Cross-Species Transfer Learning of Genetic Regulatory Networks

Elizabeth Silver
Carnegie Mellon University
Department of Philosophy
silver@cmu.edu

Introduction

Goal: learn Genetic Regulatory Network (GRN) from observational data, using *transfer learning*

- Causal network discovery methods applied successfully to learn GRN,[1] using a compendium of gene expression profiles for yeast [2]
- However: For most species, little public data exists
- Idea: leverage information from *related species*

Difficulties:

- General problems for GRN discovery:
  - High dimension: e.g. 4,300 genes in *E. coli*
  - Causal system includes feedback cycles, unobserved confounders, non-linear mechanisms, non-Gaussian distributions
  - Background knowledge is unreliable
  - Gold standard incomplete: we do not know the whole GRN for any species
- Adapting high-dimensional discovery algorithm for transfer learning
  - Other transfer learning method for GRNs [3] only covers a small # of genes

Data

M3D Many Microbes Microarrays Database (M3D) [4]: manually curated, uniformly normalized, whole-genome microarray data on *E. coli* and *S. oneidensis*

RegulonDB Regulon Database (RegulonDB) [5]: Expert-curated database of known regulatory relationships in *E. coli*

Strategy: Learn GRN of *E. coli* using data from both *E. coli* and *S. oneidensis*; evaluate using RegulonDB.

Data Preprocessing:

- Excluded data from gene manipulation experiments (knockouts, over-expression, plasmids, etc.) as these alter the causal network
- Excluded auto-regulatory relationships from RegulonDB as these are undetectable by causal network discovery algorithms
- OMA Browser provided list of homologous genes between *E. coli* and *S. oneidensis*

Method: Two rounds of greedy search

- Greedy Equivalence Search (GES) [6]
  - Score-based search (score is usually Bayesian Information Criterion)
  - GES **starts from an empty graph**, has two search phases:
    1. Add edges that improve score, until score stays constant; then
    2. Delete edges that improve score, until score stays constant; end.
  - Asymptotically consistent, but with small n, can get stuck in local optima
- Transfer Learning Idea (based on [7]): run GES on pooled data, **then use this graph as a starting point for 2nd round of GES on target species data**
- Large sample size in first round may help GES get close to global optima. Unbiased data in second round may help GES reach the optimum.

References

[1] Marloes H Maathuis, Diego Colombo, Markus Kalisch, and Peter Bühlmann. Predicting causal effects in large-scale systems from observational data. *Nature Methods*, 7(4):247–248, 2010.

[2] Timothy R Hughes, Matthew J Marton, Allan R Jones, Christopher J Roberts, Roland Stoughton, Christopher D Armour, Holly A Bennett, Ernest Coffey, Hongyue Dai, Yudong D He, et al. Functional discovery via a compendium of expression profiles. *Cell*, 102(1):109–126, 2000.

[3] Zaher Dawy, Elias Yaacoub, Marcel Nassar, Rami Abdallah, and Hady Ali Zeineddine. A multiorganism based method for bayesian gene network estimation. *BioSystems*, 103:425–434, 2011.

[4] Jeremiah J. Faith, Michael E. Driscoll, Vincent A. Fusaro, Elissa J. Cosgrove, Boris Hayete, Frank S. Juhn, Stephen J. Schneider, and Timothy S. Gardner. Many microbe microarrays database: uniformly normalized affymetrix compendia with structured experimental metadata. *Nucleic Acids Research*, 36(Database Issue):D866–D870, doi:10.1093/nar/gkm815 2008.

[5] H Salgado et al. Regulondb (version 8.0): Omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Research*, doi: 10.1093/nar/gks1201 PMID: 23203884 PMC: PMC3531196, November 2012.

[6] David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002.

[7] Kathleen M. Gates and Peter C. M. Molenaar. Group search algorithm recovers effective connectivity maps for individuals in homogeneous and heterogeneous samples. *NeuroImage*, 63:310–319, 2012.

