Causality - Perspective

This manuscript (permalink) was automatically generated from slobentanzer/causality_perspective_2023@5a5c663 on October 27, 2023.

Authors

- **Sebastian Lobentanzer**
  0000-0003-3399-6695 · slobentanzer · slobentanzer
  Heidelberg University, Faculty of Medicine and Heidelberg University Hospital, Institute for Computational Biomedicine, Heidelberg, Germany

- **Julio Saez-Rodriguez**
  0000-0002-8552-8976 · saezrodriguez · saezlab
  Heidelberg University, Faculty of Medicine and Heidelberg University Hospital, Institute for Computational Biomedicine, Heidelberg, Germany

✉ — Correspondence possible via GitHub Issues

Abstract

Introduction

Correlation is not causation. As simple as this widely agreed-upon statement may seem, scientifically defining causality and using it to drive our modern biomedical research is immensely challenging. Since being described by Aristotle approximately 2500 years ago, causal reasoning (CR) remained virtually unchanged [1] until significant formal and mathematical advancements in the last decades [2,3,4]. In parallel, biomedicine has made major leaps in the past century, in particular in the development of high-throughput and large-scale methods.

Randomised clinical trials show that, in a lower-dimensional context, we can achieve the high levels of confidence needed to satisfy the ethical requirements of modern medicine. However, translating this mode of reasoning into the high-dimensional space of modern omics is met with enormous challenges. The dramatically increased parameter space of models at the molecular level leads to problems in the performance of methods and the identifiability of results, as well as in model explainability.

With this perspective, we want to encourage and guide the use of CR to inform biomedical problems and vice versa. We will elaborate on three main points:

- biases and what they mean for CR, particularly in the context of biomedical data
- the role of prior knowledge in CR and how to translate prior knowledge into suitable biases
- the role of foundation models in molecular systems biology and their relationship to CR

Motivation

Should we have this?

Background

To ultimately explain biases, we must briefly touch on the background of CR. The field of CR distinguishes between *causal discovery* - the process of building hypotheses from data on how agents interact causally - and *causal inference* - the process of predicting how a specific situation will turn out given data and the causal relationships known about the system. In the scientific process from unknowingness to inference on a specific event, the process of causal discovery is more data-intensive than the process of inference, which almost always relies on the prior knowledge from the discovery stage. As a result, most inference mechanisms perform better when including prior knowledge at some point of the process. This has also been observed in biomedical research, for instance in the DREAM challenges [5].

Causal discovery is computationally and statistically very expensive because it needs to account for the variability in data generation while isolating generalisable relationships between single measured species (cite). For modern systems biology, this means that methods for causal discovery typically require large amounts of measurements. Highly parameterised models such as neural networks increase this requirement even further. As such, many regard causal discovery in molecular biomedicine as a scaling problem.

Causal inference, on the other hand, only requires sufficient measurements (replicates) to confidently account for the state of measured species in any condition (which can still be expensive, given the many technical and biological parameters that can influence molecular biology measurements). However, inference is also very sensitive to the completeness of the prior knowledge that is applied; most biomedical prior knowledge is far from complete. For instance, the function of more than 95% of all the known phosphorylation events that occur in human cells is currently unknown [6,7]. In contrast to causal discovery, scaling therefore plays a smaller role in causal inference; here, the main problem is incompleteness and identifying the "right" biases to apply.

The Ladder of Causality

Orthogonally to the distinction between causal discovery and inference, we can also distinguish between different levels of causality. The framework of the *ladder of causality* [] roughly distinguishes three types of CR in increasing order of power: observation, intervention, and counterfactuals. While the inferences we wish to make in biomedical research are often of the counterfactual type (e.g., "would this patient have survived if they had received this treatment?"), the data we have available is typically observational (e.g., "this patient received this treatment and survived") and sometimes interventional (e.g., clinical trials or perturbation screening). To generate interventional or even counterfactual inferences from observational data is a major challenge at least, and impossible at most, depending on the characteristics of the system under study.

There are approaches to delineate interventional inference from observational data, such as the 'natural experiments' framework []. However, these approaches are by their nature even more data-hungry than when using interventional data, as they necessarily discard information that is not relevant to the intervention. Therefore, in biomedical research, there has been a push towards generating large-scale interventional data, for instance through the use of CRISPR/Cas9 screens with single-cell resolution []. Current developments of CR in the biomedical field therefore mostly focus on these types of data.

