Model-free estimation of higher-order interactions in single cell expression data

L Del Debbio [2], A Jansma[1 2], A Khamseh[1 2 3], C P Ponting[1]

[1]MRC Human Genetics Unit, Institute of Genetics & Molecular Medicine, University of Edinburgh, Edinburgh EH4 2XU, United Kingdom.
[2]Higgs Centre for Theoretical Physics, School of Physics Astronomy, University of Edinburgh, Edinburgh EH9 3FD, United Kingdom.
[3]School of Informatics, University of Edinburgh, Edinburgh EH8 9AB, United Kingdom

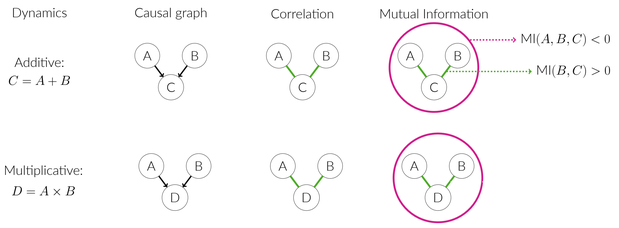
1 Motivation

Molecular cell biology is often understood in terms of pathways of causal interactions between molecules. These networks guide our understanding of biological processes, but identifying these causal effects from observational data is non-trivial. Existing methods suffer from model-misspecification, ignoring higher-order interactions and dependent variables.
We propose a new way to reconstruct genetic networks that suffers from none of these issues, and generalises to interactions that involve more than 2 genes, potentially revealing higher-order, combinatorial gene regulation. **In four different cell types, we find hundreds of 3-point interactions.**

2 Traditional methods hide dynamics

Non-parametric associations

Consider two binary transcription factors A and B separately affecting a target gene C, and a gene D that is only affected by a bound complex of A and B. Estimating the causal graph, correlation structure, and mutual information gives the networks below:



That is, none of these methods are able to distinguish between these biologically different scenarios - they hide the underlying dynamics.

Model bias

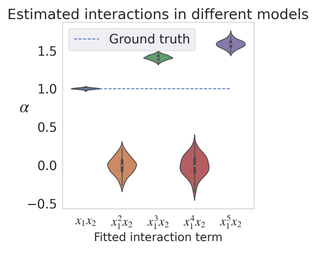
By fitting a model, the dynamics become explicit, but at a high cost. Consider fitting a model of the form:

$$\hat{y} = x_1 + x_2 + \alpha x_1^n x_2$$

to data generated as

$$y = x_1 + x_2 + x_1 x_2$$

where $X_1, X_2 \sim \mathrm{Unif}(-1, 1)$.



The best fits across multiple simulations are shown for $1 \leq n \leq 5$. It can be seen that all models except the ground truth ($n = 1$) are biased, and even powers find no interaction at all.
Even worse: If there is a hidden X_3 such that

$$\begin{aligned} y &= x_1 + x_2 + \alpha x_1 x_2 + \beta x_1 x_2 x_3 \\ &= x_1 + x_2 + (\alpha + \beta x_3) x_1 x_2 \end{aligned}$$

Then the 3-point interaction starts mixing with the 2-point.

3 Model-free interactions

The **effect** I_i of a gene $X_i \in X$ on an outcome Y is the extent to which Y changes when the expression of X_i changes, all other genes ($\underline{X}$) being fixed:

$$I_i = \frac{\partial Y}{\partial X_i}\Big|_{\underline{X}=0}$$

We say that two genes X_i and X_j **interact** when the level of X_j changes the effect of X_i on Y:

$$I_{ij} = \frac{\partial I_i}{\partial X_j}\Big|_{\underline{X}=0} = \frac{\partial^2 Y}{\partial X_j \partial X_i}\Big|_{\underline{X}=0}$$

Similarly, a third gene X_k might modulate this interaction, which results in a **3-point interaction**:

$$I_{ijk} = \frac{\partial I_{ij}}{\partial X_k}\Big|_{\underline{X}=0} = \frac{\partial^3 Y}{\partial X_k \partial X_j \partial X_i}\Big|_{\underline{X}=0}$$

If we take the most general outcome Y possible - the (log of) the joint distribution of all genes X - we get the definition of interaction from [1]:

