Inari EDA: Exploratory *Direction* Analysis

**Goal:** Analyze the direction of the entire project, it’s scope, and a definition of an MVP. We are writing a statement-of-work to formalize this currently, and want to clarify our understanding of the problem proposed by Inari, the desired outcomes, and elucidate the shape of what a successful and interesting project looks like. This need arises acutely in light of a couple of key concerns we have formed regarding the initial project specification, and whether or not there is room for adjusting the direction.

Introduction

**Problem Statement and an Analogy**

We have a dataset of gene expression values of ~40,000 genes, across 10 tissue samples, on 26 individuals of corn. The dataset is currently purely a representation of the expression levels, without metadata regarding ontology, underlying sequences, etc. We are proposed to first create a gene expression network, and then from the problem statement directly:

“Using the resulting expression network or networks, a graph neural network or graph convolutional network could be trained and used to predict the effect of specific perturbations in the network including removing specific nodes or increasing/decreasing the expression of genes at specific nodes. Prediction accuracy for constructed models can be validated on a holdout set of the original expression data.”

To understand this problem, we consider an analogous one: suppose that we are trying to figure out a language, and are given 26 books in this language ostensibly on the subject of corn. Each book is made up from a full set of ~40,000 words, and each book has 10 chapters. We actually don’t have the books themselves, but are instead given a word count of each word across each chapter and each book.

Interestingly enough, this analogy seems quite salient in the sense that the word counts themselves are a few levels removed from understanding what they actually mean in the context of something as complex as a language, just like how gene expression levels are very meaningful, but won’t ever give us the full picture of the complex dynamics of the complete genome. Each “individual” or book can certainly vary from author to author, but given that they have the same structure, and are on the same subject, there are a lot of commonalities. And finally, with just the word count, we can certainly begin to piece together patterns like “generic and common words show up more in the introduction chapters, while more technical words are more expressed in the middle chapters in the meat of the details”.

With this analogy, the problem statement is now something like: suppose we then commission a new book on corn from an author called, say, Chris Perendez, whose friends call him Chris-Per. We tell Chris that he is now not allowed to use a particular book throughout his entire book. The question is, what do the word counts look like for this new book? And more generally, how does the banning of that word affect the rest of the book?

Concerns and Directions

**Preface/Disclaimer:** My (Eagon) background is in computational neuroscience, where I worked a lot with information theory and statistics, but not so much in a genomic or biological context. This analysis initially began as a large brainstorming session purely from this perspective, resulting in the initial analogy that I used as a thought experiment to better formulate my concerns, thoughts, and potential solutions. I intend this document to be an open discussion, so feel free to chime in (comments, etc.) wherever something seems off, or where there’s more opportunity for exploration!

In addition, after completing the first outline of this analysis *without* doing any literature review to keep my perspective fresh, I went back and thoroughly read the two excellent review papers you guys graciously provided (Serin et al. 2016, Delgado et al. 2019). It was fairly gratifying to see that point-for-point, many concerns/suggested approaches were listed in these articles, with explicit discussion on problematic approaches and references to existing approaches. I have excerpted relevant sections from these two reviews to bolster each concern/approach - would love to know more if there are good leads elsewhere on any of them!

Each of the following points are structured as

1.) **Concern**: the problem or discussion point at hand,

2.) **Direction**: my immediate reflections on potential solutions or the bigger picture, and 3.) **Discussion:** where I return to flesh out my thoughts *after* doing the literature review, with relevant citations and excerpts that elucidate out how the field has approached the problem, and strong evidence for the necessity of discussing the concern w.r.t. our project scope.

**Concern 1**: What are we really asking from our model, and is our question way too broad?

In particular, we have non-specific data that is certainly large in absolute size, but if we think about the kinds of questions we can ask, the data size could also be orders of magnitudes smaller than necessary. Consider asking someone to train a CNN model differentiating between 26 different animals with just 26 images, but each is 100 megapixels large. You have a tremendous *amount* of data (look at all these pixels!) but most of us would be fairly confident saying that it wouldn’t be enough as the sole data set, without really deliberating much (this particular example is somewhat facetious, the parallels to the current problem are not as salient, it is only illustrative - the rest of this document is far more targeted).

