

From correlation to causality: statistical approaches to learning regulatory relationships in large-scale biomolecular investigations

Robert O. Ness,^{†,‡} Karen Sachs,[¶] and Olga Vitek^{*,‡}

Purdue University Department of Statistics, West Lafayette, College of Science, College of Computer and Information Science, Northeastern University, Boston, and School of Medicine, Stanford University, Palo Alto

E-mail: o.vitek@neu.edu

Abstract

Causal inference – the task of uncovering regulatory relationships between components of biomolecular pathways and networks - is a primary goal of many high throughput investigations. Statistical associations between quantitative measurements can suggest an enticing number of hypotheses regarding the underlying causal interactions. But when do such associations reflect the underlying causal biomolecular mechanisms? The goal of this perspective is to provide suggestions for causal inference in large scale experiments, which utilize high throughput technologies such as mass spectrometry-based proteomics. We describe in non-technical terms the pitfalls of inference in large datasets, and suggest methods to overcome these pitfalls and reliably find regulatory associations.

*To whom correspondence should be addressed

[†]Purdue University

[‡]Northeastern University

[¶]Stanford University

Introduction

Modern high-throughput technologies such as mass spectrometry simultaneously quantify hundreds or even thousands of biomolecular analytes. Statistical associations (e.g. Pearson correlation, Spearman correlation, and mutual information) between the analytes can suggest an enticing number of hypotheses regarding the underlying causal biomolecular mechanism. However, associations do not imply causation. Formal methods of causal inference^{1,2} are required to probe these statistical associations for causality, and infer the underlying regulatory mechanisms.

There is an increasing interest in causal inference in proteome research. It has previously been used to learn the directed structure of signal transduction networks, e.g. in flow cytometry investigations in Sachs et al. 2005,³ antibody/bead-based XMAP technology from Luminex in Saez-Rodriguez et al. 2009⁴ and participants in the DREAM4 Predictive Signaling Network Challenge.⁵ However, the successful examples of causal inference in proteomics are very few.⁶ In this perspective we demonstrate that this is not an accident. The difficulty stems from the fact that large numbers of quantified analytes, combined with small sample size, lead to more spurious pairwise associations, obfuscate the true signal, and increase the false discoveries of putative causal events. Below we describe in non-technical terms the process of elucidating causal associations from high-throughput data, and suggest practical approaches for causal inference in large scale proteomic datasets. Specifically, we suggest that the task of causal inference can be facilitated by refining the biological question, and by improving experimental design in terms of selection of (1) the subset of analytes, (2) the number of biological replicates, and (3) the type of biological conditions and stresses.

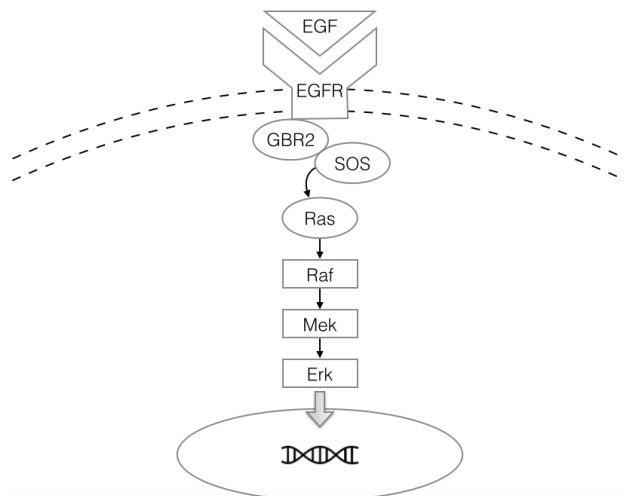


Figure 1: EGFR MAPK signaling pathway, an example of a pathway containing the phosphorylation cascade from Raf to Mek to Erk. The binding of ligand EGF to EGFR initiates a signal that leads to the cascade, which in turn regulates transcription. This cascade implies two direct causal relationships, namely $\text{Raf} \rightarrow \text{MEK}$, and $\text{Mek} \rightarrow \text{Erk}$. Raf and Erk have an indirect causal relationship, through Mek.

I. Small-scale statistical inference of causal relationships: conditional independence and interventions

Consider, e.g. the MAPK signaling cascade in Figure 1, which is part of several signaling pathways such as the EGFR MAPK pathway.⁷ In this cascade Raf causally affects the level of active (i.e., phosphorylated) Mek, while Mek causally affects Erk. Imagine these causal relationships were unknown: could they be detected from quantitative measurements on these phosphoproteins?

To illustrate the process of causal inference in this context, we simulated artificial data using the computational Huang-Ferrell model⁸ of this cascade. The model represents the key binding, phosphorylation, and dephosphorylation reactions of the cascade with mass action kinetics, and replicates the MAPK key signaling behavior observed in nature. We used the model to simulate an experiment with 50 replicate biological samples, and measurements of concentration (umol) of phosphorylated Raf, and doubly phosphorylated Mek and Erk in each sample.

