

Computational causal inference methods and single cell proteomics enable the estimation of interventions from purely observational experiments

Devon Kohler

Khoury College of Computer Sciences, Northeastern University, Boston, MA

Presentation Outline

- Problem statement
 - Single cell MS-based proteomics enables estimating the effect of perturbations from observational studies
- Background
- Case studies Targeted vs Discovery
 - Targeted experiment (simulation)
 - Discovery experiment (biological)

Single cell MS-based proteomics enables estimating the effect of perturbations from observational studies

- Understanding the proteome response to perturbations is an important step towards understanding the protein function
- Causal inference methods allow us to estimate this response from purely observational data (i.e., without performing the perturbation)
- Single cell proteomics allows us to do this better than traditional bulk proteomics

Why do we need single cell?

- Remove confounding cell effects in bulk experiments
- Differentiate behavior of cellular subpopulations





Specht and Slavov (2018). JPR 17(8)

Presentation Outline

- Problem statement
- Background
 - Causal inference
- Case studies Targeted vs Discovery
 - Targeted experiment (simulation)
 - Discovery experiment (biological)

Standard statistical analysis

- Infer parameters of a distribution using samples from the distribution
- Use parameters to infer associations among variables
- Calculate P(Erk | Ras)





Causal inference

- Infer parameters under changing conditions, e.g., in the presences of treatments or external interventions
- Calculate P(Erk | do(Ras))



How to infer causality?

- Randomization
 - Test if inhibiting Ras impacts Erk
 - Split replicates into two groups: one who receives intervention, and one does not
 - Estimate probability distribution in the presence of an intervention
 - P(Erk | do(Ras))



What if randomization is not possible?

- Only observational data available
- Can we still estimate the effect of the intervention
 - Yes! (Under certain conditions)



Estimate the effect of interventions using purely observational data

- Requirements for estimating the effect of interventions
 - Observational experimental data
 - Causal network (in the form of a directed acyclic graph (DAG))

Not doing causal discovery

• Correct combination of graphical topology and measured proteins



Causal network – IGF signaling pathway

- Insulin-like growth factor (IGF) or epidermal growth factor (EGF) trigger an event including the MAPK signaling pathway
- Well studied with dynamics characterized in ODE/SDE models



- Latent (unmeasured)



Zucker J. et al. (2021). IEEE Trans. Big Data

IGF signaling system

- Interested in the causal effect of Ras on Erk
- Latent confounder between SOS and PI3K
- P(Erk | do(Ras), SOS) is *identifiable*





Zucker J. et al. (2021). IEEE Trans. Big Data

IGF signaling system

- Assume SOS was not measured
- Now we cannot close the "backdoor" path
- P(Erk | do(Ras)) is not identifiable





Zucker J. et al. (2021). IEEE Trans. Big Data

Presentation Outline

- Problem statement
- Background
- Case studies Targeted vs Discovery
 - Targeted experiment (simulation)
 - Discovery experiment (biological)

Application of causal inference is dependent on the experimental design and biological question of interest

• Targeted Experiment

Construct graph of pathway/network of interest

Decide what proteins need to be measured Run experiment to measure proteins of interest (SRM/PRM) Application of causal inference is dependent on the experimental design and biological question of interest

• Targeted Experiment



• Exploratory experiment

Run experiment to measure many proteins (DIA/DDA) Build a graph around the proteins that are measured Determine what causal queries are possible

Presentation Outline

- Problem statement
- Background
- Experimental goals and design
 - Targeted experiment (simulation)
 - Bulk vs single cell
 - Replication (e.g., the number of cells)
 - Protein-level imputation
 - Discovery experiment (biological)

Targeted experiments allow us to answer a specific question of interest





Use latent variable model (LVM) to leverage information from unmeasured proteins



Presentation Outline

- Problem statement
- Background
- Case studies Targeted vs Discovery
 - Targeted experiment (simulation)
 - Bulk vs single cell
 - Replication (e.g., the number of cells)
 - Protein-level imputation
 - Discovery experiment (biological)

Simulate bulk-MS data

- Linear relationships between proteins
- Peptide ions are missing with some probability being missing at random and missing not at random
- Multiple cell types are mixed

 $\begin{array}{l} \underline{System \ of \ linear \ relationships}} \\ Ras &= \beta_{Ras_0} + \beta_{Ras_1} * SOS + \epsilon_{Ras} \\ Raf &= \beta_{Raf_0} + \beta_{Raf_1} * Ras + \beta_{Raf_2} * Akt + \epsilon_{Raf} \\ Mek &= \beta_{Mek_0} + \beta_{Mek_1} * Raf + \epsilon_{Mek} \\ Erk &= \beta_{Erk_0} + \beta_{Erk_1} * Mek + \epsilon_{Erk} \end{array}$



20.0

17.5

15.0

12.5 논 山 10.0

7.5

5.0

2.5

10

8

12

14

Ras

16

18

20

Simulate bulk-MS data

- Linear relationships between nodes
- Peptide ions are missing with some probability of being missing at random and missing not at random
- Multiple cell types are mixed





Ground truth average causal effect (ACE)

Average effect of increasing the log₂ intensity of Ras by 10 on Erk is 5.85

$$P(Erk \mid do(Ras = 20)) - P(Erk \mid do(Ras = 10)) = 5.85$$

22

Interventional results are very different from true effect

- Compare two interventions
 - P(Erk | do(Ras = 10)
 - P(Erk | do(Ras = 20)
- Average causal effect (ACE)

• 1.3





Why is the estimation incorrect?



