Inferring Causality from Finite Data using Conditional Independence

Daniel Speyer
Columbia University

Abstract

We propose techniques for inferring causal structure from limited observational data. These techniques produce less extensive results than those which require exact joint probabilities, but are more practical. The general principle is to describe multiple possible causal models in sufficient detail to have with specific consequences and test those consequences. After developing these techniques abstractly, we test them in simulation and on real world Crohn's Disease data.

1 Introduction

Inferring causal graphs from observational data is a widely-sought goal in statistics. It is especially important in medicine, where finding the cause of a disease may provide a way to cure or prevent it, but finding an effect of a disease is of little practical value. Medicine is also a context in which observational data is plentiful and finding correlations is easy, but running intervention studies is slow, expensive, regulated, and potentially dangerous.

The most common tool for causal inference is conditional independence. The specifics of a causal graph determine which node are independent conditioned on which others by the "bayes ball" rule, and therefore it should be possible to observe the independences and work backward to the causal graph.

This has proved to be more difficult in practice. In particular, observing independence is not as simple as it sounds. [Pearl 2008] suggests we test joint conditional probability distributions for equality. Leaving aside the curse of dimensionality, this assumes we have the exact distributions. If all we have is a finite random sample from those distributions, we can construct posteriors, but we cannot perform an equality test. [Spirtes, Glymour & Scheines 2001] encourage us to assume that "the statistical decisions required by the [causal inference] algorithms are correct for the population", while admitting that this requirement "is often not met in practice". Others have described this as "possessing an independence oracle" [Peters *et al.* 2015; Kalisch *et al.* 2012]. Many simply include a "test if $a \perp\!\!\!\perp b|s$" step in their algorithms, assuming the reader will already know how.

In practice, the usual solution [Kelleher 2017; Kalisch *et al.* 2012] is to treat independence as a null hypothesis, try to reject it at some p threshold, and treat any failure as establishing it. Needless to say, this is incorrect. [Murphy 2007] punts the responsibility to the caller instead, which, while not actually wrong, is unhelpful

Dealing with finite data means the possibility of dealing with too little data. The elegant solution is to give some numerical expression of confidence which becomes "I don't know" when the data gets too small. This solution has another benefit: in the biomedical context, it is routine to test thousands of equally plausible hypotheses at once. A flat accuracy of 99% is unhelpful in the face of this, but a calibrated bayes factor allows compensation.

Traditional inference algorithms such as [Pearl 2008] generally construct the entire DAG by a series of logical deductions that depend on each other. Such approaches do not translate easily to uncertainty. If we took the posterior probability for each successive test in a traditional algorithm and multiplied them to get a posterior for the entire graph, the resulting posterior would likely be too low for any use, and might not be the overall plurality. Furthermore, the number of possible causal graphs is roughly exponential in the number of variables, so finding probabilities for all of them is intractable outside of toy problems (even *storing* those probabilities is beyond the reach of real-world hardware for most interesting problems in medicine!).

It may be possible to usefully approximate entire-DAG inference, but we will attempt a less ambitious solution: to infer, as accurately as possible, the direction of a single arrow. Specifically, one pointing from a variable we can easily intervene on (a "cure") to one we care about (a "disease").

1.1 Motivating Problem: Crohn's Disease

This paper is optimized around a specific practical problem: untangling the microbiome's role in Illial Crohn's Disease. It is well established that there are many differences in the intestinal bacteria of healthy people and of people with the disease [Hofer 2014], but it is not established which (if any) of the differences of bacteria *cause* the disease. There have been attempts to determine this by randomized controlled trial, but early results are discouraging [Prantera 2016] and the number of species, combined with other relevant variables, make exhaustive RCTs impractical.

Data is available on this problem from [Li, Frank & Sartor 2014], including genetic, microbiome and health information, but for only 58 patients (plus another 24 with microbiome and health information, but not genetic). The microbiome information is a series of 16S reads, but can generally be described as "present" or "absent", with very low concentrations of a species rounded off to "absent". This comes much closer to fitting the empirical distributions than any simple scalar formula, which is why this paper uses binary variables.