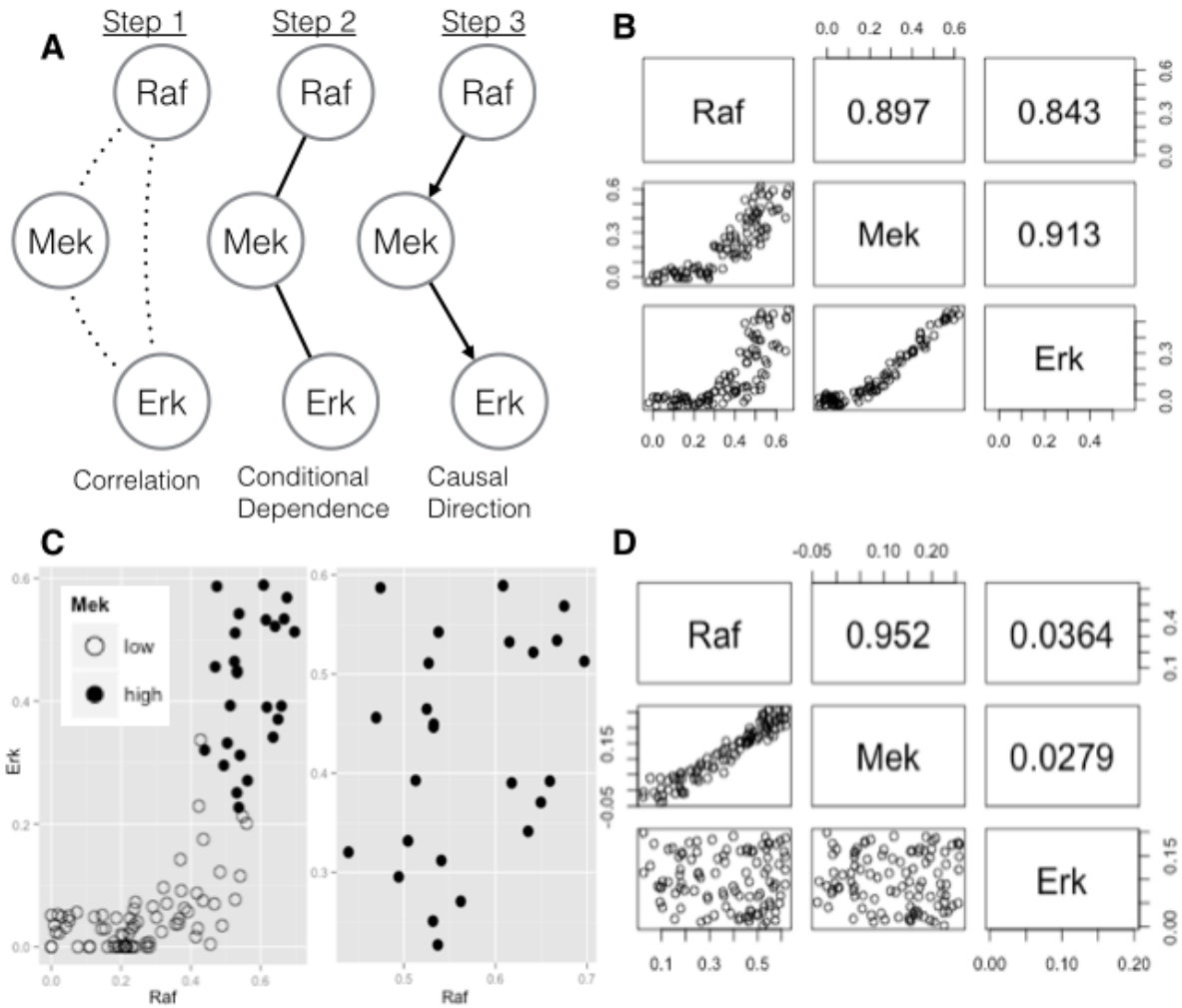


Figure 2: A: Overview of the 3 steps of causal inference, illustrated for the MAPK signaling cascade. B, C, and D feature an experiment simulated from the Huang-Ferrell computational model of the phosphorylation cascade. B: Pairwise plots of concentration values of phosphorylated (doubly phosphorylated for Mek and Erk) forms of each protein, and observed Spearman correlation values. The Raf – Erk correlation is high, despite the fact that Raf does not directly regulate Erk. C: Concentrations of Raf versus Erk, where samples corresponding to high Mek (here, set to the top quartile) are highlighted with filled circles. The right panel of C shows the subset of samples with high Mek (i.e. conditional on Mek being high). In these samples the association between Raf and Erk disappears, and we infer that Raf and Erk are *conditionally independent* given Mek. D: After Mek is inhibited, the observed association between Raf and Mek remains, while the association between Mek and Erk disappears. This reveals the causal flow from Raf to Mek.

Figure 2A demonstrates the causal inference workflow starting with analysis of statistical associations in the data. In step 1, a correlation graph between cascade components Raf, Mek, and Erk is assembled from the measurements of protein concentration. Step 2 reduces the correlation graph to a [sparse graph of inferred](#) underlying conditional dependencies (Raf–Mek, and Mek–Erk). Step 3 interrogates this graph to find putative directions of causal relationships (Raf→Mek, and Mek→Erk). While step 1 has little requirements, step 2 requires multiple samples or replicates, and step 3 requires systematic interventions (e.g. with protein inhibitors).