[8] Jeremiah J. Faith, Boris Hayete, Joshua T. Thaden, Ilaria Mogno, Jamey Wierzbowski, Guillaume Cottarel, Simon Kasif, James J. Collins, and Timothy S. Gardner. Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. *PLOS Biology*, 5(1):0054–0066, 2007.

Evaluation

- Several searches performed:
  1. G1 (**single species search**): 1-round GES on all of the *E. coli* data, regardless of strain ($n = 424$, $p = 4297$)
  2. G2 (**two-species search**): 1-round GES on pooled data from *E. coli* & *S. oneidensis* (excluding non-homologous genes) ($n = 635, p = 1672$)
  3. G3 (**cross-species transfer**): Starting from G2, 2nd round of GES on only *E. coli* data ($n = 424$, $p = 4297$)
  4. G4 (**cross-strain transfer**): Starting from G1, 2nd round of GES on only *E. coli* MG1655 strain data ($n = 239, p = 4297$)
- Also compared with **absolute marginal correlation**, and **random guessing**
- Each output graph compared against RegulonDB in terms of adjacencies
- If # of nodes = p = 4,297, then # of possible adjacencies = $\binom{p}{2}$ = 9,229,956
- RegulonDB only has 4,106 edges and is likely to be very incomplete
  - Only 2,345 edges supported by strong evidence
  - A “false positive” could be a true-but-unknown edge
- Best outcome measure is “Number Needed to Test” (NNT): expected # of experiments performed to discover one new transcriptional regulator

Results

RegulonDB: all 4,106 edges	**# Edges**	**TPR**	**FPR**	**TDR**	**NNT**
Guessing (95% quantile)[a]	14,381	0.268%	0.156%	0.0765%	1307
Marginal correlation[b]	14,381	1.76%	0.155%	0.501%	200
1-round GES (all *E coli*)	14,381	2.33%	0.155%	0.661%	151
1-round GES (*E. coli* + *S. on.*)	6,143	0.857%	0.0662%	0.570%	175
2-round GES (*E. coli* + *S. on.* → *E. coli*)	20,263	2.72%	0.218%	0.548%	182
2-round GES (*E. coli* → *E. coli* MG1655)	17,322	1.79%	0.187%	0.421%	237

Table 1: Adjacencies compared to RegulonDB (all edges)

RegulonDB: 2,345 strong edges	**# Edges**	**TPR**	**FPR**	**TDR**	**NNT**
Guessing (95% quantile)	14,381	0.299%	0.156%	0.0487%	2054
Marginal correlation	14,381	2.19%	0.187%	0.294%	277
1-round GES (all *E coli*)	14,381	3.13%	0.155%	0.508%	197
1-round GES (*E. coli* + *S. on.*)	6,143	0.987%	0.0663%	0.374%	267
2-round GES (*E. coli* + *S. on.* → *E. coli*)	20,263	3.56%	0.219%	0.410%	244
2-round GES (*E. coli* → *E. coli* MG1655)	17,322	2.19%	0.187%	0.294%	340

Table 2: Adjacencies compared to RegulonDB (edges with strong evidence)

[a] Choosing 14,381 edges at random, the # of true positives is distributed hypergeometrically

[b] Assuming same density as graph produced by 1-round GES

Conclusion

- Unfortunately, vanilla GES outperformed 2-round GES: **transfer learning doesn't help!**
- GES does a little better than marginal correlation (using GES, researcher must perform only 75% as many experiments as when using marginal correlation).
- Open question: Are results driven by weird data set, or problems with algorithm?

Planned extensions

- Simulation studies (eliminate weird data)
- Incorporate background knowledge into search
  - Faith et al. [8] only allowed edges out of genes known to be Transcription Factors
  - Many methods restrict search to a small subset of genes
  - Use computational predictions to feed GES a structured prior
- Use more closely related species &/or more homogenous data
  - Need another convenient database like M3D
- Tweak edge-deleting phase of GES so it is more aggressive (to get sparser graphs in 2nd phase)