In fact in our book analogy the problem can potentially be viewed as even worse, where depending on our performance metric, we have an input of just one potential word to ban, and need to train a model to give us the expression change (how do we even quantify this? top 10 words affected?) *across the entire rest of the language.* With just 26, or perhaps if we’re willing to say that all chapters/tissues are similar enough, 260 samples, we are many, many orders of magnitude away from expecting to have enough variation in the data to capture anything truly useful at the end of the day. That’s not to say we can’t say useful things - only that information theoretically, if we’re expecting to *train* a model and *evaluate* that model on any genuine metric across the entire potential space of 40,000 expression levels, there simply isn’t enough information.

**Directions:** There are plenty of ways to utilize what we do have however, and more specific questions (e.g. training a whole dataset with a performance metric on a specific, much smaller candidate selection of genes to perturb) where the ratio of available variation to outcome space is much more favorable are immediately more tangible. In fact, this just gets down to what kind of problem or analysis are we even trying to do - a useful broad ML distinction here is, are we looking at an unsupervised setting, or a supervised setting?

A supervised approach pares down this ratio greatly by reducing the dimensionality of the output space massively, allowing a model to derive its own notions of equivalence sets across the data distributions. In exchange, it requires labeled data and more modalities of data, but provides interpretable evaluations of the model in the context of well-defined success conditions.

An unsupervised approach is at its core inherently an implicit step of a larger pipeline in more applicable settings - we transform data to data, but by definition there is no output space the model is privy to in discerning importance. We *can* evaluate metrics in settings like autoencoders, where we learn a lower dimension representation of the data, and evaluate the model’s ability to preserve information, yet there is no inherent predictive functionality. As a trade off, unsupervised approaches have no strict ratio of “data to performance” - just having 26 books is good enough to get an idea of word counts and their correlations, more data only increases our confidence that we have a good picture. Yet this is only sufficient *in the context* *without* a strict predictive task with performance goals.

**Discussion:** In Serin et. al (2016), this particular categorization was succinctly described as non-targeted/global vs. targeted/gene-guided approaches. This section (Data Selection for Co-Expression Network Analysis) basically listed out a lot of the above concerns, and mentions several times that in order to actually interpret results or make useful predictions, data from other modalities (or even better, experimental setups) is almost necessary. Given that our dataset is currently strictly in the condition-independent, or non-targeted setting, a summary of this concern is that we have a dataset for a global approach, but are trying to ask a predictive question. Obviously this may not have been the intent in the problem description (clarification on this point would be greatly appreciated!), but the explicit line from the given problem statement “*Prediction accuracy for constructed models can be validated on a holdout set of the original expression data*” suggests that there is a performance metric in mind.

Again, to fully reemphasize this point, we are not concerned that the dataset is small or insufficient to say useful things - only that it will be very important to be specific on exactly what the metric is that we are evaluating by, and what the real goal is here.

**Concern 2:** What useful things *can* we say, and what are tangible approaches for breaking the problem apart?

This one boils down to the fundamental capabilities of our model, and exactly why it can say useful things. An interesting theoretical insight that we can generalize for a lot of ML models and approaches is that at some point, we are fundamentally reducing the dimensionality of our problem. A lower dimension representation of our data, or our solution, allows the model to combat combinatorial explosion and get at the essence of the problem. Here I’d like to brainstorm about potential approaches that utilize this kind of thinking in order to make the problem more tangible. The discussion following also highlights a few examples other people in the field have used, along with lots of caveats and important points highlighting the necessity of getting the problem statement/performance metric correct.