Figure 2B illustrates Step 1 of the causal inference, and shows 2-way plots of the protein concentrations across the biological samples, and Spearman correlations to quantify the extent of the associations. [We chose Spearman correlation instead of Pearson correlation because it best quantifies nonlinear relationships.](#) The correlation values are high, and would meet most reasonable cut-off thresholds for constructing the correlation network in the left part of panel A. The Raf–Mek and the Mek–Erk correlation edges match the Raf→Mek, Mek→Erk known causal edges. What about the non-causal Raf–Erk edge? Despite the high Raf–Erk correlation, there is no direct causal mechanism between them (aside from the one via Mek, which is already accounted for via the Raf→Mek and Mek→Erk edges). In causal inference, our goal is to eliminate this “nuisance” edge. How is this done?

[To describe Step 2 of causal inference, we introduce some terminology. If concentrations of two proteins vary between the biological samples in a coordinated manner, such that knowing the concentration of one protein provides information on the concentration of the other, the two proteins are called *dependent*. Otherwise they are called *independent*. *Conditional independence* is a special case that is important to causal inference. Two *dependent* proteins are *conditionally independent* if, after knowing \(in probability language, *conditioning on*\) the concentration of third-party biomolecules, the two proteins become independent. In other words, after considering the information from the third party, the behavior of one protein provides no additional information on the behavior of the other..](#)

While statistical associations and correlations are properties of the observed data, dependence and conditional independence are properties of the underlying processes that generate the data. Step 2 relies on statistical inference^{1,2} to infer from the observed data pairs of proteins that are conditionally independent. The conditionally independent pairs are ignored, and the remaining pairs are kept as hypothesized causal relations. This sparse inferred conditional independence graph is often desirable, as it reduces the many pairwise associations to a smaller number of hypothesized causal relations.

Let's see how this applies to the MAPK signaling cascade. Since Raf regulates the concentration of Erk by way of regulating Mek, Raf and Erk are dependent. However, if we know the concentration of Mek, then the concentration of Raf provides no additional information about the concentration of Erk. Therefore, even though Raf and Erk are dependent, they are also conditionally independent given Mek. Figure 2C illustrates the process of statistical inference by comparing the concentrations of Raf and Erk. When we subset the measurements to only the samples with high Mek, we can no longer see the association between Raf and Erk. Formally, the algorithm tests the null hypothesis of conditional independence between Raf and Erk given the full range of values of Mek (and not just high values of Mek, as was shown for the purposes of illustration), and evaluates evidence against the null hypothesis.⁹ In this example the test did not reject the null hypothesis, and resulted in removing the edge between Raf and Erk as in the middle graph of Figure 1A.

At Step 2 the direction of the regulation remains unknown. Inference of the direction of the chain of events requires the experimental design, which involves external interventions or stresses. Figure 2D illustrates the results of Step 3, in the case where an intervention targeted Mek with an inhibitor. The intervention does not affect the concentration of Mek, however it blocks its ability to phosphorylate other proteins. After this intervention the Raf–Mek relationship is unchanged, while Erk drops to a low level. From this we can infer that Mek has causal influence on Erk. Since Raf was unaffected by the intervention, Mek does not have a causal influence on Raf, and therefore the direction of causal influence in

this edge goes from Raf to Mek. This intervention is required to transform the undirected graph in panel A - Step 2 to the causal graph in panel A - Step 3.

In the general case, computational methods for causal inference follow the workflow in Figure 2A, while scaling it to characterize multiple inter-related proteins. Step 1 creates a dense network of pairwise associations. Step 2 identifies cases of conditional independence, and removes the corresponding edges to create a much sparser network. Finally, Step 3 uses the experimental design, specifically the information regarding the interventions, to evaluate these edges as evidence for potential causal events. See Koller and Friedman¹⁰ for a detailed description of these methods and their theoretical underpinnings. Numerous implementations are available in statistical software, e.g. in the R package `bnlearn`.¹¹ Depending on the biological system and on the experimental setting, the strength of causal evidence may vary. For example, Sachs, Itani *et al.* 2013³ highlight conditions in phospho-proteomic experiments where it may be infeasible to disentangle causality using perturbations.

II. Large-scale statistical inference of causal relationships: challenges of scaling up

Experimentalists can choose to quantify more or less analytes, and collect more or less replicates such as proteins. A typical high-throughput experiment includes [Removed: a small number of interventions,] a small number of biological replicates, and quantifies a large number of analytes [Removed: such as proteins]. [Removed: This creates challenges in each of the 3 steps of causal inference above.] This experimental approach is an obstacle to causal inference.

To understand why, recall that we are interested in paring down the correlation graph in step 1 to the sparse conditional independence graph in step 2. In Step 1, the challenge is in quantifying statistical associations between each pair of the analytes across the biological samples. A large number of analytes yields a large number of spurious statistical associations,

which arise without any biological justification, and are purely an artifact of random chance. Systematic pairwise relationships such as between Raf, Mek and Erk in the MAPK pathway will be obscured by the many spurious relationships that they will each form with causally unrelated proteins.

