Graphical Causal Discovery from Big Biomedical Data

AMIA Workshop on Data Mining for Medical Informatics (DMMI) – Causal Inference for Health Data Analytics
November 4, 2017

Gregory F. Cooper
Department of Biomedical Informatics
University of Pittsburgh
Outline

• Why focus on causal discovery and why now?
• What is the essence of graphical causal discovery?
• What is Center for Causal Discovery and how can it help you?
• What are some open problems in graphical causal discovery?
Science is centrally concerned with the discovery of causal relationships in nature.

- Understanding mechanisms
- Predicting the results of intervention
- Controlling events

Examples:

- Determining the genes and cell signaling pathways that cause breast cancer
- Discovering the clinical effects of a new drug
- Uncovering the mechanisms of pathogenicity of a recently mutated virus that is spreading rapidly in the population
Methods for Graphical Causal Discovery Have Been Advancing Rapidly
Methods for Graphical Causal Discovery Have Been Advancing Rapidly Why?
Methods for Graphical Causal Discovery Have Been Advancing Rapidly
Why?

Algorithmic Advances
+
Availability of Big Biomedical Data
Algorithmic Advances

- In the past 25 years, there has been tremendous progress in the development of computational methods for representing and discovering causal networks from a combination of data and knowledge.
- These methods are often applicable to biomedical data.
Availability of Big Biomedical Data

The variety, richness, and quantity of biomedical data have been increasing very rapidly.
- High-throughput molecular data (e.g., whole-genome sequencing)
- Clinical EMR data (>75% of U.S. physicians use some form of EMR)
- Population health data from social media and mobile sensors

The appropriate analysis of these data has great potential to advance biomedical science.
An Example of Causal Network Discovery from Biomedical Data

A Portion of a Cell Signaling Network
(and Points of Experimental Intervention)

Overview of Experimental Design and Data Analysis

Results of Causal Network Analysis for the Example

Causal Data Analysis

Causal Bayesian network (CBN)
- Directed acyclic graph
- Nodes represent variables
- Arcs represent causal influence
- Specify $P(X \mid \text{parents}(X))$ for each node $X$
Several Key Causal Relationships

<table>
<thead>
<tr>
<th>Graphical Representation</th>
<th>Causal Relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>X → Y</td>
<td>Direct cause</td>
</tr>
<tr>
<td>X → H → Y → Z</td>
<td>An endogenous latent variable (H)</td>
</tr>
<tr>
<td>H</td>
<td>A latent confounding variable (H)</td>
</tr>
<tr>
<td>X → Y</td>
<td>Selection</td>
</tr>
</tbody>
</table>
Basic Causal Discovery Workflow
Basic Causal Discovery Workflow

- Prior Knowledge
- Data
- Causal Analysis
- Causal Networks
- Causal Hypotheses Generated by Biomedical Scientists
The main goal is to suggest causal hypotheses that are novel, significant, and valid.
Basic Causal Discovery Workflow

Experiments

Causal Hypotheses
Generated by Biomedical Scientists

Prior Knowledge

Data

Causal Analysis

Causal Networks
Basic Causal Discovery Workflow

Experiments

Prior Knowledge

Data

Causal Analysis

Causal Hypotheses
Generated by Biomedical Scientists

Causal Networks
Types of Data Include ...

- Experimental data – controlled manipulation of some variables and observation of the others
- Observational data – observation only of the variables with no manipulation
- Both
Types of Data Include ...

- Experimental data – controlled manipulation of some variables and observation of the others
- Observational data – observation only of the variables with no manipulation
- Both
Traditional Method for Learning Causal Relationships from Observational Data

• Question: Does $Y$ cause $Z$?
Traditional Method for Learning Causal Relationships from Observational Data

- Question: Does $Y$ cause $Z$?
- Identify all confounders of $Y$ and $Z$
- Measure them
Traditional Method for Learning Causal Relationships from Observational Data

- Question: Does $Y$ cause $Z$?

- Identify all confounders of $Y$ and $Z$

- Measure them

- Condition on them
Traditional Method for Learning Causal Relationships from Observational Data

- Question: Does $Y$ cause $Z$?
- Identify all confounders of $Y$ and $Z$
- Measure them
- Condition on them
Traditional Method for Learning Causal Relationships from Observational Data

• Question: Does $Y$ cause $Z$?

