# Bayesian network models of biological signaling pathways

by

Karen Sachs

Submitted to the Department of Biological Engineering
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Biological Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

July 2006



Author .....
Department of Biological Engineering
July 13, 2006

Certified by .....
Douglas A. Lauffenburger
Whitaker Professor of Biological Engineering, Chemical Engineering,
and Biology; Director, Biological Engineering Division
Thesis Supervisor

Accepted by .....
Alan J. Grodzinsky
Chairman, Department Committee on Graduate Students

## Thesis Committee

Approved by.....

Bruce Tidor

Professor of Biological Engineering

Professor of Computer Science

Chair of Thesis Committee

Approved by.....

Douglas A. Lauffenburger

Whitaker Professor of Biological Engineering

Professor of Chemical Engineering

Professor of Biology

Thesis Supervisor

Approved by.....

Christopher B. Burge

Professor of Biology

Thesis Committee Member

Approved by.....

Peter K. Sorger

Professor of Biology

Thesis Committee Member

Approved by.....

Steven R. Tannenbaum

Professor of Biological Engineering

Thesis Committee Member

**Bayesian network models of biological signaling pathways**

by

Karen Sachs

Submitted to the Department of Biological Engineering
on July 13, 2006, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Biological Engineering

## Abstract

Cells communicate with other cells, and process cues from their environment, via signaling pathways, in which extracellular cues trigger a cascade of information flow, causing signaling molecules to become chemically, physically or locationally modified, gain new functional capabilities, and affect subsequent molecules in the cascade, culminating in a phenotypic cellular response. Mapping the influence connections among biomolecules in a signaling cascade aids in understanding of the underlying biological process and in development of therapeutics for diseases involving aberrant pathways, such as cancer and autoimmune disease. In this thesis, we present an approach for automatically reverse-engineering the structure of a signaling pathway, from high-throughput data. We apply Bayesian network structure inference to signaling protein measurements performed in thousands of single cells, using a machine called a *flow cytometer*. Our *de novo* reconstruction of a T-cell signaling map was highly accurate, closely reproducing the known pathway structure, and accurately predicted novel pathway connections. The flow cytometry measurements include specific perturbations of signaling molecules, aiding in a causal interpretation of the Bayesian network graph structure. However, this machine can measure only ~4-12 molecules per cell, too few for effective coverage of a signaling pathway. To address this problem, we employ a number of biologically motivated assumptions to extend our technique to scale up from the number of molecules measured to larger models, using measurements of overlapping variable subsets. We demonstrate this approach by scaling up to a model of 11 variables, using ~15 overlapping 4-variable measurements.

Thesis Supervisor: Douglas A. Lauffenburger
Title: Whitaker Professor of Biological Engineering, Chemical Engineering, and Biology; Director, Biological Engineering Division

# Acknowledgments

I have often thought, that if I ever did finish my thesis, I would have *many* people to thank.

To my advisor, Doug Lauffenburger, thank you for being my mentor during my time at MIT. I am really grateful that you were willing to take a chance and give me the opportunity to follow my interests (even though it may have seemed somewhat risky). I have always appreciated your confidence in my ability to figure out this project, even (perhaps especially) when I was not so sure of this myself. I want to thank you for your persistent patience and support, which have been tremendously helpful (and inspiring), and have defined a standard for me which I will try to emulate when I supervise students. I hope that I am lucky enough in my future work to find others with your enthusiasm, brilliance, and appreciation of the intricacies of systems biology. They will have a hard act to follow.

To my thesis committee, thank you for being always focused and engaged when pondering my project. I have been fortunate to receive guidance, feedback and suggestions which have challenged me to think about my project from different perspectives, and I feel that I am a better scientist due to my interaction with my committee over the years.

To Bruce Tidor, head of my committee, thank you for being tolerant of summer-scheduling, for sticking with me till the end, and for many challenging discussions. Your perspective, as one doing related but not immediately similar work, has been very valuable. To Chris Burge, thank you for being willing to fill in at a late stage, and thank you for your thoughtful feedback and specific suggestions, particularly regarding CPD visualization. To Steve Tannenbaum, I want to thank you for making me feel so welcome when I first arrived at MIT, for being the first to make BE (then BEH!) seem like a home, and for all your help, especially while I was still looking for a lab. I have not forgotten your concern for my welfare, I have always appreciated it and have always been glad to see you, in committee meetings or just around MIT. To Peter Sorger, I have benefited tremendously from your brilliant and imaginative

insight not only in my committee meetings but also in my many interactions with you throughout the years, through BIM/CDP related meetings. Finally, to David Gifford, thank you for adopting me into your group, for always going out of your way to make me feel welcome at group events, meetings and retreats, and for many useful suggestions during my thesis work. The affiliation with you and your group has been an invaluable resource for me in all my years here.

I had the incredible fortune to obtain an additional mentor for my thesis project. I stumbled across Dana Pe'er at a conference in St. Louis, and we soon developed a wonderful collaboration. Dana and I spent many pleasant hours at Peets coffee, discussing, among other things, Bayesian networks and their applications to signaling pathways. Thank you for so many stimulating conversations, for teaching me and looking out for me, and for investing your energy, enthusiasm and brilliance into my project. Myself and my project benefited *tremendously* from your contributions.

To members of the Lauffenburger and Griffith groups, and members of CDP, the interaction with you all has helped to make MIT a place of amazing intellectual energy and excitement. It has also made doing science a fun and stimulating experience. I want to thank former Lauffenburger lab members, including Kirsty Smith, Bart Hendriks, Stas Shvartsman, Dan Kamei, Lilly Koo, Keith Duggar, Ann Dewitt and Casim Sarkar for a supportive lab environment early on, and Peter Woolf for illuminating discussions on Bayesian networks and numerous other topics. To CDP members and more recent lab members, particularly Birgit Schoeberl, Lucia Wille, Sampsa Hautaniemi, Suzzane Gaudet, John Albeck, Bree Aldridge, Kevin Janes, Melissa Kemp, Megan Palmer, John Burke and Pam Kreeger, I feel like this group has been an integral part of my thesis project. Thank you for the wonderfully open, collaborative and friendly environment, and for countless useful discussions. Finally, to JoAnn Sorrento, thank you for your friendly attitude, patience, and willingness to help with everything from room reservations to gardening.

To members of the Gifford group, past and present, thank you for many technical explanations, advice with coding, useful discussions, and help with the specifics of the likes of unix and LaTex. Special thanks to Alex Hartemink, for getting me

started with Bayesian networks via many discussions and patient explanations. To Alex Phillip Rolfe, Matt Rasmusen and Ken Takusagawa, thank you for millions of snippets of invaluable technical advice (especially Alex, who has been around more and therefore has answered more questions); to Georg Gerber for many helpful discussions and generally supportive attitude; to Tim Danford for discussions on many topics including graphical models and coding principles; to Yuan (Alan) Qi for your willingness to answer difficult Bayes nets questions; to Robin Dowell for your persistent warmth and advice on post doc issues, and to Kenzie MacIsaac for being an overall marvelous officemate. Last but not least, to Jeanne Darling, thank you for fixing a thousand problems for me over the last several years, and for being an inspiring person. To all of you, thank you for sharing your (seemingly infinite) expertise and providing a friendly environment in the group.

To other members of the community, who taught me much and enriched my experience at MIT: Tommi Jaakkola, Ernest Fraenkel, Joachim Theilhaber, James Sherley, Bevin Engelward and John Essigman. Also thanks to Teresa Wright, who introduced me to BE and helped me get started in the department, and to Rohit Singh, for lending me his expertise on several occasions.

To my fabulous friends and classmates, Hiroko Sudo, Yelena Margolin, Helen Banava, Sam Boutin, Maya Said, thank you for being friendly, fun and supportive. To the people who were so important to me in (and since) my early years at MIT: Heelo Sudo, Chrissy Guth, Nati Srebro, Thandi Muno, Maxim Kalashnikov; to the people who arrived slightly later: Alexandria Sams, Jenn Cheng, Tanyel Kiziltepe, Peter Bermel, Claire Monteleoni, Ulrike Spaete, Martin Zalesak; to the people who (unfortunately) I only met in my last year or so: Leah Henderson, Patrick van der Wel, Tammy Shoham, Solomon Itani, Hariharan Rahul, Jennifer Carlisle, Gregory Marton; and to those who I have known for ages: Masha Ishutkina, Misha Davidson, Irit Orgil, Vered Greenberg: thank you for the hikes, the talks, the coffee breaks, dinners, bike rides, hours at the gym, trips to the Walden, thanks for help with research, for help with life, for giving me perspective when I needed it and for being incredible friends. My time at MIT has been wonderful because of all of you; along

the thesis process, finding people like you is the greatest accomplishment. (I have much more to say, but I had better stop here, for fear of having a 400-page thesis)

Finally, to my family, your boundless love and support has always been my foundation in everything I do. Thank you to my parents, who solved physics problems for me and explained imaginary numbers at the first hint of curiosity. To my father, Rimmon Sachs, thank you for your infinite dedication, and for investing hours in solving math problems and deciphering buggy code whenever I needed help. To my loving mother, Ella Sachs, for being thrilled at everything I do and am. To my sister Carrie, for always reminding me that she is proud of me and always being eager for my success. To my sister Zohar, for guiding me along the grad school process and helping me, as always, in countless ways. Grad school was a million times better because you were there. To my brother Shai for your calm, supportive ways, for always thinking about others, and for helping me with coursework. To my littlest sister Noga, thank you for the very real help you provided on several occasions and for your willingness to do so, thank you for always encouraging me when I was down, and thanks for being a wonderful and supportive sister. And to my nephew Adam, thank you just for being your adorable self.

# Contents

# List of Figures

# Chapter 1

# Introduction

Survival of organisms depends on the ability of cells to communicate with their environment, often in response to a set of cues. To this end, cells have evolved signaling pathways, in which extracellular cues trigger a cascade of information flow, causing signaling molecules to become chemically, physically or locationally modified, gain new functional capabilities, and affect subsequent molecules in the cascade, culminating in a phenotypic cellular response. Mapping of signaling pathways typically has involved intuitive inferences arising from aggregation of studies of individual pathway components from diverse experimental systems. Although pathways are often conceptualized as distinct entities responding to specific triggers, it is now appreciated that inter-pathway crosstalk and other properties of networks reflect underlying complexities that cannot be explained by consideration of individual pathways or model systems in isolation. To properly understand normal cellular responses and their potential dysregulation in disease, a global, multivariate approach is required [32]. Bayesian networks [63], a form of graphical models, have been proffered as a promising framework for modeling complex systems such as cell signaling cascades as they can represent probabilistic dependence relationships among multiple interacting components [16, 17, 77]. Bayesian network models illustrate the effects of pathway components upon each other (that is, the dependence of each biomolecule in the pathway on other biomolecules) in the form of an influence diagram. These models can be automatically derived from experimental data through a statistically

founded computational procedure termed network inference. Although the relationships are statistical in nature, they can sometimes be interpreted as causal influence connections when interventional data are used, for example with the use of kinase specific inhibitors [66, 64].

There are several attractive properties of Bayesian networks for the inference of signaling pathways from biological datasets. They can represent complex stochastic nonlinear relationships among multiple interacting molecules, and their probabilistic nature can accommodate noise inherent to biologically derived data. They can describe direct molecular interactions as well as indirect influences that proceed through additional, unobserved components, a property crucial for discovering previously unknown effects and unknown components—including crosstalk between pathways. Therefore, very complex relationships that likely exist in signaling pathway architectures can be modeled and discovered. They can also incorporate prior biological knowledge when available, by assigning increased or decreased likelihoods to particular inter-molecular connections. The Bayesian network inference algorithm constructs a graph diagram in which nodes represent the measured molecules and arcs (drawn as lines between nodes) represent statistically meaningful relations and dependencies between these molecules. When inferring a Bayesian network from experimental data, the network inference algorithm aims to discern a model that is as close as possible to the observations made. The algorithm finds the most likely models by traversing the space of possibilities, via single arc changes that improve the score. There is a trade-off between simple models and those that accurately capture the empirical distribution observed in the data. The employed Bayesian scoring metric captures this trade-off, thus a high scoring model is a both simple and accurate representation of the data[65]. Bayesian networks have been applied to gene expression data for the study and discovery of genetic regulatory pathways [17, 66, 23]. However, due to the probabilistic nature of the Bayesian modeling approach, effective inference requires many observations of the system. Thus, such studies have often been limited by the insufficient size of datasets, comprising for instance measurements based on averaged samples derived from heterogeneous cell populations (a necessary limitation

when using lysates from large numbers of cells [77, 102]).

In contrast to lysate-based methods, intracellular multicolor flow cytometry [28, 68] allows more quantitative, simultaneous observation of multiple signaling molecules in many thousands of individual cells, and hence is an especially appropriate source of data for Bayesian network modeling of signaling pathways. Importantly, it allows for measurement of biological states in more native contexts. Flow cytometry can be used to quantitatively measure a given protein's expression level, and can also include measures of protein modification states such as phosphorylation [68, 59, 39]. Because each cell is treated as an independent observation, flow cytometric data provide a statistically large sample that could enable Bayesian network inference to accurately predict pathway structure.

In this dissertation, we present work in which Bayesian network structure inference is used to learn the structure of a signaling pathway, using high-throughput single cell measurements of phosphoproteins in CD4+ T-cells. In the remainder of the introduction, we discuss previous and related work (Section 1.1) and state the significance of this work (Section 1.4). We then give an abbreviated background on flow cytometry (Section 1.2) and on CD4+ T-cell signaling (Section 1.3).

## 1.1 Previous and related work

Our work on computational elucidation of signaling pathways is unique in a number of ways. However, others have used high-throughput data sources and/or computational methods to analyze signaling pathways employing a variety of methods. Although not directly focused on signaling pathways, the pioneering work of Pe'er and Friedman [17], and Hartemink *et al.* [23], were some of the first efforts to learn biological regulatory pathways directly from high throughput data. In these studies, the authors applied Bayesian network structure learning to gene expression data, in order to learn genetic regulatory pathways. These efforts suffered from two main problems: a severe data shortage (though thousands of molecules were measured, each one was only measured hundreds of times), and an insufficient amount of interventional data.

For these reasons, they were generally not able to learn specific pathway structure, but rather pathway features with respect to certain genes. Both of these groups found ways to address these problems: Hartemink *et al.*by also employing binding data, which indicates which activating gene (*transcription factor*) binds upstream of which potentially regulated gene [24]. Segal *et al.* [82] extend the earlier work by Pe'er with an elegant solution to both problems: they pre-identify and select potential regulators, reducing the need for interventional data, and they define gene modules, enabling them to pool data from multiple genes, greatly reducing the data shortage problem. In this dissertation, we find a way to solve these two problems in the related but separate domain of signaling pathways. Our work can be considered a direct extension of these earlier studies.

### Protein–protein interaction data

Most of the work aimed at elucidating signaling pathways fits the category of elucidation of protein–protein interactions and protein–protein interaction networks. A subset of such interactions constitute signaling pathways. The first large scale experimental efforts were presented in 2000 [94, 34, 33]. These employed yeast two-hybrid screens, which are designed to detect both transient and stable interactions. Assays of protein co-immunoprecipitation coupled to mass spectrometric identification of proteins have also been used; however, these focus primarily on finding components of protein complexes, as they are more suitable for detection of more stable interactions [30, 20]. Although these methods suffer from high false-positive and false-negative rates, these networks can be thought of as a noisy super-set, containing some of the interactions present in signaling pathways (other true interactions detected include, e.g. protein association for complex formation). von Mering *et al.*were among the first to address the task of increasing accuracy of these noisy datasources, proposing to use the intersection of high-throughput experiments [97]. This work resulted in a low false positive rate, but a high false negative rate, finding just 3% of known interactions. Technological improvements have helped to increase the confidence of protein–protein interaction datasets [10].

More recently, indirect biological data has been integrated with data from the high-throughput interaction screens. In these studies, weak signals from various sources are merged, aiding in detection of real interactions. Such studies use coexpression, Gene Ontology (GO) annotations, localization, transcription factor binding data, and/or sequence information, in addition to high throughput protein–protein interaction data. [37, 47, 72, 107, 110, 35, 44, 4, 103] The various information sources are used as input to a classifier, which classifies interactions as true or false. The classification methods include decision trees, naive Bayes, random forests, logistic regression, k-nearest neighbor, kernel method, and others. Aside from the classifier used, these studies differ in the dataset they use for training and testing their classifier, in the specific features (data types/sources) that they integrate, and in the encoding of the data (whether similar types of experiments are grouped together and merged or summarized, or each used as a separate experiment). Although these studies are primarily data-driven, a few groups ([84, 11, 98]) have utilized a data-independent approach, predicting interactions based on sequence motifs. More recently, Lu *et al.* [49] and Singh *et al.* [84] have introduced structure based methods, first predicting structures via homology modeling and, using these structures, predicting the likelihood of interaction based on energy considerations. In general the data-independent approaches incorporate data when available (though they have the advantage of being useable even in extremely data poor domains).

With various sophisticated computational approaches, these studies have helped greatly improve the accuracy of interaction prediction in an otherwise noisy compendium of possible interactions. Such studies are focused on general protein interactions, rather than those specific to signaling pathways. Information gleaned from these approaches may be useful to incorporate into our approach: first defining a set of protein–to–protein connection possibilities based on the interaction data supplemented by various other datsets, then increasing the likelihood of potential graph arcs in the Bayesian network according to their likelihood of interaction. Such *prior knowledge* over graph substructures is straightforward to incorporate in the Bayesian network formalism, and constitutes an interesting direction for future work.

### Cell signaling interactions

A sub-category of protein interaction prediction is that of signaling interaction prediction, a field that is more closely related to the work presented here. Programs such as Scansite use predicted or known modular signaling domains (e.g. a kinase domain) and protein sequence motifs to predict specific signaling interactions (either a specific modification, such as phosphorylation or dephosphorylation, or a binding event) [105, 58]. Modular domains are distinct and large enough to be recognized directly from protein sequences. The sequence motifs with which they interact, however, are more difficult to detect (due to their small size).

These have been identified primarily via experimental binding information from oriented peptide library screening [87, 106, 104] and phage display experiments [31], combined with information from biochemical characterization. More recently, a purely computational approach has been described for this task, in which protein–protein interaction data is used to identify a small set of binding partners for a particular domain [57]. Because the search space is greatly constrained by the binding information, the signals from the motifs' short sequences are detectable.

Like the protein–protein interaction studies, these approaches attempt to find *all potential interactions* (for the latter studies, all potential *signaling* interactions), rather than specific interactions occurring in a particular cell or cell type, in response to specific stimuli or conditions. In this way, they differ markedly from our approach, which extracts interactions in a particular dataset and can be catered to a specific biological condition, cell type, disease or other cell state. The set of influence connections in a particular dataset (from our work) and the set of potential signaling interactions (from the studies described above) can be extremely complementary datasets. Not only can we take advantage of predicted signaling interactions to aid in the selection of an optimal Bayesian network (using priors on potential edges, as described above), and to select an initial set of molecules to model (based on which molecules are thought to interact), but we can also use potential signaling interaction data to help us elucidate the signaling events that underlie newly discovered connections in

the Bayesian network. This is because many connections in the Bayes net structure are *indirect*—in other words, they do not include an enzyme and its direct substrate, but rather an upstream regulator and a molecule that is eventually affected via several other intermediaries. This will occur whenever the intermediate molecules that mediate the affect of the upstream regulator on the downstream target are not measured as part of the dataset, a common problem in this data-limited domain. (See for example the arc PKC→Jnk in Chapter 4 Figure 4-6. In this connection, the influence of the upstream regulator, PKC, on the downstream molecule, Jnk, is likely mediated by the unmeasured molecules, MAPKKK and MAPKK.)

One way to overcome this problem is to include more molecules in the dataset, a direction we are pursuing (see Chapter 5). However, in cases with insufficient prior biological knowledge, we may not have a good guess regarding which molecules may act as intermediates in an interaction (or indeed if the interaction is indirect or direct). In such cases, predicted signaling interactions that connect the upstream molecule to the downstream one can be used to complete the hypothesis of pathway structure, and to direct experiments for future Bayesian network analyses (with data which includes candidate intermediate molecules) or for direct wet lab verification. This too is an exciting direction for future extensions of these approaches.

### Analysis of signaling pathways using flow cytometry data

We now take the perspective of work relating to our datasource, and briefly discuss work in which flow cytometry data is used to analyze signaling pathways. This is a new field, involving a fairly limited number of studies, as described below. We also present a brief background on flow cytometry in Section 1.2.

The flow cytometer was invented by 1974 by Len Herzenberg and colleagues [28]. It was at the time a single-parameter machine, used to sort immune cells with a particular cell surface protein from those which lacked the protein. Immune system cells contain many distinct subpopulations, often distinguished by a particular set of cell surface markers. Once these markers are defined, it is possible to select a specific subpopulation by identifying cells with these specific markers. As the flow

cytometer's detection capabilities grew, it was used to identify increasingly specific subsets of cells. Thus, its primary purpose is for cell *identification*, via detection of cell surface markers. Today, it is possible to detect 17 distinct cell surface markers simultaneously, enabling the isolation of rare and distinct cellular populations [70].

The extension of flow cytometry to intracellular molecules required additional advances, as intracellular staining of molecules entails a number of technical challenges (see Section 1.2). This was first accomplished in the mid-1990's [14], opening new possibilities for analysis of interacellular events, rather than just cellular identification, using flow data. Since then, there have been a few dozen studies using intracellular staining of phospho-epitopes citePerez06, most of which include no more than 3 intracellular molecules. Thus, the analysis of signaling pathways through flow data is a field that is truly in its infancy. Generally, these studies simply examined the phosphoprotein histogram in order to detect a shift in the response of the signaling protein to varying experimental conditions. In those studies in which a larger number of signaling proteins were profiled, data analysis became more challenging. The flow community traditionally used *gating*, a process in which the entire cell population is partitioned, and only cells with parameters (i.e. quantitiy of proteins) between certain values are considered for further analysis. Thus, for example, a particular study may examine the level of protein **B**, in cells in which proteins **A** and **C** are both above a certain threshold. In this way, dimensionality reduction was performed manually, allowing the researcher to analyze specific differences in signaling protein distribution, in a context-dependent manner. These studies often examine the distribution of two proteins simultaneously (by gating on the remaining proteins, and then visualizing the data in a 2-dimensional plot), however, high-dimensional information could not be analyzed in an integrated fashion [41, 69]. To confront this problem, several studies attempted to examine all the different measured signaling proteins simultaneously, by clustering *samples* according to their profiled signaling molecules. [68, 39] To do this, these studies *first collapsed the distribution information into a single number*, a geometric mean, before clustering samples. Thus, though this approach did enable high-dimensional analysis, it could not also include the distribution information. Our

approach was the first to use information from each dimension while taking advantage of distribution information, by examining correlations among all dimensions in individual cells.

## 1.2 Flow cytometry

In this dissertation, we focus on learning signaling pathway structure by examining data from single cells. We acquire single cell data using a technique called flow cytometry. In flow cytometry, molecules of interest in or on cells are bound by antibodies attached to a fluorophore. Cells thus stained pass in a thin stream of fluid through the beam of a laser or series of lasers, and absorb light, causing them to fluoresce. The emitted light is detected and photomultiplier tubes convert this light to electric signals, which are recorded by the flow cytometer, providing a readout of fluorescence, and, therefore, of the abundance of the detected molecules [29].

Flow cytometry was classically used to measure cell surface markers in order to distinguish functionally distinct cellular subpopulations. More recently, methods have been developed to detect intracellular epitopes (such as e.g. signaling molecules) in order to characterize the cellular response to various conditions and ligands [14, 55]. In intracellular flow cytometry, an antibody is often raised to the phosphorylated form of the molecule, under the assumption that this is the active form (or at least the form of interest). However, it is equally appropriate to use antibodies specific to any other form of interest, such as a molecule phosphorylated on an alternate site, or a cleaved form of a molecule (such as caspase 3). To stain intacellular epitopes, it is necessary to fix and permeabilize the cells, in order to allow the antibodies to penetrate the plasma membrane and bind their targets. The procedure involves acquiring a biological sample (this can be cells from a cell line, or, as in this work, primary cells from a human or animal donor), the cells are (typically) treated with various stimuli, then fixed using a cross-linking agent (such as formaldehyde), permeabilized with a detergent or with alcohol (such as Triton or saponin, methanol or ethanol, respectively), then stained with antibodies that are each conjugated to a different fluorophore, and analyzed with the flow cytometer. The flow cytometer determines the relative abundance of each fluorophore in each cell, providing a relative measure of the signaling protein abundance. [41]

A number of issues arise when dealing with intracellular molecules, which mostly

tend to lead to false negative results with respect to the presence of a protein (or at least to decrease its apparent abundance). An antibody may not be able to bind its target antigen due to lack of antigen accessibility, if the epitope phosphosite happens to be buried in a protein–protein interaction, or if the protein exists in a cellular compartment that is not permeabilized by standard methods. Phosphoepitope stability is a concern, particularly with respect to treatment of samples prior to fixation. It is necessary to optimize protocols for different specific applications, including, importantly, careful selection of the antibodies. Antibodies that work well in Western blots may not work in the nondenatured, fixation condition of cells in a flow cytometer, and in general control experiments must be carried out to ensure that an antibody is working in a flow context.[42, 41]

Another major issue in mutlicolor flow cytometry is fluorophore selection. As discussed above, each antibody is conjugated (directly or indirectly) to a distinct fluorophore. Larger fluorophores may physically interfere with an antibody's binding characteristics or permeability, and thus can be considered another contributor to potential false negative results. In order to quantify each antigen, it is necessary to detect the emmission of each antibody separately from the others. It is also necessary to ensure that each fluorophore's absorption spectrum is included in the range of laser light used in the flow cytometer. To accomplish this, it is often necessary to incorporate multiple lasers. A 3-laser flow cytometer may be able to handle as many as twelve distinct fluorophores, hitting a different portion of the excitation spectrum of each (it is not necessary to include the highest point of the excitation spectrum). Although purchasing additional lasers can be expensive, once purchased, it is straightforward to ensure that the absorption spectrum of all the colors (up to the limit possible) is covered. Separating the signal from multiple overlapping emission spectra can be more difficult. [56]

Separation of signals from multiple dyes requires *compensation*, an adjustment of the measured fluorescence values based on the amount of spillover from one color to another (the amount of emission spectral overlap). We measure fluoresence emissions by selecting an optical filter for each color's detector, that only transmits certain

wavelengths of light, thus creating a *channel* for that color. Although a color may be primarily green (measured in the green channel), it may also have some component of its emmission spectrum that is yellow; thus, it spills over into the yellow channel. This will tend to inflate the value reported for the yellow dye. Furthermore, the yellow dye may also spill over into the green channel, causing a similar (but not equivalent in magnitude) inflation effect. For a given fluorophore, the proportion that will be emitted in each channel will always be the same for a particular instrument/ instrument setting. For example, green may emit 75% into the green channel and 25% into the yellow, while yellow emits 90% into the yellow channel and 10% into the green. Therefore, by determining this channel ratio for each fluorophore, and measuring the signal in each channel, it is possible to determine how much must be subtracted for correct compensation. In this example, the true green and yellow values are determined by solving the equations $M_G = T_G + 0.1T_Y, M_Y = T_Y + 0.25T_G$ for $T_G$ and $T_Y$, where $M_G$ and $M_T$ denote the measured values of green and yellow, respectively, and $T_G$, $T_Y$ denote the true values. In principle this could be expanded to an n-color system. In matrix notation, we solve for $T$ in the equation $M = CxT$, where $M$ is an nx1 vector of the measured values in each channel, $T$ is an nx1 vector of the true values of each color for each channel, and $C$ is an nxn matrix containing the percent spillover from each color into each other channel. $C$ is determined by measuring each color alone, and determining how much of the total signal spills over into other channels. Since $C$ is known, the system of n equations and n unknowns can be solved for $T$.

In principle, this approach can be used to include a large number of colors, but in practice, 'crowding' the colors in the emission spectrum leads to noisy data. This is because there is error in each measurement (from each channel), and, since compensation involves using measurements from multiple channels to determine a single true abundance, the noise for each color is additive noise from all the channels that contribute to it. This is particularly a problem because some dyes are far brighter than others, so some true measurements are much larger than others. (This can also happen if certain molecules are in far greater abundance or certain antibodies more

efficient) When subtracting out the overlapping effect of a bright dye out of a dim channel, even a small measurement error in the bright dye can affect the dim measurement significantly. Therefore, scaling up to increased number of colors must be done carefully. When controls indicate that indeed a bright signal is causing increased spread in the variance of a dimmer signal, it may be preferable to reduce the

number of colors in an experiment, rather than contend with noisy data. [74]

## 1.3 T-cell signaling

The primary model system used in this thesis is signaling in human CD4+ T-lymphcytes. We present here a brief overview of signaling in T-cells, with emphasis on the specific molecules profiled in this work.

CD4+ T-cells, also known as T-helper cells, are immune system cells crucial for their role in stimulating and activating other immune system cells upon encountering a foreign, non-self protein known as an *antigen* (which may be, e.g., virally derived). In general, each T-cell is specific to a different antigen, and there exists a remarkable heterogeneity in the repertoire of T-cell antigen specificity, allowing the T-cell population to respond to many (if not all) possible invasions that the host may encounter.

In order for the T-cell to begin activating other cells, it must first itself be activated, by binding to its specific antigen in the context of an *antigen presenting cell* (APC). The APC is another immune system cell, which, as its name implies, is able to present foreign antigen to the T-cell on the occasion of a viral or bacterial infection. When the APC presenting a T-cell's specific antigen binds the T-cell receptor, a series of steps occur in which signaling pathways are activated, culminating in the production and secretion of cytokines (which then stimulate and activate other immune system cells), and in the modification and proliferation of the T-cell itself. [2]

It is the T-cell receptor (TCR) on the T-cell surface, along with associated receptors such as CD28 that initially bind the antigen presenting cell. The antigen is presented as a short peptide in the context of an APC-surface protein called the major histocompatibility complex class II (MHC). Binding of the TCR to the MHC-

peptide complex leads to the activation of the Src family protein tyrosine kinase, Lck [27, 83]. Lck phosphorylates specialized motifs called immunoreceptor tyrosine-based activation motifs (ITAMs) on the receptor-associated CD3 molecules, creating docking sites for the protein tyrosine kinase Zap-70, leading to its recruitment and subsequent phosphorylation [40].

Activation of Zap-70 leads to the phosphorylation of a protein called linker for activation of T cells (LAT) [101], a scaffold protein that binds a variety of proteins with adaptor or enzymatic functions. Among these is the protein Grb2, which functions as an adaptor by binding to LAT via its SH2 domains, and to other other proteins via its SH3 domains. Grb2 mediates the translocation of the guanine nucleotide exchange factor son of sevenless (Sos), an activator of the G protein Ras [6]Egan. LAT also leads to the activation of PLC$\gamma$ which cleaves phosphatidylinositol 4,5-bisphosphate (PIP2) into inositol triphosphate (IP3) and diacylglycerol (DAG). IP3 leads to calcium release which results in several events, among them the activation of certain Protein Kinase C (PKC) isoforms. DAG leads to the activation of several molecules, including PKC$\theta$. PKC is an important molecule in T-cell signaling, with a number of downstream targets that eventually contribute to such processes as actin reorganization and cytokin production. [90] PKC$\theta$ further contributes to Ras activation. [13] Active Ras recruits Raf, a mitogen-activated protein kinase kinase kinase (MAPKKK- an activator of the activator of the mitogen-activated protein kinase), to the membrane, where it is phosphorylated and activated. Raf activates Mek (a MAPKK) and Mek activates the MAPKs Erk1/2. [21] Erk activates transcription factors that regulate cytokine production and T-cell differentiation. The MAPKs Jnk and p38 are similarly important for their modification of transcription factors during T-cell activation, particularly for cytokine production. These MAPKs also have roles in T-cell development. [73]

Binding of the CD28 receptor to its receptor on the APC (B7-1/CD80 or B7-2/CD86) results in its phosphorylation, enabling its interaction with and activation of phosphatidylinositol-3 kinase (PI3K). [75, 71, 61] PI3k phosphorylates PIP2, producing phosphatidyl inositol-3,4,5-triphosphate (PIP3), which recruits proteins to the

plasma membrane via its pleckstrin homology (PH) domains. This leads to the phosphorylation of the PH-containing serine/threonine kinase Akt (also known as protein kinase B) via the kinase phosphatidylinositol-dependent kinase-1 (PDK1). [88] Akt phosphorylates a number of substrates, influencing decisions of cell fate and cytokine production.

The common secondary messenger, cyclic AMP (cAMP), is transiently increased in response to TCR engagement. It is generally a negative regulator of T-cell activation. [85, 89] cAMP binds to and activates cAMP-dependent protein kinase (PKA). PKA modulates immune function in crucial ways; its hyperactivation is has been implicated in T-cell dysfunction in HIV, while it hypoactivity is implicated in systemic lupus erythematosus. Among its targets are the MAPKKK raf, and Src kinase Csk, a negative regulator of Lck. [1, 91]

## 1.4 Significance

In this dissertation, we describe an approach for learning molecule-to-molecule connections, and pathway structure, in biological signaling pathways. Our approach unifies a datasource which requires (and which lacked) a computational method able to extract information from a multitude of datapoints in many dimensions, with a computational technique which requires (and which also lacked, in this domain) many datapoints in order to effectively extract statistical dependencies.

Previous methods to learn molecular networks were primarily gene expression based (Section 1.1). These had key problems, such as a lack of sufficient data and measurements of a proxy to signaling events; as a consequence the resulting networks suffered from lack of accuracy. In this thesis we apply network learning techniques, specifically Bayesian networks, to single cell measurements of phosphorylated proteins, overcoming shortcomings of previous approaches. Previous approaches to analyzing flow cytometry data focused on cellular populations rather than single cells. They employed gating techniques to extract specific insights gleaned from a particular restricted dimension of the multidimensional distribution, or used the population mean, collapsing the distribution information into one metric. We take a mulit-parameter approach, revealing the correlations within single cells, and extract a landscape of protein to protein statistical dependencies from this data. Correlations within single cells are particularly useful because they provide a statistically robust sample size, avoid confounding affects of population averaging, and provide access to specific cellular subsets.

The approach we present is not dependent on the specifics of the implementation: Flow cytometry data could be replaced by automated microscopy, other computational tools could be employed (most obviously Markov random fields), and the application to learning signaling pathway structure could be modified, for example, for elucidating difference in model parameters, under different experimental conditions. Our contribution lies in the proffering of single cells as the smallest unit of 'biological computation', and in the subsequent use of information from single cells to learn

inter-molecular relationships.

# Chapter 2

# A Bayesian Networks Tutorial

In this thesis, we chose a modeling framework called Bayesian networks in order to model signaling pathways. Bayesian networks come from a family of models called graphical models, a family of flexible and interpretable models, in which probabilistic relationships among variables are represented in a graph. In our context, the Bayesian network represents relationships among variables in a signaling pathway, where the variables can represent signaling molecules, small molecules, lipids, or any biologically relevant molecule.

Bayesian networks can uncover statistical relationships among variables from a set of data. Revealed relationships are not restricted to pairwise or linear functions and, in fact, can be arbitrarily complex. Because statistical relationships may imply a physical or functional connection, we can use Bayesian networks to refine existing knowledge or uncover potential relationships in signaling pathways. When interventional data is available (i.e. data in which specific biomolecules have been manipulated or perturbed), we can begin to add a causal interpretation to our model.

In this work, to model signaling pathways, we analyze single cell data from flow cytometry, using the Bayesian network approach. The single cell data is (relatively) abundant but noisy, and the underlying biological processes are noisy as well, so a probabilistic approach is particularly suitable in this domain. The probabilistic nature of the Bayesian network enables it to extract signals from noisy data and to naturally handle uncertainty that arises in the modeling of biological processes.