We illustrate this problem with a computer simulation, inspired by Fan et. al¹² but translated to our context. First, we simulated an experiment that quantifies the abundances of 60 proteins in 60 biological samples. Second, we simulated another experiment where the number of proteins was increased from 60 to 6000. In both experiments the proteins are completely independent from each other, and each protein in each replicate is assigned a value randomly drawn from a Gaussian distribution. In other words, we do not expect any biologically meaningful associations in these data. We repeated each of these simulations 500 times. Figure 3 shows for each experiment the histogram of the highest Pearson correlation across any pair of proteins in the 500 instances of the simulation. As can be seen, the increase to 6000 proteins produces relatively large maximum pairwise correlations, demonstrating that an increase in the number of proteins leads to an increase in spurious correlation. This is clearly a problem when high Pearson correlation is used as an initial evidence of biological function.

Similarly, the increased incidence of spurious correlations impedes the performance of statistical methods in Step 2, which elucidate conditional independences in the data. If we think of this step as [attempting to detect which pairs are not conditionally independent](#), then the spurious correlations result in more false positives [detections](#). To illustrate, [we performed a second simulation](#), this time we included some dependence between proteins so we could evaluate the effect of dimensionality of the data on true positives (pairs of proteins that are not conditionally independent) as well as false positives.

[To build dependence into the simulated data, we created multiple sets of 3-proteins "triplets" A, B, and C, where all three are dependent but A and B are conditionally independent given C. Thus perform perfectly on the \$i^{th}\$ triplet, the algorithm would need two](#)

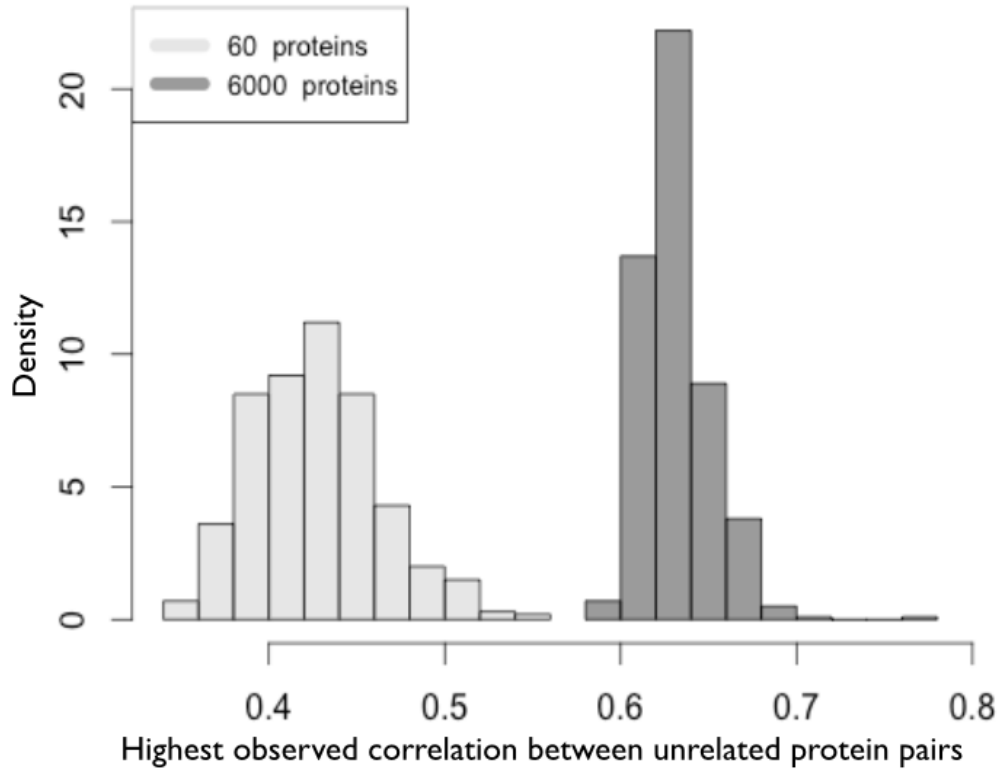


Figure 3: Histogram of maximum observed correlation values between unrelated protein pairs in a simulation of 500 experiments. In each iteration, first 60 proteins are quantified, after which throughput is increased to 6000 proteins. Increasing the number of proteins increases spurious correlation. This is a problem because correlation is used as initial evidence for a biological relationship.

true positives (rejecting conditional independence between $A_i - C_i$ and $B_i - C_i$), find the A_i and B_i pair conditionally independent, and find any pair containing a node in the i^{th} triplet and any node not in the i^{th} triplet to be independent. We made it such that the proteins in each triplet generate measurements with Gaussian values with mean 0 and variance 1, and that the covariance between A-C and B-C was .3. We simulated data from a multivariate Gaussian.