2 Collider Finding

Let us begin with the simplest case. Suppose we have a known cause, a known effect, and a variable which connects to the effect in an unknown way. For Crohn's Disease, the known cause is mutations in the NOD2 gene [Philpott *et al.* 2014], the known effect is the disease, and the unknown is each of 222 species of plausibly relevant bacteria, taken independently. To help keep our example in mind, we will refer to these factors as G ("gene"), D ("disease") and B ("bacterium"). So our initial graph is $G \rightarrow D - B$. We do not observe a correlation between G and B, but that might only mean our test is underpowered.

For now, we will assume that there is no direct causal link between G and B, and furthermore that there are no unobserved confounders or selection effects. We will consider these later. This leaves us only two models, $G \rightarrow D \rightarrow B$ or $G \rightarrow D \leftarrow B$. We can call these "chain" and "collide" models, or m_{ch} and m_{co} for short. Can we distinguish between them?

Yes. Let us consider $p(G, B)$:

$$\begin{aligned} p(G, B|m_{ch}) &= \sum_D p(G)p(D|G)p(B|D) \\ p(G, B|m_{co}) &= p(G)p(B) \end{aligned}$$

We do not actually know the terms on the right side of those equations, so let us parameterize both models with θ and rewrite the equations as:

$$p(G, B|m_{ch}) = \int \left(\sum_D p(G|\theta)p(D|G, \theta)p(B|D, \theta) \right) p(\theta) d\theta$$

$$p(G, B|m_{co}) = \int p(G|\theta)p(B|\theta)p(\theta) d\theta$$

We can learn $p(\theta)$ from available data using dirichlet priors. The integrals would be difficult algebraically, but they can be adequately approximated with monte-carlo sampling.

Once we have these, let $n_{g,b}$ be the count of datapoints with G=g,B=b and we can use:

$$p(n_*|m) \propto \prod_{g,b} p(g, b|m)^{n_{g,b}}$$

This is proportional instead of equal because there is a combinatoric term, but it cancels when taking odds ratios so it can be safely ignored. From here, we can apply standard bayesian updating.

2.1 Multiply-Connected Graphs

What if there is a causal effect $G \rightarrow B$? It must be weak enough not to be detected, but that doesn't say much. Let us consider it by cases.

2.1.1 False Colliders

Can a true graph of $G \rightarrow D \rightarrow B$ produce a $p(G, B)$ more similar to a graph $G \rightarrow D \leftarrow B$? This would require the direct and indirect influence of G on B to cancel out rather precisely. If the indirect influence is stronger, we will pick the correct model, albeit underconfidently. If the direct influence is too strong, we will see a correlation between G and B. And if the direct influence is in the same direction as the indirect, the correct model will remain the better fit (though both models will be worse).

2.1.2 False Chains

Similarly, can a true graph of $G \rightarrow D \leftarrow B$ produce a $p(G, B)$ more similar to a graph $\overbrace{G \rightarrow D \rightarrow B}$? Again, this is possible, but there is no reason for $p(B|G)$ to resemble $\sum_D p(B|D)p(D|G)$. If $p(B|G) - p(B|\bar{G})$ is too large, $G \not\perp\!\!\!\perp B$ will be detected initially, whereas if it's too small, the correct model will be chosen. Furthermore, what shrinks the upper bound is for the predicted $G - B$ relation in the chain model to be small, which also means any erroneous bayes factor will be small.

2.2 Unobserved Confounders

This technique is vulnerable to confounders. Specifically, if the $D-B$ link is the product of a common cause, that will generate no $G - B$ dependence, exactly like a collider. That is, it will suggest we ought to use that bacterium as a treatment, though in fact doing so would be ineffective.

2.3 Without a Known Cause

Suppose we do not have a link of known direction. If we can find two variables B_1 and B_2 such that both correlate to D but not to each other, then the only simple causal graph that could generate this is $B_1 \rightarrow D \leftarrow B_2$. This finds two causes.