The probabilistic approach determines dependencies and conditional independencies among variables; for this reason, it is able to include edges that represent meaningful relationships, but exclude those edges that are not necessary, leading to a relatively sparse representation of the underlying dependencies. Thus, given sufficient data, a Bayesian network can provide a first order map of a signaling pathway, and serve as an in silico generator of testable hypotheses.

This chapter contains a tutorial on Bayesian networks, intended to provide the reader with an understanding of these models, including what they are and how we use them in the context of this dissertation. It is primarily intended to confer an intuitive understanding. For a more technical review, we recommend [65] and [26]. We note that notation in this tutorial is borrowed from (and therefore consistent with) [65]. In this tutorial, we start with a stand-alone section that provides a quick and intuitive overview of Bayesian networks (Section 2.1). We then introduce Bayesian networks and explain what they are and what are some of their applications (Section 2.2), followed by an explanation of how Bayesian networks are used in this work (Section 2.3). Finally, we discuss model properties and causal interpretations of the models (Section 2.4).

## 2.1 Bayesian networks in a nutshell

This section is included for the reader who wants a very quick overview of Bayesian networks and their application to signaling pathways. It is meant as a stand-alone section; therefore, it is not necessary reading for those who intend to read the remainder of the tutorial. It was originally published in a slightly modified form in the Biomedical Computation Review [76]

Advances in technology have brought to molecular biology datasets that are bigger, more sophisticated, and, unfortunately, more difficult to interpret than ever before. In this thesis, we employ a computational analysis approach called Bayesian networks, a machine learning tool that is able to automatically discover networks of dependencies and causal interactions among biomolecules of interest.

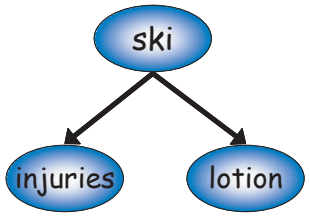



Figure 2-1: **A Bayesian network depicting statistical dependencies among variables.** The variables **injury** and **lotion** are statistically dependent upon the variable **ski**.

Bayesian networks are a form of graphical modeling, in which dependencies among variables are depicted in a graph, with the nodes representing variables (e.g., biomolecules such as proteins), and the edges representing dependencies. Dependencies are statistical in nature, so an edge from **A** to **B** indicates that knowing **A** can help us predict **B**. This may or may not indicate a causal relationship, i.e. one in which **A** (directly or indirectly) affects **B**. Interventional data, in which biomolecules are specifically manipulated, can be used to discover causal connections.

The Bayesian network inference algorithm takes data in which biomolecules were measured (and, ideally, also manipulated), and automatically reconstructs the underlying network of protein to protein influences that may have created the data. How does this process work? Consider a ski resort, with skiers and nonskiers (hot-tub sitters). A study discovers a strong statistical correlation between sunscreen lotion use and skiing injuries. To further investigate this statistical dependency, a manipulation is performed on the **lotion** variable: all sunscreen lotions are secretly replaced with an ineffective placebo. This fails to affect the number of skiing **injuries**, and so it is determined that **lotion** use does not causally affect skiing **injuries**. The variable **ski** is also well correlated. When the ski variable is manipulated (the ski slopes are closed for a day), both **lotion** use and **injuries** are greatly reduced or eliminated, thus implicating **ski** as the variable causally responsible for the other two (Figure 2-1).

The study now expands to include a tropical island. The correlation between skiing and sunscreen use is weakened; however, when tropical **sports** are included in

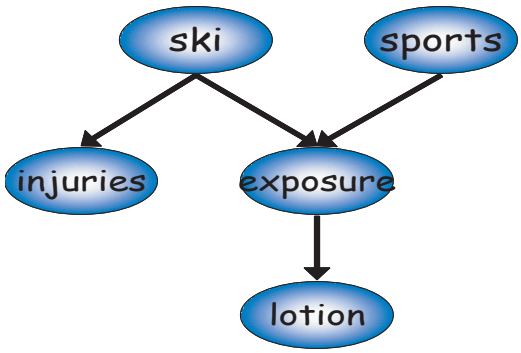



Figure 2-2: **The Bayesian network including additional variables. Ski** and **sports** together predict sun **exposure**, which is a good predictor of **lotion** use.

the study, the **ski** and **sports** variables together are able to predict sunscreen **lotion** use. If sun **exposure** is also included, it is found to be well predicted by **ski**ing and tropical **sports**, and is itself a good predictor of **lotion** use (Figure 2-2).

The Bayesian network works much like this example, examining correlations, determining which variable(s) can be used to predict which other variables, and relying heavily on interventional data to determine causal connections and the directionality of node to node connections. It is able to find complex relationships beyond simple correlations, it can handle indirect relationships (e.g **ski**→**lotion**, when exposure is not measured), and it can eliminate unnecessary edges (**ski**→**lotion** when exposure is measured). Therefore, it is potentially able to automatically construct a network much like the canonical pathways sketched out in biology text books. In this thesis, we show an application of this approach to signaling proteins measured in single cells, demonstrating the ability of Bayes nets to find a first order map of a signaling pathway, and serve as an *in silico* generator of testable influence hypotheses.

## 2.2 An introduction to Bayesian networks

This introduction is intended to serve as a basis for understanding Bayesian networks in their general context, including typical applications, in order to provide a context for their use in the analysis of signaling pathways. It is primarily intended to

provide an intuitive understanding, readers intending to pursue the use of Bayesian networks are referred to more in-depth references. [63, 26, 65] In this introduction, we assume only knowledge of basic concepts from probability, including such concepts as Bayes rule, marginalization, and conditional independencies. A reader who is not comfortable with these concepts can read about them in any basic probability text.

### 2.2.1 Model semantics

In a Bayesian network, probabilistic relationships are represented by a qualitative description- a graph ($\mathcal{G}$), and a quantitative description- an underlying joint probability distribution. In the graph, the nodes represent variables (in our case, these are biomolecules, usually signaling molecules) and the edges represent dependencies (or more precisely, the lack of edges indicate a conditional independency).[63] The graph must be a $DAG$- a directed acyclic graph. By *directed* we mean that the edges must be single-headed arrows, originating from one node (the *parent* node) and ending in another (called the *child* node). *Acyclic* indicates that the graph must not include directed cycles, so it should not be possible to follow a path from any node back to itself. (This constraint is a serious limitation in the biological domain, a point which we discuss later.)

For each variable, a conditional probability distribution ($CPD$) quantitatively describes the form and magnitude of its dependence on its parent(s). This conditional probability distribution is described by a vector of its parameters, $\theta$. The CPD must be consistent with the conditional independencies implied by $\mathcal{G}$. In general, variables in a Bayesian network may be continuous or discrete, and joint probability distributions may take on any form that specifies a valid probability distribution. However, in this dissertation, we handle only discrete variables, and use primarily multinomial distributions. When discrete variables are used, each variable may take on one of a finite set of states. (E.g. a protein variable may be in state low, medium or high.)

In general, a Bayesian network represents the joint probability distribution for a finite set $X = \{X_1, \ldots, X_n\}$ of random variables where each variable $X_i$ may take on

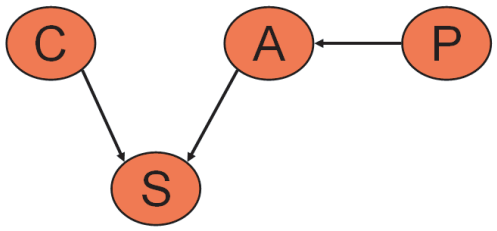



Figure 2-3: A Bayesian network structure for the variables **pollen**, **allergy**, **cold** and **sneezing**.

a value $x_i$ from the domain $ValX_i$. In our notation, we use capital letters, such as $X, Y, Z$, for variable names and lowercase letters $x, y, z$ to denote specific values taken by those variables. Sets of variables are denoted by boldface capital letters $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$, and assignments of values to the variables in these sets are denoted by boldface lowercase letters $\mathbf{x}, \mathbf{y}, \mathbf{z}$. We denote the parents of $X_i$ in $\mathcal{G}$ as $ParX_i$. $\theta$ describes the CPD, $P(X_i|ParX_i)$, for each variable $X_i$ in $X$. For illustrative purposes, consider a toy example in which a Bayesian network represents the joint probability distribution between a **cold** virus (**c**), an **allergy** to pollen (**a**), the presence of **pollen** (**p**), and the occurrence of **sneezing** (**s**). We may represent the dependencies among these variables in a graph such as (figure). In this graph, **allergy** is the child of **pollen**, **sneezing** is the child of **cold** and **allergy**, and the nodes **cold** and **pollen** have no parents (such nodes are called *root* nodes). Assume each variable can take on the value 0 ('absent') or 1 ('present').

While the graph appears to represent variable dependencies, its primary purpose is actually to encode conditional independencies, critical for their ability to confer a compact representation to a joint probability distribution. In our toy example, **sneezing** is dependent upon **cold** and **allergy**. **Sneezing** is dependent upon **pollen** as well; however, when the value of **allergy** attack is known, **sneezing** and **pollen** become independent. If we already know that an **allergy** attack is occurring (**a**=1) (or not occurring- **a**=0) knowing something about the presence of **pollen** will not help

us determine the value of **sneeze**. Therefore, **sneezing** is conditionally independent of **pollen** given **allergy**.

Formally, We say that $X$ is *conditionally independent* of $Y$ given $Z$ if

$$P(X|Y,Z) = P(X|Z)$$

and we denote this statement by $(X \perp Y \mid Z)$.

The graph $\mathcal{G}$ encodes the *Markov Assumptions*: Each variable $X_i$ is independent of its non-descendants, given its parents in $\mathcal{G}$.

$$\forall X_i (X_i \perp \mathit{NonDescendants}_{X_i} \mid \mathbf{Pa}_{X_i})$$

As a consequence of the Markov assumption, the joint probability distribution over the variables represented by the Bayesian network can be factored into a product over variables, where each term is local conditional probability distribution of that variable, conditioned on its parent variables:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|\mathbf{Pa}_{X_i}) \tag{2.1}$$

This is called the chain rule for Bayesian networks, and it follows directly from the chain rule of probabilities, which states that the joint probability of independent entities is the product of their individual probabilities.

A key advantage of the Bayesian network is its compact representation of the joint probability distribution. With no independence assumptions, the joint probability distribution over the variables pollen, allergy, cold and sneeze is $P(P,C,A,S) = P(P)P(C|P)P(A|C,P)P(S|C,P,A)$[1]. For binary variables, this is 1+2+4+8=15 parameters. Employing the conditional independencies, the joint probability distribution becomes $P(P,C,A,S) = P(P)P(C)P(A|P)P(S|A,C)$, or $1+1+2+4 = 8$ parameters. The more sparse a graph structure, the fewer edges it contains and the

[1]Or, equivalently, $P(A,C,P,S) = P(A)P(C|A)P(P|A,C)P(S|A,C,P)$, or any other order. If this representation of the joint probability distribution is not familiar, derive it by starting with the more familiar $P(A,B) = P(A|B)P(B)$, then expanding to more variables.

more conditional independencies it encodes, yielding a greater savings in parameters.

What about the joint probability distribution representation? In the case of the multinomial distributions, each CPD can be presented in a conditional probability table ($CPT$), in which each row in the table corresponds to a specific joint assignment $\mathbf{pa}_{X_i}$ to $\mathbf{Pa}_{X_i}$, and specifies the probability vector for $X_i$ conditioned on $\mathbf{pa}_{X_i}$. If we assume these variables are binary (each can be in one of two states, either 0 or 1), then in order to specify all the parameters of the CPD for variable $X_i$ with parent(s) $\mathbf{Pa}_{X_i}$, each CPT must have $2^k$ entries, where $k$ is the number of parents. In general, for a variable $X_i$ with k parents, if each variable takes on one of $s$ states, the number of entries in the CPT will be $s^k * (s-1)$.

Returning to our toy example, we will assume (for now) that the parameters of the CPTs are known:

| $P(p=0)$ | $P(p=1)$ |
|---|---|
| 0.7 | 0.3 |

| $P(c=0)$ | $P(c=1)$ |
|---|---|
| 0.9 | 0.1 |

| $p$ | $P(a=0)$ | $P(a=1)$ |
|---|---|---|
| $p=0$ | 0.999 | 0.001 |
| $p=1$ | 0.4 | 0.6 |

| $a$ | $c$ | $P(s=0)$ | $P(s=1)$ |
|---|---|---|---|
| $a=0$ | $c=0$ | 0.99 | 0.01 |
| $a=0$ | $c=1$ | 0.4 | 0.6 |
| $a=1$ | $c=0$ | 0.5 | 0.5 |
| $a=1$ | $c=1$ | 0.1 | 0.9 |

These indicate that, for instance, the probability of having a cold is 0.1 (this is the 'general' or *a priori* probability, not taking into account any information), the probability of pollen being present is 0.3, the probability of having an allergy attack when pollen is present is 0.6 (the person is only somewhat allergic), and the probability of sneezing when the person has a cold but no allergy attack is 0.6. While the graph reveals the conditional independencies, the CPTs demonstrate the strength

of dependencies. For instance, although both **allergy** and **cold** can cause **sneezing**, **cold** alone has a stronger affect than **allergy** alone ($P(s = 1|c = 1, a = 0) = 0.6$ while $P(s = 1|c = 0, a = 1) = 0.5$). Notice that, as expected, each row in the CPT sums to 1, so the second column of probabilities is redundant. If we consider only the first column, we can count exactly 8 parameters that are specified for this Bayesian network.

### 2.2.2 Inference

The most common task for which Bayesian networks are used is *inference*- reasoning from factual knowledge or evidence. This is a 'classic' use of Bayesian networks which is applicable whenever (often incomplete) information must be used to evaluate a probabilistic system (examples include assessment of the likelihood of an accident by a car insurance company or a doctor's assessment of the likelihood of a particular patient diagnosis). In an inference task, we wish to know what is the value of a particular node, but we do not have access to that information. Instead, we use the Bayesian network to get an answer in the form "variable $X_i$ is 0 with probability y". Usually we have *evidence* that we take into account. Evidence takes the form of assignments to some other variables. For example, if we wish to know if a person is sneezing, and we know that the person is having an allergy attack, we can use this information to assess $P(sneezing = 1|allergy = 1)$. We say that the variable allergy is *instantiated*, and we call this information *evidence*. When evidence is available, we may reason about a cause of the instantiated variable (this is diagnostic or 'bottom up' reasoning), or we may have evidence on the cause, and instead reason about the effect (this is causal, or 'top down' reasoning). An example of the former would be $P(p = 1|s = 1)$; $P(s = 1|c = 1)$ is an example of the latter.

To clarify these concepts, lets consider a number of examples.

Lets say we want to know $P(c = 0)$ - the probability that the person does not have a cold. With no available evidence, we need to sum the joint probability distribution over all possible values of the other variables:

$P(c = 0) = \sum_{pollen,allergy,sneeze} P(p)P(a|p)P(c)P(s|a, c)$

$$
\begin{aligned}
&= P(p=0)P(a=0|p=0)P(c=0)P(s=0|a=0,c=0)\\
&+P(p=0)P(a=0|p=0)P(c=0)P(s=1|a=0,c=0)\\
&+P(p=0)P(a=1|p=0)P(c=0)P(s=0|a=1,c=0)\\
&+P(p=0)P(a=1|p=0)P(c=0)P(s=1|a=1,c=0)\\
&+P(p=1)P(a=0|p=1)P(c=0)P(s=0|a=0,c=0)\\
&+P(p=1)P(a=0|p=1)P(c=0)P(s=1|a=0,c=0)\\
&+P(p=1)P(a=1|p=1)P(c=0)P(s=0|a=1,c=0)\\
&+P(p=1)P(a=1|p=1)P(c=0)P(s=1|a=1,c=0) =\\
&= 0.7*0.999*0.9*0.99 \ +0.7*0.999*0.9*0.01\\
&+0.7*0.001*0.9*0.5 \ +0.7*0.001*0.9*0.5\\
&+0.3*0.4*0.9*0.99 \ +0.3*0.4*0.9*0.01\\
&+0.3*0.6*0.9*0.5 \ +0.3*0.6*0.9*0.5 = 0.9
\end{aligned}
$$

This result is, as expected, the same as the *a priori* probability of c=0, since we did not take any evidence into consideration. In the case of cold, it is not a very useful calculation. But we can use the same approach to calculate an overall probability for variables that are not root as well. For example, lets say that we want to know what is the overall probability of a person sneezing, when the allergy, pollen, and cold state are not known. We can do a similar calculation:

$$
\begin{aligned}
&P(s=1) = \textstyle\sum_{pollen,allergy,cold} P(p)P(a|p)P(c)P(s|a,c)\\
&\quad = P(p=0)P(a=0|p=0)P(c=1)P(s=1|a=0,c=1)\\
&\quad +P(p=0)P(a=0|p=0)P(c=0)P(s=1|a=0,c=0)\\
&\quad +P(p=0)P(a=1|p=0)P(c=1)P(s=1|a=1,c=1)\\
&\quad +P(p=0)P(a=1|p=0)P(c=0)P(s=1|a=1,c=0)\\
&\quad +P(p=1)P(a=0|p=1)P(c=1)P(s=1|a=0,c=1)\\
&\quad +P(p=1)P(a=0|p=1)P(c=0)P(s=1|a=0,c=0)\\
&\quad +P(p=1)P(a=1|p=1)P(c=1)P(s=1|a=1,c=1)\\
&\quad +P(p=1)P(a=1|p=1)P(c=0)P(s=1|a=1,c=0) =\\
&\quad = 0.7*0.999*0.1*0.6 \ +0.7*0.999*0.9*0.01\\
&\quad +0.7*0.001*0.1*0.9 \ +0.7*0.001*0.9*0.5
\end{aligned}
$$

$+0.3 * 0.4 * 0.1 * 0.6 \; +0.3 * 0.4 * 0.9 * 0.01$

$+0.3 * 0.6 * 0.1 * 0.9 \; +0.3 * 0.6 * 0.9 * 0.5 = 0.15411$

So the overall probability of sneezing is fairly low; not surprising, since sneezing depends on two variables which occur at low probability.

Usually an inference task involves evidence. Lets consider an example in which one or more variables are instantiated. For instance, perhaps we notice that the person is sneezing ($s = 1$), and we wonder if this is due to a cold or an allergy attack. We might examine the CPTs and see that cold alone is more likely than allergy alone to cause sneezing (since $P(s = 1|c = 1) = 0.6$ while $P(s = 1|a = 1) = 0.5$). Can we guess that a cold is likely the cause of the sneezing? We cannot, because we must also take into account the different *a priori* probabilities of both cold and allergy. If we do the calculation:

$P(c = 1|s = 1) = \sum_{pollen,allergy} P(p)P(a|p)P(c)P(s|a, c)/P(s = 1)$

$= [P(p = 0)P(a = 1|p = 0)P(c = 1)P(s = 1|a = 1, c = 1)$

$+P(p = 0)P(a = 0|p = 0)P(c = 1)P(s = 1|a = 0, c = 1)$

$+P(p = 1)P(a = 1|p = 1)P(c = 1)P(s = 1|a = 1, c = 1)$

$+P(p = 1)P(a = 0|p = 1)P(c = 1)P(s = 1|a = 0, c = 1)] \; /P(s = 1) =$

$0.7 * 0.001 * 0.1 * 0.9 \; +0.7 * 0.999 * 0.1 * 0.6$

$+0.3 * 0.6 * 0.1 * 0.9 \; +0.3 * 0.4 * 0.1 * 0.6 \; /0.15411$

$= 0.065421/0.15411 = 0.42451$

$P(a = 1|s = 1) = \sum_{pollen,cold} P(p)P(a|p)P(c)P(s|a, c)/P(s = 1)$

$= [P(p = 0)P(a = 1|p = 0)P(c = 0)P(s = 1|a = 1, c = 0)$

$+P(p = 0)P(a = 1|p = 0)P(c = 1)P(s = 1|a = 1, c = 1)$

$+P(p = 1)P(a = 1|p = 1)P(c = 0)P(s = 1|a = 1, c = 0)$

$+P(p = 1)P(a = 1|p = 1)P(c = 1)P(s = 1|a = 1, c = 1)] \; /P(s = 1) =$

$0.7 * 0.001 * 0.9 * 0.5 \; +0.7 * 0.001 * 0.1 * 0.9$

$+0.3 * 0.6 * 0.9 * 0.5 \; +0.3 * 0.6 * 0.1 * 0.9 = 0.097578/0.15411 = 0.63317$

So given only the information that sneezing is occurring, the more likely cause is

actually an allergy attack[2]. Notice that when we perform inference with evidence, we must rule out all possibilities in which the evidence is refuted- in this case, we cannot take into consideration any possibilities in which s=0, since we are given that s=1. The probability distribution over the space of possibilities must sum to 1. Since we are constraining the space of possibilities, in order to maintain the constraint that the probabilities sum to 1, we must renormalize our probability distribution; we do this by dividing by the overall probability of the evidence (in this case, P(s=1), which we already calculated above). This is an application of Bayes rule: $P(a|s) = \frac{P(a,s)}{P(s)}$. It is this use of Bayes rule for inference tasks that gives Bayesian networks their name (and, in fact, Bayesian ***methods*** are not necessarily employed in the context of Bayesian networks. It is common to use Bayesian networks with non-Bayesian methods of, e.g. parameter estimation. We discuss Bayesian methods below.).

Note also that the probability of cold increased when this evidence was taken into account (we say that our *belief* in c=1 was increased), because the presence of sneezing made a cold more likely (i.e. $P(c = 1|s = 1) > P(c = 1)$). Although we have not calculated the a priori probability of allergy, it is safe to assume we would similarly find that $P(a = 1|s = 1) > P(a = 1)$.

Finally, consider the inference task in which we know the person is sneezing and we also know the person is having an allergy attack. What happens to the probability of cold? How does it compare to the probability of a cold, given only that the person is sneezing, but with no information regarding allergy? If we do the calculation:

$$
\begin{aligned}
&P(c = 1|a = 1, s = 1) = \textstyle\sum_{pollen} P(p)P(a|p)P(c)P(s|a, c)/P(s = 1, a = 1) \\
&= \textstyle\sum_{pollen} P(p)P(a|p)P(c)P(s|a, c)/ \sum_{pollen,cold} P(p)P(a|p)P(c)P(s|a, c) \\
&= [P(p = 0)P(a = 1|p = 0)P(c = 1)P(s = 1|a = 1, c = 1) \\
&+P(p = 1)P(a = 1|p = 1)P(c = 1)P(s = 1|a = 1, c = 1)] \\
&/[P(p = 0)P(a = 1|p = 0)P(c = 0)P(s = 1|a = 1, c = 0) \\
&+P(p = 0)P(a = 1|p = 0)P(c = 1)P(s = 1|a = 1, c = 1)
\end{aligned}
$$

[2]Note that $P(a = 1|s = 1)$ puts no constraint on the value of c, so the probability includes the case in which a=1 *and* c=1. Similarly, $P(c = 1|s = 1)$ includes the possibility that a=1.

$$+P(p=1)P(a=1|p=1)P(c=0)P(s=1|a=1,c=0)$$
$$+P(p=1)P(a=1|p=1)P(c=1)P(s=1|a=1,c=1)] =$$
$$[0.7*0.001*0.1*0.9 +0.3*0.6*0.1*0.9] /[0.7*0.001*0.9*0.5$$
$$+0.7*0.001*0.1*0.9 +0.3*0.6*0.9*0.5 +0.3*0.6*0.1*0.9] =$$
$$= 0.016263/0.097578 = 0.16667$$

We see that the presence of an allergy attack reduces our belief that the person has a cold. So our belief in c=1 increases when we know s=1, but then is reduced when we also discover that a=1. This is because the presence of sneezing makes us think the person may have a cold that is the cause of the sneezing. However, when we discover that the person has an allergy attack, we reason that the allergy attack may be the cause of the sneezing, and so we become less convinced that the person has a cold. This process is called *explaining away* (because the allergy attack explains away the sneezing), and it is one way in which a Bayesian network is able to reason.

### 2.2.3 Parameter estimation

So far, we have discussed Bayesian networks where both the graph structure and the CPD parameters are known. In this section, we discuss the situation in which the parameters are unknown, and must be estimated from data. Suppose we have a set of samples $\mathcal{D} = \{\mathbf{x}[1], \ldots, \mathbf{x}[M]\}$, containing M measurements of the variables in the Bayesian network . We would like to find an estimate for $\theta$, denoted $\hat{\theta}$. A standard approach to parameter estimation is called *maximum likelihood estimation*, sometimes called *MLE*. In MLE, first we define a likelihood function, which expresses the likelihood of the data as a function of the (unknown) parameter(s). Then we find the parameter value that maximizes the value of the likelihood, and we use this value as our parameter estimate. Consider an example in which we have a coin, and we wish to estimate $\theta = P(heads)$. We flip the coin three times, and get two heads and one tail. The likelihood function is $L(\theta : \mathcal{D})\alpha\theta^2(1-\theta)$. Now we want to find the estimator that will maximize this likelihood:

$$\hat{\theta} = \max_{\theta} L(\theta : \mathcal{D}) \tag{2.2}$$

For this likelihood function, we can maximize by taking the derivative and setting to zero:

$\frac{d}{d\theta}\theta^2(1-\theta) = 2\theta - 3\theta^2 = 0 \Rightarrow \hat{\theta} = 2/3$

Notice that we get a $\hat{\theta}$ that is consistent with our intuition regarding what the estimate for $P(heads)$ might be.

In our case, we define a likelihood function for the parameters of our Bayesian network and the data:

$$L(\theta : \mathcal{D}) = \prod_{m=1}^{M} P(\mathbf{x}[m] \mid \theta)$$

This optimization problem is greatly facilitated by the decomposability of the Bayesian network likelihood function, which allows us to consider each local probability function (each variable given its parents) separately. For each variable $X$ with its parents $\mathbf{Pa}_{X_i}$, there exists a parameter $\theta_{x|\mathbf{u}}$ for each combination of $x \in Val(X)$ and $\mathbf{u} \in Val(\mathbf{Pa}_{X_i})$, that we find by maximizing the likelihood function. To express the estimated parameters, we group together all the instances in which $X = x$ and $\mathbf{U} = \mathbf{u}$. Denote $M[x, \mathbf{u}]$ to be the number of instances in which $X = x$ and $\mathbf{U} = \mathbf{u}$ and $M[\mathbf{u}] = \sum_{x \in X} M[x, \mathbf{u}]$, then the paramters can be expressed as:

$$\hat{\theta}_{x|\mathbf{u}} = \max_{\theta_{X|U}} L(\theta_{X|U} : \mathcal{D}) = \frac{M[x, \mathbf{u}]}{M[\mathbf{u}]} \tag{2.3}$$

The counts $M[x, \mathbf{u}]$ and $M[\mathbf{u}]$ are called *sufficient statistics*. The sufficient statistics summarize all the information from the data that is needed in order to calculate the likelihood, so given the counts, the dataset itself is no longer needed.

**Bayesian approach**

As in the two cases shown above, the MLE consistently gives a logical estimate: if a coin yields heads 2/3 of the time, we might guess that P(heads)=2/3. However,

if we revisit the coin example, we might discover that we are not convinced by this estimate, because it was based on just three total coin flips. Such a dataset could easily bias our estimate. Until we see a much larger sample size- perhaps 100 coin flips- we might believe that the coin is probably a fair one, with P(heads)=0.5. When parameters are not known, and especially when the dataset from which we estimate the parameters is limited, we must contend with the uncertainty of our parameters. The MLE fails to take uncertainly into account, but the Bayesian approach handles it naturally. In the Bayesian approach, uncertainly is handled in two ways: First, in parameter estimation, we incorporate a prior distribution that encodes our belief in the possible values of a parameter, prior to observing the data. Second, in the inference task, we integrate over all possible values of the parameter, rather than considering one particular point estimate.

The prior, $P(\theta)$, is a distribution that encodes our belief regarding the parameter value, based on everything except the data. For example, we might base it on our general knowledge about the physical world. In the coin example, we might assume that P(heads) is most likely 0.5, that P(heads)=0.4 or 0.6 is slightly less likely, and so on, incorporating our bias that the coin is likely to be a fair one. Alternatively, we can assume that any P(heads) is equally likely- this unbiased prior is called a uniform prior, since it uses the uniform distribution. The data is then used to update the prior distribution, yielding the *posterior distribution*, $P(\theta|D)$. The update is achieved using Bayes rule:

$$P(\theta \mid D) = \frac{P(\mathcal{D} \mid \theta)P(\theta)}{P(\mathcal{D})}. \tag{2.4}$$

Note that the denominator, $P(D)$, is a normalizing factor that is independent of $\theta$. For this reason, it is sometimes disregarded.

The posterior distribution is different from the MLE in several ways. First, it is not a point estimate that assigns a particular value to $\hat{\theta}$, but instead is a distribution that assigns a probability density to each possible value of $\theta$. Second, it always takes the prior distribution into account. This tends to have a smoothing affect on the estimation. For example, if the prior assumed that any value of $\theta$ between 0 and 1

is possible, then a data sample of 2 heads, 1 tail would assign a (possibly small, but) nonzero probability to *all* values of $\theta$, including, say, $\hat{\theta} = 0.5$. How heavily the data is considered versus the prior depends on the size of the dataset, and on the 'weight' of the prior, specified within the prior itself. If the dataset is very small and the prior is heavily weighted, the prior will influence the posterior substantially, however, as the size of the dataset grows relative to the weight of the prior, the affect of the prior fades. [3]

For our coin example, we might choose to express the prior distribution as a beta distribution:

$$P(\theta) = \frac{1}{c}\theta^{\alpha}(1-\theta)^{\beta} \tag{2.5}$$

where c is the normalizing factor, $c = \int_0^1 \theta^{\alpha}(1-\theta)^{\beta}$ and $\alpha, \beta > -1$. $\alpha$ and $\beta$ are called the *hyperparameters* of the beta distribution.

The beta distributionconveniently allows us to express a belief that the coin is biased towards heads (simply set $\alpha > \beta$, the bigger the difference, the more heads becomes likely), or a belief that that tails is more likely ($\beta > \alpha$), or, perhaps the more realistic case, that both are equally likely ($\alpha = \beta$). The magnitude of $\alpha$ and $\beta$ allow us to express how strongly we believe in our prior [4], with a larger magnitude indicating a stronger belief in the prior, as we shall see below. If we have no prior belief about the fairness of the coin, and we wish all values of $\theta$ to be equally likely a priori, we can use a uniform prior, which is simply the beta with $\alpha = \beta = 0$.

Our goal is to find $P(\theta|\mathcal{D})$ and the beta prior gives us $P(\theta)$. According to Eq. (2.4), we still need to express $P(\mathcal{D}|\theta)$ and $P(\mathcal{D})$ in order to find the posterior. $P(\mathcal{D}|\theta)$ is the likelihood function, which we express using the binomial distribution:

$$P(\mathcal{D}|\theta) = \frac{(h+t)!}{h!t!}\theta^{h}(1-\theta)^{t} \tag{2.6}$$

[3]The prior is never completely eliminated from the posterior- so if the prior assigns a nonzero probability to $\hat{\theta} = .5$, then the posterior will always assign a nonzero probability to $\hat{\theta} = .5$, even if we observe a dataset of 100,000 heads and zero tails. However, in this case values of $\hat{\theta}$ that are $< 1$ will likely be very small.

[4]This holds true for positive $\alpha$,$\beta$.

where h=number of heads, t=number of tails,

and $P(\mathcal{D})$ is the normalizing factor:

$$P(\mathcal{D}) = \int_0^1 P(\mathcal{D}|\theta)P(\theta)d\theta = \int_0^1 \frac{(h+t)!}{h!t!}\theta^h(1-\theta)^t\frac{1}{c}\theta^\alpha(1-\theta)^\beta d\theta \tag{2.7}$$

This gives a posterior of the form:

$$P(\theta|\mathcal{D}) = \frac{\frac{(h+t)!}{h!t!}\theta^h(1-\theta)^t * \frac{1}{c}\theta^\alpha(1-\theta)^\beta}{\int_0^1 \frac{(h+t)!}{h!t!}\theta^h(1-\theta)^t * \frac{1}{c}\theta^\alpha(1-\theta)^\beta d\theta} = \frac{1}{d}\theta^{h+\alpha}(1-\theta)^{t+\beta} \tag{2.8}$$

where $d = \int_0^1 \theta^{h+\alpha}(1-\theta)^{t+\beta}d\theta$.

Notice that the prior and the posterior are of the same form. We say that the beta is the *conjugate* to the binomial distribution, since if we start with a beta prior, the posterior is also a beta distribution. [5] The form of the posterior provides an intuitive interpretation for the hyperparameters- we view them as pseudocounts, corresponding to imaginary coin flips. This explains how the hyperparamters can be varied to give more or less weight to the prior, as mentioned above. Large hyperparamters correspond to many pseudocounts, while small ones correspond to fewer pseudocounts, which have little impact on the posterior, especially if the dataset is large.

Given the posterior distribution, we can now perform inference using a fully Bayesian approach. Say we had a dataset of size M (containing M measurements of all n variables). For the sake of comparison, we will work out the non-Bayesian (frequentist) approach first. Recall that in the frequentist approach, we used MLE to get parameter estimates for our Bayesian network. For each variable, we had a set of parameter estimates, expressing the estimated probability of that variable, given its parent(s). Returning to our toy Bayesian network example, the set would include $\hat{\theta}_{c=0}$, $\hat{\theta}_{c=1}$, $\hat{\theta}_{p=0}$, $\hat{\theta}_{p=1}$, $\hat{\theta}_{a=0|p=0}$, and so on, one estimator for each parameter. In our inference task, we wish to find the probability of a future datapoint (not one from the dataset), in which $c = 1, p = 0, a = 1$ and $s = 0$. In the frequentist approach, this would be:

[5] Note that the prior and posterior are distributions over $\theta$, and so they use the continuous distribution beta, while the likelihood is a distribution over the data, in this case coin flip counts, and so it uses the discrete binomial distribution. As is demonstrated by this example, it is mathematically very convenient to use conjugate distributions.

$P(X[M+1]) = P(c=1, p=0, a=1, s=0) = \hat{\theta}_{c=1} * \hat{\theta}_{p=0} * \hat{\theta}_{a=1|p=0} * \hat{\theta}_{s=0|c=1,a=1}$

In contrast, in the Bayesian approach, we replace $\hat{\theta}$ with the distribution $P(\theta|\mathcal{D})$, then we integrate over all possible setting of the parameters:

$$P(X[M+1] \mid \mathcal{D}) = \int P(X[M+1] \mid \mathcal{D}, \theta) P(\theta \mid \mathcal{D}) d\theta \tag{2.9}$$

We expand on parameter estimation in Bayesian networks below.

**Parameter estimation in Bayesian networks**

In this thesis, we use the multinomial distribution for our Bayesian network parameters,and Dirichlet priors. Just as the multinomial is the multivariate extension of the binomial distribution, the Dirichlet is the multivariate extension of the beta. Analogously to the beta distribution, the Dirichlet is characterized by a set of hyperparamters $\alpha_{x^1|\mathbf{u}}, \ldots \alpha_{x^K|\mathbf{u}}$, one such hyperparameter corresponding to each $x^j \in Val(X)$.

The Dirichlet distribution is specified by:

$$P(\theta) = Dirichlet(\alpha_{x^1|\mathbf{u}}, \ldots \alpha_{x^K|\mathbf{u}}) \sim \prod_j \theta_{x^j|\mathbf{u}}^{\alpha_{x^j|\mathbf{u}}-1} \tag{2.10}$$

where the hyperparamters once again can be thought of as pseudocounts (as we shall see), and $\alpha_* = \sum_j \alpha_{x^j|\mathbf{u}}$ can be considered the *effective sample size* of the prior.