Again we run the simulation 500 times, in each iteration we performed step 1 by calculating pairwise Pearson correlations, and then Step 2 by applying the Grow-Shrink algorithm¹³ for causal inference, an instance from a class of algorithms called constraint-based network inference algorithms⁹ that proceed with a series of conditional independence tests that control the false positive rate. To assure we weren't observing algorithm specific effects, we also repeated the simulation other algorithms from this class^{14,15}. The results did not differ (results not shown).

These algorithms have stricter data requirements than correlation calculations, so in this simulation we increase from 60 proteins to only 99. The starting 60 proteins contains 20 triplets, and thus perfect performance of the algorithm would yield 40 true positives (and of course 0 false positives). In a high-throughput experiment, the investigator may arbitrarily select proteins for quantification, hoping the more proteins are quantified, the higher the chance of uncovering biology. Alternatively, he or she may apply their prior knowledge about the system to select proteins that are strong candidates for being involved in the biology. We respectively term these approaches "uninformed addition" and "informed addition" of proteins. We simulate "uninformed addition" but adding 39 proteins to the original 60 that are completely independent. In this case perfect performance is still 40 true positives. In the second "informed addition" case, we add 13 new triplets, increasing the perfect performance standard from 40 to 66. In addition to comparing these two strategies, we compare the 60 replicate case to the 100 replicate case, to show what role sample size has in affecting performance.

Figure 4 shows the counts of false positive edges in the resulting conditional independence graph, i.e. the number of protein pairs where the null hypothesis of conditional independence was unduly rejected. The results demonstrate that an increase in the number of proteins leads to an increase in false positive rejections of conditional independence, and to a denser conditional independence graph. This occurs because the number of false positives depends on the total number of pairwise comparison – a 65% increase from 60 to 99 proteins results in a 175% increase in the number of comparisons. In other words, the computational methods for causal inference are more likely to fail in a large-scale experiment, because they cannot reliably distinguish the potential causal relationships from noise.

Comparing the panels figure 4 reveals a noteworthy result; informed addition of proteins offers only a slight reduction in false positives on average. Increasing sample size doesn't help much either, paradoxically in this case it hurt, because even at 100 replicates there are still more pairwise comparisons than replicates, so the algorithm uses every last degree of freedom at its disposal. Increasing the number of replicates from 60 to 100 therefore allowed more conditional independence tests to take place, and some of these additional tests added to the false positive count.

So what exactly is the use of having more replicates? How does an informed selection of proteins for the panel make a difference relative to random selection? These factors come to play in detection of true positives. Figure 5 is the true positive counterpart to figure4. Parts A and B illustrate the 40 true positives upper bound on performance with the dashed line, and in parts C and D, where "informed addition" of proteins is applied, the increase upper bound of 66 is pictured with the double-dashed line. Comparing parts A and B show that increasing the number of replicates increases the amount of true positives. Parts C and D show that when there are more total relationships to detect, having more replicates is essential – when more relationships were added without increasing replicates, the count of true positives actually decreased slightly on average as the algorithm tries to do more with little. However, when comparing parts C and D in figure 5 to C and D in figure 4, we see