• Identify all confounders of $Y$ and $Z$
• Measure them
• Condition on them

• Concern: What if we haven’t identified all significant confounders?
Causal Network Approach to Learning Causal Relationships from Observational Data

1. Mathematically define a causal network representation
2. Determine the causal networks that are the most likely, given data and background knowledge
Causal Network Methods for Learning Causal Relationships from Observational Data

- Constraint-based
- Bayesian
- Other
Causal Network Methods for Learning Causal Relationships from Observational Data

- Constraint-based
- Bayesian
- Other
A Simple Example of Learning Causal Relationships from Observational Data

• Three binary variables $X$, $Y$, $Z$
• Time ordering is known:
  
  $X$ occurs before $Y$ occurs before $Z$

• For instance
  
  • $X$: a disease present
  • $Y$: a medication taken
  • $Z$: a symptom

• Question: Does $Y$ cause $Z$?
A Simple Example

• Suppose statistical testing yields the following constraints:
  \( \text{dep}(X, Y), \text{dep}(Y, Z), \text{dep}(X, Z), \text{ind}(X, Z \mid Y) \)

• Consider the consistency of these constraints with respect to the following causal models:
A Simple Example

- Suppose statistical testing yields the following constraints:
  \[ \text{dep}(X, Y), \text{dep}(Y, Z), \text{dep}(X, Z), \text{ind}(X, Z \mid Y) \]
- Consider the consistency of these constraints with respect to the following causal models:
\[ \begin{align*}
X & \rightarrow Y \\
Y & \rightarrow Z \\
H & \rightarrow X, Y, Z
\end{align*} \]
• Given Fever = present, if Pn. = present then App. is unlikely and therefore Abd. Pain is unlikely.

• Given Fever = present, if Pn. = absent then App. is likely and therefore Abd. Pain is likely.
A Simple Example

• Suppose statistical testing yields the following constraints
  \( \text{dep}(X, Y), \text{dep}(Y, Z), \text{dep}(X, Z), \text{ind}(X, Z \mid Y) \)

• Consider the consistency of these constraints with respect to the following causal models:

---

X Y Z
X Y Z
X Y Z
X Y Z
X Y Z
X Y Z

90 additional models
A Simple Example

- Suppose statistical testing yields the following constraints
  - $\text{dep}(X, Y)$, $\text{dep}(Y, Z)$, $\text{dep}(X, Z)$, $\text{ind}(X, Z \mid Y)$

- Consider the consistency of these constraints with respect to the following causal models:

  ![Causal Models]

- In general, only causal models with $Y \rightarrow Z$ (without unconfounding) are consistent with the statistical tests above.
Some Applications of this Simple Causal Analysis Method


They find general patterns of statistical dependency among the measured variables that are consistent with the causal networks that they output.
General Constraint-Based Causal Discovery Algorithms

- They find general patterns of statistical dependency among the measured variables that are consistent with the causal networks that they output.
- They make explicit assumptions.
General Constraint-Based Causal Discovery Algorithms

• They find general patterns of statistical dependency among the measured variables that are consistent with the causal networks that they output

• They make explicit assumptions
  • Causal Markov Condition: Causality is local.
General Constraint-Based Causal Discovery Algorithms

• They find general patterns of statistical dependency among the measured variables that are consistent with the causal networks that they output

• They make explicit assumptions
  • Causal Markov Condition: A node is independent of its non-effects given its direct causes.  \( A \rightarrow B \rightarrow C \)
General Constraint-Based Causal Discovery Algorithms

- They find general patterns of statistical dependency among the measured variables that are consistent with the causal networks that they output.