Aside from lending itself naturally to our tabular multinomial CPTs, the Dirichlet prior is convenient for a number of reasons. It satisfies *parameter independence*, meaning that it decomposes into a product of independent terms,[6] it provides an intuitive approach to specifying the strength of our prior, and it is the conjugate distribution to the multinomial; therefore, the posterior will also be Dirichlet:

**Proposition 2.2.1:** *If $P(\theta)$ is Dirichlet$(\alpha_{x^1|\mathbf{u}}, \ldots, \alpha_{x^K|\mathbf{u}})$, then the posterior $P(\theta \mid \mathcal{D})$ is Dirichlet$(\alpha_{x^1|\mathbf{u}} + M[x^1, \mathbf{u}], \ldots, \alpha_{x^K|\mathbf{u}} + M[x^K, \mathbf{u}])$ where $M[x, \mathbf{u}]$ is the sufficient statistics derived from $\mathcal{D}$.*

[6] In the Dirichlet prior, the parameters for each variable given its parents in $\mathcal{G}$ are independent of the parameters for each other variable, given its parents (*global parameter independence*), *and* the parameters for each variable given a parent are independent of the parameters for the same variable given other parent(s) (*local parameter independence*).

The posterior can be used in Eq. (2.9) to calculate the probability of future samples.

## 2.3 Structure learning

In the entire tutorial so far, we have considered Bayesian networks in which the structure, $\mathcal{G}$, is known. However, in this thesis, our goal is to discover $\mathcal{G}$, by searching for a structure that is consistent with the dependencies present in a dataset. Finding the graph structure that may have generated (is most likely to have generated) the observed data is called *structure learning*.

Structure learning is conceptually different from what we have focused on so far, but our methods are consistent with what we have already seen. Our strategy for structure learning is as follows: First, we define a Bayesian score which indicates, for a given graph structure, how well the structure reflects the dependencies (and conditional independencies) present in the data. This score allows us to assess individual structures. Armed with this ability, we can now search over possible model structures, until we find the best one (i.e. the one with the highest score), or more accurately, we find a set of high-scoring model structures. Finally, we take our set of high scoring models and average them, in order to avoid overfitting and remain consistently Bayesian in our approach.

### 2.3.1 Bayesian score

To formulate the Bayesian score, we start with a probability metric:

$$\text{score}_{\mathcal{B}}(\mathcal{G} \ : \ \mathcal{D}) \ \alpha \ P(\mathcal{G}|\mathcal{D}) = \frac{P(\mathcal{D} \mid \mathcal{G})P(\mathcal{G})}{P(\mathcal{D})} \tag{2.11}$$

This score expresses the posterior probability of the the graph given the data. Because the marginal likelihood, $P(\mathcal{D})$, will be the same for any structure considered,

we are able to neglect it; thus, the Bayesian score becomes:

$$\text{score}_{\mathcal{B}}(\mathcal{G} \ : \ \mathcal{D}) = P(\mathcal{D} \mid \mathcal{G})P(\mathcal{G}) \tag{2.12}$$

This is not quite accurate, because, for convenience of computation, the score is actually the log probability:

$$\text{score}_{\mathcal{B}}(\mathcal{G} \ : \ \mathcal{D}) \ = \ \log P(\mathcal{D} \mid \mathcal{G}) + \log P(\mathcal{G})$$

$P(\mathcal{G})$ is the *structural prior*, expressing the a priori probability of each graph structure. Typically, we employ a uniform prior, so we are not biased towards one structure or another. However, when we have constraints, we can incorporate them. For example, if our model is dynamic, we may have a probability of zero for any graph in which a node representing a later timepoint is influencing an earlier timepoint. Additionally, one could use the prior to incorporate biological knowledge.

As before, we have a dataset, but are uncertain about the value of our parameters. Therefore, we average over all possible parameter assignments when we calculate $P(\mathcal{D}|\mathcal{G})$:

$$P(\mathcal{D} \mid \mathcal{G}) = \int P(\mathcal{D} \mid \mathcal{G}, \theta)P(\theta \mid \mathcal{G})d\theta \tag{2.13}$$

We can see from this expression that the Bayesian score incorporates a complexity penalty, ensuring that it will prefer simpler models, unless a more complex model is supported by a sufficiently large dataset. The complexity penalty follows from the fact that the Bayesian score integrates over all possible parameters. A more complex model will have more parameters, leading to integration over a larger-dimensional space. For the parameter prior, $P(\theta|\mathcal{G})$, this means that every point in the density function is smaller (since the distribution must integrate to 1). For $P(\mathcal{D}|\mathcal{G}, \theta)$, the distribution must be strongly peaked over the true parameters to overcome the increase in parameter dimensionality. This can happen (assuming the true underlying

graph is indeed complex) when the dataset is large. Otherwise, there is not sufficient data to learn a complex distribution (even if it is the true one), and the Bayesian score will select a simpler graph structure. Because of this complexity penalty, the Bayesian approach avoids overfitting- a process by which we fit our model to noise in the data, rather than a true signal. This is a significant strength, particularly in a data-limited domain.

The integral in Eq. (2.13) can in general be a hard integral to solve; however, due in part to the decomposability of the Bayesian network and the Dirichlet prior, this integral has a closed form solution. Assuming Dirichlet priors with hyperparamters $\{\alpha_{X_i^j|\mathbf{u}}\}$, the Bayesian score can be expressed as:

$$\text{score}_{\mathcal{B}}(\mathcal{G} \ : \ \mathcal{D}) = \sum_i \log \prod_{\mathbf{u} \in Val(\mathbf{Pa}_{X_i})} \frac{\Gamma(\alpha_{x_i|\mathbf{u}})}{\Gamma(\alpha_{x_i|\mathbf{u}} + M[\mathbf{u}])} \prod_{x_i^j \in Val(X_i)} \frac{\Gamma(\alpha_{x_i^j|\mathbf{u}} + M[x_i^j, \mathbf{u}])}{\Gamma(\alpha_{x_i^j|\mathbf{u}})} \tag{2.14}$$

where $\Gamma$ is the Gamma function ($\Gamma(n) = (n-1)!$ if $n$ is a natural number) and $\alpha_{x_i|\mathbf{u}} = \sum_{j \in Val(X_i)} \alpha_{x_i^j|\mathbf{u}}$.

As before, the hyperparameters are incorporated as pseudocounts, so it is straightforward to vary the influence of the smoothing prior. Note that the score sums over each variable separately, so the contribution of each variable to the overall score can be considered individually (a feature crucial for efficiency, as we discuss below). Furthermore, each local contribution depends only on the sufficient statistics- we need only know the counts from the data and the prior hyperparamters to calculate the score.

### 2.3.2 Searching the space of possible graph structures

For our next step, we have to consider possible graph structures and assess them using the Bayesian score, until we find one (or ones) that score well. We cannot exhaustively examine every graph, because the number of possible graph structures is super-exponential in the number of variables. In fact, finding the highest scoring

graph is known to be NP-complete (therefore computationally infeasible). Instead, we employ a heuristic search to find high scoring models. There are a number of ways to formulate such an algorithm. In this thesis, we employ a greedy random search, specifically, a variation called Metropolis. We begin with a random graph. Then, at each iteration, we select at random an edge to add, delete, or reverse (while keeping within necessary constraints of the graph, such as acyclicity). We score the new graph structure, and keep the change if we find that the score improves. Otherwise, we revert to the previous structure. We then go to the next iteration, and again make a random change, repeating the process.

This procedure can get stuck in a local maximum- because we change only one edge at a time, we may encounter a situation in which any single edge change yields a lower score, but a, say, two edge change may lead to a better score. We can never get to this two edge difference, because the first edge change would be rejected. Thus, we are stuck in a score that is locally good (better than any structure that is one edge different) but not very good overall. We avoid this problem in two ways. First, we allow a 'bad' edge change (one that leads to a lower score) with some probability $p$. $p$ starts very high, then becomes smaller as the search continues. This is a form of the simulated annealing algorithm. As in simulated annealing, after the 'temperature' (actually $p$) is reduced, it is temporarily increased late in the search (reannealing). As a second measure, we also repeat the search multiple times, each time starting from a different point in the search space (a different random graph). This is called *random restart*. The random graph can be generated simply by starting with an empty graph and running the search procedure for several iterations with a very permissive temperature (i.e. very high $p$). In this thesis, we typically employ at least 100 random restarts (sometimes up to 500).

Each search iteration requires a calculation of the new model score. As we discussed before, the score requires only the sufficient statistics from the data (which are simply counts of the form, "How many times is a variable in configuration x while its parents are in configuration y"). For a large Bayesian network , compiling the sufficient statistics may be computationally burdensome. Fortunately, as we saw in

Eq. (2.14), the score decomposes, so that the score contribution of each variable can be considered in isolation. This is very useful for our search procedure, because it means that we need to gather all the sufficient statistics only *once*, (the first time we score). After that, we change the score only locally, by updating the sufficient statistics only for the one variable which has gained (or lost) a parent as a result of the new edge change. For example, if the edge $A \rightarrow B$ is added, then $B$ has gained a parent, and so will need new counts for its sufficient statistics. In the case of an edge reversal, say $A \rightarrow B$ becomes $B \rightarrow A$, then in one step both $A$ and $B$ have had a change in their parent sets. In this case, we update the sufficient statistics for both $A$ and $B$. Thus, because of the decomposability of the Bayesian score, all search iterations (except for the first) will require at most two recalculations of sufficient statistics.

### 2.3.3 Model averaging

At the end of our search procedure, we have a collection of high-scoring models. We could simply select the highest scoring model from this collection. However, because our data is noisy, we have uncertainty with respect to our dataset. Furthermore, because our dataset is limited in size (sometimes very limited), we are concerned about choosing the one highest scoring model that happens to best reflect possibly spurious signals in this dataset. In other words, we are concerned about overfitting. For this reason, rather than select the single highest scoring model, we consider the collection of all high-scoring models, and choose those features that are common to many of them. To do this, we perform *model averaging*. Notice that this is consistent with the Bayesian approach that we have seen so far: Each time we encounter uncertainty, we average over all possibilities. In this case we are uncertain that one high scoring model is necessarily better than another.

In model averaging, we consider all high scoring models, and calculate a score for each feature (each edge) based on the number of high scoring models in which it appears, as well as the scores of those models. We do this with a weighted average:

$$Score_{Edgei} = \frac{\sum_{Modelscontainingedgei} Score_{Model}}{\sum_{AllModels} Score_{Model}} \tag{2.15}$$

The score of the edge is approximately equal to the probability of the edge. In order to get the actual probability, it is necessary to sum over all possible graphs, something we cannot do due to the enormous size of the set of all graphs. However, because we sum over numerous very high-scoring models, we do capture most of the probability density in our sum, making the probability metric more or less accurate. Other variations on model averaging can be considered. We investigate some of these in secrefRobustness analysis.

## 2.4 Model properties and Causality

So far, we have considered Bayesian networksin their general context, and discussed how to learn the Bayesian network structure from data. In this section, we will delve a bit more into the model structure, to examine how different structures can be extracted from data, even when they represent similar dependencies, what kinds of dependencies are represented, and when a Bayesian network structure can be interpreted as a causal structure.

### 2.4.1 Dependencies and independencies in the graph structure

#### Explaining away in substructures

As previously stated, a Bayesian network is a representation of dependencies and conditional independencies among a set of variables. We saw before that conditional independencies are particularly useful. In general, variables may be dependent; however, they may become independent *conditionally*, when we consider another variable or set of variables. We have touched on this topic before briefly; here, we revisit it for a more thorough treatment. Returning to our toy example (and negelcting, for a moment, the *cold* variable), we have the structure $p \rightarrow a \rightarrow s$ (such a structure is

sometimes called a *chain*). In this structure, *pollen* and *allergy* are dependent, as are *allergy* and *sneeze*. *pollen* and *sneeze* are dependent as well: if we know that there is *pollen* in the air, our belief about the possibility of *sneezing* increases. However, *pollen* and *sneezing* are *independent* given *allergy*: If we know the person has no allergic reaction, then the absence or presence of *pollen* does not change our belief in the possibility of *sneezing*. Thus, *allergy* renders *pollen* and *sneeze* independent.

Consider this from the structure learning perspective. Examining the variables $a$, $p$, $s$, we may note that all the variables are correlated. We may connect $p$ to $a$ and $a$ to $s$, but why should we stop there? Afterall, we see that $p$ and $s$ are statistically dependent, so we could also connect $p$ to $s$. The correct Bayesian network structure will *not* connect $p$ to $s$, however, because the dependence between $p$ and $a$, and between $a$ and $s$, *explains away* the dependence between $p$ and $s$ (assuming the data are consistent with the dependencies we describe in our example).

How will the conditional independence implied by the ability of $a$ to explain away $s$'s dependence on $p$ manifest itself in the Bayesian score? Consider the CPTs for this structure: In the correct structure, for the variable $s$, the CPT will specify $P(s|a)$. If we also connect $p$ to $s$, the CPT must now specify $P(s|a, p)$, so it will contain twice as many parameters. Recall that the Bayesian score penalizes complexity. For the score to allow the extra parameters, the parameters must be more peaked (as they would be if *pollen* did in fact contribute additional information towards predicting *sneeze*, for example, if *pollen* also affected *sneezing* in some other way, other than by causing an allergy attack). If $p$ affects $s$ *only* via its affect on $a$, the complexity penalty will result in the more connected model scoring more poorly; thus, it will ensure that the structure is appropriately sparse.

What other structures encode conditional independencies? Consider the structure $A \leftarrow B \rightarrow C$, sometimes called a *fork*, in which $A$ and $C$ depend upon $B$. We have not seen such a structure before, so let us devise an example: say, puddles, rain and umbrellas. In the model $puddles \leftarrow rain \rightarrow umbrellas$, we see that *puddles* and *umbrellas* depend on *rain*. Because they share a parent, we expect *puddles* and *umbrellas* to also be dependent: seeing people with umbrellas might lead us to suspect

that there are puddles on the ground, and seeing puddles on the ground may increase our belief that people will be carrying umbrellas (the puddles may have an alternate cause, such as a sprinkler nearby, which does not induce people to carry umbrellas, so while our belief in umbrellas is increased, it is not necessarily certain). What happens when we *know* that it is raining? As in the chain structure, knowing the value of $B$ ($rain$)renders $A$ ($puddles$) and $C$ ($umbrellas$) independent, because if we know it is raining, then the presence of umbrellas no longer informs us as to the possibility of puddles. In other words, the dependence of each variable on $rain$ explains away their dependence on each other, rendering them conditionally independent. As before, the Bayesian score will prefer this structure to one in which an additional edge connects $puddles$ to $umbrellas$ directly.

Let us examine one final graph structure, an important one called a *v-structure*, which has this configuration: $A \rightarrow B \leftarrow C$ (with no edge between $A$ and $C$). We saw such a structure in our toy example: $cold \rightarrow sneeze \leftarrow allergy$. The v-structure is quite unique because in a v-structure, in contrast to the other structures we have seen, two otherwise *independent* variables may become *dependent*. Consider $cold$ and $allergy$. Both affect $sneeze$, but this will not confer a dependence between them. In fact, they are completely independent: the presence of a cold virus does not affect the possibility of an allergy attack and vice versa (to confirm this, return to secref(ref section here) and calculate $P(cold|allergy)$ and $P(allergy|cold)$; these will be equal to $P(cold)$ and $P(allergy)$, respectively). What happens if we know the person is $sneezing$? Suddenly, $allergy$ and $cold$ become dependent; because we know one must be the cause of the sneezing, knowing the value of one helps us to determine the value of the other. We saw an example of this in secref: If $sneeze = 1$ and $cold = 1$, our belief in $allergy = 1$ decreases; similarly, if $sneeze = 1$ and $allergy = 1$, our belief in $cold = 1$ decreases; one cause of sneezing explains away the other cause.

**D-separation**

In the three structures we examined, having information (evidence) for certain variables confers or eliminates dependencies between other variables. In a link structure, $A \rightarrow B \rightarrow C$ and in a fork, $A \leftarrow B \rightarrow C$, $B$ makes $A$ and $C$ independent;

we say that $B$ blocks the flow of information between $A$ and $C$ when it is observed. In contrast, in a v-structure, $A \rightarrow B \leftarrow C$, $B$ confers dependence upon $A$ and $C$, so $B$ blocks the flow of information between $A$ and $C$ when it is **not** observed. For two variables $X$ and $Z$ with observed evidence $E$, we say that an active trail exists between $X$ and $Z$ if information is allowed to flow between them, indicating that for every connection in the form of a v-structure, the 'middle' variable (the child in the v-structure) [7] is part of the evidence $E$, and that no other node between them is part of $E$. If no active trail exists between $X$ and $Z$ given $E$, we say that they are *d-separated* given evidence $E$. The notion of d-separation helps us determine the presence of conditional independencies in large and convoluted Bayesian networks, in which it is harder to pick out conditional independencies by inspection.

**Selection of model structures**

The task of finding the correct model structure for a set of variables may initially appear straightforward, but it can actually be quite tricky. This is because often, many dependencies will exist in the domain that are indirect, so the challenge is to find conditional independencies, rather than dependencies. Making the graph sparse can be harder than finding accurate connections. Consider a biological domain, in which three kinases affect each other: kinase $A$ activates kinase $B$, which then phosphorylates kinase $C$. From this description, the accurate underlying Bayesian network structure is $A \rightarrow B \rightarrow C$. If we attempt to learn this structure from data, we will likely see a correlation between $A$ and $B$, $B$ and $C$, and $A$ and $C$. So how can we decide how to connect them? $A \rightarrow B \rightarrow C$ is one valid structure, but what about $B \rightarrow A \rightarrow C$? What about $A \leftarrow B \rightarrow C$? What about a fully connected model (an edge between each two variables)? The answer is that sometimes we can tell which of these models are better, but sometimes we cannot (unless we have interventional data- see next section).

Consider $A \rightarrow B \rightarrow C$ (the true model) vs. $B \rightarrow A \rightarrow C$ or $A \rightarrow C \rightarrow B$. All of these represent dependencies that we do in fact see in the data. Will the Bayesian

[7] Information can flow (and the trail can remain active) also if a descendent of the child node is instantiated

score favor the true model (assuming sufficient data)? In this case, the Bayesian score *can* tell the correct model from the other two. How does this work? If we imagine a binary dataset, with 0 representing 'kinase off' and 1 representing 'kinase on', we would anticipate that if $A$ is off, $B$ will be off and $C$ will be off; and the opposite situation if $A$ is on. Therefore, we anticipate a dataset containing rows of '0 0 0' and '1 1 1'- i.e. one in which all the kinases are always either on or off. From this dataset, it is not possible to tell which model is correct- each is equally supported. Thankfully, the Bayesian network is actually *helped* by 'noise' in the system- in some cases, while $A$ has turned on $B$, $B$ has not yet turned on $C$ (perhaps due to the action of some unknown other protein). In other cases, $A$ has already been dephosphorylated, but $B$ and $C$ are still 'on'. Because of these situations, the dependence of $C$ is greater upon $B$ than upon $A$; and $B$ and $A$ are more closely linked than $A$ and $C$. Therefore, the correct model will be preferred from the three above.

What about the model $A \rightarrow B \leftarrow C$, also consistent with the closer correlations in the data? Because this model is a v-structure, the Bayesian network can take advantage of the difference in conditional independencies to tell this model apart from the true model. In this model, when $B$ is not known, $A$ and $C$ are independent (as we discussed previously). In the true model, $A \rightarrow B \rightarrow C$, $A$ and $C$ are only independent if $B$ is known. Since in the data the latter would be true, the correct model can be selected.

**Equivalence classes**

Finally, what about the models $C \rightarrow B \rightarrow A$ or $A \leftarrow B \rightarrow C$? They have the more closely correlated variables linked to each other, so we cannot use system 'noise' to tell them apart from the true model. What about conditional independencies? In these two models, $A$ and $C$ are independent given $B$- just as in the true model, so conditional independencies cannot be used to distinguish them. Therefore, a Bayesian network using only observational data *cannot* tell these apart from the true model. The Bayesian score for all three will be exactly the same. This is an example of an *equivalence class*. Models with the same set of conditional independencies form an equivalence class. Such models will always receive the same Bayesian score and,

therefore, are impossible to tell apart using observational data. The simplest such example is the class consisting of $A \rightarrow B$ and $B \rightarrow A$. Because of equivalence classes, when we perform structure learning using observational data, we search for the best equivalence class rather than the best model. Typically, this means we can find a graph with only some of the edges directed (called a PDAG, *partially directed graph*).

### 2.4.2 Interventional data

Equivalence classes are a big problem in practice, but there is a (conceptually) easy way to get around them: using interventional data. Interventional data refers to data in which we intervene experimentally in the biological system, by perturbing the values of individual molecules in defined ways. For example, we may knock out a protein (thus set it equal to zero), we may use RNAi, or, as in this thesis, we may use pharmacological activators and inhibitors. (This is in contrast to *observational* data, in which the system may be generally stimulated, but no specific variable is forced on or off) The power of interventional data is easy to understand. Assume we have two variables, $A$ and $B$. $A$ and $B$ are highly correlated, so if we attempt structure learning, we will end up with the equivalence class consisting of $A \rightarrow B$ and $B \rightarrow A$. Now assume we have interventional data in which $A$ is set to zero. In the observational data, generally $A$ and $B$ were both 0 or both 1. Now $A$ is forced to be 0, and we see that $B$ is also zero. This makes us suspect that $A$ might be affecting $B$, i.e. $A \rightarrow B$ is the correct model. If we also have interventional data on $B$, and we see that when $B$ is set to zero, $A$ is sometimes 0 and sometimes 1, we can select the model $A \rightarrow B$ with high confidence, because the data demonstrate that $B$ is not affecting $A$. To forshadow the next section, notice that interventional data help us go from a model of statistical dependency between $A$ and $B$ to one that can have a causal interpretation.

How do we translate our intuition regarding interventions into a Bayesian score that incorporates interventions? When we assess a graph structure with the Bayesian score, we are asessing if the dependencies in the data are well represented by those in the graph- in other words, does the child variable (in each case) show a statistical de-

pendence on its parent(s) in the structure. When dealing with interventions, we have a slightly different situation: In those datapoints with intervention, the perturbed variable cannot be dependent on its parent(s), because its value has been externally set (by the experimentor). Therefore, for those datapoints only, we sever the ties between the perturbed variable and its parent(s). The remainder of the data is treated as before.

Recall that the Bayesian score sums over the contribution of each variable given its parents, and that this contribution amounts to counts extracted from the data (plus hyperparameters from the prior). Severing the ties of the perturbed variable from its parents amounts to *skipping the datapoints with intervention* when tabulating the counts for the variable that has been perturbed (only when it is the child). Note that failure to do this would weaken the dependence of the variable on its parents, since they cannot affect the variable's value in those datapoints.

The intuition behind this alteration of the Bayesian score is as follows. Consider again a two variable dataset, with well correlated variables, yielding the equivalence class $A \rightarrow B$ and $B \rightarrow A$. Assuming the true model is $A \rightarrow B$, a perturbation on $A$ would result in interventional data in which the $B$ variable is fixed according to $A$'s value. Thus, when $A \rightarrow B$ is scored, it will score well, as $A$ will appear as a good parent of $B$. When $B$ is scored, because the perturbed variable, $A$, is now the child, those interventional datapoints will be skipped, yielding an effectively smaller dataset in which to score the $B \rightarrow A$ edge. The effectively smaller dataset yields less peaked parameters, so the true model, $A \rightarrow B$, will have a higher score. A perturbation on $B$ would lead to interventional data in which $B$ is fixed, but $A$ is not fixed in response, since it does not depend on $B$. With this data, the dependence between $A$ and $B$ is weakened in the $B \rightarrow A$ (in which the interventional data is included), but not in the $A \rightarrow B$ model (in which the interventional data is disregarded), so the $A \rightarrow B$ model will score higher.

### 2.4.3 Causality and model interpretation

The Bayesian networks we have been discussing so far have been models of statistical dependencies among variables. Because statistical dependencies often accompany causal, mechanistic relationships, these relationships are often mistaken for each other. In fact, while a causal relationship may result in a statistical dependency, the presence of a statistical dependency does *not* imply causality. Puddles and umbrellas may be well correlated, but not because one causes the other; rather, because a third variable is the parent of both. In general, a statistical dependency implies *either* a causal relationship *or* a shared (set of) parent(s) (or both). The presence of equivalence classes ties in to this fact: since a statistical dependence is merely statistical, it does not necessarily have directionality. Recall from the discussion above that under some conditions, we *are* able to determine the directionality of graph edges (even when only observational data is available). In those cases, given certain assumptions, we can interpret the Bayesian network models as causal models.

A causal interpretation of a Bayesian network implies that the parents of a variable in the graph are its immediate causes.[8] It follows that in a causal graph, an intervention on a parent variable will affect the child. To interpret a Bayesian network structure as causal, two assumptions must be made. The first is the *Causal Markov Assumption*, (similar to the Markov assumption we defined earlier) which states that given its immediate causes in the graph (i.e. its parents), a variable is independent of its earlier causes. This is a reasonable assumption in our domain. Consider a simple signaling pathway in which kinase A activates kinase B which then activates kinase C: $A \rightarrow B \rightarrow C$. Assuming that $B$ is activated (or inactivated), the value of $C$ no longer depends upon the value of $A$. The second assumption states that there are no *hidden variables* (variables which are not measured) which affect the model variables. This assumption is often violated in our domain. For instance, there are many signaling molecules which we do not measure. Furthermore, even if we could measure

[8]Immediate in the sense that their effect is more direct than the effect of any nonparent variable in the graph; in the causal graph $X \rightarrow Y$, it is still possible that an unobserved variable $Z$ is an intermediary between $X$ and $Y$.

all signaling molecules, the state of the signaling molecules might also be affected by other factors, such as the concentration of ATP, $Ca^{+2}$ or other small molecules, cell cycle stage, and countless other contributors which our model typically does not encompass. Therefore, our results must be interpreted with caution. [9]

If we do make these modeling assumptions, then we can interpret as causal all edges that are directed in our model. Recall that some edges are not directed, due to the presence of equivalence classes. An edge can be directed in one of three ways: it may be the result of conditional independencies in the domain (this applies even when only observational data is available), it may be the result of an intervention, or it may be compelled by a combination of other edges that are directed, and/or our modeling assumptions.

**Conditional independency**

Conditional independencies can orient edges due to the presence of v-structures, which exist only when two variables (the parents) become dependent given a third variable (the child). Because this type of conditional independence happens only with a v-structure, it automatically orients the model edges. (See our discussion of v-structures, above). Considering that no interventional data is employed, it is interesting that a causal interpretation may be proposed for these edges. This topic has received thorough treatment in the literature [64].

**Interventions**

Edges coming out of or going into a perturbed variable are directed. Either the perturbed variable is found to influence another variable (these are the arrows going out of the perturbed variable), or it is found to not influence it (the arrows going in)[10] See above for discussion of directing edges using interventional data.

**Compelled edges**

The last type of directed edge is one that we have not yet considered: edges that

[9]Note that there is another modeling assumption that we make implicitly, with respect to all Bayesian network models (not only the causal ones). That is that the true underlying graph is acyclic. This assumption is also often violated in our domain. While we must make these modeling assumptions if we wish to use the Bayesian network tool, it is crucial that we then cautiously regard our results as ones emerging from an imperfect model.

[10]Or its influence is not consistent with the background observational distribution, leading to an arrow going into the perturbed variable.

are compelled. Edges are typically compelled by a combination of edges that are already directed, and possibly also by modeling constraints. There are two basic ways to compel edges. One is by avoiding v-structures, and the other is by avoiding cyclicity. Consider a graph containing one edge that is already directed, and one that is not: $A \rightarrow B$–$C$. If we direct the $B$–$C$ edge as $B \leftarrow C$, we will create a v-structure. We know a v-structure must not exist, because if it did, that edge would have already have been directed by the conditional independencies in the data. Therefore, we conclude that the structure is *not* a v-structure, leading us to orient the edge as $B \rightarrow C$. Now consider the model with the directed edges $A \rightarrow B \rightarrow C$ that also contains the undirected edge $A$–$C$. Orienting that edge as $A \leftarrow C$ creates a cycle, so, in order to avoid the cycle, the edge is oriented as $A \rightarrow C$. In more complex models, edges may be oriented due to a combination of acyclicity constraints and the avoidance of v-structures.

# Chapter 3

# Preliminary work

This thesis has focused on Bayesian network models for analysis of signaling pathways. However, because such models were uncharted territory, it was not clear from the beginning what data should be used, so part of the thesis work involved evaluation of various forms of data. In the preliminary work described below, we explore the use of population-level protein abundance assays (e.g. western blots), protein activity assays, and, finally, low-dimensional flow cytometry data, to construct Bayesian network models of signaling pathways. These models are limited in scope and very data-limited, and therefore are limited in their biological relevance. They are included in this thesis for two reasons. First, they illustrate the potential of Bayesian network applications to the analysis to signaling pathways through various small-scale examples (e.g. see Section 3.1.2 below for an example of dynamic Bayesian networks). Second, they show the progression of data types considered, that led to the eventual selection of single cell data. The first portion of this chapter was previously published in [77].

## 3.1 MAPK cascade models using western blot and protein activity assay data

In the realm of signaling molecules, measurements can be time-consuming and expensive, and datasets limited in size. In this modeling attempt, we take advantage of a relatively large dataset, to model the mitogen-activated protein kinase (MAPK) cascade, a ubiquitous pathway vital for such processes as the survival, proliferation, differentiation, and migration of eukaryotic cells. In particular, we address the activation of focal adhesion kinase (FAK) and extracellular signal-regulated kinase (ERK) that results from the interaction between the extracellular matrix protein fibronectin (fn) and the integrin $\alpha_5\beta_1$. There has been literature controversy about the pathway(s) relating FAK and ERK activation, in terms of the relative directions of influence of FAK activation on ERK activation and vice versa; for example, see [22, 48, 81, 80, 99]. We consider several possible models of the signaling events. To score them, we take advantage of an existing, relatively large-scale, quantitative dataset of this system. [3] Asthagiri et al. cultured CHO-B2 cells transfected with human $\alpha_5$ on fibronectin coated plates. Two levels of integrin and five levels of fibronectin were employed. A time course of ERK2 activity measurements were made for each of the ten possible fibronectin/integrin combination using a high-throughput kinase activity assay. FAK activation time courses were measured via Western blots. We utilize two metrics for FAK and ERK2 from Asthagiri et al.: an overall activation level, and the initial rate of activation under the ten integrin/fibronectin combinations. The data is summarized in Figure 3-1.

### 3.1.1 Model selection

Model selection can be an important stage in hypothesis testing, when a single model out of a number of equally favored candidates must be singled out for wet-lab verification, in particular when parallel verification of all candidate models is infeasible or prohibitively expensive. Because the Bayesian scoring metric determines the relative

| Model selection – static models | |
|---|---|
| Dataset I:<br>Initial rate of activation | 8 datapoints each for FAK and ERK<br>Expressed as the slope between an early timepoint and the zero timepoint. |
| Dataset II:<br>Overall activation | 10 datapoints each for FAK and ERK.<br>For FAK this is the steady state level achieved by the 90 minute timepoint, for ERK it is the integral of the timecourse curve. These metrics are presented in Asthagiri et al. |
| Model discovery – dynamic models | |
| 8 datapoints each of ERK at 5, 10, 15 and 20 minutes, and FAK at 7.5 and 90 minutes. | |

Figure 3-1: Data summary. For model selection, data sets I and II are used to score static models. Data set I is the initial rate of activation from Asthagiri et al, 1999; data set II is overall activation. Model discovery examines dynamic models and therefore employs the unprocessed time course data.

probability of each model structure, it can be applied to rank competing models according to their ability to explain available data. This is a form of model selection in which the data set defies interpretation through classic application of biological intuition and instead necessitates a computational tool to determine which hypothesis the data support best. In the case of Bayesian networks, that selected hypothesis is simply the model with the highest relative probability, given the data.

We pose four candidate models (Figure 3-2). Model 1 (M1) shows a link from fn and integrin to FAK, and one from FAK to ERK2. Such a dependence structure is expected if fn and integrin activate (activate is used to indicate regulation in the general sense, and may include repression or combinatorial regulation) FAK, which activates ERK2. Model 2 (M2) is consistent with a pathway in which fn and integrin activate ERK2, which activates FAK. Model 3 (M3) shows ERK2 and FAK both linked to fn and integrin but conditionally independent of each other. Finally, Model 4 (M4) is a mixture of the previous models and has no conditional independencies. Our goal was to score the models against data from Asthagiri et al. to discover which model might be best supported by these data. In addition to the four candidate models, we also consider the independence model, M0, as a control. Because some dependence is known to exist between fn-integrin signaling and FAK and ERK2 activity, this model lacks at least one dependence relationship, so it is intentionally designed to score poorly. A total of 8 measurements of initial activation rate and 10 measurements of overall activation for FAK and ERK2 were used. Data were grouped to two or three levels using a discretization method that strives to preserve mutual information between pairs of variables while reducing the number of discretization levels [25]. Resulting model score comparisons are presented (Figure 3-3). M0, the control, scores poorly, as anticipated. Of the four hypothesized structures, data set I (initial activation rate) best supports M1, which scores  3 times higher than M2 and 4 times higher than M3 and M4. Data set II (overall activation level) supports M4 100 times better than M1,  70 times better than M2, and  10 times better than M3. These differences indicate to what degree the high-scoring model explains the data set better than the other models. These results support distinct dependence structures

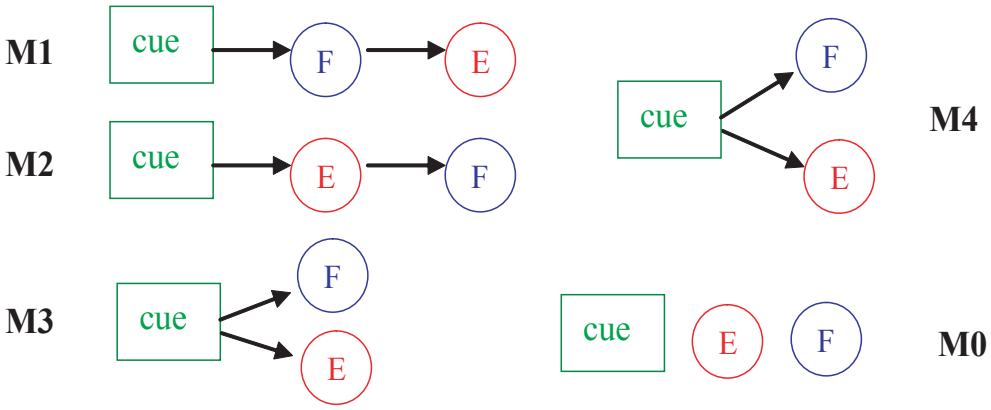



Figure 3-2: Candidate and control models. "Cue" represents the signal from the interaction betwen fn and integrin, F is FAK, and E is ERK2. M0 is the control model. An additional model identical to M4, but with the edge from F to E reversed, is not represented because it is in the same equivalence class as M4 and will, therefore, always score the same.

for the processes resulting in initial activation rates and overall activation of FAK and ERK2. Although M1 and M4 emerged as the most likely model structures of the candidates tested under data sets I and II, respectively, they are superior in likelihood to the remaining models by only small or moderate margins (large or decisive margins may be on the order of 1000-fold or more). Therefore, based on these results, we cannot select M1 or M4 with high degrees of confidence. A useful conclusion is that we will need a larger data set, acquired through a similar systematic, quantitative, dynamic, and multivariable approach.

