Model studies aimed toward constructing PI3K pathways in leukemia T-cells from shRNA knockdowns

by Heather M. Sweeney

[S.B., 6-7, MIT, 2016]

Submitted to the Department of Electrical Engineering and Computer Science
In Partial Fulfillment of the Requirements for the Degree of

Master of Engineering in Computer Science and Molecular Biology

at the

Massachusetts Institute of Technology

June, 2017

© 2017 Massachusetts Institute of Technology. All rights reserved.
Model studies aimed toward constructing PI3K pathways in leukemia T-cells from shRNA knockdowns

by Heather M. Sweeney

Submitted to the Department of Electrical Engineering and Computer Science on May 25, 2017 in Partial Fulfillment of the Requirements for the Degree of Master of Engineering in Computer Science and Molecular Biology

Abstract

Understanding the networks that regulate cellular processes, and how they are mis-regulated, is a major area of study. This is especially key in leukemia, where the complexity of the pathways involved and the heterogeneity between different leukemia cell strains makes it difficult to design therapies that are less vulnerable to deleterious side effects or compensatory mutations. However, despite their potential, methods to reconstruct networks are often specific to experimental technique and their exact ability to accurately capture networks is unexplored. Motivated by a shRNA knockdown experiment on leukemia cell lines treated with a PI3K inhibitor, we sought to devise and test a method to reconstruct the relevant network, which could be used to gain a deeper understanding of the networks involved in PI3K signaling in leukemia cells. We simulated a shRNA knockdown experiment on a known small network and tested an inverse Potts model for reconstructing the network based on correlations from the simulation output. We found the method did not infer the small network properly, but did produce several patterns that indicate it may have potential for revealing something of interest about larger systems. Understanding what the patterns convey about the system and devising a better method to see if that works are future steps that will allow us to see if we can apply this method to in vitro data.
# Table of Contents

Abstract ................................................................................................................................. 2

Introduction ........................................................................................................................... 4

The adaptive immune response and T-cells ........................................................................ 4

RNAi and shRNA knockdowns ............................................................................................ 5

The PI3K network .................................................................................................................. 6

Graphical Models of Biological Systems ............................................................................ 7

ACE as a possible route for network reconstruction ............................................................ 12

Our study ............................................................................................................................... 13

Methods ................................................................................................................................. 13

shRNA knockdown data ...................................................................................................... 13

Models and Simulations ...................................................................................................... 16

Correlation matrix .............................................................................................................. 18

ACE and Potts models ......................................................................................................... 19

Results and Discussion ...................................................................................................... 20

SSC Simulation Results ...................................................................................................... 20

Eigenvalue Analysis ............................................................................................................ 27

ACE and Network Reconstruction ....................................................................................... 29

Conclusion ............................................................................................................................. 34

Future Steps .......................................................................................................................... 35

Acknowledgements ............................................................................................................. 38

Bibliography .......................................................................................................................... 38

Appendix ............................................................................................................................... 41
Introduction

*The adaptive immune response and T-cells*

The human immune response can be divided into two parts: the innate immune response and the adaptive immune response. The innate immune response is a general response to evolutionarily conserved regions of pathogens, whereas the adaptive immune response targets specific pathogens. T lymphocytes, or T cells, are key orchestrators of the adaptive immune response. Due to stochastic recombination of the genes that code for the receptor on the T cell surface, each T-cell is likely to express a receptor that is distinct from that on most other T cells. When a new pathogen invades the body, T-cells that are able to bind to that antigen are selected for and rapidly replicated, thus allowing for immune responses to specific pathogens. Some of these T-cells later differentiate into memory cells that can mount rapid responses upon reinfection with this pathogen. Different types of T-cells serve a variety of functions, including clearing intracellular pathogens and tumors, driving B-cell proliferation, memory, and suppressing T-cell responses to prevent autoimmunity [1]. These functions as well as T-cell growth and differentiation in turn are governed by complex networks of intra- and intercellular signals, as seen in Figure 1. Mutations in key genes in these networks can lead to autoimmune diseases and cancer, yet because of the complexity of these networks, untangling the interactions and genes involved in the disease phenotype is an ongoing challenge. However, understanding the structure of these pathways and how they are perturbed in the case of disease will help with developing therapies for diseases such as leukemia.
RNAi and shRNA knockdowns

A useful tool for studying gene function is RNA interference (RNAi), where short antisense nucleic acids are designed to bind to and thus inactivate and induce degradation of target mRNA. By specifically knocking down a particular gene and observing the effects, we can gain a better understanding of the gene’s function. This can be accomplished through short interfering RNAs (siRNAs), or short hairpin RNAs (shRNAs). shRNAs are double stranded DNA with a loop structure that can be more easily transfected into the cell through bacterial or viral vectors, where the cellular machinery processes the shRNA into siRNAs that can then target the desired mRNAs. This allows shRNAs to be stably integrated into the cell and be controlled using inducible
promoters, whereas siRNAs can only be transiently expressed as cell division will result in dilution of the siRNAs. The ability to target specific genes, including those that don’t have easily targetable protein products, makes shRNA a promising therapeutic tool as well as an experimental probe [6-7].