- They make explicit assumptions:
  - Causal Markov Condition: A node is independent of its non-effects given its direct causes. $A \rightarrow B \rightarrow C$
  - Causal Faithfulness Condition: The only independence among nodes is due to the Causal Markov Condition.
Example: FCI*

- **Representation:** Partial Ancestral Graphs (PAGs)
- **Assumptions**
  - Latent confounders possible
  - No feedback cycles
- **Search:**
  - Selectively apply tests of conditional independence to the measured variables to construct an undirected skeleton graph
  - Apply orientation rules to the skeleton graph to construct a PAG
- **Guaranteed to find a PAG consistent with the data generating CBN in the large sample limit, if the Causal Markov and Faithfulness conditions hold**

Greedy FCI (GFCI)*

- A hybrid of a Bayesian and a constraint-based algorithm
- Applies FGES (see below) to construct a Bayesian network
- Uses that network to construct a draft skeleton graph
- Applies a modified version of FCI to that skeleton graph to generate a PAG

Evaluation of GFCI

• Generated more than 100 random CBNs
  – 1,000 nodes and 2,000 edges
  – Continuous variables with linear Gaussian relationships
• Sampled each CBN to generate 2,000 cases
• Provided cases to GFCI and measured its performance

<table>
<thead>
<tr>
<th>% Latent Nodes</th>
<th>Average Directed Arc Precision</th>
<th>Average Directed Arc Recall</th>
<th># Processors</th>
<th>Average Learning Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>5%</td>
<td>92%</td>
<td>93%</td>
<td>1</td>
<td>15 seconds</td>
</tr>
<tr>
<td>20%</td>
<td>71%</td>
<td>78%</td>
<td>1</td>
<td>19 seconds</td>
</tr>
</tbody>
</table>

Graphical Methods for Learning Causal Relationships from Observational Data

• Constraint-based
• Bayesian
• Other
Bayesian Model Evaluation

Causal model $S_1$

Causal model structure

$A \rightarrow B \rightarrow C$

Causal model parameters $\theta_1$

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(A=t)=0.6$</td>
<td>$P(B=t \mid A=t)=0.8$</td>
<td>$P(C=t \mid B=t)=0.9$</td>
</tr>
<tr>
<td>$P(A=f)=0.4$</td>
<td>$P(B=f \mid A=t)=0.2$</td>
<td>$P(C=f \mid B=t)=0.1$</td>
</tr>
<tr>
<td></td>
<td>$P(B=t \mid A=f)=0.1$</td>
<td>$P(C=t \mid B=f)=0.3$</td>
</tr>
<tr>
<td></td>
<td>$P(B=f \mid A=f)=0.9$</td>
<td>$P(C=f \mid B=f)=0.7$</td>
</tr>
</tbody>
</table>

Data

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>$P(A, B, C)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td></td>
</tr>
</tbody>
</table>
Bayesian Model Evaluation

Causal model structure $S_1$

Causal model parameters $\theta_1$

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(A=t)=0.6$</td>
<td>$P(B=t \mid A=t)=0.8$</td>
<td>$P(C=t \mid B=t)=0.9$</td>
</tr>
<tr>
<td>$P(A=f)=0.4$</td>
<td>$P(B=f \mid A=t)=0.2$</td>
<td>$P(C=f \mid B=t)=0.1$</td>
</tr>
<tr>
<td>$P(B=t \mid A=f)=0.1$</td>
<td>$P(C=t \mid B=f)=0.3$</td>
<td></td>
</tr>
<tr>
<td>$P(B=f \mid A=f)=0.9$</td>
<td>$P(C=f \mid B=f)=0.7$</td>
<td></td>
</tr>
</tbody>
</table>

Data

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>P(A, B, C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>0.6x0.8x0.9=0.432</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td></td>
</tr>
</tbody>
</table>
Bayesian Model Evaluation

Causal model structure $S_1$

Causal model parameters $\theta_1$

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(A=t)=0.6$</td>
<td>$P(B=t \mid A=t)=0.8$</td>
<td>$P(C=t \mid B=t)=0.9$</td>
</tr>
<tr>
<td>$P(A=f)=0.4$</td>
<td>$P(B=f \mid A=t)=0.2$</td>
<td>$P(C=f \mid B=t)=0.1$</td>
</tr>
<tr>
<td>$P(B=t \mid A=f)=0.1$</td>
<td>$P(C=t \mid B=f)=0.3$</td>
<td></td>
</tr>
<tr>
<td>$P(B=f \mid A=f)=0.9$</td>
<td>$P(C=f \mid B=f)=0.7$</td>
<td></td>
</tr>
</tbody>
</table>

data

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>$P(A, B, C)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>$0.6 \times 0.8 \times 0.9 = 0.432$</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>$0.4 \times 0.9 \times 0.3 = 0.108$</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td></td>
</tr>
</tbody>
</table>
Bayesian Model Evaluation