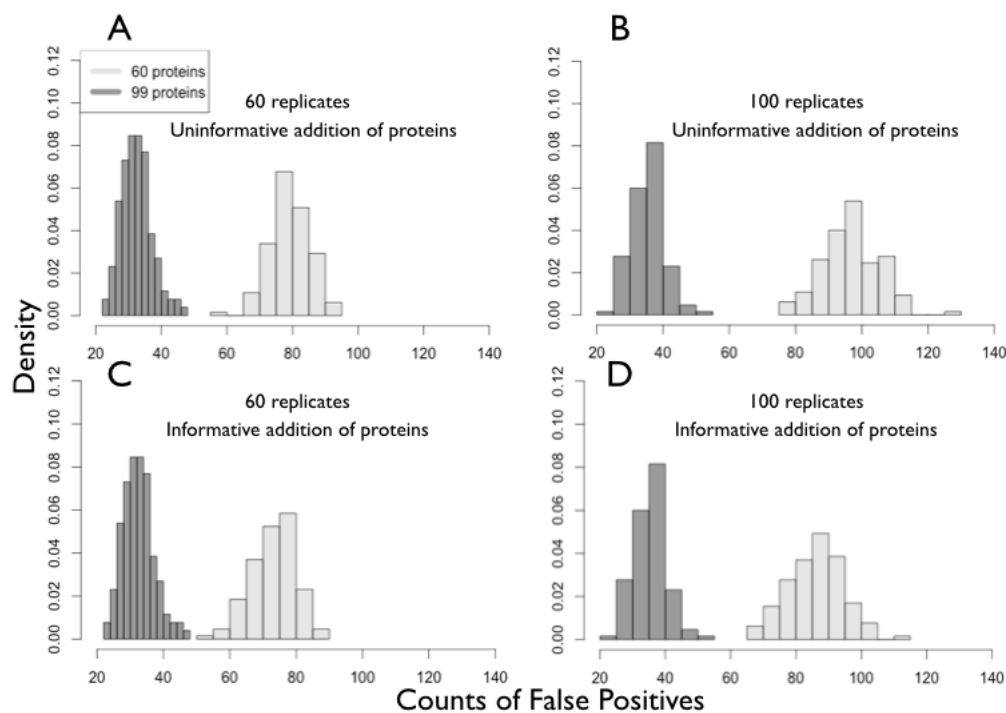


Figure 4: Histograms of false positives detection of relationships between proteins (rejections of tests for conditional independence) over 500 simulated experiment. Simulated experiments start with quantifying 60 and then expand to 99. The initial 60 proteins contains 40 pairwise relationships that the algorithm is based with detection. In parts A and B, 39 proteins are added that have no relationship with any other protein. In parts C and D, 39 proteins are added that have 26 additional pairwise relationships. The left column parts A and C show results with 60 replicates, the right column B and D for 100 replicates. Increasing the number of proteins results in more false positives without any biological justification. Informed addition of proteins mitigates this only slightly. Adding replicates paradoxically increases false positives. In both cases there are less replicates than pairwise comparisons, so the extra replicates add degrees of freedom which enables more tests, some of which add to the total false positives.

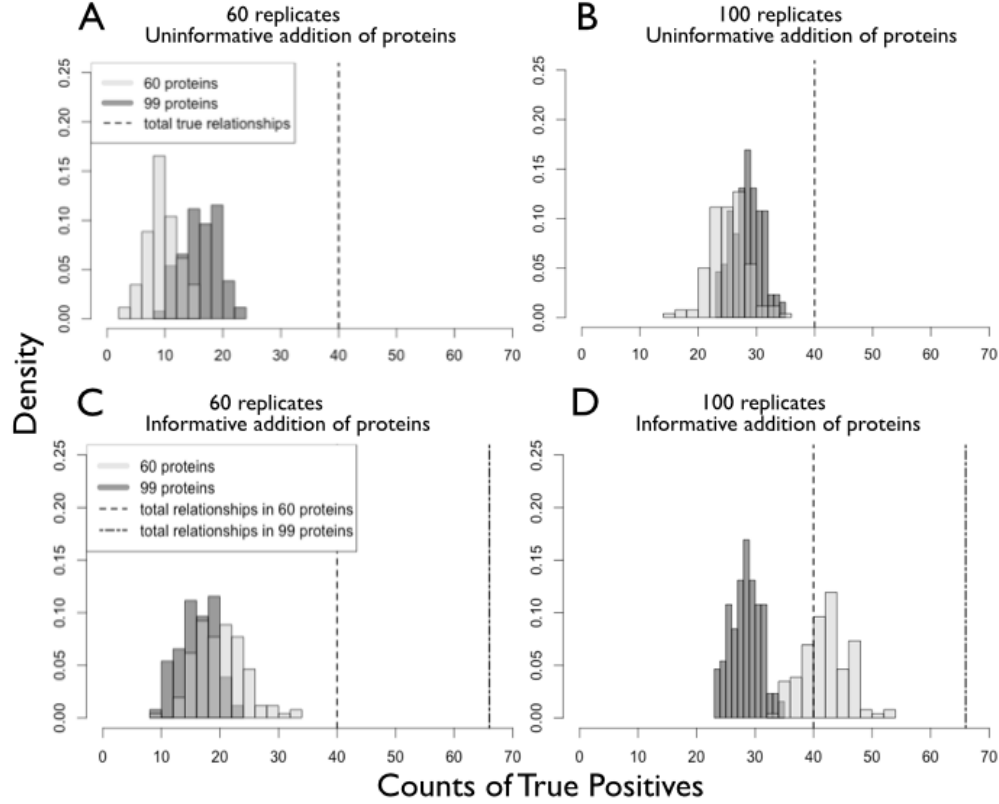


Figure 5: Simulated experiments quantifying 60 and 99 unrelated proteins in 60 and 100 biological samples. The histograms display the counts of true positives in conditional independence tests between pairs of proteins, calculated over 500 repetitions of the simulation. The top row parts A and B illustrate uninformed addition of 39 proteins to the starting amount of 60, the bottom parts C and D illustrate informed addition. The left column parts A and C show results with 60 replicates, the right column B and D for 100 replicates. Dashed lines indicate the total amount of positives the algorithm is tasked with discovering. Increasing the number of replicates increases true positives and is indeed essential when increasing the number of proteins increases the number of relationships to be detected.

that the scale of the increase in total true positives gained from informed addition of proteins was less than that of the increase of false positives. The spurious correlation and resulting false positives in high throughput experiments quickly outweigh the benefit that informed selection of proteins and greater sample size has on discovery.

[Removed: A related problem with high-throughput experiments in Step 2 is a relatively small number of biological replicates as compared to the number of analytes. Once again the purpose of step 2 is to test conditional independence, and detect relationships between proteins that remain even when conditioning on other proteins. While measuring too many analytes causes an explosion in false positives, collecting too few samples results in insufficient power to detect these relationships. While Step 1 (evaluating the set of pairwise associations) can be carried out with even a small amount of replicates, Step 2 relies on having many replicates relative to the number analytes. Note that even when the experiments demonstrated in Figure 3 had 100 replicates to its 100 proteins, and still false positives were high. The quality of causal inference approaches decreases drastically as the numbers of replicates fall.]