### 3.1.2 Dynamic models

We searched for high-scoring dynamic graphs using time series data from Asthagiri et al. We used a dynamic Bayesian network graph expanded in time, in such a way that each time point of each variable is represented as a node [18]. This representation confers a number of advantages, including time resolution of dependencies. The search is constrained to forbid dependence of early time points on later ones and of the cue node on downstream signals. We sample from the high-scoring region of the model space, thus the search algorithm will find a subset of the highest-scoring

| Model | Data set I scores | Approximate fold difference from MI | Data set II scores | Approximate fold difference from M4 |
|---|---|---|---|---|
| M0 | –50.5 | 2150 | –60.8 | 1700 |
| M1 | –42.8 | 1 | –58.0 | 104 |
| M2 | –43.8 | 3 | –57.6 | 73 |
| M3 | –44.3 | 4 | –55.6 | 10 |
| M4 | –44.3 | 4 | –53.4 | 1 |

Figure 3-3: Model scores. Data set I is initial rate of activation from Asthagiri et al; data set II is overall activation. Columns 3 and 5 indicate to what degree the top-scoring model explains the data better than the indicated model. This fold difference is equal to e to the power of (difference in model scores).

models. (Results shown here represent 200 runs of 1000 or more iterations each. The high-scoring graph in each run was typically found by approximately iteration 500). The resulting graph comprises a weighted average of high-scoring models visited over 200 runs (Figure 3-4). Darker arcs represent edges with higher posterior probability (0.86 to 1); lighter arcs indicate a posterior probability of 0.50 to 0.85. The data appear to support a dependence between the signaling cue (the signal generated by fn binding integrin) and FAK levels at the early time point, as well as ERK levels at 5, 10, and 15 min (Figure 3-4, arcs 1 through 4). An arc connects later time points with their respective predecessor(s) for ERK nodes (Figure 3-4, arcs 8 through 10). The graph suggests an interesting dynamic between FAK and ERK: A dependence is seen between ERK at 5 and 10 min and FAK at 90 min, and between FAK at 7.5 min and ERK at 15 min. Thus, the model predicts a bidirectional dependence between the two molecules. This cyclic relationship could not be elucidated with static Bayesian networks, which do not permit cycles. It is interesting to appreciate that the apparent complexity of these dynamic ERK-FAK interactions could be responsible for the difficulty in determining clear "upstream" versus "downstream" influence relationships by means of standard molecular cell biology methods.

### 3.1.3 Discussion

We have applied Bayesian networks as a tool for performing hypothesis selection and formulation, in the context of a simple preliminary problem in the realm of cell signaling pathways for illustration purposes. This problem is the influence relationship between FAK and ERK after fn-mediated integrin activation, and we used two dynamic data sets incorporating multiple levels of integrin expression and concentrations of fn [3]. Of the static graphs considered for model selection analysis (Figure 3-2), data set I (initial activation rate) best supports M1, in which FAK is dependent on the fn-integrin cue and ERK is dependent on FAK, but conditionally independent of the fn-integrin cue. Data set II (activation level) supports M4, in which there are no conditional independencies (that is, M4 indicates dependence among all the variables). This result implies that the dependence structure leading to initial activa-

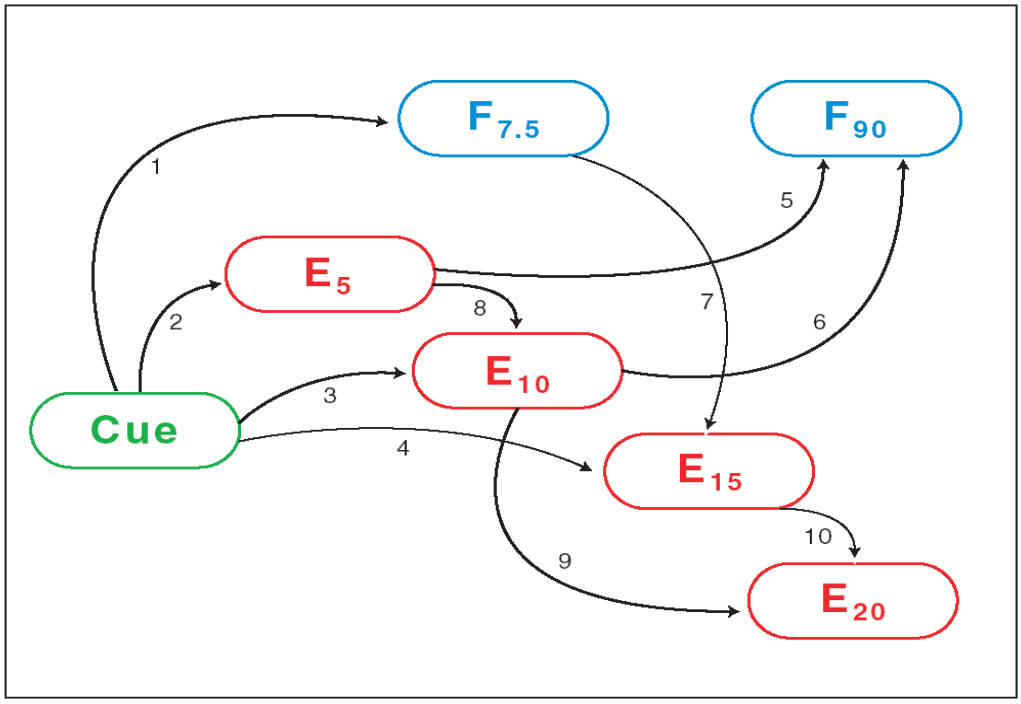



Figure 3-4: Features common to high-scoring graphs. The model presented comprises a weighted average of high-scoring graphs from 200 runs of the search algorithm. "Cue" represents the signal from fn and integrin, E is ERK2, and F is FAK. Subscripts indicate time in minutes. Light arrows are features with a posterior probability of 0.5 to 0.85; dark arrows represent consensus arcs or arcs with posterior probability 0.85. Arcs 1 through 10 are numbered for convenience.

tion rate is distinct from that resulting in overall activation levels of FAK and ERK. We have no a priori reason to doubt this result, which could in fact implicate an interesting mechanism for short-term versus long-term effects of fn-integrin signaling. However, in this case we cannot determine with certainty that M1 and M4 are the most probable dependence structures for initial and overall FAK and ERK activation, respectively, for a number of reasons. All the noncontrol models are defeated by a small or moderate margin, which may be attributed to noise. Moreover, the data sets are very small in probabilistic terms, in particular in comparison to the number of model parameters. A larger data set (5 times larger at least) would be expected to yield larger differences between scored models. Thus, we conclude that we currently have insufficient data to select M1 or M4 definitively. The use of dynamic Bayesian networks enables the elucidation of regulatory loops and dynamic dependencies not evident in static networks. The data used to search for dynamic graphs are a measurement of the level of active FAK and ERK at each time point, in contrast to the rate of initial activation and overall activation levels used for scoring the static models Figure 3-1. Because the nodes physically represent different measurements, comparing the static and dynamic models is not completely straightforward. Nevertheless, the link between the cue and early time points of FAK and ERK and later dependence between FAK and ERK in the dynamic graph Figure 3-4 are consistent with M4, a top-scoring static graph, suggesting dependence among all the variables. However, the early dependence of both FAK and ERK on the cue alone lends support, for the initial activation data set, to static M3, showing conditional independence between ERK and FAK, yet M3 scores about four times lower than the top-scoring model, M1. This discrepancy may point to inaccuracies in either model. The search results (Figure 3-4) suggest an interesting model of FAK-ERK dependence, combining a number of hypotheses previously proposed [22, 48, 81, 80, 99]. However, this has resulted from an average of high-scoring models rather than the definitive best model, and it is very likely overfitting to the small data set. (Indeed, the number of model parameters in some places exceeded the number of available data points.) Although the data set we studied here is relatively large by traditional molecular cell

biology standards, it nonetheless appears to be too limited to compellingly advance the current state of understanding regarding FAK and ERK activation. We thus emphasize again that this example was provided for illustration purposes only. The effective utility of Bayesian network approaches to modeling cell signaling pathways may depend on higher-throughput measurement methods, such as flow cytometry, protein arrays, and microfabricated analytical devices. Because the systems biology field generally expects rapid developments in these measurement techniques, we anticipate that Bayesian networks will become increasingly useful in this problem area, as they are in the analysis of gene expression problems.

## 3.2 Apoptosis models using 2-color flow cytometry

To overcome the problem of small dataset size, we next turned to single cell data, specifically, flow cytometry measurements. In this initial attempt, two-color flow is used to measure molecules involved in tumor necorsis factor (TNF) induced apoptosis. We obtained 16,000 measurements, 2000 each of 8 different conditions, of two apoptosis pathway components (caspase 3 and cytokeratin) from John Albeck (Sorger lab). The eight conditions consist of three apoptotic activators/repressors (TNF, Pi3k inhibitor, caspase 9 inhibitor) that were each either included or left out of each experiment, for a total of eight conditions. Our intent was to use the data and the information about inhibitors and inducers in the media to search for the best dependency model among these five apoptosis pathway components (2 measured, 3 manipulated molecules). This dependency structure is already known (see Figure 3-5 for a diagram showing these molecules in the context of TNF signaling)

For each manipulated molecule, we assumed the molecule was 'on' when its inhibitor was absent (or when the molecule was added to the media), and 'off' when its inhibitor was present (or molecule was not added to the media). This assumption may be violated at times, such as when an inhibitor is absent but the molecule is turned off through normal activity of other pathway components. However, we felt it was a reasonable assumption since the exogenously added inhibitors can be added

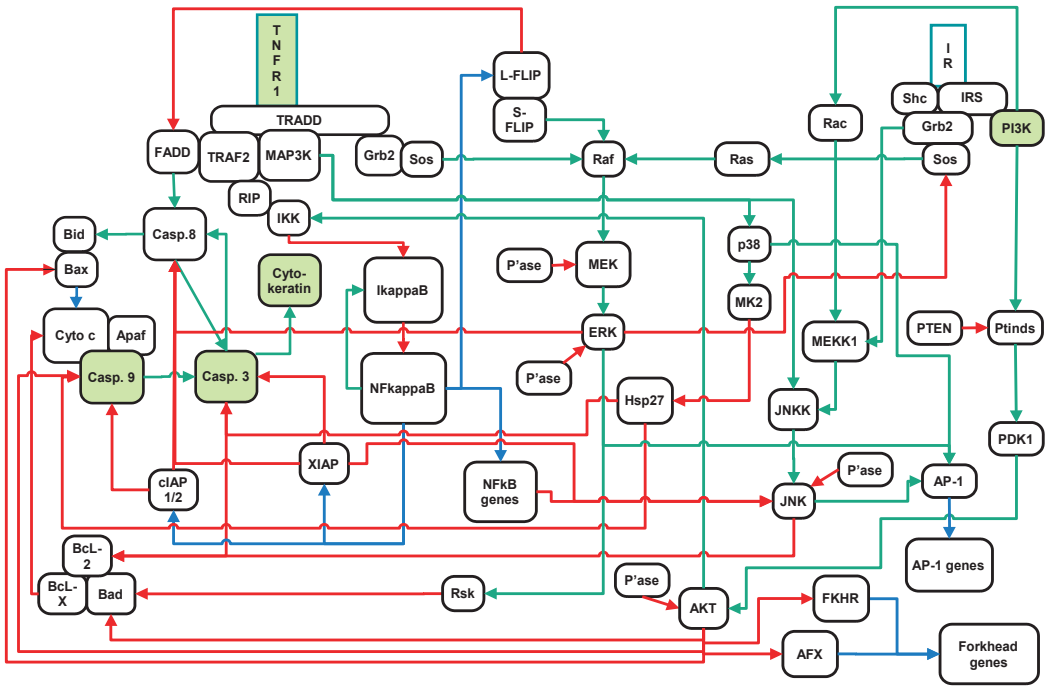



Figure 3-5: Diagram of TNF-induced signaling. Molecules either measured or manipulated in this study are shown in green. Figure source: Suzzane Gaudet, Kevin Janes.

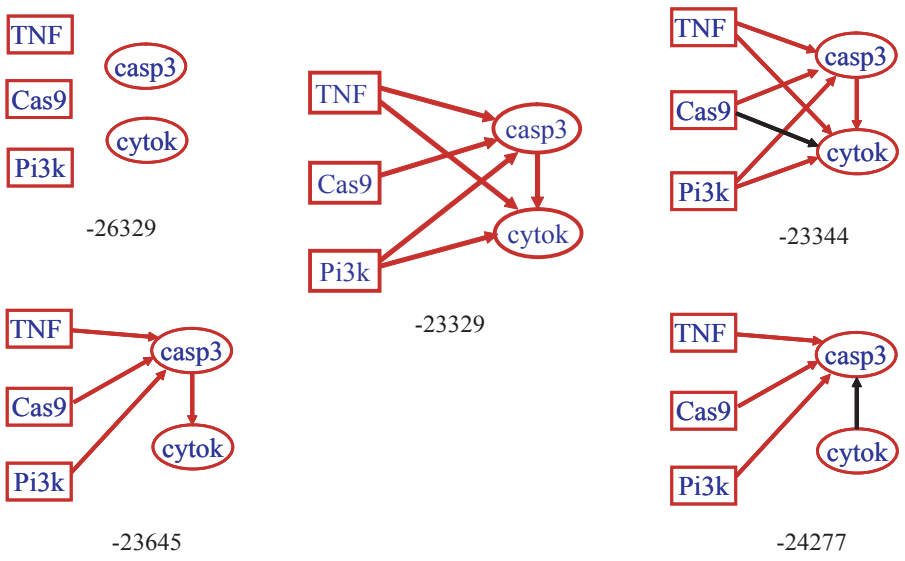



Figure 3-6: Proposed models and their scores. The scores as reported as the natural log of the relative probabilities. Therefore, the difference in score between models is $\exp^{Score_{Mi}-Score_{Mj}}$ (i.e. e to the power of (Difference in score between the two models)). The correct model scores higher than the others by a factor of at least 1000.

at a saturating dose. We included the constraint that no directed edge can point to the manipulated molecules, since these are set externally by the experimenter. Under these conditions, the known dependence structure between the measured and manipulated molecules was found as the highest scoring model in our search, scoring as at least 1000 times more probable than any other model (Figure 3-6). This work, though small in scope, demonstrates the power of addressing biological problems with high throughput (in this case, single-cell level resolution) data and probabilistic modeling tools. To our knowledge, this is the first time a large (16,000 sample), single-cell level biological dataset has been analyzed with our approach to address a question of dependencies among cell signaling molecules.

# Chapter 4

# Models of multidimensional flow cytometry data

At the core of this dissertation is an approach we developed to elucidate the influence connections in protein signaling pathways. This process involves the extraction of correlations revealed in individual cells, an approach we arrived at after consideration of various other data approaches (see Chapter 3). After turning to single cell correlations, we were fortunate to establish a collaboration with a lab that had developed a capability to measure multidimensional data in single cells. The Nolan lab at Stanford has helped to push the capability of the flow cytometer to the point where it is able to measure 12 distinct intracellular signaling molecules (simultaneously), assuming available reagents. We use this data to reverse engineer a first order map of the influence connections in the underlying signaling pathway.

This work was previously published in [78]. In this thesis text, we elaborate on the technical details (see Section 4.3.2) and add a section on robustness analysis that was not included in the published work (Section 4.4).

## 4.1 Introduction

In this proof of principle study, we applied Bayesian network analysis to multivariate flow cytometry data, in order to learn influence connections among signaling molecules

involved in T-cell signaling. Data were collected after a series of stimulatory cues and inhibitory interventions (see Table 1A), with cell reactions stopped at 15 minutes post-stimulation by fixation, to profile the effects of each condition on the intracellular signaling networks of human primary naive CD4+ T cells, downstream of CD3, CD28, and LFA-1 activation (see Figure 4-1 for an overview of the approach, and Figure 4-2 for a currently accepted consensus network).

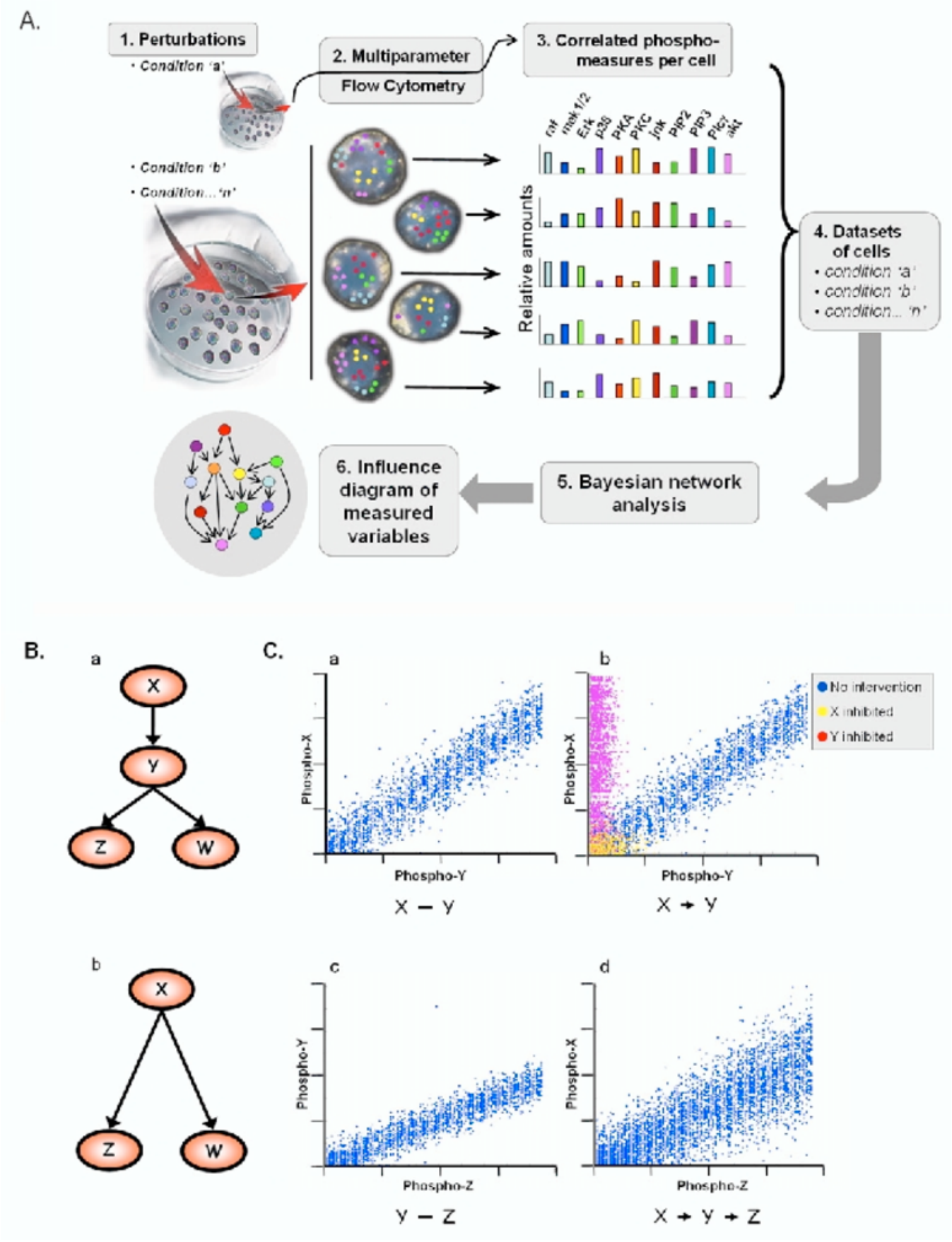

A.
1. Perturbations
• Condition 'a'
• Condition 'b'
• Condition... 'n'
2. Multiparameter
Flow Cytometry
3. Correlated phospho-measures per cell
raf
mek1/2
Erk
p38
PKA
PKC
jnk
PIP2
PIP3
Plcγ
akt
Relative amounts
4. Datasets of cells
• condition 'a'
• condition 'b'
• condition... 'n'
5. Bayesian network analysis
6. Influence diagram of measured variables
B.
a
X
Y
Z
W
b
X
Z
W
C.
a
b
No intervention
X inhibited
Y inhibited
Phospho-X
Phospho-Y
X — Y
X → Y
c
d
Phospho-Y
Phospho-Z
Y — Z
X → Y → Z

Figure 4-1: **Bayesian Network Modeling with Single Cell Data** A. Schematic of Bayesian network inference using multidimensional flow cytometry data. Nine different perturbation conditions were applied to sets of individual cells (see Table 1A). A multiparameter flow cytometer simultaneously recorded levels of 11 phosphoproteins and phospholipids in individual cells in each perturbation dataset (see Table 1B). This data conglomerate was subjected to Bayesian network analysis, which extracts an influence diagram reflecting dependencies and causal relationships in the underlying signaling network. B. Bayesian networks for hypothetical proteins X, Y, Z, and W. (a): In this model X influences Y which, in turn, influences both Z and W. (b): Same network except Y was not measured in the dataset. C. Simulated data that could reconstruct the influence connections in Figure 4-1, B (this is a simplified demonstration of how Bayesian networks operate). Each dot in the scatter plots represents the amount of two phosphorylated proteins in an individual cell. (a) Scatter plot of simulated measurements of phosphorylated X and Y show correlation. (b) Interventional data determine directionality of influence. X and Y are correlated under no manipulation (blue dots). Inhibition of X affects Y (yellow dots); inhibition of Y does not affect X (red dots). Together this indicates that X is consistent with being an upstream parent node. (c) Simulated measurements of Y and Z. (d) A noisy but distinct correlation is observed between simulated measurements of X and Z.

We made flow cytometry measurements of 11 phosphorylated proteins and phospholipids (Raf phosphorylated at position S259, mitogen activated protein kinase Erk1 and Erk2 phosphorylated at T202 and Y204, p38 MAPK phosphorylated at T180 and Y182, JNK phosphorylated at T183 and Y185, AKT phosphorylated at S473, Mek 1 and Mek2 phosphorylated at S217 and S221 (both isoforms of the protein are recognized by the same antibody), phosphorylation of PKA substrates (CREB, PKA, CAMKII, CASPASE 10, CASPASE 2) containing a consensus phosphorylation motif, phosphorylation of PLC on Y783, phosphorylation of PKC on S660, and phosphor-inositol 4,5 bisphosphate [PIP2] and phosphoinositol 3,4,5 triphosphate [PIP3]) (Figure 4-5 and see Materials and Methods). Each independent sample in this dataset consists of quantitative amounts of each of the 11 phosphorylated molecules, simultaneously measured from single cells. For purposes of illustration, examples of actual FACS data plotted in prospective co-relationship form are shown in Figure 4-4. In most cases, this reflects the activation state of the kinases monitored, or in the

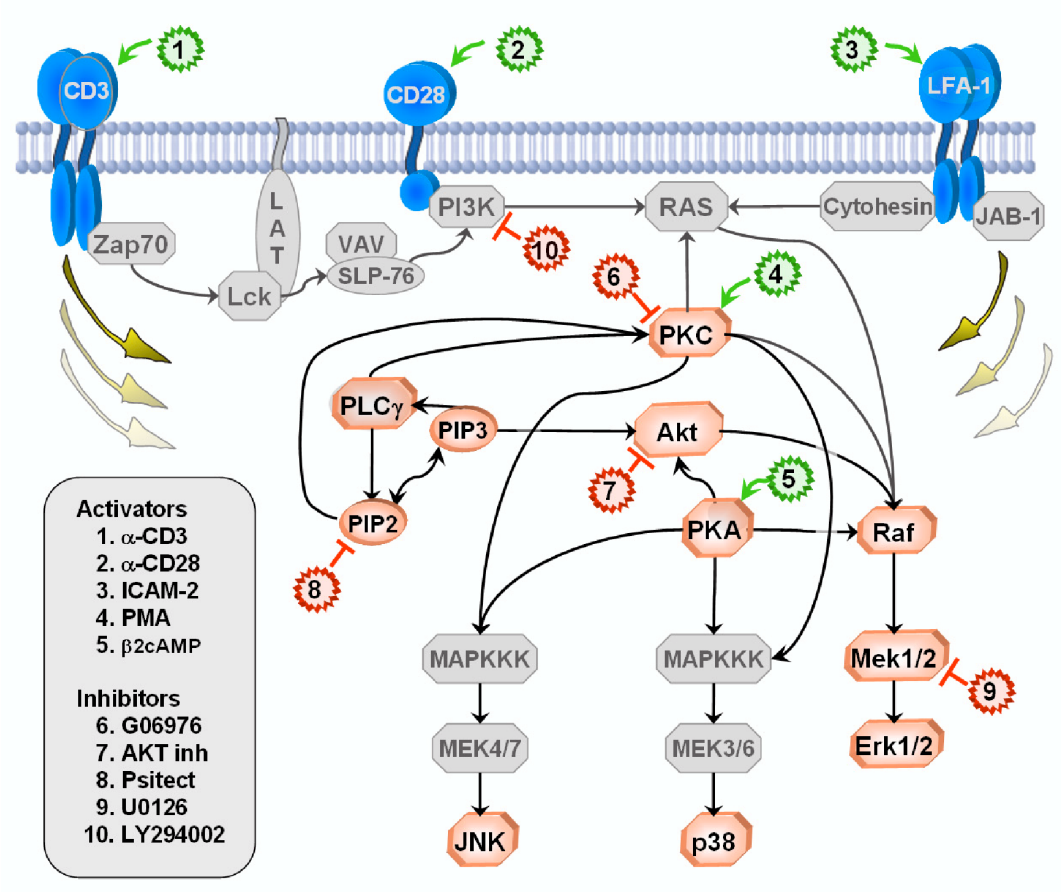



Figure 4-2: Classic Signaling Network and Points of Intervention. Graphical illustration of the conventionally accepted signaling molecule interactions, the events measured, and the points of intervention by small molecule inhibitors. Signaling nodes in color were measured directly. Signaling nodes in gray were not measured but are presented to place the signaling nodes that were measured within the context of contextual cellular pathways. Interventions classified as activators are color-coded green and inhibitors are color-coded red. Intervention site of action is indicated in the Figure. Arcs are used to illustrate connections between signaling molecules; in some cases the connections may be indirect and may involve specific phosphorylation sites of the signaling molecules (see Figure 4-7 for details of these connections). Note that this figure contains a synopsis of signaling in mammalian cells and is not representative of all cell types, with inositol signaling co-relationships being particularly complex.

| 9 Perturbations | Reagent | Reagent class |
|---|---|---|
| 1. anti-CD3 + anti-CD28 | anti-CD3/CD28 | General perturbation: Activates T cells and induces proliferation and cytokine production. induces signaling through the TCR, activated ZAP70, Lck, PLCγ, Raf, Mek, ERK, PKC. TCR signaling that converge on transcription factors NFKB, NFAT, and AP-1 to initiate IL-2 transcription. |
| 2. anti-CD3/CD28 + ICAM-2 | ICAM-2 | General perturbation: Induces LFA-1 signaling and contributes to CD3/CD28 signaling that converge on AP-1 and NFAT transcriptional activity. |
| 3. anti-CD3/CD28 + U0126 | β2cAMP | Specific perturbation: cAMP analog that activates PKA. PKA can regulate NFAT activation and T cell commitment processes. |
| 4. anti-CD3/CD28 + AKT-inhibitor | AKT-inhibitor | Specific perturbation: Binds inositol pleckstrin domain of AKT and blocks AKT translocation to the membrane where normally AKT it becomes phosphorylated and active. ($IC_{50}$ = 5 μM). Activation of of AKT and phosphorylation of AKT substrates needed to enhance cell survival. |
| 5. anti-CD3/CD28 + G06976 | U0126 | Specific perturbation: Inhibits MEK1 ($IC_{50}$ = 72 nm) and MEK2 ($IC_{50}$ = 58 nm) in a noncompetitive manner (ATP and ERK substrates). Inhibits activation of ERK, arresting T cell proliferation and cytokine synthesis. |
| 6. anti-CD3/CD28 + Psitectorigenin | PMA | Specific perturbation: Phorbol myristate acetate that activates PKC, initiates some aspects of T cell activation. |
| 7. anti-CD3/CD28 + LY294002 | G06976 | Specific perturbation: Inhibits PKC isozymes ($IC_{50}$ < 8 nM).Inhibits PKC, arrests T cell activation. |
| 8. PMA | Psitectorigenin | Specific perturbation: Inhibits phosphoinositide hydrolysis. Inhibits PIP2 production, disrupts phosphoinositol turnover. |
| 9. β2cAMP | LY294002 | Specific perturbation: PI3K inhibitor. Inhibits PI3K and subsequent activation of AKT. |

Figure 4-3: **Conditions used and biological effect.** Left hand column outlines the conditions used in this study. Middle column lists the specific reagents used in each perturbation condition and the right hand column classifies the reagent class into either a general perturbation that overall stimulated the cell or a specific perturbation that acts on a defined set of molecules.

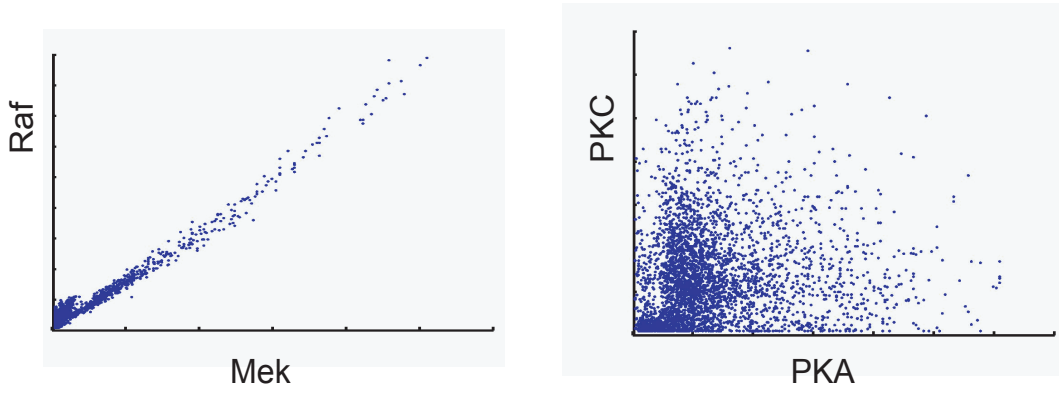



Figure 4-4: Example scatterplots of the multicolor flow cytometry data used. Each dot in the scatter plots represents the amount of two phosphorylated proteins in an individual cell. A. Scatterplot of phosphorylated proteins Raf and Mek shows a clear correlation, similar to the simulated data presented in Figure 4-1, panel a. B. Scatterplot of PKC and PKA displays a far noisier dependency that is not apparent by eye. The data used contains the entire range between the two examples in this figure. Given sufficient data, the Bayesian network is able to overcome the noise and extract these relationships.

cases of PIP3 and PIP2 the levels of these secondary messenger molecules in primary cells, under the condition measured. Nine stimulatory or inhibitory interventional conditions were used (see Table 1A, Materials and Methods). The complete datasets were analyzed with the Bayesian network structure inference algorithm as detailed below (see also, Chapter 2).

## 4.2 Results

### 4.2.1 A Human Primary T cell Signaling Causality Map

The resulting de novo causal network model was inferred (Figure 4-6) with 17 high-confidence causal arcs between various components. To evaluate the validity of this model, we compared the model arcs - and absent potential arcs-with those described in the literature. Arcs were categorized as: [i] 'expected,' for connections well-established in the literature, that have been demonstrated under numerous conditions in multiple model systems; [ii] 'reported,' for connections that are not well known, but for which

| Measured Molecule | Antibody specificity |
|---|---|
| **Raf** | Phosphorylation at Serine 259 |
| **ERK1 and ERK2** | Phosphorylation at Threonine 202 and Tyrosine 204 |
| **p38** | Phosphorylation at Threonine 180 and Tyrosine 182 |
| **JNK** | Phosphorylation at Threonine 183 and Tyrosine 185 |
| **AKT** | Phosphorylation at Serine 473 |
| **MEK 1 and MEK2** | Phosphorylation at Serine 217 and Serine 221 |
| **PKA substrates** | Detects proteins and peptides containing a phospho-Ser/Thr residue with arginine at the -3 position |
| **PKC** | Detects phosphorylated PKC alpha, beta I, beta II, delta, epsilon, eta and theta isoforms only at carboxy-terminal residue homologous to serine 660 of PKC beta II. |
| **PLC$\gamma$** | Phosphorylation at Tyrosine 783 |
| **PIP2** | Detects phosphoinositol 4,5 bisphosphate |
| **PIP3** | Detects phosphoinositol 3,4,5 triphosphate |

Figure 4-5: Molecules measured and antibody specificity. In the left hand column are shown target molecules measured in this study. These were assayed using mAb to the target residues (site of phosphorylation or phosphorylated product as described).

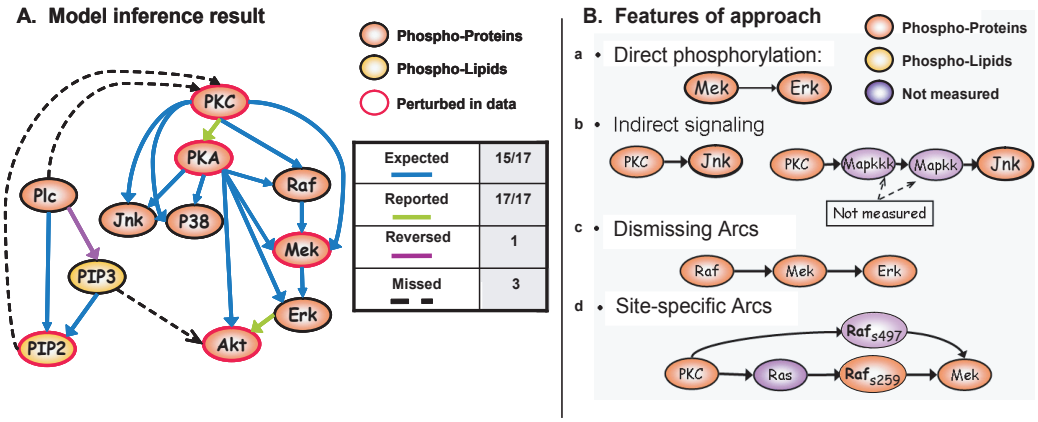



Figure 4-6: Bayesian Network Inference Results. A. Network inferred from flow cytometry data represents expected outcomes. This network represents a model average from 500 high-scoring results. High-confidence arcs, appearing in at least 85molecules are used to represent the measured phosphorylation sites, (See Figure 4-5). B. Inferred network demonstrates several features of Bayesian networks. (a) Arcs in the network may correspond to direct events or, (b) indirect influences. (c) When intermediate molecules are measured in the dataset, indirect influences rarely appear as an additional arc. No additional arc is added between Raf and Erk because the dependence between Raf and Erk is dismissed by the connection between Raf and Mek, and between Mek and Erk (for instance, see Figure 4-1). (d) Connections in the model contain phosphorylation site-specificity information. Since Raf phosphorylation on S497 and S499 was not measured in our dataset, the connection between PKC and the measured Raf phosphorylation site (S259) is indirect, likely proceeding via Ras. The connection between PKC and the undetected Raf phosphorylation on S497 and S499 is seen as an arc between PKC and Mek.

we were able to find at least one literature citation; [iii] 'unexplained,' indicates that though the arc was inferred from our model, no previous literature reports were found; and [iv] 'missing' indicates an expected connection that our Bayesian network analysis failed to find. Of the 17 arcs in our model, 14 were expected, 16 were either expected or reported, 1 was not previously reported (unexplained), and 4 were missed (Figure 4-6) [7, 52, 51, 111, 19] (see also references in Figure 4-12). Figure 4-7 enumerates the probable paths of influence corresponding to model arcs determined by surveying published reports.