Just as we looked at $p(G, B)$ in the previous case, here we look at $p(B_1, B_2)$. The competing hypothesis $B_1 \rightarrow D \rightarrow B_2$ is the same as before, but we must now consider another competing hypothesis: $B_1 \leftarrow D \rightarrow B_2$ (let us call this the "V" model). This presents no difficulty, as we can use the exact same formula:

$$\begin{aligned}
p(B_1, B_2|m_v) &= \sum_D p(B_1|D)p(B_2|D)p(D) \\
&= \sum_D \frac{p(B_1, D)}{p(D)} \frac{p(B_2, D)}{p(D)} p(D) \\
&= \sum_D \frac{p(B_1, D)p(B_2, D)}{p(D)} \\
&= \sum_D \frac{p(B_1)}{p(B_1)} \frac{p(D, B_1)}{1} \frac{p(B_2, D)}{p(D)} \\
&= \sum_D p(B_1)p(D|B_1)p(B_2|D) \\
&= p(B_1, B_2|m_{ch})
\end{aligned}$$

There is also the possiblity of $B_1 \rightarrow B_2 \rightarrow D$, but in this case the $B_1 - B_2$ link will be stronger than the $B_1 - D$ link, so there is no risk of a false collider.

More worrying is the possibility of multiple-connectedness. For boolean D, if there are many variable $B_{1..n}$ that strongly influence it, they *must* correlate to each other. Otherwise they would be responsible for more than all of the variation in D. Fortunately, the common case here will result in false negatives.

3 D-Separation

Ideally, we'd like to answer questions of the form $A \overset{?}{\perp\!\!\!\perp} B|S$, not just find individual colliders. This would allow us to apply ordinary causal inference algorithms, at least so far as confidence permits. It is probably impossible for a formula to say that $A \not\perp\!\!\!\perp B$ in full generality, but it may be possible to say $A \not\perp\!\!\!\perp B|S$ sufficiently for our purposes.

Suppose A, B and S are variables already known to correlate. The simplest associated models are $S - A - B$, $A - S - B$ and $A - B - S$, where S severs only in the $A - S - B$ case (the directions on the arrows are not important at the moment, except that there can be no colliders). For the first two models:

$$\begin{aligned}
p(A, B|S, m_{\cancel{sever}}) &= p(A|S)p(B|A) \\
p(A, B|S, m_{sever}) &= p(A|S)p(B|S)
\end{aligned}$$

And there is a symmetric formula for the last one. Since there are two alternative hypotheses, we take whichever produces the greater probability. Somewhat surprisingly, there is no advantage here in using monte-carlo posteriors instead of simple maximum likelihood estimators for the conditional probabilities.

There is another simple causal graph that can cause three variables to correlate: all could be influenced by a common confounder H (for "hidden"). This case is impossible to test for in full generality, because S could follow H so closely as to be an acceptable proxy, and therefore controlling for S effectively controls for H. In theory, it should be possible to infer H and its conditionals usings a Gibbs-sampler-like system. In practice, there is more room to overfit this more complex model and no straightforward way to compensate. What does work surprisingly well is to ignore this case and apply the exact same test as before.

3.1 Applying D-Separation

If we are willing to trust the d-separation test, can we use it to find causes of a variable of interest? Deducing the entire causal graph is a brittle endevour, but there is a local solution.

Suppose there exist three bacteria known to have a common (hidden) cause, and one of those bacteria causes the disease, that is $H \rightarrow B_1 \rightarrow D$ B_2 B_3 The role of B_1 can be identified because it severs B_2 and B_3 from D. No other simply connected graph has these properties.

Note that this does not attempt to find *all* causes of D, only those for which suitable H, B_2 and B_3 exist.

In theory, the trio $B_{1,2,3}$ can be identified without domain knowledge because they all correlate and remain correlated when controlling for any of them. In practice, this is highly brittle. If the connections are too strong, controlling for one will sever the others because the one is an adequate proxy for the cause, but if the connections are two weak, the correlations will not be discernible.

3.1.1 Application to Genomic Studies

The causal graph mentioned above is a common one in genome-wide association studies, where the hidden factor is population structure, the three correlating variables are genetic features and the variable of interest is a phenotype. The goal, once again, is to determine *which* generic feature directly affects the phenotype. One answer is that it is the one which best separates the others from the phenotype, and that this can be measured with the same formulas as here.

