This is not a standalone post. Check out the intro to this series.

What is identifiability, and why do I care?

When we construct quantitative models of biological systems, they usually have unknown parameters. These might be fertility rates in a population model or distance decay rates in a model of random polymer looping. In modeling gene regulation, the unknown parameters might be binding constants, transcription rates, and isoform ratios.

Sometimes, two different settings of the parameters will give rise to identical or highly similar predictions. In statistics, we would call this an *identifiability* problem; a common-English synonym might be a *model distinctiveness* problem. This is a big problem, and it can really dead-end your project. Suppose, for instance, that your model cannot tell whether gene A is regulated by gene B or gene C. If you aren't prepared to acknowledge this uncertainty, you may end up designing future experiments around perturbation of gene B, when really it's gene C that matters. Even if you are prepared to handle the uncertainty in an honest way, it can render your results unusable, because there are often more than two models compatible with your data. There could be hundreds or zillions (depending on how you count them), and it could turn out that your data actually have very little information about the mechanisms you are interested in.

In this post, I'll discuss systems biology identifiability problems stemming from two sources: missing data and insufficient cell-state diversity. Both of them greatly affect how to plan research and interpret results.

This post is very RNA-centric. If you are interested in chromatin state, I will discuss that in some other posts.

Missing data

In one common type of experiment, we measure gene activity by sampling RNA transcripts, reading them out, and counting them. In this common type experiment, we can only guess at:

- protein levels.
- protein modifications. (For example, phosphorylation or cleavage).
- protein complexes. (Is the protein bound to DNA? Is it bound to other proteins?)
- Hormones or signals not entering from outside. (For example, testosterone, insulin, or retinoic acid.)
- tiny little RNAs not captured well by vanilla RNA-seq.
- circular RNAs or other RNA's that might be there in the data, but haven't yet been catalogued and "tamed" by our processing techniques.
- different isoforms of each gene. (Some techniques can see this, but many widely popular or commercial single-cell RNA technologies cannot see it).
- chromatin state. (Is the DNA at each locus is accessible? How is it packaged, marked, and folded in 3D?)
- subcellular localization of each molecule. (Is it in the cytoplasm or the nucleus? MERFISH or similar might be able to answer this for mRNA but widespread high-throughput techniques circa April 2019 cannot.)

This list is long enough that one might be inclined to just give up. But, certain encouraging signs suggest that regulation of cell state can be predicted from RNA alone. For example, the stunning success of iPSC reprogramming indicates that four commonly measured factors are sufficient to radically transform the state of a cell. Our models may be able to predict cell state even when glossing over many important distinctions.

But, the details of this glossing-over will be key to interpret our models correctly. There are some examples in the technical appendix. The tl;dr is that sometimes the models will be wrong, period, until you measure more types of molecules. Other times, the models will capture direct influence but not direct physical interaction. This influence can operate in multiple ways to produce similar effects, so it is a living example of an identifiability problem.

Direct binding, direct influence, and indirect influence

I want to take a brief detour to convey what exactly is the goal of these models. I am interested in how stem cells respond to stimuli. So, I want my models to capture causal influence, even if there is no direct physical contact. The model should still distinguish between direct and indirect influence, though: if the effect of A on C can be explained entirely by `A==>B==>C`, then I don't want `A->C` in the model. The diagram below explains more.

```
 A  ==> (B) ==> C    # Desired inference: A -> C (best option since B is unseen)
 A  ==>  B  ==> C    # Desired inference: A -> B and B -> C but not A->C
 A  without () means A is measured
(A) means A is not measured.
 A ==> B means "A directly binds B, influencing its function and/or quantity."
 A -> B means A is listed as a regulator of B in the model.
```

Cell-state diversity

Most models I have seen do not attempt to explicitly represent missing data. For example, they will not have a count for Foxn1 RNA and a separate count for Foxn1 protein. They just have a single count of "Foxn1", usually taken to mean the RNA. This greatly simplifies the concepts and the computation, and I'll follow their lead in this section.

Under this simplification, everything is observed, and I can tell you what most GRN models are doing, at least in spirit.

- Step 1: formulate a complete description of how gene B is regulated. It doesn't have to be right or even close; it just has to be specific. For example, your hypothesis might state "Genes A and C together completely determine the transcription rate of B with the following formula: `dB/dt = 2*A if C>0`."
- Step 2: check if this model is compatible with your observations. If it is, put it on the "maybe" list for gene B. Otherwise, discard it.
- Step 3: make your herd of computers do this over and over until you have tested all hypotheses for all genes. (In practice, people don't test all models -- rather, they implement a variety of computational shortcuts.)

At the end of this process, check the "maybe" list for each gene.

- If it has length 0, you need to consider more flexible models (or better quality data). Try again.
- If it has length 1, congratulations! By process of elimination, and subject to the limitations discussed below in the technical appendix, you've cracked a tiny piece of the human regulatory code! Pop the champagne.
- If it has length greater than one, welcome to the club: you have an identifiability problem!

The last outcome is most interesting, and it's a key challenge for the field right now. Often, there's so much garbage on the list that you can't do much of what you hoped to do with these models.

This begs the question: what kind of data are needed to get past this issue? And how much? I will discuss a couple of relevant papers in the appendix.

Technical appendices

What happens when you ignore everything but the RNA?

When your model ignores everything but the RNA, it can really affect the interpretation. Here are some examples. Sometimes the model is flat out wrong, and we have to hope cells are simple enough that won't occur too often (example 1). Other times the model is correct, even in predicting perturbation outcomes, but the detailed physical interpretation is still unclear (examples 2-5).