Causal model structure $S_1$

Causal model parameters $\theta_1$

Data

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>P(A, B, C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>0.6×0.8×0.9=0.432</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>0.4×0.9×0.3=0.108</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
<td>0.6×0.2×0.7=0.084</td>
</tr>
</tbody>
</table>
| T   | T   | T   | }
Bayesian Model Evaluation

Causal model $S_1$

Causal model parameters $\theta_1$

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(A=t)=0.6</td>
<td>P(B=t</td>
<td>A=t)=0.8</td>
</tr>
<tr>
<td>P(A=f)=0.4</td>
<td>P(B=f</td>
<td>A=t)=0.2</td>
</tr>
<tr>
<td>P(B=t</td>
<td>A=f)=0.1</td>
<td>P(C=t</td>
</tr>
<tr>
<td>P(B=f</td>
<td>A=f)=0.9</td>
<td>P(C=f</td>
</tr>
</tbody>
</table>

Data

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>P(A, B, C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>0.6×0.8×0.9=0.432</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>0.4×0.9×0.3=0.108</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
<td>0.6×0.2×0.7=0.084</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>0.6×0.8×0.9=0.432</td>
</tr>
</tbody>
</table>
### Bayesian Model Evaluation

**Causal model $S_1$**

![Causal Model Diagram]

- **Causal model parameters $\theta_1$**

|          | A     | B (| A=t) | C (| B=t) | C (| B=f) |
|----------|-------|-------|--------|--------|--------|
| $A=t$    | 0.6   | 0.8   | 0.9    |        |        |
| $A=f$    | 0.4   | 0.2   | 0.1    | 0.3    | 0.7    |

**Data**

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>$P(A, B, C)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T$</td>
<td>$T$</td>
<td>$T$</td>
<td>0.6x0.8x0.9=0.432</td>
<td></td>
</tr>
<tr>
<td>$F$</td>
<td>$F$</td>
<td>$T$</td>
<td>0.4x0.9x0.3=0.108</td>
<td></td>
</tr>
<tr>
<td>$T$</td>
<td>$F$</td>
<td>$F$</td>
<td>0.6x0.2x0.7=0.084</td>
<td></td>
</tr>
<tr>
<td>$T$</td>
<td>$T$</td>
<td>$T$</td>
<td>0.6x0.8x0.9=0.432</td>
<td></td>
</tr>
</tbody>
</table>

$P(data \mid causal\ model_1) = 1.7 \times 10^{-3}$
Bayesian Model Evaluation

**Causal model structure $S_2$**

**Causal model parameters $\theta_2$**

**Data**

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>$P(A, B, C)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>$0.6 \times 0.7 \times 0.6 = 0.25$</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>$0.4 \times 0.6 \times 0.6 = 0.14$</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
<td>$0.6 \times 0.4 \times 0.4 = 0.10$</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>$0.6 \times 0.7 \times 0.6 = 0.25$</td>
</tr>
</tbody>
</table>

$P(\text{data} \mid \text{causal model}_2) = 8.8 \times 10^{-4}$
Bayesian Model Evaluation: Apply Bayes Rule

\[
P(\text{causal model}_1 \mid \text{data}) = \frac{P(\text{data} \mid \text{causal model}_1) \times P(\text{causal model}_1)}{P(\text{data} \mid \text{causal model}_1) \times P(\text{causal model}_1) + P(\text{data} \mid \text{causal model}_2) \times P(\text{causal model}_2)}
\]

\[
= \frac{1.7 \times 10^{-3} \times 0.5}{1.7 \times 10^{-3} \times 0.5 + 8.8 \times 10^{-4} \times 0.5}
\]

\[
= 0.66
\]
Bayesian *Causal Structure Evaluation in General*

\[
P(S_i \mid \text{data}) = \frac{\int_{\theta_i} P(\text{data} \mid S_i, \theta_i) \cdot P(\theta_i \mid S_i) \cdot P(S_i) \, d\theta_i}{\sum_j \int_{\theta_j} P(\text{data} \mid S_j, \theta_j) \cdot P(\theta_j \mid S_j) \cdot P(S_j) \, d\theta_j}
\]
Example Bayesian Algorithm: GES