Whether this will produce useful results in practice is a question for a later study.

3.1.2 Unobserved Confounders

This technique is robust against unobserved confounders. If the links between H and B_{1-3} in the preceding graph are actually mutual effect from a confounder, then that confounder can be thought of as a part of H (which is itself unobserved). If the link between B_1 and D is via confounder, then there will be no $B_{2,3} - D$ correlations to seperate.

3.1.3 Non-Simply-Connected Graphs

These techniques depend on simply-connected graphs. Both the d-separation test and its application can say nothing of significance in the face of multiple connections.

4 Parsimony

Given that it will be necessary to assume either "no unobserved confounders" or "simple connection", can we at least say that these are the most parsimonious explanations for our data?

From a graphical model perspective, they are not. It is the absence of a link which is a statement about the world.

From a biochemical perspective, they are. Each link asserts an interaction, and the default for organic molecules or micro-organisms is not to interact with each other. As for unobserved confounders, the presence of unobserved variables in biology is a given, but for them to be confounders requires *two* interactions that relate to each other.

Fields other than biology will need to reconsider this.

5 Testing in Simulation

To test these tools, we created causal networks with the structures we sought to distinguish, filling in the parameters at random. Then we used each net to create a dataset, and used the dataset to infer the shape of the net. We used the fraction of nets of each shape as a prior and used that to convert the bayes factor into a posterior, which we compared to the true shape of the net.

We calculated the mean log error, that is, we took the probability the system assigned to the correct shape, took its negative natural log, and took the mean of those over all generated nets. Always being certain and correct would be an error of 0, whereas ever being certain and wrong would be an error of infinity.

We also drew calibration graphs. For these, we gathered test nets by their posteriors (in buckets 5 percentage points wide) and calculated the actual percentage of each shape for nets in a bucket. A perfectly calibrated test is one such that of all the nets it assigns a 85% probability of being colliders, 85% of them actually are.

5.1 Collider Detection

This test is well-calibrated over random parameters, as shown by figure 1, but with only 58 datapoints, the vast majority of nets return bayes factors close to 1. Only 8% of nets produce bayes factors outside the [0.1, 10] range. The average log error is 0.62, roughly the same as if it had returned a posterior of 0.5 for everything.

This is not terribly surprising. 58 datapoints is not very much data. Any statistical test would be hard-pressed to report useful conclusions from that except in the presence of strong trends, which randomly-drawn parameters rarely produce. That the test correctly identifies what it does not know is good, but not very useful.

Increasing the size of the dataset helps (see figure 2). At 1000 datapoints, 38% of results are outside that range, and the average log error drops to 0.42. Obtaining 1000 patients' microbiome data might be expensive (though