Multiple cell types mixed confounded the inference





Using single cell data, we can observe cell type



P(Erk | do(Ras), Cell Type) is identifiable

Using single cells show the true relationship



Splitting models up by cell result in a more accurate ACE estimate



Presentation Outline

- Problem statement
- Background
- Case studies Targeted vs Discovery
 - Targeted experiment (simulation)
 - Bulk vs single cell
 - Replication (e.g., the number of cells)
 - Protein-level imputation
 - Discovery experiment (biological)

Even when using single cell data, we still need sufficient replicates



Even when using single cell data, we still need sufficient replicates



In the presence of latent confounders (e.g., bulk proteomics) no number of replicates can recover the



In the presence of latent confounders (e.g., bulk proteomics) no number of replicates can recover the



Presentation Outline

- Problem statement
- Background
- Case studies Targeted vs Discovery
 - Targeted experiment (simulation)
 - Bulk vs single cell
 - Replication (e.g., the number of cells)
 - Protein-level imputation
 - Discovery experiment (biological)

When observations are MNAR the true correlation/causal effect is masked



35

When observations are MNAR the true correlation/causal effect is masked



36

Causal imputation correctly recovers causal effect in the presence of missing data



Presentation Outline

- Problem statement
- Background
- Case studies Targeted vs Discovery
 - Targeted experiment (simulation)
 - Discovery experiment (biological)

Single Cell experiment - Leduc et al, 2022

- 1556 single cells prepared by nPOP method and acquired with TMT 18-plex
- Melanoma and monocyte cell types
- 2844 proteins identified and quantified with MaxQuant
- Data processed with methods in MSstatsTMT



Leduc A, et al. (2023) Exploring functional protein covariation across single cells using nPOP, Genome Biol

Building a causal network around measured proteins

- Creating a hand tailored network across thousands of proteins is very challenging
- Leverage biological databases to extract causal relationships between proteins in the system
- We use the INDRA database, which includes causal information between proteins

Naïve network extraction results in unusable and uninterpretable network

- Many types of connections may not be relevant
- Some edges have very low evidence
- Not all edges are applicable to the biological question of interest



Thoughtful queries results in reasonable networks

- Focus on abundance events
- Look at specific pathways of interest that show correlation in data
- Filter for edges with high confidence
- Filter for biologically relevant questions



Final causal model

Targeting CDK2 overcomes melanoma resistance against BRAF and Hsp90 inhibitors

Alireza Azimi, Stefano Caramuta, Brinton Seashore-Ludlow, Johan Boström ⁽²⁾, Jonathan L Robinson, Fredrik Edfors, Rainer Tuominen, Kristel Kemper, Oscar Krijgsman ⁽²⁾, Daniel S Peeper ⁽²⁾, Jens Nielsen ⁽²⁾, Johan Hansson, Suzanne Egyhazi Brage, Mikael Altun ⁽²⁾, Mathias Uhlen ⁽³⁾, Gianluca Maddalo ⁽³⁾ ⁽²⁾

Author Information

Molecular Systems Biology (2018) 14: e7858 | https://doi.org/10.15252/msb.20177858

FASEB Journal

RESEARCH ARTICLE

Discovery of novel CDK2 inhibitors using multistage virtual screening and in vitro melanoma cell lines

Lihong Yang, Mukuo Wang, Beibei Li, Shangqin Xu, Jianping Lin 🔀

First published: 24 March 2023 | https://doi.org/10.1096/fj.202201217RR

Fas-Mediated Apoptosis of Melanoma Cells and Infiltrating Lymphocytes in Human Malignant Melanomas

Tetsuo Shukuwa M.D. 🗠, Ichiro Katayama M.D. & Takehiko Koji Ph.D.

Modern Pathology 15, 387–396 (2002) Cite this article

924 Accesses | 17 Citations | Metrics

Melanoma Cell Expression of Fas(Apo-1/CD95) Ligand: Implications for Tumor Immune Escape

M. HAHNE, D. RIMOLDI, M. SCHRÖTER, P. ROMERO, M. SCHREIER, L. E. FRENCH, P. SCHNEIDER, T. BORNAND, A. FONTANA, [...], AND J. TSCHOPP +2 authors Authors

Info & Affiliations



Pathway of interest shows correlation in the data





Protein-level imputation

Conclusions

- Estimation of the effect of interventions is possible given observational single cell data
- Targeted and exploratory studies possible, depending on the goal of the experiment

Existing challenges

Near term (computational)

- More work to be done on building causal networks
- Data processing of single cell experiments

Long term (experimental)

- Post-translational modifications
- Temporal information

Acknowledgements

Northeastern University OLGA VITEK LAB

Statistical Methods For Studies Of Biomolecular Systems

Ben Gyori Northeastern

Northeastern

Jeremy Zucker PNNL

Karen Sachs NextGen Analytics

Lab Members

Kylie Bemis Sara Mohammad Taheri Ritwik Anand Vartika Tewari Sai Srikanth Lakkimsetty Sarah Szvetecz Yinyue Zhu Mateusz Stankiak