The potential for off-target effects exists, including shRNAs having partial sequence complementarity to non-target mRNAs, or non-specific effects like cellular toxicities due to the nucleotide construct. The latter is less of a concern in our study since it is likely similar across all shRNAs [6-7].

shRNAs and siRNAs been applied to identify undiscovered oncogenes and tumor suppressors in various cancers, including leukemia. RNAi was used to identify genes involved in T-cell polarization and migration [8], regulating development of Tfh, Th1, and CTL T-cells [9], and cytokine secretion in T-cells [10]. Other studies have identified essential genes in acute myeloid leukemia (AML) [11-13]. Despite these advances, there are a lack of studies focusing on T-cell leukemia in particular, and on the PI3K signalling network.

**The PI3K network**

One major player in leukemia and various cancers is phosphoinositide 3-kinase (PI3K), a family of intracellular signal transducers with a major role across various cell types. PI3K transduces upstream signals from cytokines and growth factors, and phosphorylates lipids which in turn activate downstream signals such as RTKs and GPCRs [2]. PI3K has been linked to cell proliferation and survival, membrane trafficking, metabolism, and cytoskeletal reorganization [3].
In the immune system specifically it has been linked to development, function, and differentiation. For instance, PI3K is required for differentiation of naive CD4 T cells and can promote and inhibit Treg development [4]. However, the extent of the roles of PI3K across cell types and the full signaling networks tied to it are still under investigation.

Because of its central role in many cellular processes, perturbations in the PI3K network have been linked to cancer. PI3K signaling promotes cell survival and PI3K hyperactivation is a contributor to drug resistance in cancer [5]. PIK3CA, the gene for the p110α isoform of PI3K, is one of the most frequently mutated oncogenes in human cancer [4]. The tumor suppression function of PTEN is linked to how it antagonizes PI3K activity by deactivating some of the lipids it activates [3]. As a result, inhibitors of PI3K as well as known members of the PI3K pathway are under investigation as possible cancer therapies. However, PI3K inhibitors alone have a limited effect on tumor cells, and long-term dosing of PI3K inhibitors has adverse effects. This can be due to many elements in the cancer cells’ PI3K network such as compensatory mutations, feedback loops and pathway crosstalk [5] [2]. In other words, in a complicated network, another mutation in the pathway can make up for the PI3K inhibition, or blocking PI3K could allow some gene to be constitutively activated to deleterious effect. To more intelligently design combination therapies that simultaneously also target other nodes in the network, a better understanding of the PI3K network is needed.

**Graphical Models of Biological Systems**

Graphical models are a popular tool for estimating networks such as the PI3K network from biological data, and multiple studies support the idea that correlation structure in static data from
biological systems could be used to gain insight into causation between variables. Reconstructing networks can allow one to identify direct interactions, potential modules of interest associated with a biological process and/or condition [14]. Further, inferring biological networks could provide insights beyond which molecules directly interact. For instance, it could be used to identify hubs, or nodes with many connections, or high degree, that may be more conserved and more likely to cause a defect when removed [15]. Another metric of interest is contact path length, the mean shortest distance between pairs of nodes. It is used as a measure of network size as well as a way to detect important nodes (in which a node that greatly affects the contact path length when removed is hypothesized to have some importance) [16]. Other topological features such as shortest path length, clustering coefficient, and betweenness may also have biological relevance [15]. However, given the wide variety of datasets and models, the capability of graphical models to capture biological network structure has not been entirely explored [17].