Several of the known connections from our model are direct enzyme-substrate relationships (Figure 4-6): (PKA to Raf , Raf to Mek, Mek to Erk, Plc to PIP2)

| Connection | Influence path | Type | Category |
|---|---|---|---|
| PKC→Raf | PKC→Ras→$Raf_{S259}$ | indirect | E |
| PKC→Mek | PKC→$Raf_{S497/S499}$→Mek | indirect | E |
| PKC→Jnk | PKC→→MKKs→Jnk | indirect | E |
| PKC→p38 | PKC→→MKKs→p38 | indirect | E |
| PKC→PKA | PKC →cAMP →PKA | indirect | R |
| PKA→Raf | PKA →$Raf_{S259}$ | direct | E |
| PKA→Mek | PKA→$Raf_{S621}$→Mek | indirect | E |
| PKA→Erk | PKA→HePTP→Erk | indirect | E |
| PKA→Jnk | PKA→→MKKs→Jnk | indirect | E |
| PKA→p38 | PKA→→MKKs→p38 | indirect | E |
| Raf→Mek | direct phosphorylation | direct | E |
| PKA→Akt | PKA→CaMKK→$Akt_{T308}$→$Akt_{S473}$ | indirect | E |
| Mek→Erk | direct phosphorylation | direct | E |
| Plcγ→PIP2 | direct hydrolysis to IP3 | direct | E |
| Plcγ→PIP3 | recruitment, phosphorylation | reversed | E |
| PIP3→PIP2 | precursor-product | | E |
| Erk→Akt | direct or indirect | | R |

Figure 4-7: Possible pathway of influence, type of connection and category of model connections. E=Expected, R=reported, U=unexplained, see main text for further discussion. Specific phosphorylation sites are included as subscript. Unmeasured sites/molecules appear in blue. See Figure 4-12 for citations.

and one a relationship of recruitment leading to phosphorylation (Plc to PIP3). In almost all cases, the direction of causal influence was correctly inferred (an exception was Plc to PIP3, in which case the arc was inferred in the reverse direction). All the influences are contained within one global model, thus the causal direction of arcs is often compelled so that these are consistent with other components in the model. These global constraints allowed detection of causal influences from molecules that were not perturbed in our assay. For instance, although Raf was not perturbed in any of the measured conditions, the method correctly inferred a directed arc from Raf to Mek-as expected for the well characterized Raf-Mek-Erk signal transduction pathway. In some cases, the influence of one molecule on another is mediated by intermediate molecules that were not measured in the dataset. In the results, these indirect connections were detected as well (Figure 4-6B, panel b). For example, the influence of PKA and PKC on the MAPKs p38 and Jnk likely proceeded via their respective (unmeasured) MAPK kinase kinases. Thus, unlike some other approaches used to elucidate signaling networks (for example, protein-protein interaction maps [50, 5] that provide static biochemical association maps with no causal links, our Bayesian network method can detect both direct and indirect causal connections and therefore provide a more contextual picture of the signaling network.

Another important feature demonstrated is the ability to dismiss connections that are already explained by other network arcs (Figure 4-6B panel c). This is seen in the Raf-Mek-Erk cascade. Erk, also known as p44/42, is downstream of Raf and therefore dependent on Raf, yet no arc appears from Raf to Erk, as the connection from Raf to Mek, and from Mek to Erk, explains the dependence of Erk on Raf. Thus, an indirect arc should appear only when one or more intermediate molecules is not present in the dataset, otherwise the connection will proceed via this molecule. The intervening molecule may also be a shared parent. For example, phosphorylation status of p38 and Jnk are correlated (Figure 4-8), yet they are not directly connected, as their shared parents (PKC and PKA) mediate the dependence between them. Although we can not know if an arc in our model represents a direct or indirect influence, it is unlikely that our model contains an indirect arc that is mediated by any molecule observed

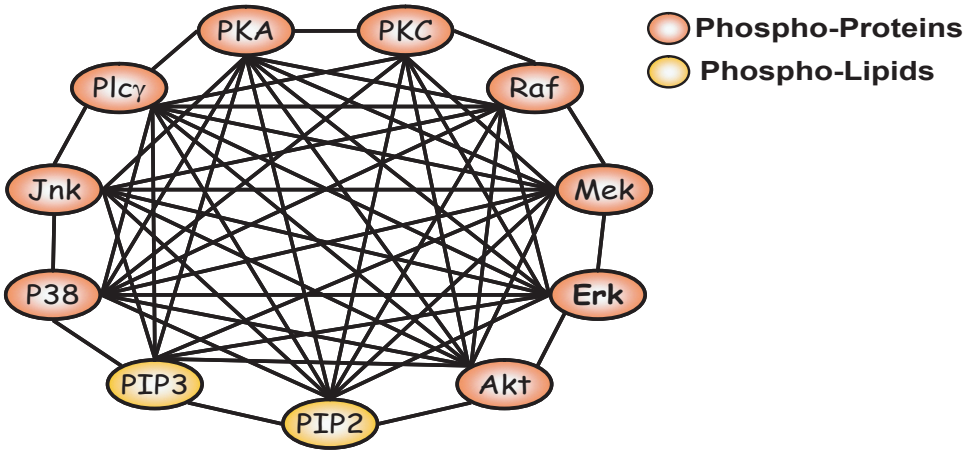



Figure 4-8: Correlation connections that pass a Bonferroni corrected p value. 52 out of 55 possible arcs appear. Only the pairs Pip3-Raf, Pip3-PKC and PKC-Jnk are not found to be significantly correlated. Note that correlations are not directed. Thus, there is a need to apply a more rigorous test (Bayesian network inference) to go beyond the simple correlations.

in our measurements. As can occur with closely connected pathways, correlation exists between most molecule pairs in this dataset (per Bonferroni corrected p value, see Figure 4-8). Therefore, the relative "lack" of arcs in our model (Figure 4-6A) contributed greatly to the accuracy and interpretability of the inferred model.

A more complex example is the influence of PKC upon Mek, known to be mediated by Raf (Figure 4-6B, panel d). PKC is known to affect Mek through two paths of influence, each mediated by a different active, phosphorylated, form of the protein Raf. Although PKC phosphorylates Raf directly at S499 and S497, this event is not detected by our measurements, as we use only an antibody specific to Raf phosphorylation at S259 (Figure 4-5). Therefore, our algorithm detects an indirect arc from PKC to Mek, mediated by the presumed unmeasured intermediate Raf phosphorylated at S497 and S499. [7] The PKC to Raf arc represents an indirect influence that proceeds via an unmeasured molecule, presumed to be Ras. [52, 51] We discuss above the ability of our approach to dismiss redundant arcs. In this case there are two paths leading from PKC to Mek because each path corresponds to a separate means of influence from PKC to Mek- one via Raf phosphorylated at S259, and the other through Raf phosphorylated at S497 and S499. Thus, neither path is redundant. This

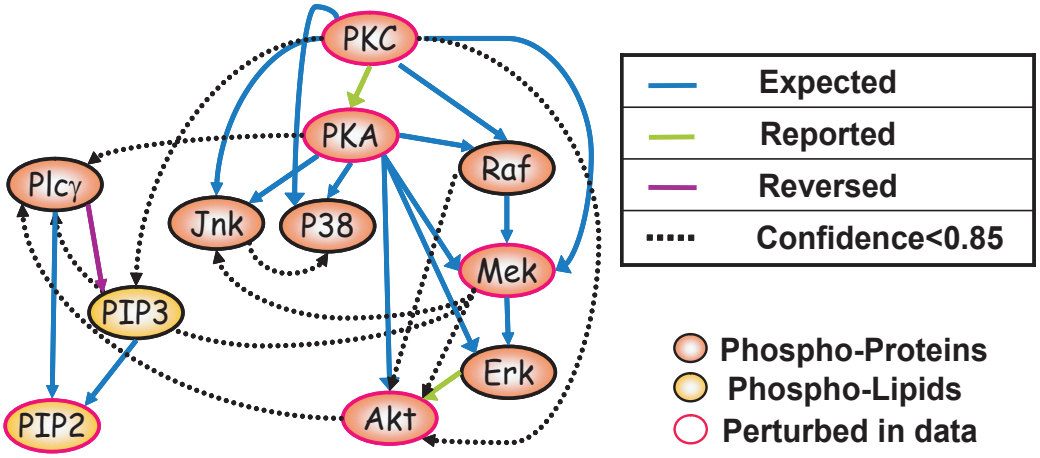



Figure 4-9: Inference results including low confidence arcs. Arcs with a confidence value of 0.5 or higher are shown. The lower confidence arcs reveal that each missing arc (from Fig. 3A) is explained by the acyclicity constraint. The missing arc Plc $\longrightarrow$PKC is precluded by the path PKC$\longrightarrow$PKA$\longrightarrow$Plc$\gamma$ , as the addition of the missing Plc$\gamma$ $\longrightarrow$PKC arc would form a cycle in the model. Similarly, the arc PIP2$\longrightarrow$PKC is precluded by the path PKC$\longrightarrow$PKA$\longrightarrow$Plc$\gamma$ $\longrightarrow$PIP2, and PIP3$\longrightarrow$Akt is precluded by the path Akt$\longrightarrow$Plc$\gamma$ $\longrightarrow$PIP3. The missing arc Akt$\longrightarrow$Raf is excluded by the (high confidence) path Raf$\longrightarrow$Mek$\longrightarrow$Erk$\longrightarrow$Akt, but it appears as a low-confidence arc in the reversed (Raf$\longrightarrow$Akt) direction. Missing arcs clearly demonstrate the limitation in the application of Bayesian network inference to biological pathways due to the acyclicity constraint.

result demonstrates the important distinction that this analysis is sensitive to specific phosphorylation sites on molecules and is capable of detecting more than one route of influence between molecules.

Four well-established influence connections do not appear in our model: PIP2 to PKC, PLC to PKC, PIP3 to Akt, and Raf to Akt. Bayesian networks are constrained to be a-cyclic, so if the underlying network contains feedback loops we cannot necessarily expect to uncover all connections (Figure 4-9). For example, in our model the path from Raf to Akt (via Mek and Erk) precludes the inclusion of an arc from Akt to Raf, due to this acyclicity constraint. Availability of suitable temporal data could possibly permit this limitation to be overcome using dynamic Bayesian networks [18, 60].

### 4.2.2 Experimental Confirmation of Predicted Network Causality

Three influence connections in our model are not well established in the literature: PKC on PKA, Erk on Akt, and PKA on Erk. To probe the validity of these proposed causal influences, we searched for prior reports in the literature. Of these 3 connections, 2 have previously been reported, the PKC to PKA connection in rat ventricular myocytes, and the Erk to Akt connection in colon cancer cell lines [111, 19]. An important goal of our work was to test the ability of Bayesian network analysis of flow cytometry data to correctly infer causal influences from unperturbed molecules within a network. For example, Erk was not acted upon directly by any activator or inhibitor in the sample sets, yet Erk showed an influence connection to Akt. Our model thus predicts that direct perturbation of Erk would influence Akt (Figure 4-10A). On the other hand, although the Erk and PKA are correlated (see Figure 4-8) the model predicts that perturbation of Erk should not influence PKA.

As a test of these predictions (Figure 4-6A), we used siRNA inhibition of either Erk1 or Erk2 and the amount of S473 phosphorylated Akt and phosphorylated PKA were then measured. In accord with the model predictions, Akt (p ¡ 9.4e-5) phosphorylation was reduced after siRNA knockdown of Erk1 but activity of PKA (p ¡ 0.28) was not (Figure 4-10B,C). Akt phosphorylation was not affected by the knock down of Erk2. The connection between Erk 1 and Akt may be direct or indirect, involving mediatory molecules yet to be understood, but the connection is supported by both the model and the validation experiment.

### 4.2.3 Enablers of Accurate Inference: Network Interventions and Sufficient Numbers of Single Cells

Three features distinguish our data from the majority of currently attainable biological datasets. First, we measured multiple protein states simultaneously in individual cells, eliminating population averaging effects that could obscure interesting correlations. Second, because the measurements are on single cells, thousands of data points

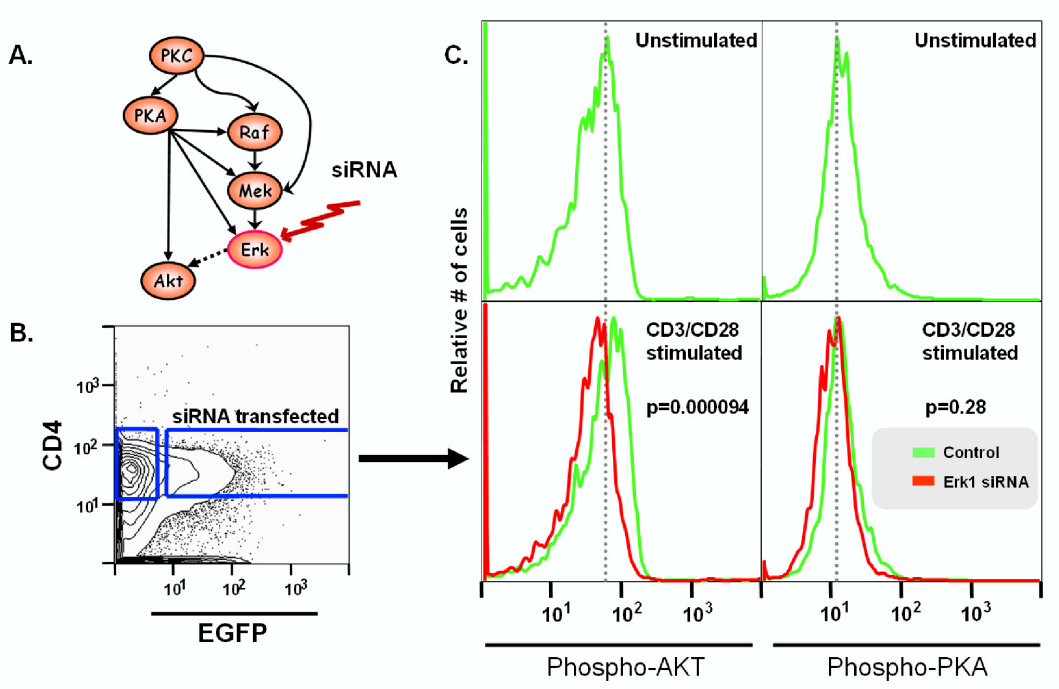


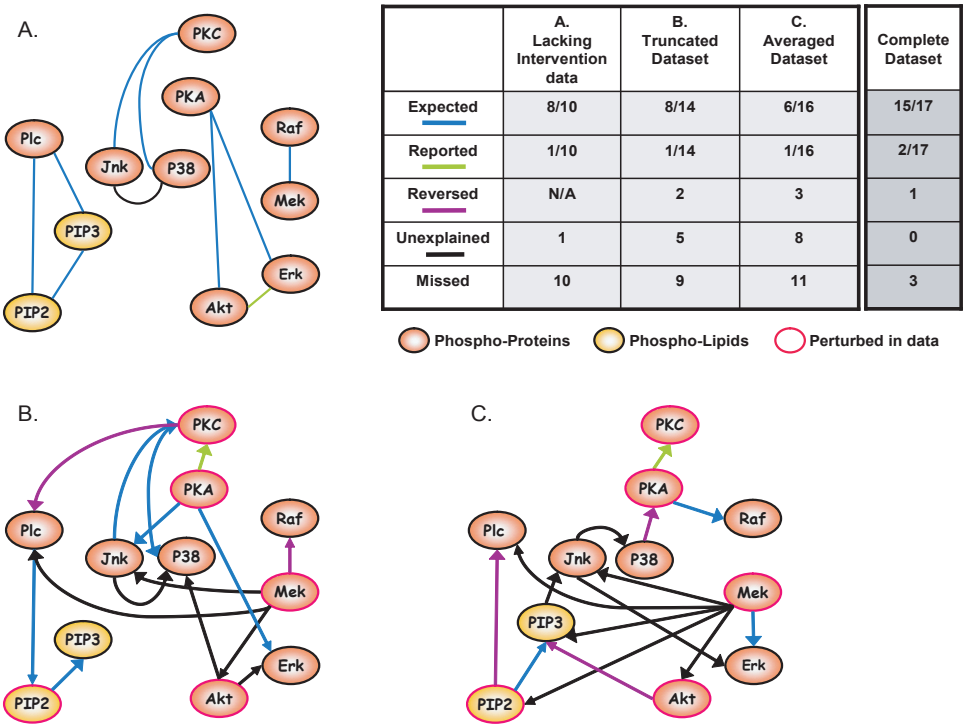

Figure 4-10: **Validation of Model Prediction.** (A) The model predicts that an intervention on Erk will affect Akt, but not PKA. (B) To test the predicted relationships, Erk1 and Erk2 were knocked down using siRNA in cells simulated with anti-CD3 and anti-CD28. Amount of Akt phosphorylation in transfected CD4+ (EGFP+ cells) were assessed, and amounts of phosphorylated PKA are included as a negative control. When Erk1 is knocked down, phosphorylated Akt is reduced to amounts similar to those in unstimulated cells, confirming our prediction (p=0.000094). PKA is unaffected (p=0.28).

were collected in each experiment. This feature constitutes a tremendous asset for Bayesian network modeling, as the large number of observations allows accurate assessment of underlying probabilistic relationships, and hence extraction of complex relationships from 'noisy' data. Third, interventional assays generated hundreds of individual data points per intervention (because flow cytometry measures single-cells in population), allowing for an increase in inferences of causality.

To evaluate the importance of these features, we created variations on our original data-set: [i] an observation-only dataset (that is, without any interventional data) of 1200 data points; [ii] a population-averaged (that is, a simulated western blot) dataset and [iii] a truncated individual-cell dataset of size comparable to the simulated western blot dataset (that is, the original dataset with most of the data randomly excluded to reduce its size, see Methods). (paragraph spacing added) Bayesian network inference was performed on each set of data. The network inferred from 1200 observational data points included only 10 arcs, all undirected, of which 7 were expected or reported, and 11 arcs were missing (Figure 4-11A). This demonstrates that interventions are critical for effective inference, particularly to establish directionality of the connections (see also Figure 4-1B). The truncated single cell dataset (420 data points) shows a large (11-arc) decline in accuracy, missing more connections and reporting more unexplained arcs than its larger (5400 data points) counterpart (Figure 4-11B). This result emphasizes the importance of sufficiently large dataset size in network inference. The network inferred from averaged data (Figure 4-11C) shows a further 4-arc decline in accuracy relative to that inferred from an equal number of single cell data points, emphasizing the importance of single cell data. The fact that population averaging destroys some of the signals present in the data may reflect the presence of heterogeneous cellular subsets that are masked by averaging techniques

## 4.3 Materials and Methods

In this section, we include the experimental materials and methods, as well as the technical details of the Bayesian network implementation, data discretization and




| | A. Lacking Intervention data | B. Truncated Dataset | C. Averaged Dataset | Complete Dataset |
|---|---|---|---|---|
| Expected | 8/10 | 8/14 | 6/16 | 15/17 |
| Reported | 1/10 | 1/14 | 1/16 | 2/17 |
| Reversed | N/A | 2 | 3 | 1 |
| Unexplained | 1 | 5 | 8 | 0 |
| Missed | 10 | 9 | 11 | 3 |

Phospho-Proteins  Phospho-Lipids  Perturbed in data

Figure 4-11: **Interventional data, large dataset size and single-cell resolution are critical for effective inference.** A. Inference results from observational data demonstrate that interventional data is crucial for effective inference. Bayesian network analysis was applied to 1200 datapoints from general stimulatory conditions. The resulting network contained only half as many expected arcs and almost three times more missed arcs than the full data counterpart (Figure 4-6A). Additionally, while it is sometimes possible to detect directed arcs with observational data alone, in this case no directed arcs were found, so the model provides no information regarding the causal direction of each link. B. Results from a truncated version of the full dataset reveal the importance of very large dataset size. Although this dataset contains all the interventions as in the full dataset, its smaller size (420 datapoints) resulted in fewer expected connections recovered and more missing arcs as compared to the result from the full dataset (Figure 4-6A). C. Results from averaged, simulated western blot data indicate the advantage of single-cell resolution. Simulated western blot data was created by averaging 20 randomly selected single-cell data points at a time, yielding a dataset of 420 points. As compared to a single-cell dataset of equal size (Figure 4-11B), this result missed more arcs and captures more unconfirmed arcs. Ten sets of truncated and averaged datasets were made; results shown in B and C represent typical results.

| Connection | Influence path | Citation |
|---|---|---|
| PKC⟶Raf | PKC⟶Ras⟶RafS259 | [96, 51, 52, 38, 53] |
| PKC⟶Mek | PKC⟶RafS497/S499⟶Mek | [7] |
| PKC⟶Jnk | PKC⟶⟶MKKs⟶Jnk | [9, 8] |
| PKC⟶p38 | PKC⟶⟶MKKs⟶p38 | [9, 8] |
| PKC⟶PKA | PKC⟶cAMP ⟶PKA | [111] |
| PKA⟶Raf | PKA⟶RafS259 | [12] |
| PKA⟶Mek | PKA⟶RafS621⟶Mek | [54] |
| PKA⟶Erk | PKA⟶HePTP⟶Erk | [79] |
| PKA⟶Jnk | PKA⟶⟶MKKs⟶Jnk | [15] |
| PKA⟶p38 | PKA⟶⟶MKKs⟶p38 | [112] |
| Raf⟶Mek | direct phosphorylation | [43, 36, 95] |
| PKA⟶Akt | PKA⟶CaMKK⟶AktT308⟶AktS473 | [100, 108, 92] |
| Mek⟶Erk | direct phosphorylation | [62] |
| Plc$\gamma$⟶PIP2 | direct hydrolysis to IP3 | [86, 45] |
| Plc$\gamma$⟶PIP3 | recruitment leading to phosphorylation | [2] |
| PIP3⟶PIP2 | precursor-product | [2] |
| Erk⟶Akt | direct or indirect | [19] |

Figure 4-12: Citations for possible pathways of influence listed in Figure 4-7.

data preprocessing. We include the experimental details for the sake of completeness and to enable the reader to repeat the experiments described. However, we stress that *all* the experimental work in this chapter was performed by Dr. Omar D. Perez (then at the Nolan lab at Stanford).

### 4.3.1 Experimental

**Reagents** Protein and chemical reagents used (and vendors) were as follows: 8-Bromo-cAMP (8-bromo Adenosine 3',5'-cyclic Monophosphate, 2cAMP), AKT inhibitor, G06976, LY294002, psitectorigenin and U0126: Calbiochem. PMA: Sigma. Recombinant human ICAM2-FC was produced as reported [67]. Alexa fluor dye series (488, 546, 568, 594, 633, 647, 680), cascade yellow, cascade blue, allophycocyanin (APC), and R-Phycoerythrin (PE): Molecular Probes; cyanine dyes (Cy5, Cy5.5, Cy7: Amersham Life Sciences. Tandem conjugate protocols for PECy5, PECy5.5, PECy7, APCCy5.5, and APCCy7 can be found at www.drmr.com/abcon. -CD3 (clone UCHT1) and -CD28 (clone 28.2): BD-Pharmingen; antibodies to phospho-

proteins Raf-259, Erk1/2-T202/T204, p38-T180/Y182, Jnk-T183/Y185, Akt-S473, Mek1/2-S217/S221, PKA substrates (a measure of PKA activation), PKC-S660, and Plc$\gamma$ -Y783: Cell Signaling Technologies; antibodies to PIP2 and PIP3: Molecular Probes; antibodies to Erk1/2-T202/T204-phycoerythrin and PKA-S114: BD-Pharmingen. Phospho-AKT-S473 in siRNA experiment was from Biosource.

**Cell culture** Human peripheral blood lymphocytes were obtained by Ficoll-plaque density centrifugation (Amersham Pharmacia, Uppsala, Sweden) of whole blood from healthy donors (Stanford Blood Bank) and depleted for adherent cells. Magnetically activated cell sorting was used to negatively isolate nave CD4+ cells (Dynal, Oslo, Norway). Human cells were maintained in RPMI-1640 supplemented with 5AB (Irvine Scientific), and 1supplemented with 2 mM L-glutamine). Cells were maintained at 5at /370C in a humidified incubator.

**Flow cytometry** Intracellular and extracellular staining was performed as described [68] Intracellular probes for active kinases were made by conjugating phospho-specific antibodies to the Alexa Fluor dye series as described and used in phosphoprotein staining [68, 67]. Briefly, purified human CD4+ T cells were dispensed in 96 wells, and treated with chemical inhibitors for 30 min, then were treated with stimulatory agents for 15 min. Analyses were performed by direct application of fixation buffer to time-synchronized 96-wells (i.e. a single 96-well plate) maintained at 370C. 2uL) was added to 0.5x106 cells (in 100 uL), stimulated as indicated. Fixation was performed for 30 min on pre-chilled 96-well metal holders at 40C. Plates were then centrifuged (1500 RPM, 5 min, 40C) and stained with pre-titred multi-color antibody cocktails. Cells were washed three times and analyzed. Flow cytometry data are representative of at least 3 three independent experiments. Data were collected on a custom-configured machine, a modified FACStar bench (Becton Dickenson) connected to MoFlo electronics (Cytomation, Fort Collins CO) [93]. This configuration allows for 11-color analysis of samples and real-time compensation for spectral overlap (plus two channels for forward and side scatter). Data was collected using Desk software (Stanford University), compensated (intra-laser and fluorophore spectral overlap demixing) and analyzed using Flowjo software (Treestar).

**siRNA inhibitions** siRNA complementary to Erk1 mRNA was purchased from Superarray Biosciences. siRNA complementary to Erk2 mRNA was purchased from Upstate Biotechnologies. siRNA oligonucleotide (100 nM) was used in primary cell transfections using the Amaxa nucleofector systems (Amaxa Biosystems) [46].

**Conditions employed** The following conditions were used for model inference: 1: (anti-CD3 and anti -CD28), 2: (anti -CD3, anti -CD28 and Intercellular Adhesion Protein-2 (ICAM-2) protein), 3: PMA (phorbol myristate acetate), 4: 2cAMP (8-bromo Adenosine 3',5'-cyclic Monophosphate), 5: (anti -CD3, anti -CD28 and U0126), 6: (anti-CD3, anti-CD28 and G06976), 7: (anti -CD3 anti -CD28 and Psitectorigenin), 8: (anti -CD3, anti -CD28 and Akt-inhibitor), and 9: (anti -CD3, anti -CD28 and LY294002). Each condition provided 600 cells, for a total of 5400 datapoints. For the simulated western blot dataset and its single-cell equivalent, the following conditions were also used: 1 (anti -CD3 anti -CD28, ICAM2 protein and U0126) 2 (anti -CD3, anti -CD28, ICAM2 protein and G06976), 3 (anti -CD3 anti -CD28, ICAM2 protein and Akt-inhibitor), 4 (anti -CD3 anti -CD28, ICAM2 protein and Psitectorigenin,) and 5 (anti -CD3 anti -CD28, ICAM2 protein and LY294002). Equal numbers of cells (600) were selected at random from each condition, to prevent biasing the network to any particular condition.

### 4.3.2 Computational

#### Processing of data

*Preprocessing* Data were preprocessed as follows: Data points that fell more than three standard deviations from the mean were eliminated. This step was intended to clean up the data by removing any suspicious datapoints (potentially debris, clumps of cells or other noise).

*Discretization* Data were then discretized to three levels (low, medium or high levels of the phosphorylated protein), using an agglomerative approach that seeks to minimize loss of pairwise mutual information among variables [25]. This algorithm is thoroughly described, including pseudocode, in [25]. Briefly, the raw measurements from each variable (across all conditions) are first binned, using either uniformly

spaced bins ('interval discretization', in which the interval is divided into equally sized subintervals), or bins which each contain an equal number of measurements ('quantile discretization'). The number of bins is an input parameter. We partitioned the measurements into 100 uniformly spaced bins, choosing the interval discretization approach as it is likely more true to the underlying biology. The quantile approach assumes that each level of a protein is equally represented in a population of cells, an assumption we wanted to avoid, whereas the interval approach does not make this assumption. [1]

Given a particular interval discretization in which the measurements are each represented as a number between 1 and 100, the algorithm then agglomerates 2 levels in each iteration, producing $100 - i$ total levels at iteration $i$, until only the desired number of levels (in our case, 3) remain. The levels to agglomerate are chosen as follows: First, the pairwise mutual information between a variable and each other variable is calculated, pre-agglomeration. Next, the mutual information is calculated for each possible agglomeration under consideration (i.e. level 1 with level 2, level 2 with level 3, level 3 with level 4, etc.), and the difference is calculated, revealing which agglomeration will result in the least total loss of pairwise mutual information. This agglomeration is performed, and the procedure iterates. In our case (because we start with 100 levels and end with 3), 97 iterations are required.

*Preprocessing of perturbed values* Under conditions of chemical intervention, more data preprocessing was often required. This is because our study uses the measured phosphorylation level of molecules as surrogates for their activity. In most cases, the inhibitors employed affected the *activity* of the molecules, but did *not* affect their *phosphorylation level*. For instance, when Mek is inhibited, its measured level actually *increases* (possibly the system responding to a lack of Mek activity), but its *activity* is inhibited. Therefore, in the presence of inhibitor, the measured level is no longer a legitimate surrogate for its activity level. To model these interventions, we assume

[1] In principle, one could use interval discretization with the desired number of levels as the final discretization, rather than as input to the agglomerative algorithm. We chose the more sophisticated approach because information preservation is crucial to the success of our network inference. See Section 4.4.3 for experiments exploring interval discretization as the discretization approach.

that the level reflects the inhibited (or activated) level, setting inhibited molecules to level 1 ('low'), and activated molecules to level 3 ('high').

Specifically, we assume that inhibition completely removes activity (setting raw values to zero) and that activation increases activity by ten fold (we multiply the raw values by 10- this value was chosen somewhat arbitrarily). While these appear to be strong assumptions, note that because of the smoothing affect of the discretization, the precise values by which we modify the raw values in the pre-processing phase should not have significant impact. This would be an interesting point to address specifically, by varying the degree of fold change assumed as a result of perturbation, but this sensitivity analysis has not been performed. Note that this preprocessing was applied only in cases in which, as a result of the inhibition, the measured value no longer reflects the activity level of the molecule. This was not always the case: under Psitectorigenin treatement, the *level* rather than the activity of Pip2 is affected. Therefore, the raw values of Pip2 were not altered.

**Simulated western blots** To create a simulated western blot dataset, the following was repeated for each condition: 20 cells were selected at random and averaged, until all the cells had been averaged (yielding 30 simulated western blot datapoints per condition). Averaging reduces the size of the dataset to 1/20th of the original size, therefore 5 additional conditions containing ICAM2 (see above) were used to create the simulated western blot dataset, for a total of 420 datapoints. For a single cell dataset of equivalent size, 30 cells were selected at random from each of the 14 conditions. This process was repeated 10 times, each with a different random seed, producing 10 different simulated western blot and truncated datasets. The Bayesian network inference procedure (see below) was independently applied to each such dataset.

**Bayesian network structure inference**

We implemented Bayesian network inference as described in [66, 109] (see also Friedman [16] for a review on the methodology). Bayesian networks and Bayesian network structure inference are described in Chapter 2. In the following, we provide a short description of network inference and of our implementation, including relevant

details.

Bayesian networks [63] provide a compact graphical representation of multivariate joint probability distributions. This representation consists of a directed acyclic graph whose nodes correspond to random variables, each representing the measured amount of a biomolecule in the dataset. An arc expresses statistical dependence between the downstream variable and the upstream (parent) variable. In certain cases, these statistical dependencies can be interpreted as causal influences from the parent upon the downstream variable (molecule) [64]. For example, Figure 4-1 panel b demonstrates how interventional data guides the inference of causality.

The goal of Bayesian network inference is to search among possible graphs and select a graph or graphs that best describe the dependency relationships observed in the empirical data. We take a score-based approach: we introduced a statistically motivated scoring function that evaluates each network with respect to the data, and search for the highest scoring network. We use the standard Bayesian scoring metric [26] that rewards relatively simple models (i.e. few arcs), that are likely to have generated the data, that is, whose underlying distribution is close to the empirical distribution of the data. Because our data were sampled under conditions that directly manipulate the amounts of the measured modified biomolecules (see Table 1B), we use an adaptation of the Bayesian scoring metric that explicitly models these interventions [66, 109]. Our modeling of interventions assumes that these are ideal, i.e. directly affect only one molecule whose identity is known. While the interventions used are not ideal, this is a reasonable first approximation. See above text for a description of preprocessing of perturbed molecule values.

Given a scoring function (the Bayesian scoring metric described above) and a set of data, network inference amounts to finding the structure that maximizes the score. The number of possible graph structures is super-exponential in the number of variables (measured biomolecules) and therefore the size of the search space prohibits an exhaustive search. Thus, we resort to a heuristic simulated annealing search. We define a search space where each state is a possible network structure and define a set of operators: addition, deletion or reversal of a single arc, that transform the

network from one structure to another. We started with an initial random structure and traversed this space using the operators, searching for high scoring networks. At each step in the search procedure, a random operator was used to change the graph, the resulting structure was rescored and the change was incorporated if it yielded an improvement in the score. To avoid local maxima, occasionally a change was incorporated even if it decreased the score. We iterated this procedure to find high-scoring graphs.

This process was initialized with different random graphs and repeated (500 times), to explore different regions of the search space. Typically, many of the resulting models explained the data almost equally well among themselves. To gain statistical robustness in our inference, instead of relying on a single high scoring structure, we performed model averaging on the compendia of high scoring networks [66]. This resulted in an averaged network, consisting of common features (arcs), on which most of the high scoring network structures agreed. The final inferred network consists of arcs of confidence 85% or greater (that is, appeared in at least 85% of the high scoring networks).

## 4.4 Robustness analysis

As with most modeling approaches, this modeling effort required a selection of "algorithm parameters": a choice of discretization algorithm, number of discretization states, type of score used in the score-based search, and other parameters. Furthermore, we recognize that our data is sampled from a distribution of possible cells and cell states, and that technically, our results could be a the result of a fortuitous sampling of data that led to familliar connections. To assess the contribution of these and other factors in the final model results, a thorough robustness analysis must be undertaken. Such an analysis is beyond the scope of this thesis, but would constitute an interesting project for future work (such results might be interesting to the Bayesian network community at large).

Although we did not do a thorough robustness analysis, we did do several analyses,

in order to be confident of our model results. In the following section, we present results showing what happens when a slightly different sample is drawn from the distribution of cells; we investigate the affect of varying the number of discretization states, and finally, we try out a simple approach to discretization, in place of the more sophisticated algorithm we originally employed.

### 4.4.1 Bootstrap analysis

Our data is a finite sample drawn from a distribution, and our modeling results are reflective of the dependencies found in this sample. It is possible then that our modeling results contain dependencies that are random fluctuations found in this particular sampling of the distribution. In other words, the results might be overfit to the data. We discourage overfiting by having a *relatively* large dataset, by using the Bayesian score, which penalizes complex models, and by averaging over high scoring models. Nevertheless, certain random signals may still emerge in our results.

To ensure that our model does not contain 'chance' connections randomly introduced by the distribution sample, we perform a *bootstrap* analysis, in which we can better estimate the sampling distribution by resampling with replacement from the original sample. To do this, we randomly sample a new dataset from the original dataset, and perform the search on it. Each sampled dataset contains 4860 datapoints, 90% of the original dataset (i.e. 10% of the cells are excluded). We created ten such sampled datasets and inspected the results to assess consistency with the original model results. Two representative results are shown in Figure 4-13. The bootsrap results show very good consistency with the original model, with the main point of sensitivity appearing to be the confidence cutoff for the edges. This point is one that appears repeatedly in our various robustness analyses.