Deduction and Induction

Lastly, in CR, we can also distinguish between *deductive* and *inductive* reasoning. This is where certain biases are pivotal to the effectiveness of the CR method. Deductive reasoning is the process of deriving a conclusion from a set of fixed and known premises. "All men are mortal, Socrates is a man, therefore Socrates is mortal" is a classic example of deductive reasoning. In biomedical research, this is typically the process of deriving a conclusion from a set of prior knowledge. For instance, knowing about the causes of stroke (which include high blood pressure) and the consequences of an angiotensin receptor-blocking drug (lowering blood pressure) allows us to deduce that the drug can be used to prevent stroke.

Inductive reasoning, on the other hand, is the process of deriving a conclusion from a set of observations; in the biomedical case, this is often measurements. For instance, we would conduct a clinical trial on the preventative effect of the angiotensin receptor-blocking drug on high-risk patients. Since we cannot feasibly test the drug on all potential patients, we instead test on a *representative* sample of the population. Given a statistically significant effect on stroke prevention, we can then perform the inductive step that the drug is generally effective at preventing stroke.

The main difference between deduction and induction is that the former is more reliable, but also more limited in scope, than the latter. In biomedical research, we often have to rely on inductive reasoning because we cannot feasibly test all hypotheses in a deductive manner. In consequence, the *inductive* biases we introduce into our models are a pivotal part of performing CR in biomedical research.

Bias

Meaning and examples of biases

Why do we need biases?

Bias from prior knowledge

Prior knowledge

OmniPath/BioCypher

Ontologies

Modelling on prior knowledge

Statistical, causal, and mechanistic models

Why modelling needs biases and how to introduce them

Causality in foundation models

Current interest in transformers

Recent foundation model benchmarks

Is attention (and large amounts of data) “all you need” to induce reliable biases in your model? (GPT “understands” language well) [8]

What is the mathematical relationship between explicit (e.g. ODE) and implicit (transformers) models?

Latent encodings of explicit prior knowledge (GEARS)

References

1. **The Organon, Or Logical Treatises, Of Aristotle, Vol. 1 Of 2**
   Aristotle
   Forgotten Books (2015)
   ISBN: 9781330267608

2. **Causality**
   Judea Pearl
   Cambridge University Press (2009-09-14) https://doi.org/ggd72q
   DOI: 10.1017/cbo9780511803161

3. **Identification of Causal Effects Using Instrumental Variables**
   Joshua D Angrist, Guido W Imbens, Donald B Rubin
   Journal of the American Statistical Association (1996-06) https://doi.org/gdz4f4
   DOI: 10.1080/01621459.1996.10476902

4. **Myth and measurement: the new economics of the minimum wage**
   David E Card, Alan B Krueger
   Princeton University Press (2016)
   ISBN: 9781400880874

5. **Inferring causal molecular networks: empirical assessment through a community-based effort**
   Steven M Hill, Laura M Heiser, Thomas Cokelaer, Michael Unger, Nicole K Nesser, Daniel E Carlin, Yang Zhang, Artem Sokolov, Evan O Paull, … Sach Mukherjee
   Nature Methods (2016-02-22) https://doi.org/f3t7t4
   DOI: 10.1038/nmeth.3773 · PMID: 26901648 · PMCID: PMC4854847

6. **Illuminating the dark phosphoproteome**
   Elise J Needham, Benjamin L Parker, Timur Burykin, David E James, Sean J Humphrey
   Science Signaling (2019-01-22) https://doi.org/gf8c3h
   DOI: 10.1126/scisignal.aau8645 · PMID: 30670635

7. **The functional landscape of the human phosphoproteome**
   David Ochoa, Andrew F Jarnuczak, Cristina Viéitez, Maja Gehre, Margaret Soucheray, André Mateus, Askar A Kleefeldt, Anthony Hill, Luz Garcia-Alonso, Frank Stein, … Pedro Beltrao
   Nature Biotechnology (2019-12-09) https://doi.org/ggd8n7
   DOI: 10.1038/s41587-019-0344-3 · PMID: 31819260 · PMCID: PMC7100915

8. **ChatGPT broke the Turing test — the race is on for new ways to assess AI**
   Celeste Biever
   Nature (2023-07-25) https://doi.org/gskd92
   DOI: 10.1038/d41586-023-02361-7 · PMID: 37491395