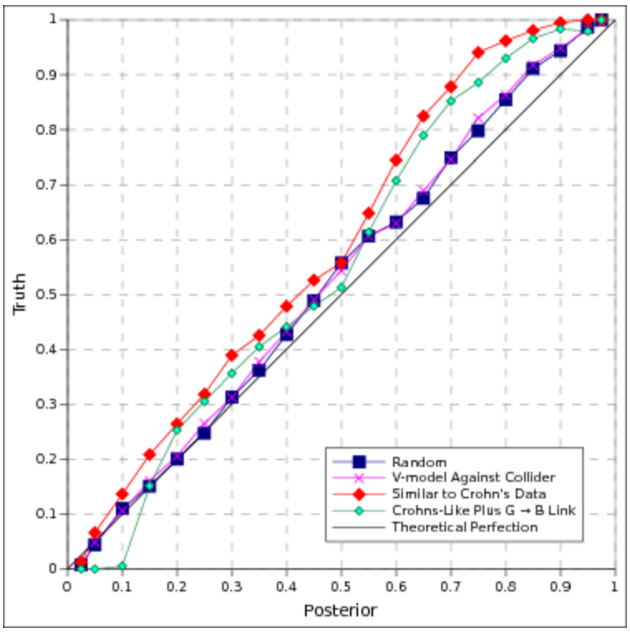

Figure 1: Simulated data using 100k chain and collide models, each generating 58 datapoints. Higher numbers indicate collider. Blue squares indicate parameters drawn from a flat distribution. Purple Xs indicate the same, except discerning colliders from v models instead of chains. Red diamonds indicate a Crohn's-like distribution ($p(G)$ and $p(D|G)$ taken from data; $G \perp\!\!\!\perp B$ and $D \not\perp\!\!\!\perp B$ 'shown' by χ^2 test, $p > 0.1$ and $p < 0.01$ respectively). Open green diamonds indicate the same, except with a $G \to B$ link in the true causal model (still $p > 0.1$).

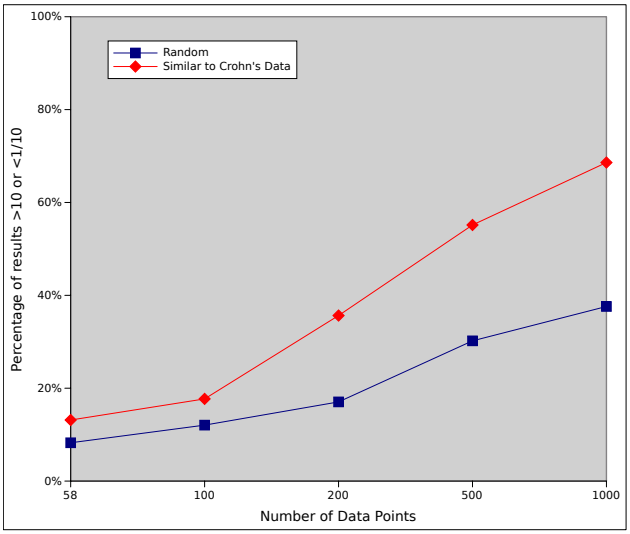

Figure 2: Fraction of bayes factors that were at least 10:1 one way or the other as a function of dataset size, for the collider-detection test.

dropping rapidly, as direct-to-consumer microbiome sequencing becomes available), but obtaining similar epidemiological data is common.

We also tested this using data more similar to the Crohn's Disease dataset. We generated nets using real-world $p(G)$ and $p(D|G)$ with random bacterial parameters. Then we filtered them to only include nets where $B \not\perp\!\!\!\perp D$ was detectable by χ^2 test $p < 0.01$, and to not include any where $G \not\perp\!\!\!\perp B$ was detectable by χ^2 test $p < 0.1$. The former is a test we would apply anyway (it is not useful to determine causality of a link that may not exist!) and the latter is a cheap test that does not rule out a significant number of species. The calibration is slightly worse as judged by the graph, though average log loss is lower 0.59 at 58 datapoints, 0.23 at 1000, and the tendency to produce useful results increases by almost a factor of 2. This is because the Crohn's-like generation rules out several scenarios about which nothing useful can be said, such a too-weak $B - D$ link, or a $p(G)$ so high or low that the other case cannot be analyzed.

Finally, we tested the system using nets that have a $G \to B$ link, albeit one that cannot be detected by χ^2 test ($p \not< 0.1$). As expected, this produced a moderate general trend to underconfidence, but still a mostly well-calibrated result.

In every case, the error at the high end was in the direction of underconfidence. As the high end corresponds to clinical applicability, this means that we will occasionally miss an effective therapy that an ideal algorithm would have found, but we will not claim an ineffective therapy is effective (any more than normal uncertainty requires).

5.2 D-Separation

The d-separation test has mean log loss of 0.41 when comparing a separator to a chain (that is, $A - S - B$ vs $A - B - S$) and 0.49 when comparing to a common, hidden cause ($A - S - B$ vs $H - A, B, S$). Both follow close to a straight line (see figure 3).

The applied severing technique does not try to find *all* causes of D, only those which happen to fit this motif and are strong enough for a χ^2 test to pick up. As such, it is unclear what it would mean to test it in simulation. Certainly with parameters chosen for the purpose, it will work. With random parameters, it will generally fail to find trios.