### 4.4.2 Impact of the number of discretization states

In this work, the data was discretized into three states. In general, employing a greater number of states provides additional flexibility in the expressivity of the model, but

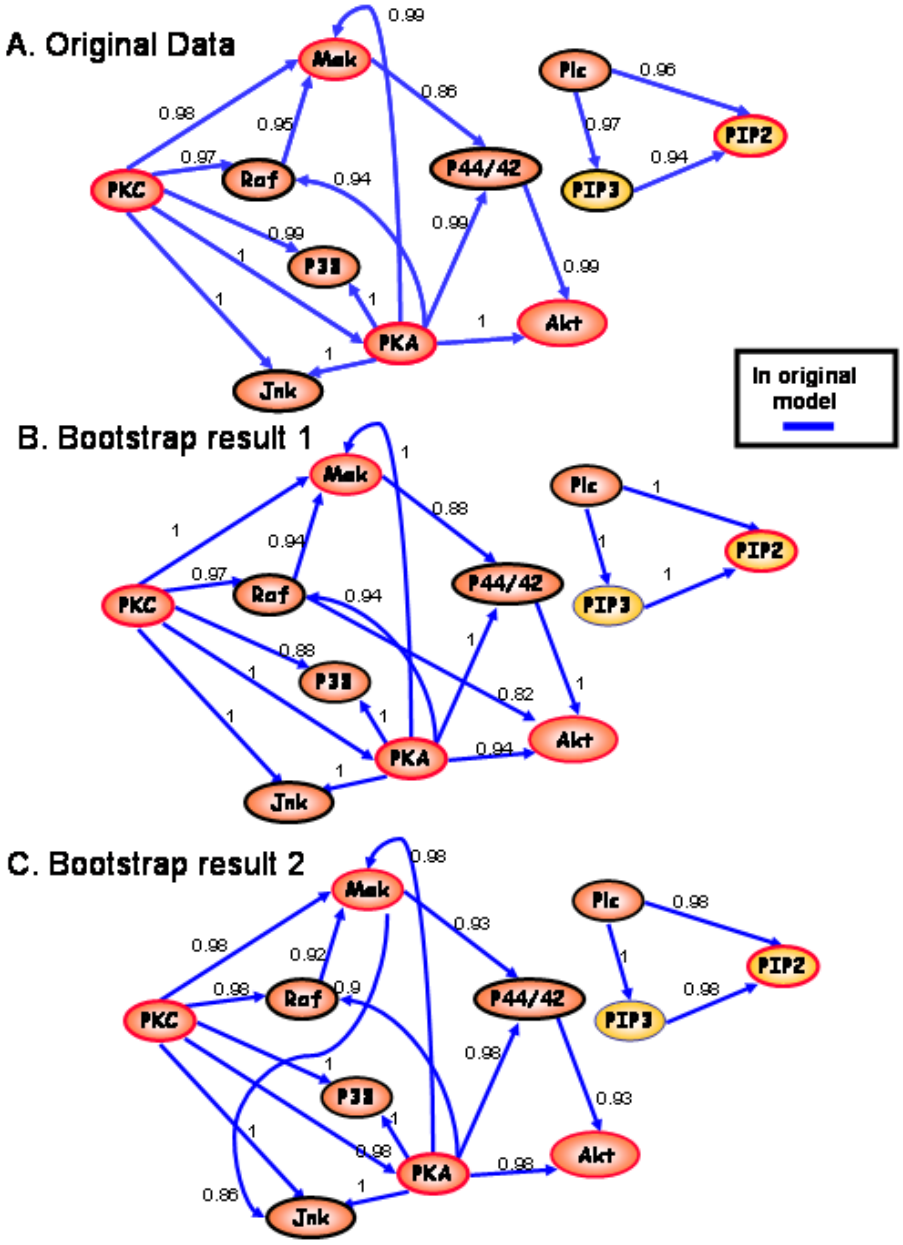



Figure 4-13: **Two representative results from ten independent bootstrap experiments** Panel A includes original results, including complete dataset, for ease of comparison. Panels B and C show the averaged search results for two independent bootstrap datasets, in which 90% of the original data is sampled randomly. Panels B and C closely resemble the original results, indicating that the results are robust to resampling of the data. In cases where an edge appears that is different from the original results, it is always one that appears in the original results, but did not make the particular confidence cuttoff employed. For example, the Raf⟶Akt connection which appears in panel B is one that appears as a lower confidence edge in the original results (see Figure 4-9).

simultaneously increases the number of parameters. One must tradeoff between the benefits afforded by a model of increased complexity, and the tendency of such a model to overfit, especially given a limited dataset (our dataset is large by biological standards, but not by general standards of datasets employed in machine learning). The choice for the number of states was partially biologically motivated- it seemed reasonable to assume that molecule states can be binned into three different levels (low, medium and high), whereas two seem insufficient, while greater than three (we hoped) were not necessary. The second motivation was a computational one. The discretization approach allowed us to assess how much mutual information would be lost with each reduction in the number of levels employed. Three levels allowed us to retain much of the mutual information in comparison to two levels; increasing to greater than three yields a relatively slight increase in the amount of mutual information retained (see [25] for a discussion of selection of the number of discretization states).

Using the same approach as that described in Section 4.3.2, we discretized the (identically preprocessed) raw data into two and four states, to assess the impact of fewer or greater number of states, respectively. Results are shown in Figure 4-14. In both cases, the model results are in good agreement with our original model results from three levels of discretization, though not as close agreement as seen in the bootstrap analysis. This is to be expected: discretizing to different number of states will establish new correlations and eliminate others.

Panel B shows the search results when the data is discretized to two levels. For a particular confidence cuttoff, it will tend to have more connections than the results from a larger number of discretization bins. This too is to be expected- a smaller number of levels will tend to encourage higher correlation coefficients. Most (but not all) of the edges in the 2-level model appear in the 3-level model, at some confidence level. A few of the edges, however, do not appear even in the low confidence edges of the low confidence model (see Figure 4-9), such as PKC$\longrightarrow$Plc$\gamma$ (interestingly, the reverse of one of our missing edges, see Figure 4-6), Mek$\longrightarrow$Plc$\gamma$ and Mek$\longrightarrow$p38. Interestingly, a couple of edges are reversed (Plc$\gamma\longrightarrow$Akt and Akt$\longrightarrow$p44), likely due

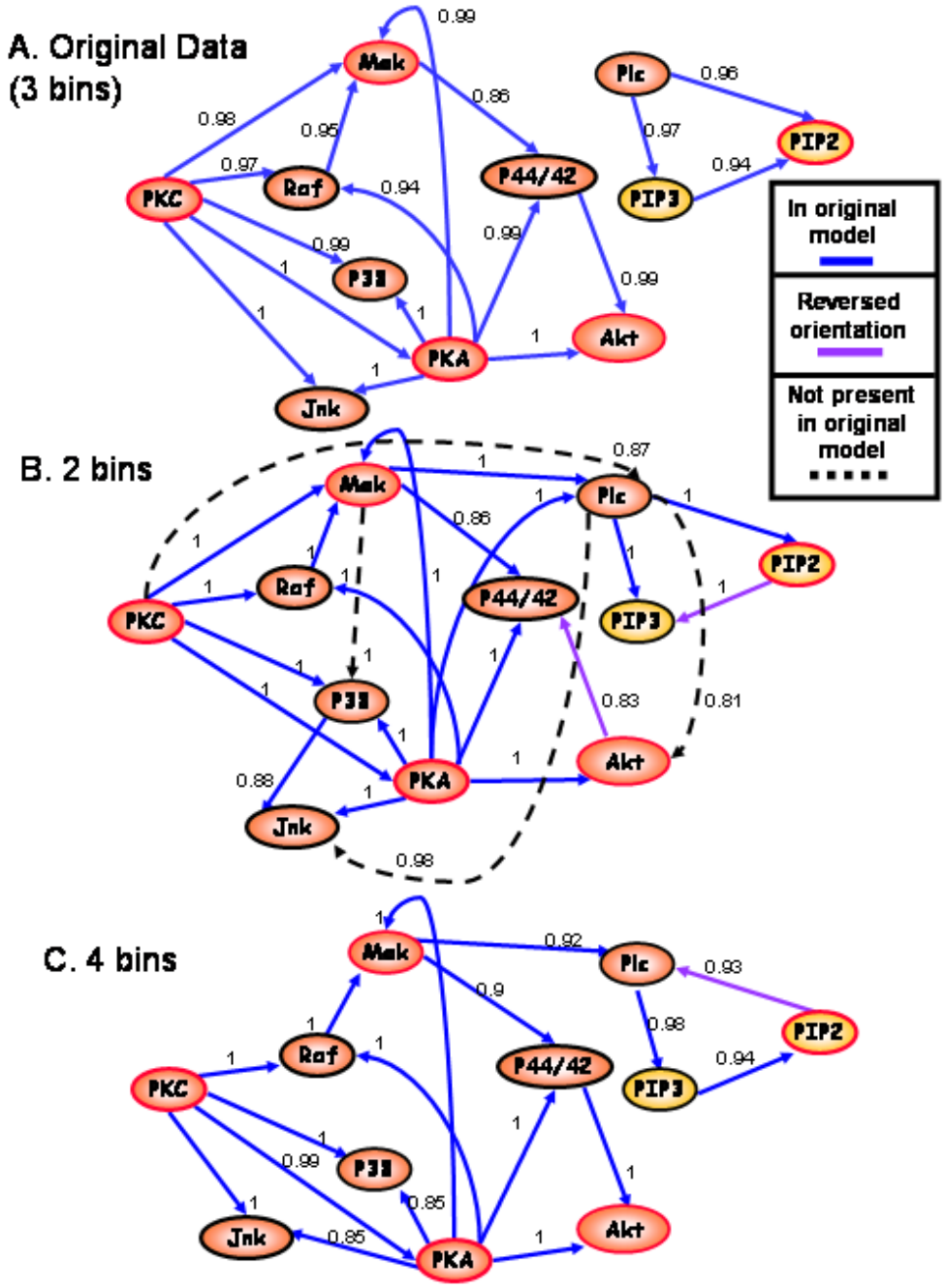



Figure 4-14: **Models resulting from varying the number of discretization states.** Panel A shows the original data for ease of comparison. In the original data, the data was discretized into 3 levels. Panel B shows the search results when data is discretized into 2 levels. Blue edges are present in original model, purple edges are present, but reversed in orientation, and dotted edges are not present in original model. Although results are similar to those in panel A, it is evident that the coarser gradiation of the data enables more numerous connections. Panel C shows search results for data in which variables are discretized to 4 levels. It too shows good agreement with the original results, though it contains fewer low confidence edges.

to a shift in the background distribution relative to the distribution under intervention (see Section 2.4.2 for an explanation of edge orientation using interventional data). As before, the choice of cuttoff value for edge confidence can have significant impact on the results, in both the 2 level and the 4 level case. The 4-level results, shown in Panel C, are closely matched to the original results. These results include very few low confidence edges when compared to the original 3-level model- reducing the confidence cuttoff would add just two more edges (not shown), whereas the 3-level model contains almost a dozen edges between the confidence values of 0.5 and 0.85.

### 4.4.3 Interval discretization

Discretization is a lossy procedure, in the sense that information is lost when the activity of a protein is expressed as 3 levels rather than the original continuous values found in the raw measurements. However, discretized data was appropriate in our application (partially because we wanted to use multinomial tables for the conditional probability distribution representation), so a discretization procedure had to be selected.

Our goal was to elucidate influence connection from the data. For this reason, we selected a discretization algorithm that sought to minimize the amount of mutual information that was lost in the discretization (see Section 4.3.2 for a description). However, as part of our robustness analysis, we wanted to assess how different our final results would be if we used a different approach. Here, we evaluate the results using simple interval discretization (with 3 levels).

In interval discretization, the interval over which the raw measurements lie is divided into 3 equal intervals, and the levels are assigned according to where a particular cell lies. (The intervals may contain a different number of cells- dividing the interval into peices which each contain the same number of cells is called quantile discretization, and it is a discretization approach which we have not tried). This is done separately for each variable.

The results of iterval discretization are shown in Figure 4-15. As in the previous analyses, the general model structure is retained. However, there is a reversal of

certain links (i.e. Raf⟶Mek and Plc$\gamma$⟶Pip2) and certain links do not appear (e.g. PKA⟶Jnk).

### 4.4.4 Discussion

We performed a series of analyses to assess the robustness of our results. We first performed a bootstrap analysis, which indicates that our results are robust to re-sampling from the distribution, and that our results are not due to a fortuitous data sample that happened to contain known dependency links. The other two analyses assess the impact of specific choices that we made in our algorithm and, importantly, give an indication of how differently our results would turn out had we chosen a different discretization approach, etc. These results- both for the variable number of discretization levels, and the interval discretization- are largely consistent with our original results, especially in the sense that the basic model structure is preserved. However, individual arcs may be lost or added, or, more commonly, reversed, as 'modeling parameters' are changed. The more coarse-grained discretization yields increased variable to variable correlations (the 2-level discretization result), increasing the number of arcs; conversely, the more fine grained discretization (4 levels) reduces correlations, yielding a more sparse model. As previously noted, the reversal of edges is likely due to a difference in the background distribtuion *relative* to the perturbation data distribtuion, which impacts the directionality of edges.

It is not surprising that altering the discretization will affect the final results- after all, the inference algorithm encounters a (potentially very) different dataset as a result of altering the number of levels or the discretization approach. A different dataset can be expected to contain different sets of dependencies, which the inference algorithm detects, yielding differences in the modeling result. Nevertheless, the existence of differences leads to doubts regarding the specifics of a particular modeling results, and begs the question: What is the *correct* discretization approach? (or the correct number of levels, etc.) The answer is that the correct approach is that which allows us to correctly and completely capture the underlying dependencies present in the data. A priori, we do not have a clear way of selecting the ideal approach. At best,

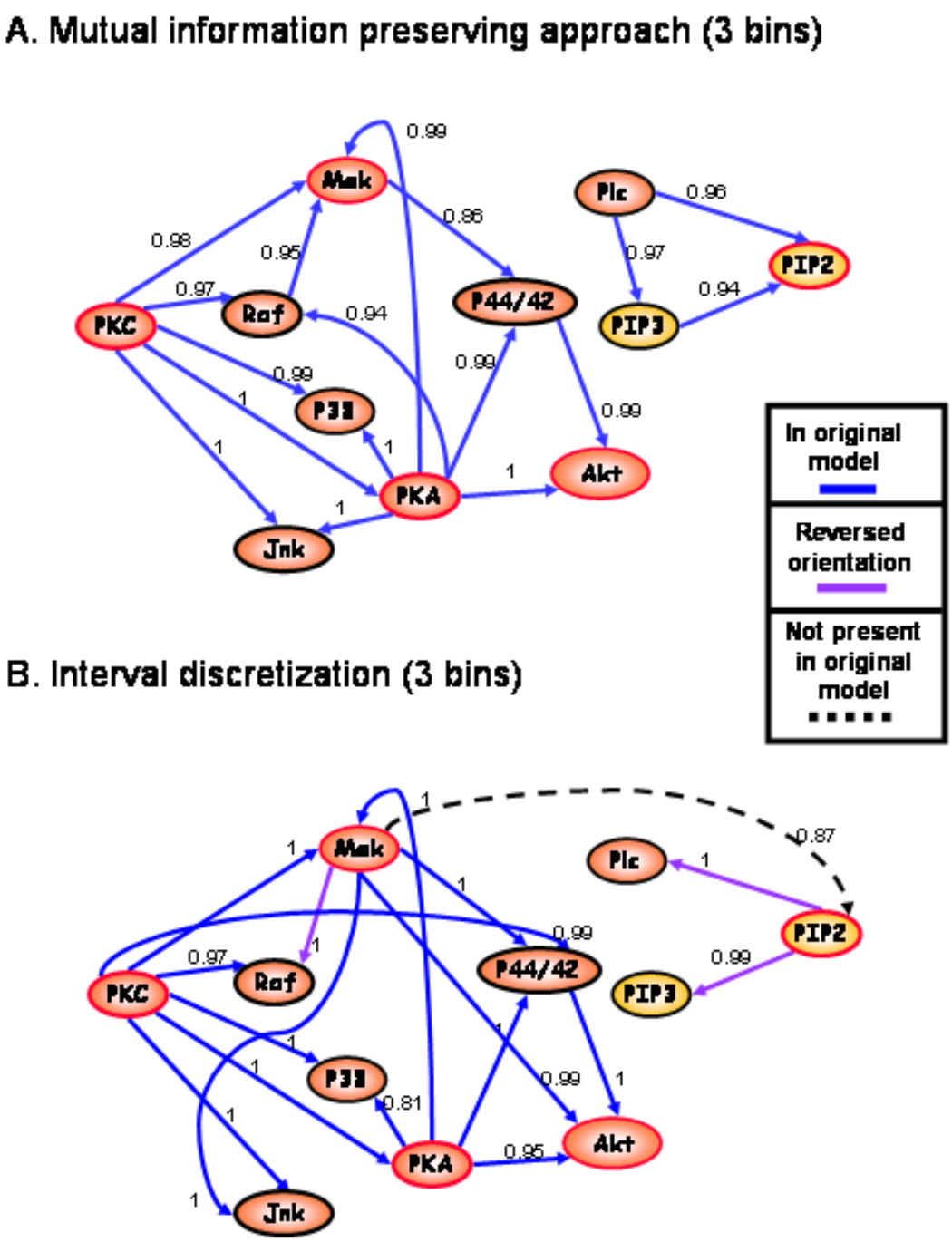



Figure 4-15: **Comparison of interval discretization with mutual-information preserving discretization.** Panel A shows the original results, in which the data were discretized using a mutual-information preserving approach (see Section 4.3.2) Panel B shows the results for data discretized into three levels using interval discretization. Blue edges are present in original model, purple edges are present, but reversed in orientation, and dotted edges are not present in original model. A comparison between them reveals several reversed arrows and other small differences. However, the basic structure of the model is retained.

we we can choose one that appears appropriate based on our biological knowledge in the domain (though this is an intuitive rather than a subjective decision) and/or that fits our computational constraints (such as, e.g., avoiding a large number of model parameters by limiting the number of levels in which each variable can exist).

Thus, although our analyses indicate good model robustness, they also indicate that given a particular model result, we can not be overly confident about each particular edge, or about the directionality of a particular edge. Furthermore, a close examination of edge confidence levels reveals that the final results are quite sensitive to the particular confidence cutoff employed, making the separation between a low confidence edge (that doesn't make the cutoff) and a slightly higher confidence edge (that does make the cutoff) appear somewhat arbitrary.

Taken together, these facts emphasize the need to regard Bayesian network inference of signaling pathway structure as in silico hypothesis generation, rather than a definitive elucidation of pathway structure. Thus, learned connections must be verified. Based on our results shown in this section, the general form of the model structure is highly robust, indicating it is a reliable source of hypotheses, hopefully meriting the effort and expense of wet lab verification.

## 4.5 Discussion and Summary

As shown, we correctly reverse-engineered and rapidly inferred the basic structure of a classically understood signaling network that connects a number of key phosphorylated proteins in human T cell signaling-a map built by classical biochemistry and genetic analysis over the last two decades. The network was automatically constructed with no a priori knowledge of pathway connectivity. Application of Bayesian networks to single cell flow cytometry has distinct advantages, including an ability to measure events in primary cells after in vivo interventions (thus measuring context specific signaling biology in tissues), inference of directed arcs and causality therein, and the ability to detect indirect as well as direct connections. This latter point is a powerful feature when the known list of participating molecules may not be exhaus-

tive, and can be especially important when networks are used to assess the effects of system perturbations (as in a pharmaceutical context). A limiting step in the experimental aspect, clearly, is the availability of suitable reagents; currently, there are about 80 antibodies to phosphorylated molecules compatible with flow cytometry, but this number is expected to rapidly increase. [70]

Application of this approach to other sets of molecules, cell types, disease states and interventions (e.g., siRNA and dominant negative screens, or pharmaceutical agents) should enhance our understanding of signaling networks, especially with respect to complex, nonlinear cross-talk between pathways. Another important experimental issue which this approach can address is the differences among specific primary cell types and even cell subpopulations. The traditional understanding of pathway structures as collated from diverse model cell types and organisms demonstrates the essential congruity of basic signaling networks, but does not easily reveal the subtle differences that exist in different primary cell subtypes. This report demonstrates that it is now possible to appreciate pathway intricacies in primary cell subsets-even with previously uncharacterized signaling molecules. Application of this approach during biochemical interrogation of cellular subset-specific signaling networks in the course of disease state or in the presence of pharmaceutical agents can potentially provide important mechanistic information of clinical relevance. This method could identify sets of signaling molecules that explain differences between responses to chemotherapy in patients with cancer. [39]

Concerning the computational aspect, a key advantage of Bayesian networks is that these are relatively robust to the existence of unobserved variables, for example their ability to detect indirect influences via unmeasured molecules. At the forefront of Bayesian network research is development of methods to automatically infer the existence and location of such hidden variables. Although the current report was restricted to 11 phosphorylated molecule measurements per cell, the number of simultaneous parameters measured by flow cytometry is steadily growing [70] . As measurement systems improve, and the ability to readily and accurately measure greater numbers of internal signaling events increases, additional opportunities to

discover novel influences and pathway structures become possible.

One of the caveats in the use of Bayesian networks for the elucidation of signaling pathways is that these are restricted to be acyclic, where as signaling pathways are known to be rich in feedback loops. Indeed, our inference missed four classic arcs, most likely for this reason. Given time series data, Dynamic Bayesian Networks could potentially capture these feedback loops. To measure the amounts of internal phosphorylated proteins, the cells must be fixed and therefore continuous real-time simultaneous, multi-parameter single cell time series data can not be collected with the current technology. Since Bayesian networks belong to a more general class of probabilistic graphical models, within the formalism of these models it is possible to develop a novel model that could handle feedback loops given a series of static time points using the current technology.

While there is much to be developed both computationally and experimentally, by extending the concepts derived here, it is clear that simultaneous, multivariate analysis of biological states in multiple discrete entities, such as cells, offers a clear advantage for rapidly deriving signaling network hierarchies and structures. Extension of this approach to biological systems involving multiple cell populations, even solid tissues and organs, or whole animal studies such as in whole body fluorescence imaging of phosphorylation states in staged *Caenorhabditis elegans* or Drosophila larva, or thin-slice tissue sections from mammalian organs, could allow automated construction of signaling network influences not only within, but also across cell boundaries in ever more physiological contexts.

# Chapter 5

# Learning larger networks using measurements of overlapping subsets

In the previous chapter, we show that a map of influence connections can be inferred from simultaneous measurements of signaling proteins in single cells. We emphasize the necessity of simultaneous measurements, crucial for finding correct edges, as well as excluding superfluous ones. However, because of the limitations of the measurement technology, as our set of variables of interest grows, it becomes increasingly hard (or impossible) to measure all the variables simultaneously. Therefore, in order to build models containing more variables than the number we can measure simultaneously, we must relax our requirement for simultaneous measurements. In this chapter, we describe a method for scaling up the number of variables in the inferred model, by measuring overlapping subsets of the full variable set in multiple experiments.

This chapter has not been previously published. Readers interested in this work should refer to the publication, which will contain the final form of this algorithm and results.

## 5.1 Introduction

A systems biology approach often relies upon an innovative, high throughput data measurement technology. In our effort to map signaling pathways, our data requirement is especially ambitious: we require simultaneous measurement of all signaling proteins of interest in each individual cell profiled, under various general and specific stimulus and inhibitory conditions. Multidimensional flow cytometry is a uniquely information rich and statistically robust data source, and its capacity to measure an increased number of variables is steadily growing. Nevertheless, the current standard capability for simultaneous measurements is ~4 molecules at a time, while the cutting edge is ~12 molecules. In either case, this is an insufficient number to enable one to address most pathways of interest, as even narrowly defined canonical pathways can include dozens of molecules.

Since we are interested in models of a biologically-relevant scale, our goal is to expand the number of variables in a model beyond the measurement capability of the technology (in this case flow cytometry, but this work applies equally well to any other technology with a limitation on the number of variables that can be measured simultaneously). Let us define the number of molecules that the technology can measure simultaneously to be $m$, and let us call the total number of molecules we wish to model $n$ (our problem is only relevant for $n > m$). Our question is, Can we scale up from measurements of $m$ variables to models of $n$?

In this chapter, we describe an approach for scaling up from a measurement capability of $m$ molecules to models of $n$ molecules. We employ an active learning method in which each experiment provides information that allows us to constrain future experiments and ultimately to constrain our search in such a way that we can relax our requirement for simultaneous measurements. The intent is to enable a ~10 color capability to scale up to models of 50-100 molecules (and, in the future, for an ever-increasing measurement capability to scale up to models of hundreds of variables). In the meantime, we use available 11-color data to demonstrate this approach on a sample problem in which $m = 4$ and $n = 11$, the problem encountered by a lab

with standard flow cytometric capability (4 colors) that wants to build an 11-variable model. To be applicable to real-world applications, our approach must contend with two issues: first, it must enable us to model more molecules then we can measure simultaneously; second, it must be able to do this while minimizing the number of required experiments. In the case of $m = 4$ and $n = 11$, measuring all possible overlapping subsets necessitates $\binom{11}{4} = 330$ experiments, clearly a requirement that is prohibitively expensive and time consuming. Therefore, this algorithm is created with two goals: to model the variables as accurately as possible, while requiring as few experiments as possible. We find that in our domain, we are able to closely reproduce the results from a dataset in which all 11 variables are measured simultaneously, using ˜15 measurements of 4 molecules each.

## 5.2 Approach

In this section, we describe the method, including a general overview as well as the specifics of the implementation. Although we apply this method to scale up from 4 molecules to 11, the same approach can be used for different values of $m$ and $n$ and, as alluded to previously, the motivating application for this algorithm is large scale models (dozens-hundreds molecules, increasing as the measurement capability increases).

### 5.2.1 Overview

A cell's complete signaling network is highly interconnected, and distant causal links (separated by numerous other molecules) probably exist between many pairs of proteins. However, as the connection between a protein and its ancestor becomes more distant, the causal effect of the ancestor fades, as does the statistical dependence between child and ancestor. Out of the complete signaling network, therefore, some variables will be largely irrelevant to the activity of certain others. We will use this fact to constrain our search. For each variable, we will find those variables with which it is more intimately connected (relative to all the molecules measured), and consider

only those as potential parents and children in our model structure search.

How does this procedure work? First, we measure each molecule with each other molecule (pairwise). We use this data to define *correlation neighborhoods*, defined as the set of those variables that have sufficiently high correlation to each other to enable us to consider them as potential parents, children, close ancestors or descendants of each other. Next, we expand this neighborhood with respect to individual molecules, by defining potential parents or (possibly distant) ancestors, based on a molecule's response to specific perturbation of other variables.

Based on these two steps, we have, for each signaling molecule, a set of variables that are in its correlation neighborhood, and a set of variables that are its potential parents (possibly more distant), based on perturbation data (these are "potential perturbation parents"). Note that the perturbation parents may or may not be in the molecule's correlation neighborhood. Together, these two sets constitute a molecule's '*extended neighborhood*'—the set of all variables that we allow as potential parents or children of the molecule. See Figure 5-1 for an illustration of correlation and extended neighborhoods, and Figure 5-2 for an overview of this approach.

We use the extended neighborhoods to constrain our search, such that in the random search, only graphs in which *permitted* parent/child sets are proposed are considered; other graphs are rejected offhand. A permitted parent/child combination is one in which the parent and child are in each others' extended neighborhood (or more precisely, one in which the proposed parents of a child are in its extended neighborhood). Note that since correlation neighborhoods are based on a symmetric metric (correlation), then if variable $X$ is in $Y$'s correlation neighborhood, then $Y$ is necessarily in $X$'s correlation neighborhood. However, it is not transitive, so for the same $X$ and $Y$, if $Z$ is in $Y$'s correlation neighborhood, this does not necessarily imply that $Z$ is also in $X$'s correlation neighborhood.

Constraining the search space limits the number of variables with which each variable must be measured simultaneously, and leads to a substantial saving in the number of necessary simultaneous measurements. However, each *permitted* parent set/child combination must still be measured simultaneously. How is this achieved?

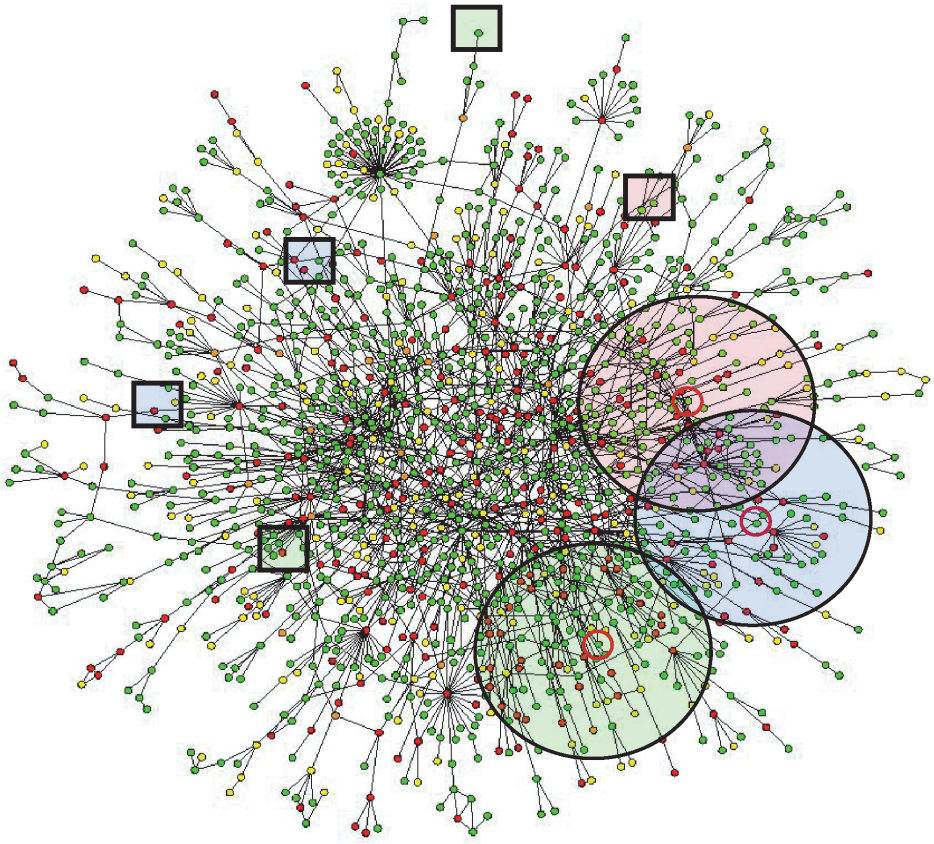


Figure 5-1: **Pictorial illustration of correlation and extended neighborhoods.** Within the complete signaling network in a cell, a particular signaling molecule (circled in red) is likely to exist within a particular neighborhood of other molecules (shaded circle). The connections in the network indicate a physical or mechanistic interaction, which may be accompanied by a statistical correlation. We assume that molecules which affect each other will show some correlation, and that closer interactions will, in general, show higher correlation than farther ones. A molecule may also have more distant, poorly correlated ancestors that may be detected using perturbations (shaded squares).

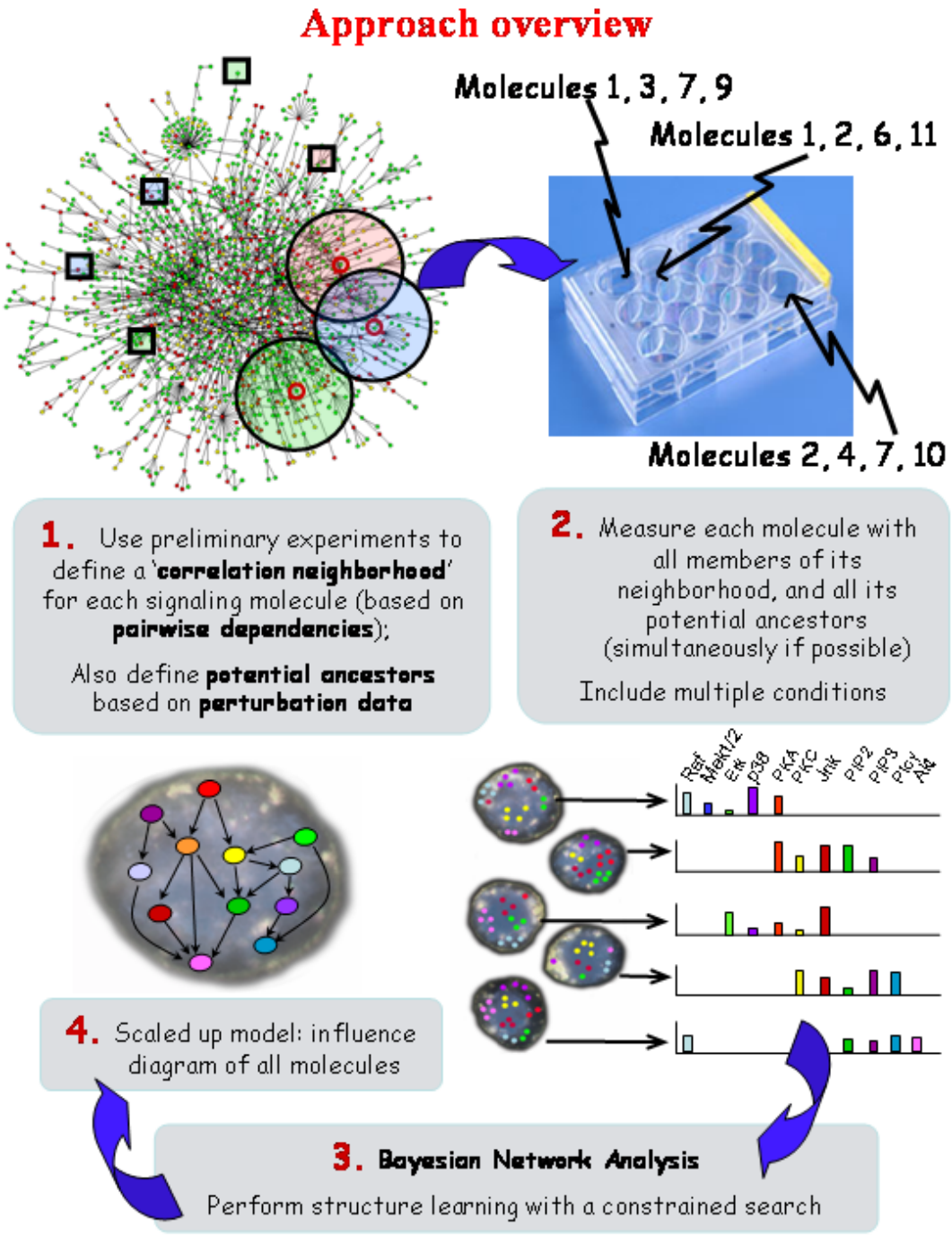



Figure 5-2: **Flow diagram of our approach.** 1. Initial experiments define correlation and extended neighborhoods. 2. Further experiments are selected based on the initial experiments. 3. Structure learning is performed with an implementation that constrains the search according to extended neighborhoods (as detailed in the text). 4. Resulting model includes all variables, even though the set was not measured simultaneously.

If the number of variables in an extended neighborhood does not exceed $m-1$, then the variable and all its extended neighborhood can be measured simultaneously. This is the simplest case. In our example, because $m = 4$, this means a variable must have no more than 3 members in its extended neighborhood, a situation which is hard to guarantee. To stretch our capabilities further, we make a biologically motivated assumption pertaining to the perturbation parents. Perturbation parents, which are *not* in a potential child's correlation neighborhood, are not well correlated with the child (by definition of correlation neighborhood). Therefore, we assume that it is necessary to observe the dependence of the child on the perturbation parent *only in that condition in which the potential perturbation parent is perturbed*. In that condition, it is actually not necessary (and sometimes not possible) to measure the potential parent anyway; rather, an inhibited parent is assumed to be at level 'low', an activated parent is assumed to be at level 'high' (see Section 4.3.2 in the previous chapter). Therefore, it is sufficient to measure the potential child in that condition, without measuring the potential perturbation parent.

