNA of a gene, italics and non-caps should be used, and when talking about a protein, all caps and no italics should be used. In other words, knocking out a will cause a perturbation in the transcriptome of b of the form:

$a^{-} = \left\{b^{\Downarrow}\right\}$

If gene a regulates b via protein-protein interactions, the knockout should be written as:

$a^{-} = \left\{\mathrm{B}^{\Downarrow}\right\}$

If gene a chemically modifies b, then every chemical modification of b should have its own transcriptome. Suppose that the protein form of b has two states, hydroxylated and non-hydroxylated and a promotes the non-hydroxylated form. Then, we write the transcriptome for each chemical species, using the protein form of the gene:

$a^{-} = \left\{\mathrm{B}^{\Uparrow}\right\} \left\{\text{B-OH}^{\Downarrow}\right\}$

Finally, different mutations may affect a protein at different levels. If a gene is knocked out by complete deletion, the gene should be written in lower-letters and italics. If a gene is knocked out by a small deletion or a single nucleotide polymorphism, the protein nomenclature should be preferred.

5 A First Attempt at the Hypoxia Pathway

The hypoxia pathway in *C. elegans* is understood to be made up of the following genetic regulatory interactions:

Next, we analyze all the predictions from all the single and double mutants we will study. First, I will write down all the genetic transcriptome equations that the model provides. I will not write the transcriptome for HIF-1-OH, since this species is widely regarded as inert. Next, I will check whether raw correlation/anticorrelation counts agree with the predictions, and we will refine the model. Finally, we will include mRNA information for each gene, and refine once more. Let's begin:

$$\textit{hif-1}^- = \left\{\textit{rhy-1}^{\Downarrow}\right\} \left\{\textit{egl-9}^{\Downarrow}\right\} \left\{\textit{hif-1}^{\Downarrow}\right\} \tag{10}$$

$$\textit{rhy-1}^- = \left\{\text{RHY-1}^{\Downarrow}\right\} \left\{\text{EGL-9}^{\Downarrow}\right\} \left\{\text{HIF-1}^{\Uparrow}\right\} \{HIF-1-OH^{\Downarrow}\}\, \textit{rhy-1}^{\Uparrow} \textit{egl-9}^{\Uparrow} \tag{11}$$

$$\textit{egl-9}^- = \left\{\textit{rhy-1}^{\Uparrow}\right\} \left\{\text{EGL-9}^{\Downarrow}\right\} \left\{\text{HIF-1}^{\Uparrow}\right\} \{HIF-1-OH^{\Downarrow}\}\, \textit{rhy-1}^{\Uparrow} \textit{egl-9}^{\Uparrow} \tag{12}$$

$$\textit{vhl-1}^- = \left\{\textit{rhy-1}^{\Uparrow}\right\} \left\{\textit{egl-9}^{\Uparrow}\right\} \left\{\text{HIF-1}^{\Uparrow}\right\} \{HIF-1-OH^{\Uparrow}\} \left\{\text{VHL-1}^{\Downarrow}\right\} \textit{rhy-1}^{\Uparrow} \textit{egl-9}^{\Uparrow} \tag{13}$$

$$\textit{egl-9}^- \textit{vhl-1}^- = \left\{\textit{rhy-1}^{\Uparrow}\right\} \left\{\text{EGL-9}^{\Uparrow}\right\} \left\{\text{HIF-1}^{\Uparrow}\right\} \{HIF-1-OH^{\Downarrow}\} \left\{\text{VHL-1}^{\Downarrow}\right\} \textit{rhy-1}^{\Uparrow} \textit{egl-9}^{\Uparrow} \tag{14}$$

$$\textit{egl-9}^- \textit{hif-1}^- = \left\{\textit{rhy-1}^{\Downarrow}\right\} \left\{\text{EGL-9}^{\Downarrow}\right\} \left\{\textit{hif-1}^{\Downarrow}\right\} \textit{egl-9}^{\Downarrow} \tag{15}$$

The model makes certain predictions about how some of these genotypes should be correlated. Let's see what information we got from our experiment (see 1). A key result from our modeling is that *egl-9* and *rhy-1* should share a large number of genes that are positively correlated, but the transcriptome associated with *rhy-1* should be negatively correlated.

Empirically, we only observe two genes that are anticorrelated between *egl-9* and *rhy-1*, and all other genes are correlated. Therefore, either *rhy-1* does not seem to have strong transcriptomic effects that are separate from *egl-9* or our model is wrong. Given the overwhelming evidence in favour of *rhy-1* being controlled by *hif-1* (including some in this paper), we reject the hypothesis that the model is wrong and favour one in which *rhy-1* has exclusively protein-level effects (which is not as crazy as it seems).

Furthermore, we predicted that *vhl-1* and *rhy-1* share extensive positively correlated genes, with the exception of genes associated with *rhy-1* and *egl-9*. Indeed, out of 431 isoforms that overlap between mutants of these genes, 411 are coexpressed, and 20 follow a pattern of antiexpression. Since previously we concluded that *rhy-1* has no transcriptome, the 20 antiexpressed genes must correspond to genes that are downstream of *egl-9* or HIF-1-OH.