6 Crohn's Disease

Returning to the original motivating problem, what do these techniques show for Crohn's Disease?

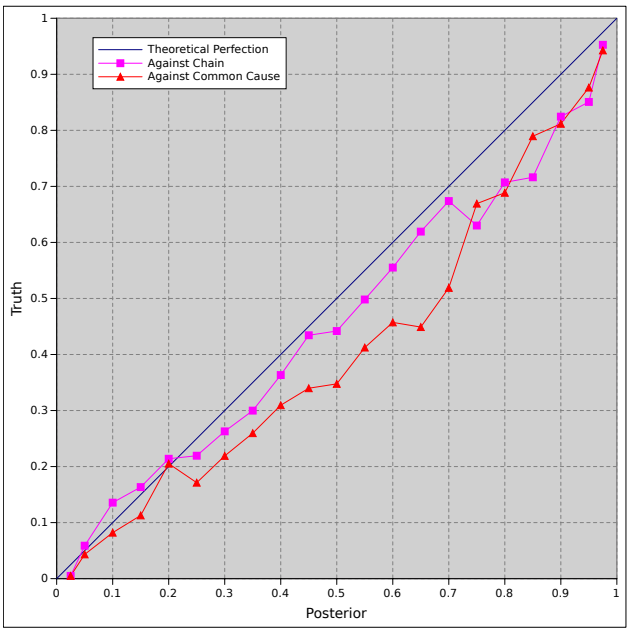

Figure 3: Calibration for the d-separation test, both in the case it was designed for and in the case of comparing severing to common cause. Higher probabilities indicate separation.

First we converted the lists of reads into boolean values. We converted each read to a species names using BLAST [Altschul *et al.* 1990] and the 16S Microbial database [NCBI 2014]. For each read, we selected the highest-matching species (an imperfect, but unbiased technique) and dropped all reads without matches. We divided by total reads to get a concentration. For each species, we chose a threshold that maximized the number of Crohn's Disease statuses that could be correctly predicted by that bacterium alone. Finally, we compared the concentrations to the thresholds to produce boolean values. These boolean values can then be thought of as "has a clinically relevant dosage of the bacterium".

Once the values were booleanized, all those with entropies less than 0.5 were dropped on the logic that a species that's always the same can't have a clinical effect. This left 222 species, of which 98 showed correlation to Crohn's Disease ($p < 0.01$). Finally we dropped 21 species which correlated with NOD2 ($p < 0.1$). While such a high p threshold is usually irrelevant, it is what we used in simulation to limit non-simple connection, so we'll apply it here too.

6.1 Gene Collision Test

Of these the collider-finding test against the NOD2 gene found 7 which show signs of impacting Crohn's Disease, though only one (Lactobacillus acidophilus) has a bayes factor greater than 10. They are shown in table 1.

Inconveniently, that species appears to be *harmful*, which is surprising given the usual role of lactobacilli but less so given that two RCTs found this same result (albeit nonsignificantly) [Prantera *et al.* 2002; Bousvaros *et al.* 2005]. It may be relevant that L. acidophilus is usually found in the *jejunum* [Walter 2008], but these samples are from *ileum* biopsies. Perhaps the cause of Crohn's Disease is not which bacteria are in the intestine, but where in the intestine they are. This result is inconvenient because removing or moving bacteria is far more difficult than adding them.

Could this be the result of a direct $G \to B$ link? In this case, the generally understood role of NOD2 as mediating *untargetted* immune responses [Philpott *et al.* 2014], makes that unlikely.

6.2 Adjusting for Multiple Hypotheses

Having considered 77 species of bacteria in order to find any that appear causal must cast some doubt on our results. Worse, considering 2926 pairs of species for the Bacterium-Bacterium Collider casts extreme doubt.