**Direction:** A network or graph representation of the full expression profile/word count is really just a “lower dimensional” representation of the full adjacency matrix. The idea of using correlations to weight important edges, either with soft or hard-thresholding, is a natural extension to use to form a pruned, lower dimension representation: indeed, this is essentially WGCNA at its heart. However, instead of Pearson correlation, we can easily come up with a host of other metrics in light of the limited nature of the data. To bring some metrics from my time in computational neuroscience, a natural extension from that would be to use Shannon information or mutual information, where we study metrics in distributional spaces in order to form ideas about the relevance of a single node. We can form hypotheses about pairwise or three-way relations under this framework, or indeed the full entropy function calculated from all possible such tuplets, and much like in Taylor expansions, we can simply trim off the rest after first and second-order relations depending on data quantity.

**Discussion:** Delgado et al. (2019) goes into depth about our class of problems, in terms of working on top of coexpression networks, where the very essence of each separate approach in the comprehensive review revolves around how each method breaks the problem down and does dimensionality reduction. As previously mentioned, WGCNA uses simple Pearson correlation, but other information theoretical approaches like ARACNE (Margolin et al., 2006) actually use pretty much what I described in terms of mutual information. The ARACNE article itself was a very interesting read and pointed out a number of interesting points to discuss about the viability of its approach with respect to dataset quality and quantity, as well as their fundamental approach.

ARACNE uses precisely mutual information as a hypothesis test to form edges in its graph, but limits its computation to pairwise interactions. An interesting result of this is that even in triplet interactions, the approach under certain assumptions will likely correctly identify the edges such that even if in pure Pearson correlation gene A and B are high, the edges instead form A-C-B if C is in truth the source of the causal relation. The paper also outlines a couple of experiments in synthetic data and in real data (human B cells) which seems to verify the veracity of the approach.

What’s really important to note though is that they repeatedly emphasize the importance of bolstering the incredible complexity of the problem with assumptions, as well as certain conditions on data size related to the problem that you want to tackle.

“Since typical microarray sample sizes are relatively small, inferring the exponential number of potential n-way interactions of Eq. (1) is infeasible and a set of simplifying assumptions must be made about the dependency structure. In fact, M > 100 is generally sufficient to estimate 2-way marginals in genomics problems, while P(g i , g j , g k ) requires about an order of magnitude more samples.” (Margolin et al. 2006)

The question for us is are we confident in grouping all tissues together to say that we exceed this very hand-wavy threshold of M > 100 (effectively 260)? Or what is our real performance metric? This is again 2-way marginals, which may not be nearly enough to fully recapitulate an expression profile, and they have already specified that third order interactions require an order of magnitude more data.

The ARACNE paper is more evidence that at datasets our size, there’s a lot of interesting work to be done and information to be gleaned - but only if you ask an appropriate question. Indeed, the ARACNE paper is only truly able to have a performance metric on the synthetic dataset by virtue of, well, being synthetic. The Mendes network generation is also operated on a scale of 2 orders of magnitude smaller, 100 genes with just 200 interactions (and a thousand samples), and the performance evaluation is interpretable in this context *because* they have the underlying probabilities:

“Although this model is a clear simplification of real biological networks, it forms a reasonably complex interaction network that captures some elements of transcriptional regulation, and an algorithm that does not perform well on this model is unlikely to perform well in a more complex case. Within this model, an interaction is unambiguously defined as a direct regulatory effect of one gene on another. Thus the performance of reverse engineering algorithms can be studied by comparing the inferred statistical interactions to the direct interactions in the model.” (Margolin et al. 2006)

Even more importantly, for the real world data set, not only are they targeting variations in a very specific cell, and not across an entire organism, they make no attempt to construct an overall prediction metric (indeed there seems to be no real sensible overall metric). On top of that, they specifically select a *single* gene c-MYC proto-oncogene, that is well known to be one of the key players in the network, and then *qualitatively* evaluate the fact that their network predicts roughly 50% of biochemically verified genes that are related from first neighbors. This is nowhere close to being an automatic performance metric on a leave-one-out dataset, as it looks only at a single gene’s interactions (one of the most well defined in the network), and can only evaluate *against* external data.