In practice, this translates to the following situation. If a variable has more than $m$ variables in its extended neighborhood (counting itself), the perturbation parent (or parents) which are *not* in its correlation neighborhood are not measured simultaneously but instead are appended on, their values taken from an independent experiment. (See Figure 5-3 for an illustration.) In other words, for each treatment condition, the values of the perturbation parent are kept consistent with its distribution under that condition, but any specific dependence between that perturbation parent and its potential child are lost or minimized, except under the condition in which the parent is perturbed. This approach will work if at least one of the following conditions hold: either the dependence of the child on the parent is only sufficiently detectable under the perturbation condition anyway (this may be the case, especially for a distantly connected ancestor), or the parent and child are each sufficiently homogenous under a particular condition, such that dependencies can be preserved even when each are sampled independently. Since we know the parent and child are not well correlated, the former condition is likely to be satisfied. The latter condition is

probably more variable, and for this reason we do not rely on it. In cases where it is true, this approximation should work particularly well.

So far we have handled the case in which an extended neighborhood includes up to $m$ members, and the case in which the correlation neighborhood includes up to $m$ members, and extended neighborhood members are appended on from independent experiments. What about the case in which a correlation neighborhood contains more than $m$ members? This is a truly problematic case, because in this case it is not possible to observe the child with all of its possible parents simultaneously, and we have no assumptions available that enable us to ease the requirement for simultaneous measurement. In this case, we have no choice but to permit parent set/child combinations which include the child with *up to* any set of $m-1$ variables out of its correlation neighborhood. In the case where the variable is in fact regulated by a complex function involving more than $m-1$ parents, our algorithm will fail to find this relationship.

There are a number of ways in which we minimize the number of required experiments. First, we only require each variable to be measured with each other *pairwise*, neglecting the measurement with each other tuple. This dramatically reduces the number of experiments required to initially assess correlations (in the $m = 4, n = 11$ case, from 330 to about 12-13). We then take one further step to reduce this number. We assume that if variable $X$ and variable $Y$ are highly correlated (correlation above some minimal threshold, higher then that required for membership in the correlation neighborhood), then a third variable $Z$ which is *very weakly* correlated with $X$ (below some maximal threshold) will not be highly correlated with $Y$. In this case, $Z$ need not be measured with $Y$, as it is automatically excluded from $Y$'s correlation neighborhood. We call this assumption *approximate transitivity of correlations*. See Figure 5-4 for an illustration. This assumption reduces the number of experiments, in our domain, to ˜7-8 (from ˜12-13).

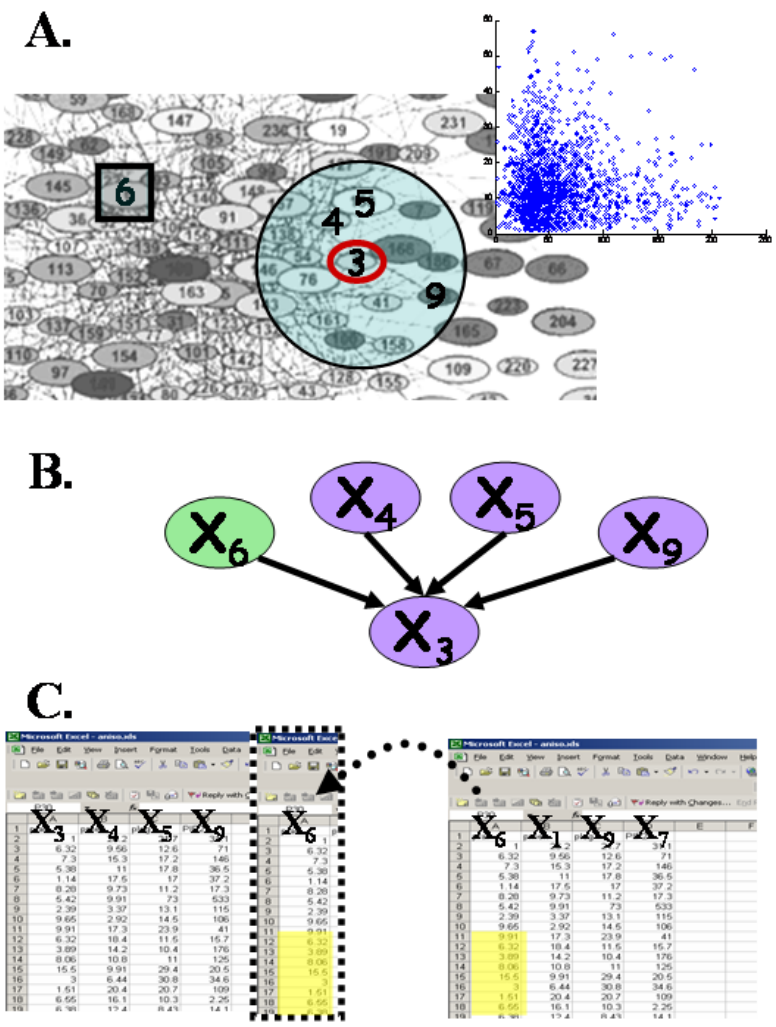



Figure 5-3: **Appending perturbation parents which cannot be measured simultaneously.** *Panel A.* Variable $X_3$ has 3 other variables in its correlation neighborhood, as well as a potential perturbation parent, $X_6$. $X_6$ is *not* in $X_3$'s correlation neighborhood, as they are not well correlated (inset). *Panel B.* Possible search query in which $X_3$ is the child of all its potential parents. For $m = 4$, it is not possible to score this local conditional probability distribution by measuring all variables involved. *Panel C.* To score this query, data in which $X_3$ and its correlation neighborhood were measured is supplemented with data *from a separate experiment*, in which $X_6$ is measured. $X_6$—$X_3$ dependence will not be preserved in the 'background' distribution, unless the levels of $X_6$ and $X_3$ are fairly homogenous within a particular experimental condition. However, in the $X_6$—perturbation condition (yellow shaded area), the level of $X_6$ is known because it is determined experimentally (by the perturbation). Therefore, under this condition, the $X_6$—$X_3$ dependence is observable. When we use this approach, we are making the assumption that this perturbation—condition dependence is sufficient for the perturbation parent to emerge as a parent in the learned Bayesian network structure.

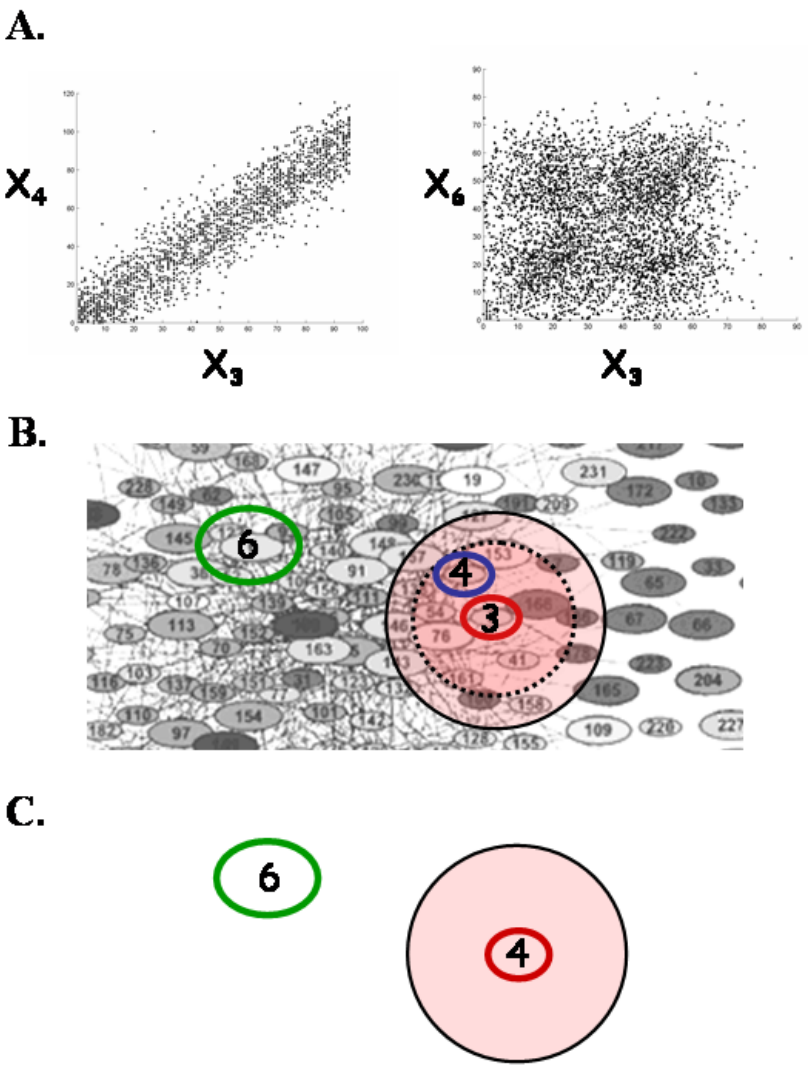



Figure 5-4: **Assumption of *approximate transitivity of correlations* is used to eliminate certain measurements.** *Panel A.* Variables $X_3$ and $X_4$ are very well correlated, variables $X_3$ and $X_6$ are very poorly correlated. *Panel B.* In accordance with their correlations, $X_4$ is in the correlation neighborhood of $X_3$, while $X_6$ is not. The dotted inner circle depicts the 'inner neighborhood' of variables that are particularly well correlated with $X_3$ (the cutoff for these highly correlated variables is more stringent than the cutoff for membership in the correlation neighborhood. See Section 5.2.3). *Panel C.* Because of the correlation between $X_3$ and $X_4$, and the lack of correlation between $X_3$ and $X_6$, we *assume* that $X_6$ is not in $X_4$'s correlation neighborhood, *without measuring* $X_4$ *and* $X_6$ *together*.

### 5.2.2 Assumptions and Limitations

In the previous sections, we made statements that encompass some assumptions of our method, and alluded directly to several others. Here, we explicitly state the assumptions used in the development of this heuristic, approximate approach. Although these are necessary to enable us to model a number of variables that exceeds our measurement capability, we attempted to keep our assumptions biologically motivated and biologically reasonable. The success of our approach will depend heavily on the legitimacy of our assumptions.

Assumption *There exists an independent component to each parent/child relationship*

Because we base the correlation neighborhoods on pairwise data, we implicitly make the assumption that the dependence of a child variable on each parent can be detected independently. This results from a pairwise dependence, and is detectable when the child is observed with the particular parent in isolation. With this assumption, we neglect any complex relationship that lacks an independent component between the child and an individual parent, such as an XOR relationship. We neglect this type of relationship out of necessity, as including it would necessitate performing far more experiments (in order to consider each variable with each other tuple). Although this assumption is not ideal, we note that most biological relationships are *not* strictly nonpairwise relationships, such as XOR. Typically, even in cases of complex, multi-parent combinatorial regulation, each parent maintains an independent influence relationship component with each child. Additionally, as the number of experiments grows, or as the number of measurable variable grows, this assumption can be relaxed.

Assumption *Correlation weakens with distance*

We constrain our search using correlation neighborhoods, which are defined based on the strength of correlations. Thus, we assume that relationships demonstrating weaker correlation are less important, in that they are less likely

to be closely linked ancestors or dependents. If we define a distance metric as the number of links connecting between two proteins (i.e. how many signaling molecules are needed to transfer the affect from the parent to the child), then we are assuming that the more distantly connected variables have a weaker correlation than closely connected ones. This assumption is not valid when a variable has a close but noisy (and therefore weakly correlated) connection. If the noisy connection is 'outranked' by other, higher correlation connections, it will not be included in the correlation neighborhood and therefore the connection will not be found. As with the previous point, as the number of experiments and/or the measurement capability increase, this limitation will be diminished.

Assumption *Signal in perturbation condition is sufficient to detect dependence of child on perturbation parent*

As explained in the previous section, when the number of variables in an extended neighborhood exceed $m$, any perturbation parents which are not in the correlation neighborhood are not measured simultaneously with the other neighborhood members; their values are taken from an independent experiment. A specific statistical dependency cannot be established in the nonperturbed condition; however, in the perturbed condition, one can be established because the value of the perturbational parent is known even though it is not measured (since it is experimentally fixed). This assumes that the contribution of the perturbation parent is felt primarily in the condition under which it is perturbed, and that the background distribution can be sufficiently estimated from the overall distribution of both parent and child, in the other conditions. It is straightforward to assume that the perturbation parent's contribution is not strong in the nonperturbed conditions: since it is not part of the correlation neighborhood, then by assumption 1, it is not a strong predictor of the child. However, this assumption can be problematic when the background (nonperturbational conditions) is altered by measuring the child and perturbation parent separately. This is because edge orientation is determined by contrasting the

conditional dependence of the child on the parent in the perturbation condition to that in the background conditions (See Chapter 2). Implicit in this assumption is the idea that for a particular condition, both the child and the parent have a sufficiently homogenous distribution, such that they can each be measured separately without dramatically changing their dependence. When this assumption is satisfied, then we are able to orient edges in the same way as we would with a complete, fully simultaneously measured dataset. Otherwise, it can be expected that edges will persist, but their orientation might be reversed.

Assumption *Node indegree does not exceed* $m - 1 + p$

Where $p$ is the number of (uncorrelated) perturbation parents. Our approach is limited by the constraints of the measurement technology. Though this limitation is important for small $m$ values, it quickly becomes insignificant as $m$ grows. Today's cutting edge technology ($m \approx 12$) allows an indegree of at least 11, certainly more than sufficient for most signaling molecules.

Assumption *Approximate transitivity of correlations*

Let us define the correlation cutoff for inclusion in a correlation neighborhood to be $c$, and let us define a higher value, $h$, as well as a lower value, $l$, such that $h > c > l$. Then for any pair of variables $X$, $Y$ with correlation greater than $h$, if there exists a third variable $Z$ with correlation less than $l$ to $X$, then it is assumed that it will not be sufficiently highly correlated with $Y$ in order to be included in its correlation neighborhood; in other words, the correlation between $Z$ and $Y$ is assumed to be less than $c$. (In Figure 5-4, Panel B, the inner dotted circle depicts variables with correlation $h$ or higher, the outer circle depicts the correlation neighborhood, or variables with correlation $c$ or higher. The variable indicated with a green circle ($X_6$) has correlation $l$ or less to $X_3$.) While this assumption appears reasonable, it too can be relaxed if the resources to perform a greater number of experiments are available.

### 5.2.3 Details of implementation

**Defining correlation neighborhoods** For each variable, its correlation neighborhood was defined to be the set of variables which were correlated with correlation coefficient $> 0.5$ under *any* condition. We selected these criteria to be nonstringent to reduce the likelihood that connections would be missed. Additionally, we wanted to ensure that a parent which had even a weak individual relationship component would have a good chance of being included. Conditions were considered separately because it is possible for a parent to be predictive of a child only under particular stimulations (e.g. when the cell is in a particular state or when certain signaling pathways are activated). For each particular domain, this parameter can be set in accordance with the correlation levels found among the variables, defining it more stringently in a highly correlated domain, and more laxly in a more loosely connected domain.

**Identifying perturbation parents for inclusion in extended neighborhoods** Each perturbation condition involved the inhibition or activation of a particular molecule (see previous chapter for details). In each such condition, we identified which molecules were affected by detecting a shift in the median of one half standard deviation when compared to the background distribution. Perturbation parents were then added to the list of correlation parents to form the extended neighborhood.

**Selecting experiments** As described above, we wish to perform experiments first in order to define the correlation neighborhoods and extended neighborhoods, and then in order to measure each variable with all members of its correlation neighborhood (and extended neighborhood), simultaneously if possible. For our first round of experiments, the variables in each 4-color set are selected randomly. The four variables in the first set are examined for levels of correlation- if any pairs are correlated with coefficient (actually the absolute value of the coefficient) exceeding 0.5, they are included in each others correlation neighborhood. If any are correlated with coefficient exceeding 0.7, they are considered 'highly correlated' and are used to eliminate future experiments (via the approximate transitivity of correlations assumption). This

works as follows: if in a future experiment a member of a highly correlated pair is found to have 'low correlation' (defined as a coefficient $< 0.2$ in *all* conditions) with another variable, then it is deemed unnecessary to perform an experiment in which the other member of the highly correlated pair is measured with the poorly correlated variable. This procedure is repeated until all pairwise combinations, with the exception of those that get eliminated in this fashion, have been performed.

After all pairs have been identified, additional experiments are performed so that all members of each correlation neighborhood are measured simultaneously, if possible (i.e. if the correlation neighborhood size does not exceed $m$). If not, each set of size $m$ from the correlation neighborhood can be measured. If this becomes cumbersome from the perspective of the number of experiments required (something that could easily happen if the size of the neighborhood is large compared to $m$), then it is possible to constrain the correlation neighborhood further, by using a more stringent cutoff for the correlation coefficient. Alternatively, it is possible to use prior biological knowledge to constrain the neighborhood. In this domain, no correlation neighborhood was larger than $m$, so we did not contend with this issue.

When possible, the perturbation parents are also measured simultaneously (i.e. the entire extended neighborhood together). If this is not possible, then perturbation parents which are *not* members of the correlation neighborhood are removed one by one, until the number of variables to be measured together no longer exceeds $m$. These perturbation parents are then appended by taking their values from a separate experiment. Removal of the parents from the measured set proceeds as follows: the parents whose perturbation caused a stronger response in the child are removed first (such that when it is only possible to measure a subset of the perturbation parents along with the correlation neighborhood, the weakest parents are those that get measured). This is because we anticipate that the stronger parents are more likely to demonstrate a sufficiently prominent affect to establish themselves as parents based only or primarily on the condition in which they are perturbed. In case a query should ask for just a subset of the perturbation parents, if the correlation set size is smaller than $m - 1$, the perturbation parents are also measured *individually* along

with the correlation neighborhood. This ensures that a query involving only a few perturbation parents will be able to use actual measured values, whenever possible.

**Constraining the search** Once the set of measurements have been performed, they are used to constrain the search. In each query of the random search (described in the previous chapter), if a parent set is proposed that has been measured with the proposed child, then that query is allowed to continue, and data from the relevant experiment is used to score this local interaction. If the query involves a parent set that has *not* been measured with the child, then it is allowed to continue only if the subset of the parent set that remains after the child's perturbation parents have been excluded has been measured with the child. That is, if the parent set includes only variables that have been measured with the child, and perturbation parents of the child, then the search is allowed to continue. When this occurs, the local interaction is scored using the measured set that includes all the nonperturbation parents, as well as as many of the perturbation parents as possible, along with values for the perturbation parents, taken from an independent experiment. Note that because we use the set of performed experiments, rather than the extended neighborhoods, to constrain the search, we are enabling connections that are not technically permitted. This could occur if, for instance, in the early experiments in which we survey each pair, a variable is measured with various other variables which are not in its correlation neighborhood. We employ this approach to enable maximal flexibility, just in case a potential parent is somehow missed, or in case our assumptions are imperfect. We also try a more constrained approach, allowing only members of the same extended neighborhood to influence each other.

**Making pseudo 4-color data** This work is designed as a proof of principle project, for which it is convenient to have a 'ground truth' model. In this case, we wish to use multiple overlapping 4-color experiments in order to build an 11-variable model. We use an 11-color dataset, on which we have already performed Bayesian network analysis (see previous chapter) to define our ground truth model. We then need to create pseudo 4-color data, in order to test our approach. We do this simply by constraining which sets of variables we permit in our search (see above). In the

case when we wish to append the values of a perturbation parent from an independent experiment, we simulate an independent experiment by taking the original values and randomly permuting them within each condition.

## 5.3 Results

In this section, we present results from three independent attempts at learning an 11-variable model from measurements of 4-colors each. We present both model results and an analysis of the number of experiments required to build the model in this particular domain.

### 5.3.1 Number of experiments required

A thorough analysis of the number of experiments required is difficult to perform, because the number depends heavily on the particular strength (and weakness) of correlations found in the domain. Strong correlations will help to reduce the number of experiments needed to cover all variable pairs, but may lead to a combinatorial increase in the number of experiments needed to cover all variables in the correlation neighborhood, unless $m$ is large relative to the correlation neighborhood size. Weak correlations in the domain do not enable a saving in the number of experiments necessary to cover all variable pairs, but may result in smaller correlation neighborhoods.

To assess approximately how many experiments would be needed, we start by asking how many experiments must be performed to measure each pair of variables in a set of size $n$, using measurement groups of size $m$. We are not aware of an exact solution for this problem. However, we can find an upper and lower bound. For a lower bound, we assume that in each measurement, we cover $\binom{m}{2}$ total pairs. Since we need to measure $\binom{n}{2}$ total pairs, this gives us a lower bound of $\frac{\binom{n}{2}}{\binom{m}{2}}$. This lower bound is not an exact solution, because after a few experiments, it is no longer possible to measure $\binom{m}{2}$ unique pairs in each experiment.

Let's see how this works. Consider a situation with $m = 4$ and $n = 10$. We start by measuring variable 1 with 3 other variables, let's say variables $2, 3$ and 4.

This gives: $Experiment1 = [1234]$. This experiment covers $\binom{m}{2} = \binom{4}{2} = 6$ pairs: 1 and 2, 1 and 3, 1 and 4, 2 and 3, 2 and 4, 3 and 4. Focusing on pairs with variable 1, we continue with $Experiment2 = [1567]$, which also covers 6 pairs, as does our next experiment: $Experiment3 = [18910]$. At this point, we have measured variable 1 with all possible other variables, and, additionally, we have also measured a substantial number of pairs consisting of other variables. If we could continue like this, we would need exactly the number of measurements in our lower bound. However, we soon run into a situation in which we cannot eliminate $\binom{m}{2}$ pairs in each experiment, because we cannot measure $m$ unique variables together (by "unique" we mean variables that have not been measured together previously). Consider our next experiment, focusing on pairs with variable 2. We have already measured 2 with $1, 3$ and 4, so our next experiment should draw from variables 5 and above. If we select $Experiment4 = [2567]$ then we are covering only 3 pairs, (2 and 5, 2 and 6, 2 and 7), because variables $5, 6$ and 7 have already been measured together (in experiment 2). We can choose more carefully, say, $Experiment4 = [2586]$, an experiment which covers 5 pairs (2 and 5, 2 and 8, 2 and 6, 5 and 8, 8 and 6), but we cannot find any combination of variables which includes 2 and covers a total of 6 unique pairs. This problem becomes more prevalent as we continue our measurements of all pairs.

We know that measuring all pairs requires at least $\frac{\binom{n}{2}}{\binom{m}{2}}$, but we also know, from the above description, that this number is too low. For an upper bound, consider dividing the space into sets of size $m/2$ (in our case, sets of size 2, say, variable 1 with variable 2, variable 3 with variable 4, etc). This yields $\frac{n}{m/2}$ subsets, each of size $m/2$ (in our case, $11/2 = 5.5$ subsets, which we round to 6). Now to make sure we include each variable pair, we measure each pair of subsets together (we can do this, because the subsets are each of size $m/2$). This requires $\binom{\frac{n}{m/2}}{2}$ measurements. For our problem, this is 15 measurements, not far from the actual number required, which appears to be 12. (This is not bad for an upper bound, especially considering we rounded the number of subsets up to 6) [1]. We thank Erik Demaine, MIT Computer

[1] We can also assess our upper bound by comparing to the lower bound. Our lower bound is $\frac{\binom{n}{2}\binom{m}{2}}{\approx} n^2/m^2$. Our upper bound is $\binom{\frac{n}{m/2}}{2} = \binom{\frac{2n}{m}}{2} = \frac{2n^2}{m^2} - \frac{n}{m}$, so it is $n/m$ less than twice the lower bound.

Science, for help with this formulation.

From this number, we can subtract the experiments that are eliminated due to the approximate transitivity of correlations assumption, which will depend heavily on the correlations found in the domain. In our domain, we were able to reduce the number of experiments needed for the pair coverage step to about 7, a savings of about 5 experiments.

Finally, it is then necessary to measure each extended correlation set, and, in cases where the correlation neighborhood exceeds the size of $m$, it is necessary to measure each set of $m$ from the correlation set. For an extended correlation set size of $s$, with a correlation set size of $c$, this step takes about $n/s$ experiments if $s < m$, but takes approximately $n/c * (s - c)$ (where $s - c$ is the average number of uncorrelated perturbation parents per extended neighborhood) if $s > m$ and when the correlation neighborhood is bigger than $m$, this step can take up to $n/c * \binom{c}{m} * (s - c)$ experiments. However, in many cases experiments become redundant due to symmetry in correlation neighborhoods and shared perturbation parents, so these are approximate upper bounds. In our case, $c$ was approximately 3 (never exceeding $m$) and $s$ approximately 5, requiring about 8 experiments.

### 5.3.2 Model results

Because we start with a random variable set for our first experiments, there is a certain degree of randomness in which specific sets of variable subsets are measured. The specific subsets can change the model results, because in the search we may considered all measured sets (see Section 5.2.3). For this reason, we present below results from two independent selections of variable subset (each starting with a different random ordering of the variables), as well as one result from a more strictly constrained search allowing only connections within extended correlation neighborhoods. Aside from the details given in Section 5.2.3, all other aspects of the implementation are as described in Chapter 4.

**Model result I: Search constrained to extended correlation neighborhoods** This model result uses just 8 experiments, the number required in our domain

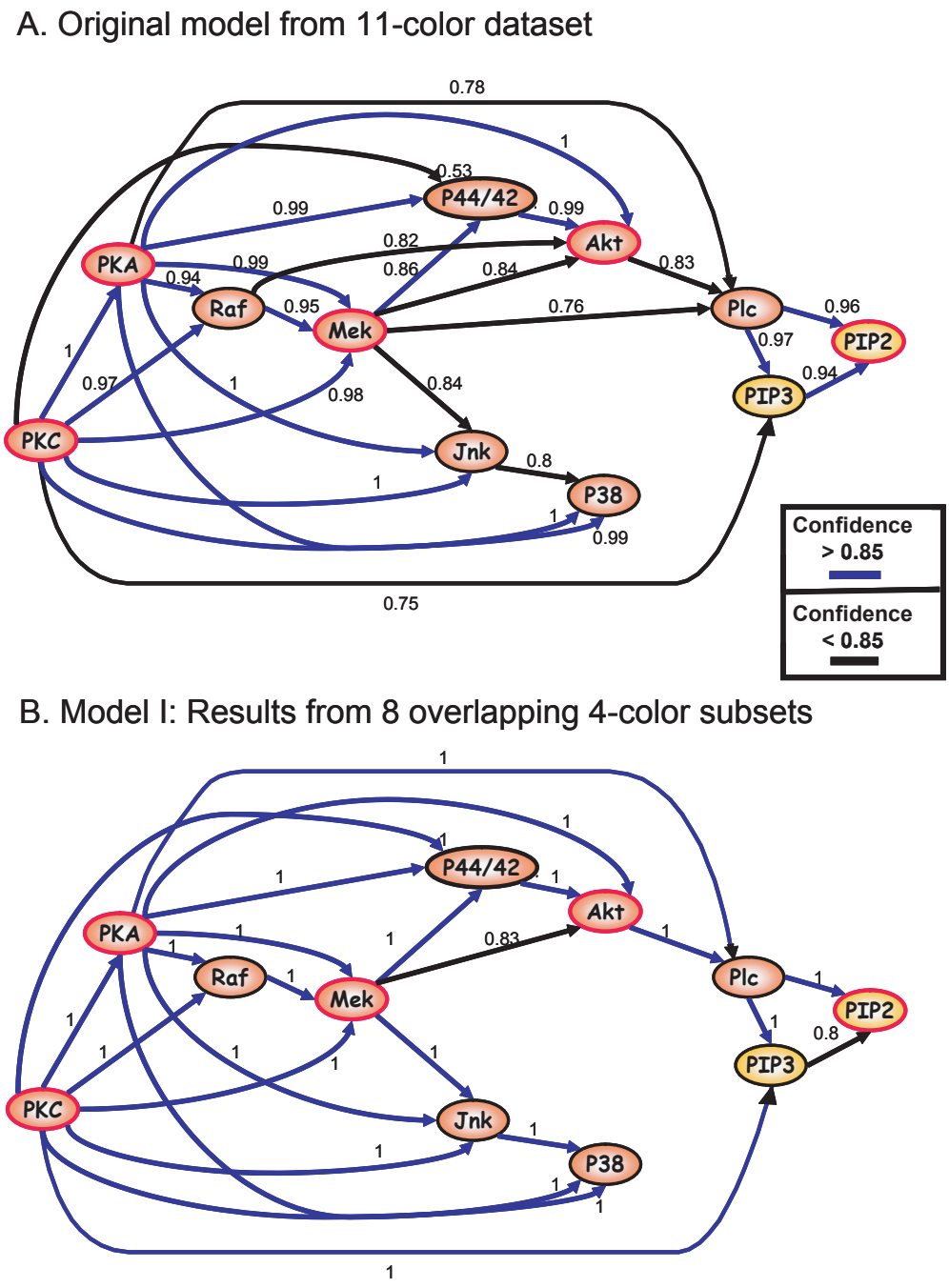



Figure 5-5: **Model I: Results from 8 4-color experiments.** Panel A shows the original model, from the full 11-color dataset. Panel B shows the model inferred using 8 4-color data subsets. Although only 8 experiments were used in the search, another 7 were necessary in order to heuristically determine which measurements to include in the search. The edges in both models are annotated with edge confidence values, obtained by averaging 100 high scoring models.

to cover all extended neighborhoods as described in Section 5.2.3. However, it still effectively requires ~15 experiments in all, as ~7 experiments are required to elucidate the correlation neighborhoods. Figure 5-5 shows the model result with edge confidence levels included. The result from the full 11-color dataset (from Chapter 4) is included for convenience. The result is identical to the original results at any confidence cutoff up to 0.8. At a higher confidence cutoff value, differences exist between the two models. In particular, this model result has mostly higher confidence values for most edges. If the original confidence cutoff of 0.85 is employed, 6 edges are above the cutoff in this model whereas in the original they are below it, and one edge makes the confidence cutoff in the original model but not in this result. Recall that the edge confidence values are based on model averaging results, essentially a weighted average of a compendium of high scoring models (see Chapter 4 for details).

**Model result II and III: Search includes all available measured subsets**
Figure 5-6 and Figure 5-7 show the results from two separate sets of measured variable subsets. The measurements covering the extended neighborhoods are identical (and identical to the ones used in Figure 5-5), but the initial measurements used to establish correlation neighborhoods are distinct. The result from the full 11-color dataset is included for convenience. The model in Figure 5-6 is nearly identical to the original model at confidence level 0.76. However, even at this cutoff value, it has an extra edge (P38→Plc$\gamma$) that is not seen in the original model. When the original confidence level of 0.85 is used, this model includes 7 edges that are lower confidence in the original model (mostly the same edges as in Figure 5-5), and one edge that is higher confidence in these results (as before, this edge is PIP3→PIP2).

The third result (Figure 5-7) is identical to the original at confidence cutoff value 0.66. At the original cutoff value, this model includes 6 additional edges (the same edges as found in the other two results) and excludes one (also the same, PIP3→PIP2).

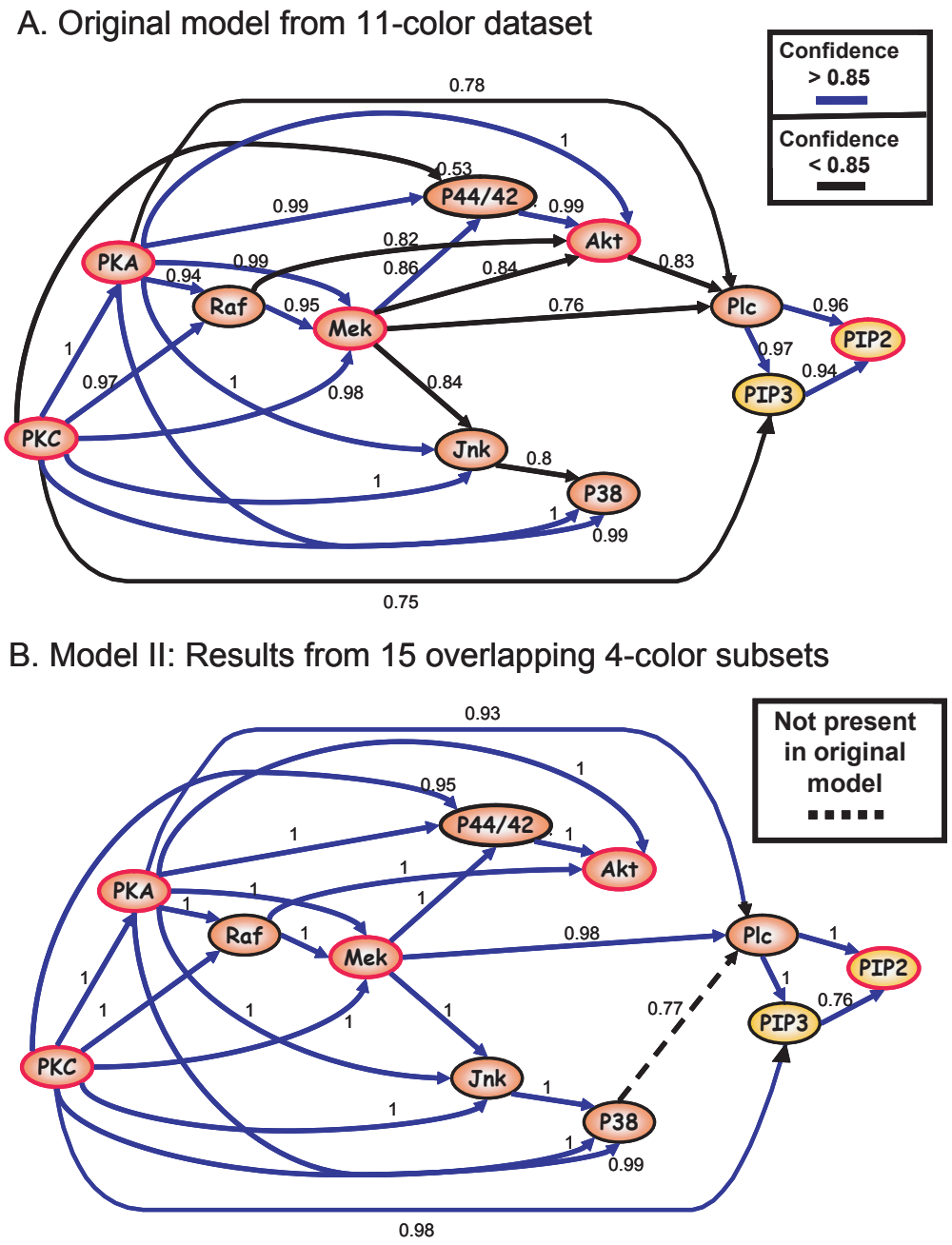



Figure 5-6: **Model II: Results from 15 4-color experiments.** Panel A shows the original model, from the full 11-color dataset. Panel B shows the model inferred using 15 4-color data subsets. The edges in both models are annotated with edge confidence values, obtained by averaging 100 high scoring models.