The classic response to this problem would be to divide our prior by the number of hypotheses, similarly to a Bonferroni correction. This would represent our a-priori knowledge that exactly one of the species under consid-

Species	Sick When	P-Value ICD link	Bayes Factor Causality
Streptococcus pseudopneumoniae	$> 6.36 \times 10^{-5}$	0.000031	5.15
Streptococcus infantis	present	0.0014	2.55
Lactobacillus acidophilus	$> 8.15 \times 10^{-5}$	0.00017	10.64
Sphingopyxis alaskensis	present	0.0004	2.43
Clostridium methylpentosum	$\leq 6.31 \times 10^{-4}$	0.00085	2.53
Roseiflexus castenholzii	present	0.000073	3.30
Ruminococcus faecis	$\leq 1.07 \times 10^{-3}$	0.0022	2.40

Table 1: Results of the direction test on 222 interesting species, using χ^2 to check a relationship to Crohn's Disease and the direction test described here to establish that it causes (or prevents) the disease.

Species 1	Sick When	p.v.	Species 2	Sick When	p.v.	B.F
Lawsonia intracellularis	$\leq$ 1.3E$-$4	0.004	Sporobacter termitidis	$\leq$ 2.5E$-$4	4E$-$6	4.5
Anoxybacillus thermarum	present	1E$-$6	Ruminococcus albus	$\leq$ 1.2E$-$4	0.004	4.5
Hespellia stercorisuis	$\leq$ 9.9E$-$5	0.002	Ralstonia solanacearum	$\leq$ 1.7E$-$2	0.004	4.4
Clostridium xylanolyticum	$\leq$ 1.2E$-$3	0.001	Ralstonia solanacearum	$\leq$ 1.7E$-$2	0.004	4.2
Anoxybacillus flavithermus	$>$ 1.7E$-$4	1E$-$6	Eubacterium siraeum	$\leq$ 9.9E$-$5	0.0002	4.1
Anoxybacillus thermarum	present	1E$-$6	Coprococcus eutactus	$\leq$ 1.6E$-$4	0.0002	5.9
Anoxybacillus flavithermus	$>$ 1.7E$-$4	1E$-$6	Coprococcus eutactus	$\leq$ 1.6E$-$4	0.0002	5.3
Anoxybacillus kamchatkensis	present	3E$-$5	Gemmiger formicilis	$\leq$ 8.7E$-$4	5E$-$6	8.5

Table 2: Results of collider-search looking for pairs of bacteria. The "sick when" column indicates when that species is associated with disease and the "p.v." column the p-value showing that association (χ^2 test). The "B.F." column gives the bayes factor for the collider model as opposed to chain or v.

eration is causal. In fact, we know nothing of the sort. Different species of bacteria routinely emit similar chemical signals, and likely play similar biological roles.

What do we know a-priori? For any given correlating species, a 50% prior credence of causality sounds reasonable, albeit debatable. A 99.9% confidence that between 20 and 50 of the species are causal most emphatically does not. A different prior does not protect us from this class of absurdity. The only reasonable conclusion is that our uncertainties regarding different species are not independent of one another. Specifically, each time we learn the causal nature of a single correlating species, we should update our belief in the others in the same direction. This makes sense biologically, since some of our uncertainty about different species stems from the same unknowns in human biology.

Sadly, this does not give us a numeric answer, but it does encourage us to look at the histogram (figure 4). We see a clear peak on the "weak evidence for disease $\rightarrow$ bacteria" side, and most of the species over the line look like the spread around the peak. L. acidophilus and S. pseudopneumoniae appear to be outliers, and therefore more likely to be real causal effects.

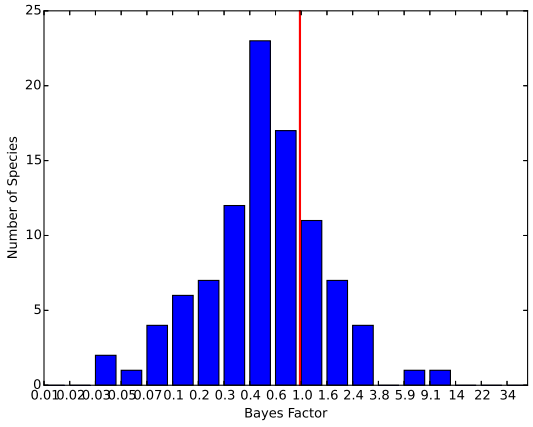

Figure 4: Number of correlating species with each bayes factor (higher means cause, lower means effect) as found by the gene collision test.