At this point it is also perfectly reasonable to bring up the point that ARACNE is only graph construction, and that the whole point of the project is to build a graph neural network that works on top of such a network to do the things we want it to do. Yet it doesn’t change the fact that performance metrics are poorly defined in this context without further data or experimental setups, or the fact that any held-out sample needs some way to be compared to the outputs of our GNN, i.e. form a graph of its own, or a metric derived from a graph. In this context, if the input to our model truly is simply subsets of genes we wish to perturb, then by definition a “good” output is a recapitulation of that entire held-out gene expression with that particular “perturbed pattern” in some shape or form. This is further complicated by the obvious issues of correlation vs. causality, exponentially compounded in the context that we simply aren’t afforded enough samples to have much of a test “set” at all.

**Concern 3:** If the problem statement is indeed more on the targeted side, and we *do* want to do predictive inference under some evaluative framework, is it requisite to have more modalities of data? Also, what kinds of evaluations are useful?

If the project statement is indeed clarified/altered to require a predictive model, I think there’s a lot of fascinating work to be done in establishing a useful evaluation metric or framework. This will require lots of communication on exactly what kinds of data to insight pathways exists, and perhaps a significant amount of work and research in establishing such an analytical framework that *enables* the training of a graph network.

**Direction:** I think this concern gets at the heart of the issue we see, and summarizes many of the points above. With the understanding that we are primarily in a condition-independent, or unsupervised setting, as well as the context that a graph network is simply dimensionality reduction done smartly with our fundamental understanding of the physiology, this last concern points at the fact that this setting seems to be poorly suited for a predictive task. The directions available immediately are either: 1.) add more modalities of data to combat the condition dependency, or 2.) pare down the problem we’re looking at to enable effective dimensionality reduction as well as establishing a much more tractable performance metric.

**Discussion:** Serin et al. (2016) in the section “Gene Prioritization” argues about the benefits of introducing *both* data modalities and more targeted questions in increasing interpretability and model tractability. In an excerpt on using these networks, particularly in similar cases with no *a priori* knowledge, they advocate for bringing in enrichment analyses or essentially additional data modalities:

“Enrichment analysis for the genes within a module is the most widely used technique to associate modules with particular functions. Under the “guilt-by-association” rule, these functional modules provide a powerful framework for the identification of new genes relevant to biological processes and their functional annotation in the absence of strong a priori knowledge. These enrichment analyses mostly rely on annotation databases (Table 1). The most popular ones are the gene ontology (GO) database (Ashburner et al., 2000) and manually curated databases for metabolite pathways such as the Kyoto Encyclopedia for Genes and Genomes (KEGG) (Kanehisa and Goto, 2000), Mapman (Thimm et al., 2004), or BioCyc (Caspi et al., 2014).” (Serin et al., 2016)

Obviously the downside of such an approach is, as mentioned in the very first meeting when I first enquired about the possibilities of pursuing just this avenue in solving the problems above, the scope of the project expands and would add more work. It seems however in light of further consideration, if the problem statement remains predictive in nature, such additions may be necessary. If not additional data, a paring down of the problem statement for specific predictive tasks would greatly increase the chance of yielding useful results as well.

Conclusion

For whoever made it down here, thanks for sticking around! My intention is entirely to start a conversation, as well as discuss possible directions to take the project or specific steps to advance. We’re all very excited to work on the project, but in the specific context of delivering a meaningful Statement of Work that is not disingenuous, we wanted to be fully sure that the problem statement (and hence MVP to deliver) is not only clear, but reasonably tangible by some existing metric in the field. Obviously we’d love to innovate and shoot for the moon, but we want to avoid steps that cannot work at all at certain scales, or produce hopelessly overfit/not meaningful results even if they do complete. I like to think that this process of being super critical of a success condition is not closing off possibilities, but actually expanding them, by leading to all sorts of interesting literature search, deep collaborative reflection, and communication at higher levels to heighten understanding of the fundamental goals of research for all parties involved.