Genes	Overlapped Isoforms	Positive Correlated Isoforms	Negative Correlated Isoforms
rhy-1 $\otimes$ *egl-9*	1098	1096	2
rhy-1 $\otimes$ *vhl-1*	431	411	20
egl-9 $\otimes$ *vhl-1*	462	448	17
hif-1 $\otimes$ *rhy-1*	184	165	19
hif-1 $\otimes$ *egl-9*	165	148	27
hif-1 $\otimes$ *vhl-1*	97	85	12

Table 1: Number of overlapped genes between single mutants in the hypoxia pathway.

Next, let us consider *egl-9* and *vhl-1*. As in the previous pair, this pair of genes the only genes that should be anticorrelated are those associated with *egl-9* or HIF-1-OH. It would be a nice check if the genes that are anti-correlated in this comparison are the ones that appear in the previous comparison. Indeed, 11 genes are common, out of a total of 27 distinct isoforms present in the union of these sets. It appears we are actually sampling from the same pool.

For the comparisons involving *hif-1*, the only genes that should be anticorrelated should be those that are under the control of *hif-1* in a normoxic state, except for the comparison with *vhl-1*, which should also include genes associated with *egl-9*. The measured transcriptome for *hif-1*$^{\Downarrow}$ should be small, given that *hif-1* is expressed only at low levels in this state. Moreover, these genes should (hopefully) be the same across all the comparisons.

Given this information, we can now re-write the model to remove the *rhy-1* transcriptome, since we find no evidence of transcriptome perturbations that are observable in this dataset.

$$\textit{hif-1}^- = \left\{\textit{egl-9}^{\Downarrow}\right\}\left\{\textit{hif-1}^{\Downarrow}\right\} \tag{16}$$

$$\textit{rhy-1}^- = \left\{\text{EGL-9}^{\Downarrow}\right\}\left\{\text{HIF-1}^{\Uparrow}\right\}\left\{\text{HIF-1-OH}^{\Downarrow}\right\}\textit{rhy-1}^{\Uparrow}\textit{egl-9}^{\Uparrow} \tag{17}$$

$$\textit{egl-9}^- = \left\{\text{EGL-9}^{\Downarrow}\right\}\left\{\text{HIF-1}^{\Uparrow}\right\}\left\{\text{HIF-1-OH}^{\Downarrow}\right\}\textit{rhy-1}^{\Uparrow}\textit{egl-9}^{\Uparrow} \tag{18}$$

$$\textit{vhl-1}^- = \left\{\textit{egl-9}^{\Uparrow}\right\}\left\{\text{HIF-1}^{\Uparrow}\right\}\left\{\text{HIF-1-OH}^{\Uparrow}\right\}\left\{\text{VHL-1}^{\Downarrow}\right\}\textit{rhy-1}^{\Uparrow}\textit{egl-9}^{\Uparrow} \tag{19}$$

$$\textit{egl-9}^-\textit{vhl-1}^- = \left\{\text{EGL-9}^{\Uparrow}\right\}\left\{\text{HIF-1}^{\Uparrow}\right\}\left\{\text{HIF-1-OH}^{\Downarrow}\right\}\left\{\text{VHL-1}^{\Downarrow}\right\}\textit{rhy-1}^{\Uparrow}\textit{egl-9}^{\Uparrow} \tag{20}$$

$$\textit{egl-9}^-\textit{hif-1}^- = \left\{\text{EGL-9}^{\Downarrow}\right\}\left\{\textit{hif-1}^{\Downarrow}\right\}\textit{egl-9}^{\Downarrow} \tag{21}$$

5.1 qPCR analysis

We have thus far refined the model to remove the *rhy-1* transcriptome. Next, we would like to verify that the mRNA for each species is reacting as we expect it to. Long story short, *rhy-1* is activated when HIF-1 is present, but also when *hif-1* is knocked out. *egl-9* only changes expression at enough level to reach statistical significance for a two mutants (*egl-9*, *egl-9*;*vhl-1*). Therefore, we remove the transcriptomic control of *egl-9* by *hif-1* from the model for some genes momentarily..