6.3 Two Bacterium Test

Looking for two-bacterium colliders, we find 8 pairs with bayes factors above 4, containing 13 species (3 of them twice). This is shown in table 2. To a degree, the number of pairs we did *not* find is disturbing. The two strongest species from the previous test show signs of colliding at a bayes factor of only 1.03.

The simplest explanation is that there are hidden environmental factors effecting multiple species, that is $H \rightarrow B_1, B_2 \rightarrow D$. Considering the wide variety of things that can be "environmental factors" and the chemical commonalities of all life, we can safely assume such factors exist. Usually these confounders will cause us to miss colliders, but it is always possible for two species that should correlate for two reasons to have those effects cancel.

The false negatives also prevent us from using a histogram approach to compensating for multiple hypotheses as before, leaving that problem unsolved.

These issues, combined with the unimpressive bayes factors, mean that this technique has not yielded useful results in this case.

6.4 The Common Cause and Severing Test

While we tried to apply the applied severing test, we were unable to produce trustworthy data. In particular, finding commonly-caused trios required too much threshold-selection that we were unable to meaningfully validate.

7 Conclusions

The techniques shown here work, both in that they give accurate results and in that they can be applied to real world data.

Dataset size is important. The 58 patients used here are pushing the limits. Larger studies will fare better.

Unobserved confounders and multiply-connected graphs cannot be ruled out by these techniques alone. Some domain knowledge must be used, and combined with a parsimony assumption.

As for Crohn's Disease, the results here are discouraging. They suggest that simple probiotic therapies are unlikely to be effective. Still, if one wishes to attempt one, the bacteria found here are a better starting place than guesswork. And demonstrating that a thing is not effective is itself useful to the field.

References

1. Altschul, Gish, Miller, Myers & Lipman. Basic local alignment search tool. *J. Mol. Biol.* (1990).
2. Bousvaros *et al.* A randomized, double blind trial of Lactobacillus GG versus placebo in addition to standard maintenance therapy for children with Crohn's disease. *Inflamm Bowel Dis* (2005).
3. Hofer. Bacterial imbalance in Crohn's disease. *Nature Reviews Microbiology* (2014).
4. Kalisch, Mächler, Colombo, Maathuis & Bühlmann. Causal Inference Using Graphical Models with the R Package pcalg. *Journal of Statistical Software* (2012).
5. Kelleher. *Causality* `https://github.com/akelleh/causality`. 2017.
6. Li, Frank & Sartor. Effect of Crohn's Disease Risk Alleles on Enteric Microbiota. *Encyclopedia of Metagenomics* (2014).
7. Murphy. *Bayes Net Toolbox* `https://github.com/bayenet/bnt`. 2007.
8. NCBI. *16SMicrobial Genetic Database* `ftp://ftp.ncbi.nlm.nih.gov/blast/db/16SMicrobial.tar.gz`. 2014.
9. Pearl. *Causality: Models, Reasoning and Inference, Chapter 2* (Springer, 2008).
10. Peters, Mooij, Janzing & Schölkopf. Causal Discovery with Continuous Additive Noise Models. *Journal of Machine Learning Research* (2015).
11. Philpott, Sorbara, Robertson, Croitoru & Girardin. NOD proteins: regulators of inflammation in health and disease. *Nature Reviews: Immunology* (2014).
12. Prantera. Probiotics for Crohn's Disease: What Have We Learned? *Gut* (2016).
13. Prantera, Scribano, Falasco, Andreoli & Luzi. Ineffectiveness of probiotics in preventing recurrenece after curative resection for Crohn's disease: a randomised controlled trial with Lactobacillus GG. *Gut* (2002).
14. Spirtes, Glymour & Scheines. *Causation, Prediction, and Search, Section 5.4* (MIT Press, 2001).
15. Walter. Ecological Role of Lactobacilli in the Gastrointestinal Tract: Implications for Fundamental and Biomedical Research. *Appl. Environ. Microbiol.* (2008).