1. Suppose activity in genes A and C marks cell type T, which contains an unobserved circular RNA called D. Suppose T is the only cell type in which gene B is active. Suppose D controls the activity of B, with A and C merely indicating the presence of D but not causally upstream of it. The model likely says A and C together upregulate B, but it's wrong, and perturbing A and C will not do anything to B.
2. Suppose gene A has isoforms A1 and A2. A1 upregulates gene B but A2 downregulates gene B. Suppose gene C promotes isoform A1, and isoform A2 is otherwise the default. Hopefully, we can infer the rules `A AND C makes B go UP` and `A and NOT C makes B go down`. It's causal, meaning if we perturb C, the rule still works. But, we should not claim or expect that C's protein product directly binds B.
3. Suppose gene A's protein product must be complexed with small molecule S in order to bind and upregulate gene B. Otherwise, gene B is not transcribed and the existing mRNA begins to decay. Gene C's protein product produces S. Hopefully, we can infer the same rule `A AND C makes B go UP` and `A and NOT C makes B go down`. It's still causal, meaning if we perturb C, the rule still works. Again, we should not claim or expect that C directly binds B's promoter.
4. Suppose gene A's protein product upregulates gene B, but only if enhancer E has been opened up. If gene C's protein product is the pioneer factor that opens up E, Hopefully, we can infer, again, the same rule `A AND C makes B go UP` and `A and NOT C makes B go down`. It's still causal, again, but we should not claim or expect that C directly binds B's promoter.
5. In T cell receptor signaling, the TCR goes to the cell surface, forms a complex with several other proteins, binds another molecule outside the cell, opening an intracellular domain to phosphorylation, which recruits a kinase, which phosphorylates a thing, which recruits things that phosphorylate things that hydrolyze things to generate things that promote transcription of B. I'm not sure what rule I want infer in this situation, but it probably contains a lot of `AND` s!

Maybe we can eventually organize these into an ontology or an algebraic grammar of sys-bio modeling screwups.

If we ignore everything but the RNA, how much data do we need?

To decide how much data we need, we should first choose a class of models to study. I will comment on some results for boolean network models. These models assume each genes is "on" or "off", with no other states allowed. The activity of gene A at time t+1 is determined by a boolean function of other genes. For a made-up example, "Ccl25 is active at time t+1 if, at time t, Foxn1 is active and Gcm2 is not active." Some papers allows randomness: "Ccl25 is active with 90% probability if at time t+1 if, at time t, Foxn1 is active and Gcm2 is not active. Otherwise, it is active with probability 0.05."

I chose boolean models because there is a clear line of reasoning to reduce the cell to a boolean system starting from physical first principles. In 1977, Daniel Gillespie reduced a cell reasonably well from a bath of chemicals to a system of differential equations. Four years earlier in 1973, Leon Glass and Stuart Kauffmann had already figured out how to reduce a set of differential equations to a discrete logical system. (Differential equations were already in use by the time they were formally justified by Gillespie.) Neither of the reductions is perfect -- for instance, Gillespie's reduction ignores stochasticity and ignores spatial organization within the cell. But, each reduction makes a convincing claim to capture much of the important behavior of the system.

> Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry*, *81*(25), 2340-2361.

> Glass, L., & Kauffman, S. A. (1973). The logical analysis of continuous, non-linear biochemical control networks. *Journal of theoretical Biology*, *39*(1), 103-129.

Using boolean networks, people have begun to study how much data is necessary to identify a unique regulatory rule for each gene. Akutsu et al consider a case where the in-degree is bounded: that is, no gene can be controlled by more that K regulators. Their answer to "how much data" is measured in "state-transition pairs": each "state-transition pair" is a measurement of cellular state at time t paired with its successor at time t+1. Akutsu et al prove that for a network with n genes, $O(\log n)$ state transition pairs are sufficient to identify the original Boolean network.

This is a wonderful and encouraging finding. It suggests that if our consortia or data resources grow at a linear rate, the size of the networks we can successfully study will explode exponentially! Unfortunately, I think most people would agree this explosion hasn't manifested yet (though it may be about to). This is partly due to the "fine print" governing Akutsu et al's results. First, the scaling with k -- the maximum number of regulators per gene -- is terrible. Increasing k from 3 to 4 or 4 to 5 leads to a massive increase in the amount of data needed. Second, Akutsu et al require that cell states are diverse, so densely sampling a single cell type or developmental trajectory is not useful. Ideally, observations would be sprinkled throughout the state space uniformly at random. In practice, the best we can do is to study many diverse cell types.

Here is the paper if you want to learn more.

> Akutsu, T., Miyano, S., & Kuhara, S. (1999). Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. In Biocomputing'99 (pp. 17-28). https://www.ncbi.nlm.nih.gov/pubmed/10380182

If you enjoy super intense boolean modeling and you want to know what has happened more recently, check out this next paper. Many empirically derived rules are simpler-than-boolean in certain useful ways: for example, they are often strictly increasing or strictly decreasing in most inputs. This paper takes advantage of that to prune models faster (computationally) and better (in terms of the required amount of data). I don't understand this very well but I think the scaling is very similar to Akutsu et al's result.

> Schober, S., Kracht, D., Heckel, R., & Bossert, M. (2011). Detecting controlling nodes of boolean regulatory networks. EURASIP Journal on Bioinformatics and Systems Biology, 2011(1), 6. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3377916/