Finally, the challenge of completing step 3 in the workflow also increases with the number of proteins quantified. Step 3 requires interventions to infer causal relationships from evidence of conditional independence. For a panel of k analytes, theoretical work by Eberhardt and Scheines *et al.*¹⁶ show that in least ideal case of the only interventions available being interventions that target a single analyte (such as with small molecule inhibitors), $k - 1$ separate interventions would be needed. With application of interventions that simultaneously target multiple analytes, this number can be lowered significantly, but this depends on several different factors^{16–18} beyond the scope of this discussion. [Removed: For example, in the Raf-Mek-Erk pathway one intervention on Mek was sufficient to infer the two causal relationships. However, scaling this approach to a large number of proteins also introduces.] But generally speaking as the number of proteins grows, so does the number of interventions required to fully infer causality, and eventually the total number of required interventions

may become infeasible. Moreover, if the step 2 results in too many false positive hypothesized causal relations, this will adversely affect the results of step 3 regardless of the size of the dataset.

III. Approaches for inferring causality from high-throughput experiments

The problems outlined in section II paint a grim picture for causal inference in large datasets. Fortunately, these can be overcome, and effective causal inference can be a reality for large scale datasets. We provide suggestions for the best practices below.

1. *Limit the number of analytes.* Reducing the number of analytes dramatically reduces the number of spurious associations and resulting false positives. Even though a list of analytes quantifiable with high-throughput technologies grows larger, only use a subset of measurements that are both biologically relevant and technologically accurate. The length of the list is not as important as the quality of measurements on the key parts of the system. If the broader biological system is well understood, it may be possible to design a targeted experiment that focuses on a specific network or pathway, and ask more specific questions of the data, such as the presence of a particular regulatory event. The more specific the question, the less data are needed to make solid causal conclusions.
2. *Profile more biological replicates.* While reducing the number of analytes reduces spurious associations (false positives), increasing the number of replicates increases statistical power (true positives). The high-throughput measurements should provide more samples from distinct biological sources, which come from a same underlying population. This fact gives advantage to technologies, such as targeted proteomics tools, that quantify fewer analytes but have a higher sample throughput. A review by Twerve

and Saez-Rodriguez provides a contrast of technologies by their content and sampling throughput.¹⁹

Single cell mass cytometry is a technology deserving of special mention, where many thousands of cells per sample provide much more data as an input to causal inference, though cell samples still need to be drawn from multiple individuals in order to make inferences about the population from which those individuals were drawn. See Sachs et al 2005 for an in depth case study of network inference with a single cell dataset.²⁰

3. *Use prior knowledge.* Prior knowledge improves the search for conditional independence and helps to determine causality. The prior knowledge can be in form of known canonical networks, extracted, e.g. from pathway databases such as KEGG. One example of such prior information is the MAPK pathway. The prior information reduces the search space of unknown associations that need to be considered, enables a more effective use of the data, and increases the confidence in newly discovered statistical associations. Work by Saez-Rodriguez, Lauffenburger, and Sorger,²¹ as well as Terfve and Saez-Rodriguez²² use prior network knowledge to build logic models that reflect causal relationships between signaling proteins using proteomics measurements, the former using antibody/bead-based XMAP technology from Luminex, and the latter using LC-MS/MS. Another example of prior knowledge is contextual information, such as spatial or temporal annotations of the quantitative measurements in the cell. The contextual information can be extracted from the literature or from other complementary (and potentially noisy) datasets. The causal inference algorithms can be extended to weigh evidence of conditional independence, depending on whether the analytes share the same spatial or temporal context.
4. *Select targeted interventions wisely.* Targeted interventions perturb individual components of the biological system. An example is siRNA knockdowns, as well as small molecule inhibitors, which block the causal influence of a specific protein on its down-

stream components. Although effective, such targeted interventions are limited in number. Therefore, a strategic experimental design would use prior information, prioritize the interventions and the targets, and apply them to parts of the biological system that have most potential for new discovery of regulatory events. For example, a graph with undirected edges can be inspected, to reveal which nodes have potential to reveal the most causality if perturbed. Such targeted perturbations can be applied iteratively, after an initial statistical analysis revealed areas of the network where causal inference would benefit from extra measurements and data.

5. *Consider broad-scale interventions.* Broad-scale interventions sacrifice specificity of targeted interventions to simultaneously perturb many proteins in a biological system. One example of broad-scale interventions is varying experimental conditions, in order to activate multiple pathways. Signals from endocrine, paracrine, and autocrine ligands elicit various signaling responses in hepatocytes, thus interventions that cover this range of signals gives the best picture of the broader causal network of hepatocyte signaling.²³ Similarly, interventions that go beyond receptor-level and perturb multiple components of the system bring cascading causal direct orientation deeper into the network. Although they do not provide specific information about the downstream effects of stimulation, broad-scale interventions can provide more causal insight. Therefore, the advantage of this approach is that it may enable elucidation of causality across the entire system.