- Representation: Causal Bayesian networks (CBNs)
- Assumption: No latent confounders
- Search:
  - Forward greedy search phase
  - Backward greedy search phase
- Scoring function: Bayesian (e.g., BDeu)
- Guaranteed to find the data generating CBN in the large sample limit, if the Causal Markov and Faithfulness conditions hold

Some Applications of GES


Fast GES (FGES)

Optimizes GES to be more efficient

• **Approach**
  – Optimized the single processor version
  – Parallelized the algorithm

• **Preliminary evaluation using simulated data**
  – Generated a random causal Bayesian network (CBN)
    o 30,000 nodes and 60,000 edges
    o Continuous-variable version: linear Gaussian relationships
    o Discrete-variable version: multinomial relationships
  – Sampled each CBN to generate 1000 cases
  – Provided cases to FGS and measured performance to learn a CBN
    o Precision and recall for finding causal arcs ($X \rightarrow Y$)
    o Run time
## Results

<table>
<thead>
<tr>
<th># Nodes</th>
<th># Edges</th>
<th># Repetitions</th>
<th>Average Arrow Precision</th>
<th>Average Arrow Recall</th>
<th># Processors</th>
<th>Average Learning Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>30,000 continuous</td>
<td>60,000</td>
<td>10</td>
<td>99%</td>
<td>84%</td>
<td>120</td>
<td>3.4 minutes</td>
</tr>
<tr>
<td>30,000 discrete</td>
<td>60,000</td>
<td>10</td>
<td>86%</td>
<td>24%</td>
<td>120</td>
<td>3.2 minutes</td>
</tr>
</tbody>
</table>

For more information see:

Center for Causal Discovery

www.ccd.pitt.edu

- NIH Big Data to Knowledge (BD2K) Center of Excellence
- A collaboration of 50+ investigators at Pitt, CMU, PSC, and Yale
- Developing and deploying free, open source software for causal analysis of data (e.g., FGES and GFCI)
- Investigating three driving biomedical projects
  - Evaluate the usefulness of causal discovery algorithms on these problems
  - Drive further the development of the algorithms
Driving Biomedical Projects (DBPs)

• Discover cell signaling networks in cancer

• Discover the mechanisms of disease onset and progression in chronic obstructive pulmonary disease and idiopathic pulmonary fibrosis

• Discover the functional (causal) connectivity of regions of the human brain from fMRI data
Center for Causal Discovery

www.ccd.pitt.edu

- NIH Big Data to Knowledge (BD2K) Center of Excellence
- A collaboration of 50+ investigators at Pitt, CMU, PSC, and Yale
- Developing and deploying free, open source software for causal analysis of data
- Investigating three driving biomedical projects
  - Evaluate the usefulness of causal discovery algorithms on these problems
  - Drive further the development of the algorithms
- Training in causal modeling and discovery
  - Tutorial videos and training modules available on website
  - Summer 2018 (June 11-15): week-long course and datathon at CMU
Open Problems in Graphical Causal Discovery Include ...

- Allow a mixture of continuous and discrete variables
- Model non-linear relationships
- Allow a rich mixture of experimental and observational data
- Output uncertainty in edge relationships
- Model causal feedback
Open Problems in Graphical Causal Discovery Include ...

- Learn the causal relationships among latent variables
- Support a variety of forms of temporal modeling
- Improve further the efficiency of causal discovery algorithms
- Perform additional evaluations using simulated data
- Perform additional evaluations using real data in which some of the causal relationships are known
Open Problems
in Graphical Causal Discovery Include ...

• Support causal discovery from case control data
• Support representing a richer variety of prior knowledge
• Explore other representations of causal systems
Summary

• Discovering causal relationships is at the core of biomedical science.

• Graphical methods exist for representing and discovering causal relationships from data, including purely observational data.

• The Center for Causal Discovery is focused on developing and making readily available easy-to-use algorithms and systems for generating plausible causal hypotheses from big biomedical data.

• Many open and important problems remain to be addressed or better addressed.
The Center for Causal Discovery is supported by grant U54HG008540 awarded by the National Human Genome Research Institute through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov).

The content of this presentation is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.
Thank you

gfc@pitt.edu

www.ccd.pitt.edu