$$\textit{hif-1}^- = \left\{\text{EGL-9}^{\Uparrow}\right\}\left\{\textit{hif-1}^{\Downarrow}\right\}\textit{rhy-1}^{\Uparrow} \tag{22}$$

$$\textit{rhy-1}^- = \left\{\text{EGL-9}^{\Downarrow}\right\}\left\{\text{HIF-1}^{\Uparrow}\right\}\left\{\text{HIF-1-OH}^{\Downarrow}\right\}\textit{rhy-1}^{\Uparrow} \tag{23}$$

$$\textit{egl-9}^- = \left\{\text{EGL-9}^{\Downarrow}\right\}\left\{\text{HIF-1}^{\Uparrow}\right\}\left\{\text{HIF-1-OH}^{\Downarrow}\right\}\textit{rhy-1}^{\Uparrow}\textit{egl-9}^{\Uparrow} \tag{24}$$

$$\textit{vhl-1}^- = \left\{\text{EGL-9}^{\Uparrow}\right\}\left\{\text{HIF-1}^{\Uparrow}\right\}\left\{\text{HIF-1-OH}^{\Uparrow}\right\}\left\{\text{VHL-1}^{\Downarrow}\right\}\textit{rhy-1}^{\Uparrow} \tag{25}$$

$$\textit{egl-9}^-\textit{vhl-1}^- = \left\{\text{EGL-9}^{\Downarrow}\right\}\left\{\text{HIF-1}^{\Uparrow}\right\}\left\{\text{HIF-1-OH}^{\Downarrow}\right\}\left\{\text{VHL-1}^{\Downarrow}\right\}\textit{rhy-1}^{\Uparrow}\textit{egl-9}^{\Uparrow} \tag{26}$$

$$\textit{egl-9}^-\textit{hif-1}^- = \left\{\text{EGL-9}^{\Downarrow}\right\}\left\{\textit{hif-1}^{\Downarrow}\right\}\textit{rhy-1}^{\Uparrow} \tag{27}$$

Notice that in equations (22, 25), EGL-9 transcriptome is denoted as going up, although the mRNA for this gene did not increase with the mutation of *vhl-1* or *hif-1*. This is because *rhy-1* mRNA levels went up in these mutants, which suggests that RHY-1 also increased. Since RHY-1 activates EGL-9, our logic dictates that EGL-9 activity likely increases.

$$\textit{hif-1}^- = \left\{\text{EGL-9}^{\Uparrow}\right\} \left\{\textit{hif-1}^{\Downarrow}\right\} \textit{rhy-1}^{\Uparrow} \tag{28}$$

$$\textit{rhy-1}^- = \left\{\text{EGL-9}^{\Downarrow}\right\} \left\{\text{HIF-1}^{\Uparrow}\right\} \left\{\text{HIF-1-OH}^{\Downarrow}\right\} \textit{rhy-1}^{\Uparrow} \textit{egl-9}^{\Uparrow} \tag{29}$$

$$\textit{egl-9}^- = \left\{\text{EGL-9}^{\Downarrow}\right\} \left\{\text{HIF-1}^{\Uparrow}\right\} \left\{\text{HIF-1-OH}^{\Downarrow}\right\} \textit{rhy-1}^{\Uparrow} \textit{egl-9}^{\Uparrow} \tag{30}$$

$$\textit{vhl-1}^- = \left\{\text{EGL-9}^{\Uparrow}\right\} \left\{\text{HIF-1}^{\Uparrow}\right\} \left\{\text{HIF-1-OH}^{\Uparrow}\right\} \left\{\text{VHL-1}^{\Downarrow}\right\} \textit{rhy-1}^{\Uparrow} \textit{egl-9}^{\Uparrow} \tag{31}$$

$$\textit{egl-9}^- \textit{vhl-1}^- = \left\{\text{EGL-9}^{\Downarrow}\right\} \left\{\text{HIF-1}^{\Uparrow}\right\} \left\{\text{HIF-1-OH}^{\Downarrow}\right\} \left\{\text{VHL-1}^{\Downarrow}\right\} \textit{rhy-1}^{\Uparrow} \textit{egl-9}^{\Uparrow} \tag{32}$$

$$\textit{egl-9}^- \textit{hif-1}^- = \left\{\text{EGL-9}^{\Downarrow}\right\} \left\{\textit{hif-1}^{\Downarrow}\right\} \textit{rhy-1}^{\Uparrow} \tag{33}$$

Now, let's consider the following. We see that *hif-1* has positive correlations with *egl-9*, *rhy-1*, and *vhl-1*. Our equations suggest that the only way this can happen is if HIF-OH has transcriptomic effects. However, if we attempt to identify the genes that are affected by the hydroxylated form of the *hif-1*, we find that the overlap is very small—we can only identify 2–4 genes depending on strict we set the filters.

Why such poor overlap? One reason may be that HIF-1-OH might not be particularly abundant in the worm. A second reason may be that this protein might not be terribly active. A third reason may be that the perturbation to this protein is small in all cases, so our measurements are noisy. Another thing to take into consideration is the shallowness of the sequencing, which will further increase noise in our estimates and decrease overlap between samples.

At this point, our assertion that HIF-1-OH has downstream targets should be viewed as contentious, although not contradictory with current literature. At least in the worm, studying the hydroxylated form of this gene has not been possible due to a lack of tools to stimulate hydroxylation in the absence of degradation, particularly without other mutants. Thus, my assertion about the hydroxylated *hif-1* transcriptome might be almost unfalsifiable. This is a problem, and the reason why I will not make this statement in the main body of the text.