This list suggests impactful approaches that can drastically improve causal inference from high-throughput experiments, by constraining the inference task, and thus allowing for accurate statistical inferences. For instance, the task of assessing which of all the possible KEGG pathways is present in a dataset will be far less error-prone than the task of assessing which of all possible combinations of my measured proteins might form a biological pathway.

How should the tools listed be used? They are most powerful when used in combination, and in fact the lines between them are somewhat arbitrary and frequently blurred. For in-

stance, using item #1 and item #2 in concert can be thought of as reducing the breadth and increasing the depth of the investigation. Items #4 and #5 call for use of interventions, but this task itself is complicated by measuring many analytes. Item #3, prior biological knowledge, can be used to prioritize what to target with that limited set of interventions. Causal inference becomes possible when using these tools in combination with a sound experimental design.

Acknowledgement

We acknowledge the participants of Dagstuhl seminar 15351 "Computational Mass Spectrometry" (December 2015) (<http://www.dagstuhl.de/de/programm/kalender/semhp/?seminr=15351>) for their contributions to the discussion on computational manuscripts.

References

- (1) Pearl, J. *Causality*; Cambridge university press, 2009.
- (2) Markowetz, F.; Spang, R. *BMC bioinformatics* **2007**, *8*, S5.
- (3) Sachs, K.; Itani, S.; Fitzgerald, J.; Schoeberl, B.; Nolan, G.; Tomlin, C. *Interface focus* **2013**, *3*, 20130019.
- (4) Saez-Rodriguez, J.; Alexopoulos, L. G.; Epperlein, J.; Samaga, R.; Lauffenburger, D. A.; Klamt, S.; Sorger, P. K. *Molecular systems biology* **2009**, *5*, 331.
- (5) Prill, R. J.; Saez-Rodriguez, J.; Alexopoulos, L. G.; Sorger, P. K.; Stolovitzky, G. *Science signaling* **2011**, *4*, mr7.
- (6) Bensimon, A.; Heck, A. J.; Aebersold, R. *Annual review of biochemistry* **2012**, *81*, 379–405.
- (7) Holbro, T.; Hynes, N. E. *Annu. Rev. Pharmacol. Toxicol.* **2004**, *44*, 195–217.

- (8) Huang, C.-Y.; Ferrell, J. E. *Proceedings of the National Academy of Sciences* **1996**, *93*, 10078–10083.
- (9) Spirtes, P.; Glymour, C. N.; Scheines, R. *Causation, prediction, and search*; MIT press, 2000; Vol. 81.
- (10) Koller, D.; Friedman, N. *Probabilistic graphical models: principles and techniques*; MIT press, 2009.
- (11) Scutari, M. *arXiv preprint arXiv:0908.3817* **2009**,
- (12) Fan, J.; Han, F.; Liu, H. *National science review* **2014**, *1*, 293–314.
- (13) Margaritis, D. Learning Bayesian network model structure from data. Ph.D. thesis, US Army, 2003.
- (14) Tsamardinos, I.; Aliferis, C. F.; Statnikov, A. R.; Statnikov, E. Algorithms for Large Scale Markov Blanket Discovery. FLAIRS Conference. 2003.
- (15) Yaramakala, S.; Margaritis, D. Speculative Markov blanket discovery for optimal feature selection. Data mining, fifth IEEE international conference on. 2005; pp 4–pp.
- (16) Eberhardt, F.; Glymour, C.; Scheines, R. *arXiv preprint arXiv:1207.1389* **2012**,
- (17) Hauser, A.; Bühlmann, P. *arXiv preprint arXiv:1205.4174* **2012**,
- (18) Hyttinen, A.; Eberhardt, F.; Hoyer, P. O. *The Journal of Machine Learning Research* **2013**, *14*, 3041–3071.
- (19) Saez-Rodriguez, J.; Terfve, C. D. A. In *Modeling Signaling Networks Using High-throughput Phospho-proteomics* ; Goryanin, I. I., Goryachev, A. B., Eds.; Advances in Experimental Medicine and Biology; Springer New York: New York, NY, 2012; Vol. 736.

- (20) Sachs, K.; Perez, O.; Pe'er, D.; Lauffenburger, D. A.; Nolan, G. P. *Science* **2005**, *308*, 523–529.
- (21) Saez-Rodriguez, J.; Alexopoulos, L. G.; Epperlein, J.; Samaga, R.; Lauffenburger, D. A.; Klamt, S.; Sorger, P. K. *Molecular Systems Biology* **2009**, *5*, 331.
- (22) Terfve, C. D. A.; Wilkes, E. H.; Casado, P.; Cutillas, P. R.; Saez-Rodriguez, J. *Nature Communications* **2015**, *6*, 1–11.
- (23) Alexopoulos, L. G.; Saez-Rodriguez, J.; Cosgrove, B. D.; Lauffenburger, D. A.; Sorger, P. K. *Molecular & Cellular Proteomics* **2010**, *9*, 1849–1865.