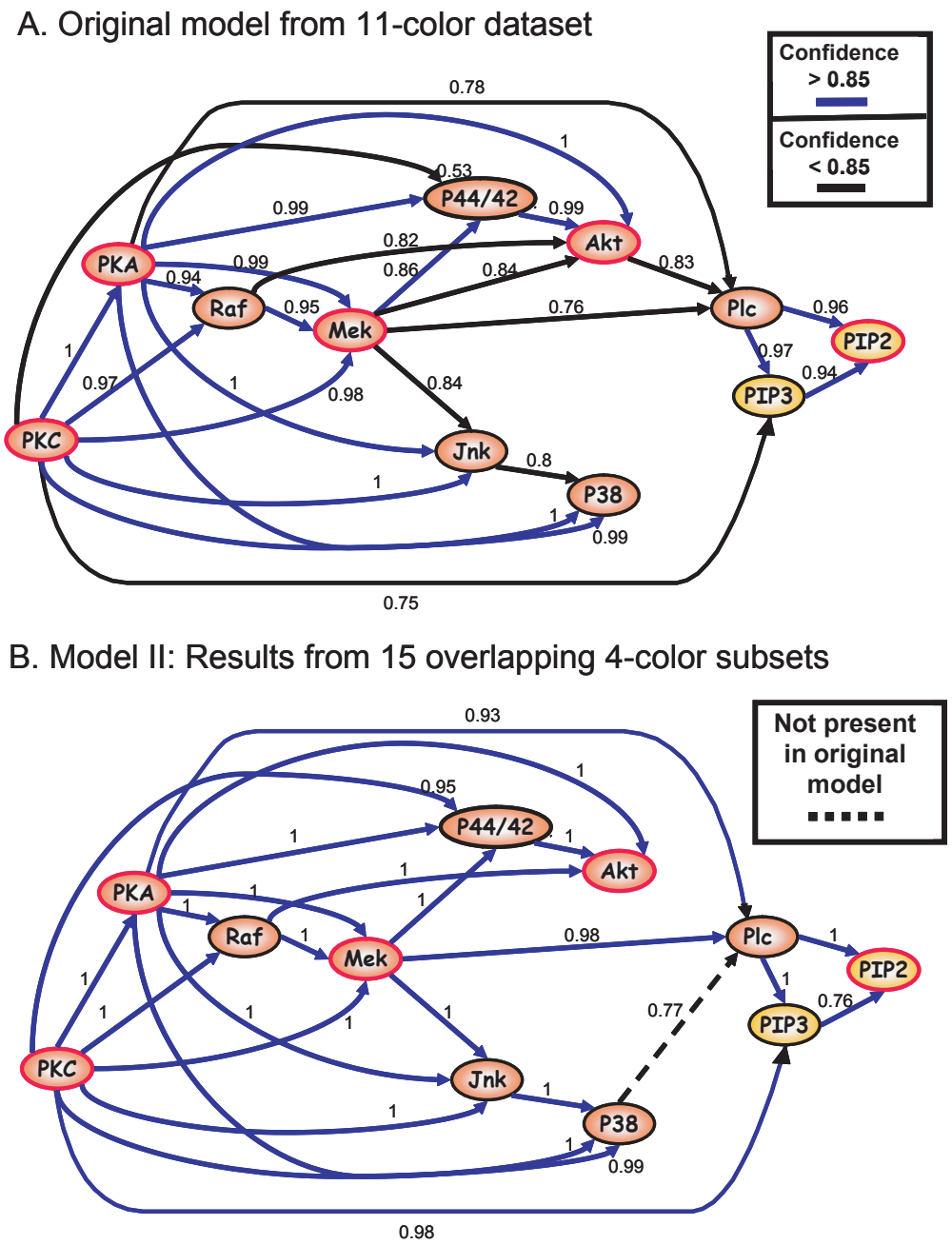



Figure 5-7: **Model III: Results from an independent set of 15 4-color experiments.** Panel A shows the original model, from the full 11-color dataset. Panel B shows the model inferred using 15 4-color data subsets (7 of which are distinct from those used in Figure 5-6). The edges in both models are annotated with edge confidence values, obtained by averaging 100 high scoring models.

## 5.4 Discussion

In this chapter, we describe an approximate, heuristic approach to apply Bayesian network structure inference to signaling pathways in a domain in which the entire variable set can not be observed simultaneously. We learn the network structure using measurements of overlapping subsets of the variable set. Our approach seeks to minimize the number of experiments measured while maximizing the accuracy of model inference, by employing a set of biologically motivated assumptions. In our domain, we find that, using approximately 15 measurements of 4-color variable subsets, we are able to infer a model structure that closely resembles the original structure inferred from a fully measured, 11-color dataset.

Models results consistently contained higher confidence edges than the original models, in all 3 results, and, as a result, consistently had ˜6 more edges than the original model, given the original cutoff value of 0.85. This result is not surprising, and stems from the fact that the search space is greatly constrained. In an unconstrained search space, each high confidence model is likely to have small variations, edges that appear in a relatively small percent of models. These are edges that express weaker dependencies, or dependencies that can be explained *equally* (or approximately equally) well by different edges, such that a different one appears in different high scoring models. These more minor edges explain some of the dependencies found in the entire domain, thus reducing the prevalence of other edges. For this reason, the edge confidence values in the original model are generally lower than in these constrained models. (The surprising result is PIP3→PIP2, which consistently has a lower confidence value. There appears to be no obvious reason for this result. It is possible that the original result simply had, due to the randomness of the search process, a particularly high confidence value for this edge).

The shift in confidence values can actually be problematic, as a dense model is harder to interpret and validate, and it is harder to determine which are the more dominant, reliable edges. An interesting point for future work would be to devise a method to rank the model edges that is distinct from model averaging. One straight-

forward approach would assess the contribution of each edge to the model score. Another point for future work includes a more thorough theoretical treatment of the number of experiments required, given in terms of the size of the correlation neighborhoods, the extended neighborhoods, $m$ and $n$.

Or approach uses just 15 4-color experiments to accurately reproduce the results of an 11-color experiment (albeit with a shift in confidence). However, these results, particularly with regard to the number of measurements needed, are relevant for our particular domain, with its set of variables and treatment conditions. A different domain may have a very different correlation landscape, leading to very different requirement for experiments; furthermore, it may violate our assumptions to varying degrees, yielding poorer results. Our work has an additional caveat: because we simulate 4-color data using an 11-color dataset (see Section 5.2.3), our data may be cleaner than data taken from actual independent experiments, which may contain more variability. We do not believe this will be a major source of variation, as the data is a moderately large sample drawn from the same distribution, and individual flow cytomtery runs are highly reproducible (data not shown). Nevertheless, the results must be interpreted with caution, as they are not conclusive regarding the number of experiments needed or about the success of the modeling effort.

Although we present an example in which we scale up from a 4 variable measurement capacity to an 11 variable model, our intent is to enable a larger measurement capability to scale to larger models, say, from 10 variables measured to models of 50-100. How would our approach differ for these larger scale models? We can expect to need a larger number of experiments in order to initially cover each pair of variables. For scaling from 10 variables to a 50 variable model, our lower bound calls for 25 experiments, and the upper bound calls for 45. This number grows for a 100 variable model: 110 experiments according to the lower bound, 190 in the upper bound. These numbers may appear prohibitive but, depending on the correlation landscape of the particular variable set, a large number of these may be eliminated using the assumption of approximate transitivity of correlations. Note that as the number of variables measured in an experiment grows, this assumption eliminates experiments

more quickly and efficiently, since the low and high correlations can be found more easily.

The large scale models have the distinct advantage that, in contrast to the smaller scale models, it is unlikely that an experiment measuring all of the members of an extended correlation neighborhood will be infeasible. This claim is based on the assumption that most signaling molecules have constrained in—degrees, so that a measurement capability of ˜10 molecules can easily accommodate each variable and its likely parents, whereas this does not hold true for a 4-color measurement capability. This fact also constrains the number of experiments that are necessary, because the combinatorial step of measuring the correlation neighborhood with each perturbation parent, or the subsets of size $m$ from the correlation neighborhood, is likely to be unnecessary. Although the details in terms of the number of experiments needed, and the success of the results will depend on various factors (prominently, the correlation landscape of the domain, and the particular dependencies among the variables, respectively), we anticipate that larger scale models will prove *more* amenable to this approach than smaller models.

This work proffers an approach for approximate model inference in a pathway with a number of molecules exceeding our measurement capabilities. Although our results are specific to just one particular variable set, they nonetheless indicate the possibility that this approach is both feasible (in terms of the number of experiments required) and effective (in terms of the modeling results). It is our hope that this method can be used to infer increasingly large signaling pathway models. Particularly as measurement capabilities improve, it may be possible to build highly integrated models involving numerous canonical pathways, in order to discover points of cross talk, enrich our knowledge of various pathways, and enable a truly systems approach even in relatively uncharted domains.

# Chapter 6

# Discussion and Future Directions

In this dissertation, we explore the use of multiparameter flow cytometry data to learn signaling pathway structure. We find that when cells are stimulated in heterogeneous ways, and specific perturbations are performed, we are able to infer a network structure that is consistent with the structure described in the literature. We further explored the problem of a partially observable set- when only a subset of the variables can be measured in each experiment, due to technical limitations- and find that when we make a number of specific, biologically plausible assumptions, more limited datasets can also be used to learn a signaling pathway structure.

We demonstrate the use of Bayesian network structure learning for the elucidation of signaling pathway models. However, as discussed in Chapters 2 and 5, this approach has a number of shortcomings and limitations, among them the nonideality of perturbations, abundance of hidden nodes in the domain (which hinder inference of causal connections), and acyclicity constraint. Due to such limitations, model results will often fall short of an accurate representation of the actual signaling pathway structure; therefore, it is necessary to interpret model results with caution, and confirm suggested connections via wet lab validation. In one view, Bayesian network analysis can be considered an *in silico* generator of testable hypotheses. We note that many of the limitations we have discussed can be addressed with technological advances and/or modification of the algorithm to render it better suited to the specific modeling domain, pointing to a useful direction for future work.

In our work, we treat single cells as biological replicates (in fact, the idea for using single cell data evolved from a search for data with many replicates). Large numbers of single cells are informative because they are not identical (or even necessarily similar) to each other. This fact is in contrast to the simplifying assumption we implicitly make when we work with lysates, in which we assume that cells respond in concert to an experimental condition. Often it is true that there is a detectable overall shift; however, this change may not be representative of the "average cell" in the population, especially if the population is in fact multimodal. Even when the population contains just one discernable distribution, great variability exists between cells. This may be due to the presence of cellular subsets, as perhaps our CD4+ cells contained 'impurities' in the form of other cell types, or perhaps there existed subsets among the CD4+ cells themselves (such as, e.g., TH1 and TH2 cells), which behave differently. Even for a very pure population, however, it is easy to imagine reasons for heterogeneity ranging from differences in cell cycle stage to variation in cell state (in terms of concentration of proteins and small molecules) to genetic and epigenetic heterogeneities. We ignore the presence of subsets, but specifically take advantage of intercellular variability, to examine how molecules in CD4+ cells covary, and how they appear to influence each other probabilistically.

It is also possible to seek out these subsets, or, more explicitly, to specifically examine different populations. We could profile cell cycle markers, and compare molecular interactions between different cell cycle stages. Alternatively, an interesting study may compare signaling pathway connections in samples from healthy subjects versus cancer patients. On the surface, there is nothing new in this idea- indeed, comparisons between healthy and unhealthy samples is a staple in disease research. Even using multidimentsional flow cytometry, in fact, Irish et al examined the response of various molecules to various conditions in various patients (and compared to a healthy control population) in an elegant study [39]. Irish and colleagues, however, collapsed the single cell measurements into a single metric—geometric mean–disregarding the molecular distributions. Our approach takes this analysis a step further, suggesting a comparison not only of the response of the molecules themselves (e.g. a shift),

but also of their *interdependencies*, in response to different conditions and among different patients. Thus, the use of distribution information translates to an added dimension in this type of analysis, enabling also an examination of changes in specific intermolecular influences.

We have discussed how single cell data enables a probabilistic analysis of molecular dependencies, due primarily to the ease in acquiring a large sample size. Yet another advantage is the accessibility of specific, rare cellular subsets, generally not amenable to standard analyses due to difficulties in isolation and insufficient sample quantities. Assuming these cellular subsets are detectably distinct, our approach easily extends to the analysis of their signaling pathway behavior. Rather then analyze their signaling pathway directly (this could be difficult, due to the relative small number of these cells), we build a signaling pathway model based on a 'background,' standard cell type, and then profile the rare cells to detect distinct differences. In this way, we may be able to assess the differences between rare immune subtypes and their 'parental' cell line, between transformed cells and their healthy counterparts, and between elusive, chemoresistant tumor cells and neighboring chemosensitive cells in the same tumor.

As these examples illustrate, the advantage of using single cells goes beyond the advantages of large sample size. Furthermore, as we alluded to above, signaling pathway analysis can go beyond an attempt at learning pathway structure, to a more in depth analysis of differences among cells, which may include subtle differences such as, e.g., a difference in model parameters, in addition to distinct changes in molecular connectivity. Existing computational methods will likely need to be modified in order to extract relevant information in a biologically meaningful way. We hope that with appropriate computational tools, the analysis of single cell data will lead to improvements in understanding of cellular biology and disease states, in accurate and specific diagnostics, and in more specifically tailored and perhaps personalized therapies for human disease.

# Bibliography


[1] H. Abrahamsen, G. Baillie, J. Ngai, T. Vang, K. Nika, A. Ruppelt, T. Mustelin, M. Zaccolo, M. Houslay, and K. Tasken. Tcr- and cd28-mediated recruitment of phosphodiesterase 4 to lipid rafts potentiates tcr signaling. *J Immunol*, 173(8), 2004.

[2] Bruce Alberts. *Molecular biology of the cell*. Garland Science, New York, 4th edition, 2002.

[3] A. R. Asthagiri, C. M. Nelson, A. F. Horwitz, and D. A. Lauffenburger. Quantitative relationship among integrin-ligand binding, adhesion, and signaling via focal adhesion kinase and extracellular signal-regulated kinase 2. *J Biol Chem*, 274(38), 1999.

[4] J. S. Bader, A. Chaudhuri, J. M. Rothberg, and J. Chant. Gaining confidence in high-throughput protein interaction networks. *Nat Biotechnol*, 22(1), 2004.

[5] Kelley BP, Yuan B, Lewitter F, Sharan R, Stockwell BR, and Ideker T. Pathblast: a tool for alignment of protein interaction networks. *Nucleic Acids Res.*, 32(Web Server issue), 2004.

[6] L. Buday, S. E. Egan, P. Rodriguez Viciana, D. A. Cantrell, and J. Downward. A complex of grb2 adaptor protein, sos exchange factor, and a 36-kda membrane-bound tyrosine phosphoprotein is implicated in ras activation in t cells. *J Biol Chem*, 269(12), 1994.

[7] M. P. Carroll and W. S. May. Protein kinase c-mediated serine phosphorylation directly activates raf-1 in murine hematopoietic cells. *J Biol Chem*, 269(2), 1994.

[8] A. Clerk, F. H. Pham, S. J. Fuller, E. Sahai, K. Aktories, R. Marais, C. Marshall, and P. H. Sugden. Regulation of mitogen-activated protein kinases in cardiac myocytes through the small g protein rac1. *Mol Cell Biol*, 21(4), 2001.

[9] A. Clerk and P. H. Sugden. Untangling the web: specific signaling from pkc isoforms to mapk cascades. *Circ Res*, 89(10), 2001.

[10] M. E. Cusick, N. Klitgord, M. Vidal, and D. E. Hill. Interactome: gateway into systems biology. *Hum Mol Genet*, 14 Spec No. 2, 2005.

[11] M. Deng, S. Mehta, F. Sun, and T. Chen. Inferring domain-domain interactions from protein-protein interactions. *Genome Res*, 12(10), 2002.

[12] A. S. Dhillon, C. Pollock, H. Steen, P. E. Shaw, H. Mischak, and W. Kolch. Cyclic amp-dependent kinase regulates raf-1 kinase mainly by phosphorylation of serine 259. *Mol Cell Biol*, 22(10), 2002.

[13] J. Downward, J. D. Graves, P. H. Warne, S. Rayter, and D. A. Cantrell. Stimulation of p21ras upon t-cell activation. *Nature*, 346(6286), 1990.

[14] D. F. Far, J. F. Peyron, V. Imbert, and B. Rossi. Immunofluorescent quantification of tyrosine phosphorylation of cellular proteins in whole cells by flow cytometry. *Cytometry*, 15(4), 1994.

[15] V. Fortino, C. Torricelli, C. Gardi, G. Valacchi, S. Rossi Paccani, and E. Maioli. Erks are the point of divergence of pka and pkc activation by pthrp in human skin fibroblasts. *Cell Mol Life Sci*, 59(12), 2002.

[16] N. Friedman. Inferring cellular networks using probabilistic graphical models. *Science*, 303(5659), 2004.

[17] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using bayesian networks to analyze expression data. *J Comput Biol*, 7(3-4), 2000.

[18] Nir Friedman, Kevin Murphy, and Stuart Russell. Learning the structure of dynamic probabilistic networks. In *Proceedings of Uncertainty in Artificial Intelligence*, 1998.

[19] R. Fukuda, B. Kelly, and G. L. Semenza. Vascular endothelial growth factor gene expression in colon cancer cells exposed to prostaglandin e2 is mediated by hypoxia-inducible factor 1. *Cancer Res*, 63(9), 2003.

[20] A. C. Gavin, M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A. M. Michon, C. M. Cruciat, M. Remor, C. Hofert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M. A. Heurtier, R. R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, and G. Superti-Furga. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868), 2002.

[21] E. Genot and D. A. Cantrell. Ras regulation and function in lymphocytes. *Curr Opin Immunol*, 12(3), 2000.

[22] F. G. Giancotti. Integrin signaling: specificity and control of cell survival and cell cycle progression. *Curr Opin Cell Biol*, 9(5), 1997.

[23] A. J. Hartemink, D. K. Gifford, T. S. Jaakkola, and R. A. Young. Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Pac Symp Biocomput*, 2001.

[24] A. J. Hartemink, D. K. Gifford, T. S. Jaakkola, and R. A. Young. Combining location and expression data for principled discovery of genetic regulatory network models. *Pac Symp Biocomput*, 2002.

[25] Alexander J. Hartemink, Massachusetts Institute of Technology. Dept. of Electrical Engineering, and Computer Science. *Principled computational methods*

*for the validation discovery of genetic regulatory networks*. PhD thesis, Massachusetts Institute of Technology Dept. of Electrical Engineering and Computer Science, 2001.

[26] D. Heckerman. A tutorial on learning with bayesian networks. Technical report, Microsoft Research, 1995.

[27] M. L. Hermiston, Z. Xu, R. Majeti, and A. Weiss. Reciprocal regulation of lymphocyte activation by tyrosine kinases and phosphatases. *J Clin Invest*, 109(1), 2002.

[28] L. A. Herzenberg, D. Parks, B. Sahaf, O. Perez, and M. Roederer. The history and future of the fluorescence activated cell sorter and flow cytometry: a view from stanford. *Clin Chem*, 48(10), 2002.

[29] L. A. Herzenberg and R. G. Sweet. Fluorescence-activated cell sorting. *Sci Am*, 234(3), 1976.

[30] Y. Ho, A. Gruhler, A. Heilbut, G. D. Bader, L. Moore, S. L. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, L. Yang, C. Wolting, I. Donaldson, S. Schandorff, J. Shewnarane, M. Vo, J. Taggart, M. Goudreault, B. Muskat, C. Alfarano, D. Dewar, Z. Lin, K. Michalickova, A. R. Willems, H. Sassi, P. A. Nielsen, K. J. Rasmussen, J. R. Andersen, L. E. Johansen, L. H. Hansen, H. Jespersen, A. Podtelejnikov, E. Nielsen, J. Crawford, V. Poulsen, B. D. Sorensen, J. Matthiesen, R. C. Hendrickson, F. Gleeson, T. Pawson, M. F. Moran, D. Durocher, M. Mann, C. W. Hogue, D. Figeys, and M. Tyers. Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. *Nature*, 415(6868), 2002.

[31] H. R. Hoogenboom and P. Chames. Natural and designer binding sites made by phage display technology. *Immunol Today*, 21(8), 2000.

[32] T. Ideker, T. Galitski, and L. Hood. A new approach to decoding lifel systems biology. *Annu. Rev. Genomics Human Gen*, 2, 2001.

[33] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A*, 98(8), 2001.

[34] T. Ito, K. Tashiro, S. Muta, R. Ozawa, T. Chiba, M. Nishizawa, K. Yamamoto, S. Kuhara, and Y. Sakaki. Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc Natl Acad Sci U S A*, 97(3), 2000.

[35] A. Jaimovich, G. Elidan, H. Margalit, and N. Friedman. Towards an integrated protein-protein interaction network: a relational markov network approach. *J Comput Biol*, 13(2), 2006.

[36] R K Jaiswal, S A Moodie, A Wolfman, and G E Landreth. The mitogen-activated protein kinase cascade is activated by b-raf in response to nerve growth factor through interaction with p21ras. *Mol Cell Biol*, 14(10):6944–6953, October 1994.

[37] R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N. J. Krogan, S. Chung, A. Emili, M. Snyder, J. F. Greenblatt, and M. Gerstein. A bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, 302(5644), 2003.

[38] M. Jaumot and J. F. Hancock. Protein phosphatases 1 and 2a promote raf-1 activation by regulating 14-3-3 interactions. *Oncogene*, 20(30), 2001.

[39] Irish JM, Hovland R, Krutzik PO, Perez OD, Bruserud O, Gjertsen BT, and Nolan GP. Single cell profiling of potentiated phospho-protein networks in cancer cells. *Cell*, 2(118), 2004.

[40] L. P. Kane, J. Lin, and A. Weiss. Signal transduction by the tcr for antigen. *Curr Opin Immunol*, 12(3), 2000.

[41] P. O. Krutzik, J. M. Irish, G. P. Nolan, and O. D. Perez. Analysis of protein phosphorylation and cellular signaling events by flow cytometry: techniques and clinical applications. *Clin Immunol*, 110(3), 2004.

[42] P. O. Krutzik and G. P. Nolan. Intracellular phospho-protein staining techniques for flow cytometry: monitoring single cell signaling events. *Cytometry A*, 55(2), 2003.

[43] C. A. Lange-Carter and G. L. Johnson. Ras-dependent growth factor regulation of mek kinase in pc12 cells. *Science*, 265(5177), 1994.

[44] I. Lee, S. V. Date, A. T. Adai, and E. M. Marcotte. A probabilistic functional network of yeast genes. *Science*, 306(5701), 2004.

[45] S. B. Lee and S. G. Rhee. Significance of pip2 hydrolysis and regulation of phospholipase c isozymes. *Curr Opin Cell Biol*, 7(2), 1995.

[46] P. Lenz, S. M. Bacot, M. R. Frazier-Jessen, and G. M. Feldman. Nucleoporation of dendritic cells: efficient gene transfer by electroporation into human monocyte-derived dendritic cells. *FEBS Lett*, 538(1-3), 2003.

[47] N. Lin, B. Wu, R. Jansen, M. Gerstein, and H. Zhao. Information assessment on predicting protein-protein interactions. *BMC Bioinformatics*, 5, 2004.

[48] T. H. Lin, A. E. Aplin, Y. Shen, Q. Chen, M. Schaller, L. Romer, I. Aukhil, and R. L. Juliano. Integrin-mediated activation of map kinase is independent of fak: evidence for dual integrin signaling pathways in fibroblasts. *J Cell Biol*, 136(6), 1997.

[49] H. Lu, L. Lu, and J. Skolnick. Development of unified statistical potentials describing protein-protein interactions. *Biophys J*, 84(3), 2003.

[50] Steffen M, Petti A, Aach J, D'haeseleer P, and Church G. Automated modelling of signal transduction networks. *BMC Bioinformatics.*, 1(3), 2002.

[51] R. Marais, Y. Light, C. Mason, H. Paterson, M. F. Olson, and C. J. Marshall. Requirement of ras-gtp-raf complexes for activation of raf-1 by protein kinase c. *Science*, 280(5360), 1998.

[52] R. Marais, Y. Light, H. F. Paterson, and C. J. Marshall. Ras recruits raf-1 to the plasma membrane for activation by tyrosine phosphorylation. *Embo J*, 14(13), 1995.

[53] C. J. Marshall. Map kinase kinase kinase, map kinase kinase and map kinase. *Curr Opin Genet Dev*, 4(1), 1994.

[54] H. Mischak, T. Seitz, P. Janosch, M. Eulitz, H. Steen, M. Schellerer, A. Philipp, and W. Kolch. Negative regulation of raf-1 by phosphorylation of serine 621. *Mol Cell Biol*, 16(10), 1996.

[55] C. Muller, J. Kremerskothen, M. Zuhlsdorf, U. Cassens, T. Buchner, A. Barnekow, and O. M. Koch. Rapid quantitative analysis of protein tyrosine residue phosphorylation in defined cell populations in whole blood and bone marrow aspirates. *Br J Haematol*, 94(3), 1996.

[56] J. M. Mullins. Overview of fluorophores. *Methods Mol Biol*, 34, 1994.

[57] V. Neduva, R. Linding, I. Su-Angrand, A. Stark, F. de Masi, T. J. Gibson, J. Lewis, L. Serrano, and R. B. Russell. Systematic discovery of new recognition peptides mediating protein interaction networks. *PLoS Biol*, 3(12), 2005.

[58] J. C. Obenauer, L. C. Cantley, and M. B. Yaffe. Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res*, 31(13), 2003.

[59] Perez OD, Mitchell D, Jager GC, South S, Murriel C, McBride J, Herzenberg LA, Kinoshita S, and Nolan GP. Leukocyte functional antigen 1 lowers t cell activation thresholds and signaling through cytohesin-1 and jun-activating binding protein 1. *Nat Immunol*, 11, 2003.

[60] Irene M. Ong, Jeremy D. Glasner, and David Page. Modelling regulatory pathways in e. coli from time series expression profiles. *Bioinformatics*, 18, 2002.

[61] F. Pages, M. Ragueneau, R. Rottapel, A. Truneh, J. Nunes, J. Imbert, and D. Olive. Binding of phosphatidylinositol-3-oh kinase to cd28 is required for t-cell signalling. *Nature*, 369(6478), 1994.

[62] D. M. Payne, A. J. Rossomando, P. Martino, A. K. Erickson, J. H. Her, J. Shabanowitz, D. F. Hunt, M. J. Weber, and T. W. Sturgill. Identification of the regulatory phosphorylation sites in pp42/mitogen-activated protein kinase (map kinase). *Embo J*, 10(4), 1991.

[63] Judea Pearl. *Probabilistic reasoning in intelligent systems : networks of plausible inference*. Morgan Kaufmann Publishers, San Mateo, Calif., 1988.

[64] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, Cambridge, UK, 2000.

[65] D. Pe'er. Bayesian network analysis of signaling networks: a primer. *Sci STKE*, 2005(281), 2005.

[66] D. Pe'er, A. Regev, G. Elidan, and N. Friedman. Inferring subnetworks from perturbed expression profiles. *Bioinformatics*, 17 Suppl 1, 2001.

[67] O. D. Perez, P. O. Krutzik, and G. P. Nolan. Flow cytometric analysis of kinase signaling cascades. *Methods Mol Biol*, 263, 2004.

[68] O. D. Perez and G. P. Nolan. Simultaneous measurement of multiple active kinase states using polychromatic flow cytometry. *Nat Biotechnol*, 20(2), 2002.

[69] O. D. Perez and G. P. Nolan. Phospho-proteomic immune analysis by flow cytometry: from mechanism to translational medicine at the single-cell level. *Immunol Rev*, 210, 2006.

[70] A. Perfetto, P. Chattopadhyay, and M. Roederer. Seventeen-colour flow cytometry: Unravelling the immune system. *Nature Reviews Immunology*, 4, 2004.

[71] K. V. Prasad, Y. C. Cai, M. Raab, B. Duckworth, L. Cantley, S. E. Shoelson, and C. E. Rudd. T-cell antigen cd28 interacts with the lipid kinase phosphatidylinositol 3-kinase by a cytoplasmic tyr(p)-met-xaa-met motif. *Proc Natl Acad Sci U S A*, 91(7), 1994.

[72] Y. Qi, J. Klein-Seetharaman, and Z. Bar-Joseph. Random forest similarity for protein-protein interaction prediction from multiple sources. *Pac Symp Biocomput*, 2005.

[73] M. Rincon, R. A. Flavell, and R. A. Davis. The jnk and p38 map kinase signaling pathways in t cell-mediated immune responses. *Free Radic Biol Med*, 28(9), 2000.

[74] Mario Roederer. http://www.drmr.com/compensation/, May 24, 2000.

[75] C. E. Rudd, O. Janssen, Y. C. Cai, A. J. da Silva, M. Raab, and K. V. Prasad. Two-step tcr zeta/cd3-cd4 and cd28 signaling in t cells: Sh2/sh3 domains, protein-tyrosine and lipid kinases. *Immunol Today*, 15(5), 1994.

[76] K. Sachs. Bayesian networks: A quick intro. *Biomedical Computation Review*, 1, Summer 2005 2005.

[77] K. Sachs, D. Gifford, T. Jaakkola, P. Sorger, and D. A. Lauffenburger. Bayesian network approach to cell signaling pathway modeling. *Sci STKE*, 2002(148), 2002.

[78] K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721), 2005.

[79] M. Saxena, S. Williams, K. Tasken, and T. Mustelin. Crosstalk between camp-dependent kinase and map kinase through a protein tyrosine phosphatase. *Nat Cell Biol*, 1(5), 1999.

[80] D. D. Schlaepfer, S. K. Hanks, T. Hunter, and P. van der Geer. Integrin-mediated signal transduction linked to ras pathway by grb2 binding to focal adhesion kinase. *Nature*, 372(6508), 1994.

[81] D. D. Schlaepfer and T. Hunter. Evidence for in vivo phosphorylation of the grb2 sh2-domain binding site on focal adhesion kinase by src-family protein-tyrosine kinases. *Mol Cell Biol*, 16(10), 1996.

[82] E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet*, 34(2), 2003.

[83] A. L. Singer and G. A. Koretzky. Control of t cell function by positive and negative regulators. *Science*, 296(5573), 2002.

[84] R. Singh, J. Xu, and B. Berger. Struct2net: Integrating structure into protein-protein interaction prediction. In *Proceedings of the Pacific Symposium on Biocomputing*, volume 11, 2006.

[85] B. S. Skalhegg, B. F. Landmark, S. O. Doskeland, V. Hansson, T. Lea, and T. Jahnsen. Cyclic amp-dependent protein kinase type i mediates the inhibitory effects of 3',5'-cyclic adenosine monophosphate on cell replication in human t lymphocytes. *J Biol Chem*, 267(22), 1992.

[86] M. V. Sofroniew, C. L. Howe, and W. C. Mobley. Nerve growth factor signaling, neuroprotection, and neural repair. *Annu Rev Neurosci*, 24, 2001.

[87] Z. Songyang, S. E. Shoelson, M. Chaudhuri, G. Gish, T. Pawson, W. G. Haser, F. King, T. Roberts, S. Ratnofsky, R. J. Lechleider, and et al. Sh2 domains recognize specific phosphopeptide sequences. *Cell*, 72(5), 1993.

[88] D. Stokoe, L. R. Stephens, T. Copeland, P. R. Gaffney, C. B. Reese, G. F. Painter, A. B. Holmes, F. McCormick, and P. T. Hawkins. Dual role of phosphatidylinositol-3,4,5-trisphosphate in the activation of protein kinase b. *Science*, 277(5325), 1997.

[89] A. Tamir and N. Isakov. Cyclic amp inhibits phosphatidylinositol-coupled and -uncoupled mitogenic signals in t lymphocytes. evidence that camp alters pkc-induced transcription regulation of members of the jun and fos family of genes. *J Immunol*, 152(7), 1994.

[90] S. L. Tan and P. J. Parker. Emerging and diverse roles of protein kinase c in immune cell signalling. *Biochem J*, 376(Pt 3), 2003.

[91] K. Tasken and A. Ruppelt. Negative regulation of t-cell receptor activation by the camp-pka-csk signalling pathway in t-cell lipid rafts. *Front Biosci*, 11, 2006.

[92] Newton AC. Toker A. Akt/protein kinase b is regulated by autophosphorylation at the hypothetical pdk-2 site. *J Biol Chem*, 12(275), 2000.

[93] J. W. Tung, D. R. Parks, W. A. Moore, and L. A. Herzenberg. Identification of b-cell subsets: an exposition of 11-color (hi-d) facs methods. *Methods Mol Biol*, 271, 2004.

[94] P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, and J. M. Rothberg. A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. *Nature*, 403(6770), 2000.

[95] R. R. Vaillancourt, A. M. Gardner, and G. L. Johnson. B-raf-dependent regulation of the mek-1/mitogen-activated protein kinase pathway in pc12 cells and regulation by cyclic amp. *Mol Cell Biol*, 14(10), 1994.

[96] L. Van Aelst, M. Barr, S. Marcus, A. Polverino, and M. Wigler. Complex formation between ras and raf and other protein kinases. *Proc Natl Acad Sci U S A*, 90(13), 1993.

[97] C. von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417(6887), 2002.

[98] H. Wang, E. Segal, A. Ben-Hur, D. Koller, and D. Brutlag. Identifying protein-protein interaction sites on a genome-wide scale. In *Advances in Neural Information Processing Systems (NIPS 2004)*, Vancouver, Canada, 2005.

[99] K. K. Wary, F. Mainiero, S. J. Isakoff, E. E. Marcantonio, and F. G. Giancotti. The adaptor protein shc couples a class of integrins to the control of cell cycle progression. *Cell*, 87(4), 1996.

[100] Soderling TR. Wayman GA, Tokumitsu H. Inhibitory cross-talk by camp kinase on the calmodulin-dependent protein kinase cascade. *J Biol Chem*, 26(272), 1997.

[101] B. L. Williams, K. L. Schreiber, W. Zhang, R. L. Wange, L. E. Samelson, P. J. Leibson, and R. T. Abraham. Genetic evidence for differential coupling of syk family kinases to the t-cell receptor: reconstitution studies in a zap-70-deficient jurkat t-cell line. *Mol Cell Biol*, 18(3), 1998.

[102] Peter J. Woolf, Wendy Prudhomme, Laurence Daheron, George Q. Daley, and Douglas A. Lauffenburger. Bayesian analysis of signaling networks governing embryonic stem cell fate decisions. *Bioinformatics*, 2005.

[103] I. Xenarios, L. Salwinski, X. J. Duan, P. Higney, S. M. Kim, and D. Eisenberg. Dip, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res*, 30(1), 2002.

[104] M. B. Yaffe and L. C. Cantley. Mapping specificity determinants for protein-protein association using protein fusions and random peptide libraries. *Methods Enzymol*, 328, 2000.

[105] M. B. Yaffe, G. G. Leparc, J. Lai, T. Obata, S. Volinia, and L. C. Cantley. A motif-based profile scanning approach for genome-wide prediction of signaling pathways. *Nat Biotechnol*, 19(4), 2001.

[106] M. B. Yaffe, K. Rittinger, S. Volinia, P. R. Caron, A. Aitken, H. Leffers, S. J. Gamblin, S. J. Smerdon, and L. C. Cantley. The structural basis for 14-3-3:phosphopeptide binding specificity. *Cell*, 91(7), 1997.

[107] Y. Yamanishi, J. P. Vert, A. Nakaya, and M. Kanehisa. Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis. *Bioinformatics*, 19 Suppl 1, 2003.

[108] S. Yano, H. Tokumitsu, and T.R. Soderling. Calcium promotes cell survival through cam kinase kinase activation of the protein kinase b pathway. *Nature*, 396(6711), 1998.

[109] Changwon Yoo and Cooper Gregory F. Causal discovery from a mixture of experimental and observational data. In *Proceedings of Uncertainty in Artificial Intelligence*, 1999.

[110] L. V. Zhang, S. L. Wong, O. D. King, and F. P. Roth. Predicting co-complexed protein pairs using genomic and proteomic data integration. *BMC Bioinformatics*, 5, 2004.

[111] W. M. Zhang and T. M. Wong. Suppression of camp by phosphoinositol/ca2+ pathway in the cardiac kappa-opioid receptor. *Am J Physiol*, 274(1 Pt 1), 1998.

[112] M. Zheng, S. J. Zhang, W. Z. Zhu, B. Ziman, B. K. Kobilka, and R. P. Xiao. beta 2-adrenergic receptor-induced p38 mapk activation is mediated by protein kinase a rather than by gi or gbeta gamma in adult mouse cardiomyocytes. *J Biol Chem*, 275(51), 2000.