CAUSAL DISCOVERY & INFERENCE FOR GUT MICROBIOME DATA

what is causality in statistics? Some basics...

- Recover the true, underlying causal structure from observed data.
  - Represent causal structure as a structural causal model (SCM)

    (def.) A structural causal model, M, consists of ...
    - a set of exogenous variables, U
    - a set of endogenous variables, V
    - a set of functions f that assigns each variable in V a value based on other variables in the model.

  - Given an SCM, M, one can build a graphical model, G, which contains a node for each variable in M. The graph is a directed acyclic graph (DAG).

- (conditional) independencies and graphical models.
  - chains: $X \rightarrow Y \rightarrow Z$. $X \perp\!\!\!\perp Z \mid \{Y\}$
  - forks: $X \leftarrow Y \rightarrow Z$. $X \perp\!\!\!\perp Z \mid \{Y\}$
  - colliders: $X \rightarrow Y \leftarrow Z$. $X \perp\!\!\!\perp Z$. If we condition on the collider, Y, this "opens" the path between X & Z; so, $X \not\perp\!\!\!\perp Z \mid \{Y\}$.

  - d(irectional)-separation: determine if a pair of nodes are d-connected (likely dependent) or d-separated (independent).

    (def.) A path, p, is blocked by a set of nodes Z iff
    (i) p contains a chain ($A \rightarrow B \rightarrow C$) or a fork ($A \leftarrow B \rightarrow C$) s.t. the middle node B is in Z, or
    (ii) p contains a collider ($A \rightarrow B \leftarrow C$) s.t. the collision node B and any descendant of B is not in Z.

    If Z blocks every path between two nodes X and Y, then X and Y are d-separated conditional on Z. Thus, $X \perp\!\!\!\perp Y \mid Z$.

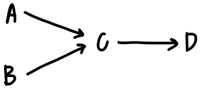
  - Assumptions.
    - causal Markov condition: Every variable in the set of variables V is independent of its non-descendants given its parents.
    - Faithfulness: The only independencies among the variables V are those entailed by the causal Markov condition.

    The above two assumptions are both required to build a causal graph from conditional independencies in the observed data.

How to build a causal graph from data?

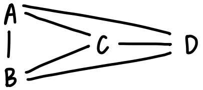
- Peter-Clark (PC) Algorithm.
  - a constraint-based causal discovery algorithm.
  - pseudoalgorithm:
    1. start w/ a complete undirected graph.
    2. Remove edges based on statistical (conditional) independence tests
    3. Identify v-structures
    4. Apply Meek's rules to orient additional edges while preserving v-structures
  - assumptions: causal Markov condition, Faithfulness, no hidden confounders.

↳ Example. Suppose the true causal structure is:

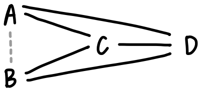


From the data, we see: $A \perp\!\!\!\perp B$, $A \perp\!\!\!\perp D \mid \{C\}$, $B \perp\!\!\!\perp D \mid \{C\}$.

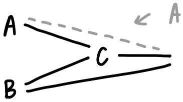
① complete undirected graph:



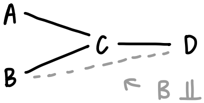
② Remove edges:



$A \perp\!\!\!\perp B$ →

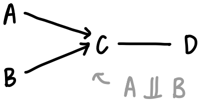


↙ $A \perp\!\!\!\perp D \mid \{C\}$



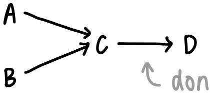
↖ $B \perp\!\!\!\perp D \mid \{C\}$

③ v-structures.



↖ $A \perp\!\!\!\perp B$

④ Meek's Rules.
(not all edges can be oriented; the final graph may have undirected and directed edges)



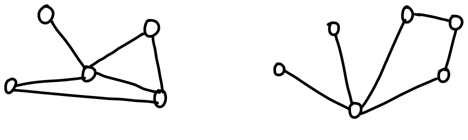
↑ don't create spurious v-structures.

- some issues:
  - choose independence tests that are appropriate for the data distribution.
  - False discovery rate. As # of variables increases, the # of conditional independence tests grows quickly. The FDR is not controlled at α.

How do we apply these methods to gut microbiome data?

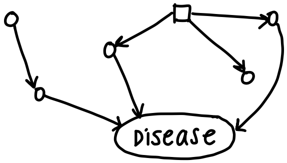
- we can construct two types of networks :

We will call these microbe-microbe interaction networks, where we have networks for each cohort and the nodes are microbes only.



Healthy Cohort Diseased Cohort

We will call this a microbe-disease interaction network, where nodes consist of microbes (○), other covariates (□) and disease status.



- Microbe-Microbe Interaction Network
  - Steps:
    ① Split the dataset into healthy & diseased cohorts.
    ② Eliminate edges using a sparse method, e.g. SparCC, graphical lasso.
      - Purpose: reduce the multiple testing burden in the PC step.
    ③ Run PC w/ a max depth (i.e. conditioning set cardinality) of 2.
  - Interpretation:
    - which microbes are common in both cohorts' networks?
      - Among these common microbes, how do their "subgraphs"/directly linked microbes differ between the two cohorts?
    - which microbes are only in one cohort's network?
    - compare these microbes w/ those in the microbe-disease network?
- Microbe-Disease Interaction Network
  - Steps:
    ① Eliminate microbes using logistic lasso (or some other feature selection procedure)
      - Purpose: reduce the multiple testing burden in the CD-NOD step; interpretability of a complicated vs. simpler network.
    ② Run CD-NOD w/ the non-microbe covariates as the heterogeneity/time index.
  - Interpretation:
    - which microbes are directly linked to disease status?
      - Estimate their causal effect using do-calculus.