Definition (Pairwise interaction between binary genes)
A pair of binary genes $\{X_i, X_j\} \in X$ has a pairwise interaction I_{ij} where

$$I_{ij} = \log \frac{p\left(X_i = 1, X_j = 1 \mid \underline{X} = 0\right) p\left(X_i = 0, X_j = 0 \mid \underline{X} = 0\right)}{p\left(X_i = 1, X_j = 0 \mid \underline{X} = 0\right) p\left(X_i = 0, X_j = 1 \mid \underline{X} = 0\right)}$$

This definition has the following properties:

- It is symmetric: $I_{ij} = I_{ji}$.
- Conditionally independent genes do not interact: $X_i \perp\!\!\!\perp X_j \mid \underline{X} \implies I_{ij} = 0$.
- If $\underline{X} = \emptyset$, the interaction reduces to a log-odds ratio.
- It is the double derivative of the joint self-information: $I_{ij} = \frac{\partial^2}{\partial X_i \partial X_j} \log p(X)\Big|_{X=0}$, which describes equilibrium interactions in statistical physics.
- It is model-independent and can be directly estimated from expression data.
- It can be naturally extended to n-point interactions by taking n'th derivatives of $\log p(X)$.

4 Interactions reveal underlying dynamics

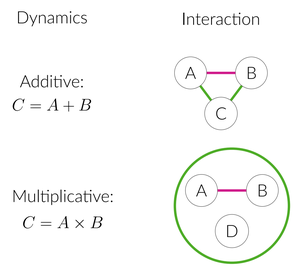
With this definition in hand, we estimate the interaction on the systems from section 2.
We now find we are able to distinguish the two systems - **D being affected by a bound complex of A and B lead to a 3-point interaction**.

This leads us to conclude:

Correlation **is not** Causation

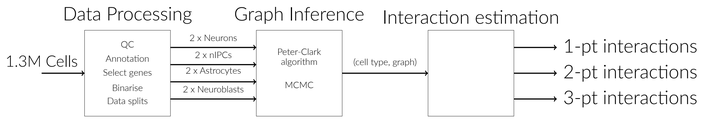
is not **is not**

Interaction

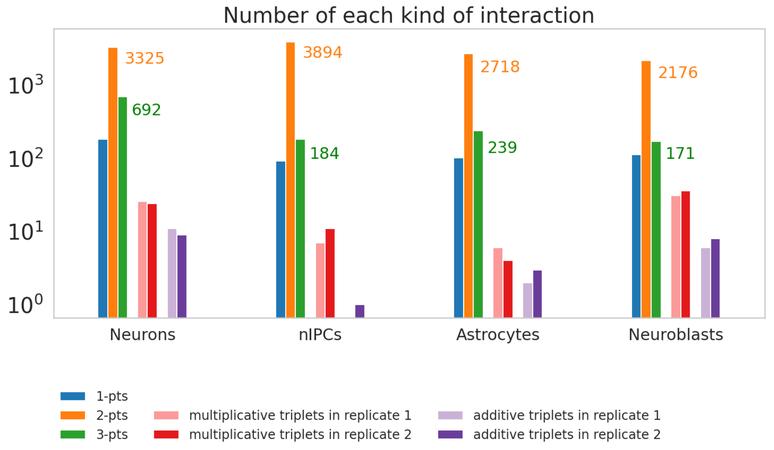


5 Mouse brain cells contain hundreds of 3-point interactions

We consider interactions in a data set of 1.3M embryonic (E18.5) mouse brain cells [2]. From each of four cell types, we construct two biological replicates of 10k cells. To specifically focus on collider triplets, we also estimate the causal graph in each cell type.



Below we show the number of 1-, 2-, and 3-point interactions that are significant in both replicates. Also shown is the number of triplets with an additive or multiplicative interaction pattern in each replicate separately.



6 Conclusion

- Model-free interactions can reveal underlying dynamics in observational data where other methods cannot.
- **We claim the existence of these replicated 3-point interactions, and propose that they reflect higher-order gene regulation.**
- We will validate this biologically and investigate the role of higher-order interactions further in future research.

References

[1] Sjoerd Viktor Beentjes and Ava Khamseh.
Higher-order interactions in statistical physics and machine learning: A model-independent solution to the inverse problem at equilibrium.
Physical Review E, 102(5):053314, 2020.

[2] 10X Genomics.
Million cell dataset, 2017.