A common approach to inferring graphical models from biological systems is to create a matrix representing how different nodes in a network correlate. The difficulty with directly translating a correlation matrix to an adjacency matrix, a matrix representing the connections between nodes in a network, is that strongly correlated nodes are not always neighbors [18]. As a result further algorithms are often required to disentangle such higher order connections and return more accurate networks. Both inferring correlations and the process to reconstruct the network vary depending on the network being analyzed, the data used, and the research question being asked of the network. Thus, a variety of algorithms have been implemented for reconstructing biological networks, depending on the exact nature of the data.
Some methods focus on graph-clustering to identify molecule sets of interest as opposed to network reconstruction. For instance, the Diffcor package uses correlations from omics data such as RNAseq to identify differentially expressed genes across a large-scale network between two experimental conditions, such as wild type and mutant or disease and control populations. It does so by detecting the pattern change in correlation networks between two experimental conditions and clustering molecules with similar profile patterns, or differential correlations. These clusters consist of putative coexpressed genes. Other graph clustering methods include DPClus, Markov clustering, hierarchical clustering, and local and density-based searches. [19]

Kotze et al. took a different approach to using correlations to understand biological networks, in which they mapped correlations onto known networks to reveal novel pathways of interest for a particular condition for metabolomic studies [18].

Other methods focus on generating a graph of direct biological interactions from experimental data. These can be in the form of many types of interactions, such as protein-protein interactions, metabolic interactions, genetic interactions, protein phosphorylation, and transcription factor binding [15]. For instance, Hageman et al proposed a Bayesian statistical method to use expression quantitative trait loci (eQTL) data from the genetic variation in segregating populations to infer causal networks of genotypes and phenotypes. eQTL data consists of genome-wide gene expression, genotypes at certain genome markers, and other phenotypes. Their algorithm uses Metropolis-Hastings, a MC method, to sample the posterior distribution of model spaces and return a set of highly probable networks. Bayesian networks are graphical models that describe joint multivariate probability distributions. This method was limited by network scale, becoming
unstable after inferring networks beyond the order of 30 nodes. They applied it to kidney gene expression from an intercross population and predicted a “previously uncharacterized feedback loop”. [20]

Batushansky et al used correlation based network analysis (CNA) to generate an undirected network of gene interactions from a large dataset of metabolomics data of human breast carcinoma cell lines under two conditions: normal and hypoxia. The calculated pairwise correlation coefficients and created difference graphs to isolate condition-specific relations by comparing constructed vertices and edges across conditions. In the network vertices represented modules of genes, which allowed the authors to reduce network complexity given the large dataset size [14]. Adourian et al created a cross compartment correlation network from high throughput data to link biomarkers to processes and better understand molecular responses to drug induced liver injury. This method revealed relevant molecules although not directionality of causation [22].

Blair et al assessed the ability of graphical models to reconstruct metabolic pathways from high-throughput mQTL data under two sources of perturbation. In particular, they used a Markov Chain Monte Carlo (MCMC) algorithm to generate multiple causal networks from mQTL data and analyzed the top scoring networks. They first applied the MCMC to simulations of dynamic models of network motifs and then to a well studied metabolic pathway in a genetic cross of plants. They found that biological pathways could lead to meaningful correlations reflecting their topology, but that it’s possible but not expected to recover pathway order with their method, and that their inference was sensitive to certain subtle patterns in correlation structure. For instance they concluded that there was a high risk of false positive edges in their inference and that
experimental design factors/experimental artifacts could also skew the inferred network. They also noted that substrate inhibition could lead to missing edges and that feedback, buffering, and cycling could mask correlations. They found that linear pathways have problems with unclear order and that branching helps elucidate order [17].

Rice et al devised a “pairwise conditional correlation method” that reconstructed networks from the conditional correlation of RNA equilibrium concentrations between two genes given one was knocked down. They simulated gene expression on known networks, and studied how reconstruction error was affected by noise, network topology, network size, sparseness, and dynamic parameters. They performed their simulations by adding different levels of noise and forcing each node to zero concentration and measuring the steady state concentrations of the other nodes. Their conditional correlation matrix represented the correlation between two genes given one was perturbed, allowing them to infer directionality. Their study found several common errors with their technique. False positives where two nodes connected through intermediates were directly linked was a common error. False negatives also occurred when the correlation between two directly connected nodes was obscured by noise, multiple inputs to a target, or nonlinearity. They found that noise decreased the number of false positives as it masked higher order connections, but at a cost of overall error rate. These errors in particular may be good to watch out for in other network reconstruction methods [23].

Further, a variety of other techniques have been proposed to infer networks from correlations, such as functional module analysis, binary network approaches, fuzzy logic analysis, Jacobian matrix methods, and conditional independence tests [20, 23]. Thus, there are a wide set of algorithms
available for reconstructing networks from correlations with varying levels of accuracy. However, many of these techniques are geared more toward high-throughput data, rendering them unfit for the experimental system in our case.

**ACE as a possible route for network reconstruction**

Compared to many of the aforementioned studies where high-throughput data under two different conditions were used for network reconstruction, the dataset that motivates our problem consists of a univariate output, the leukemia cell growth as represented by the log-fold change of barcoded shRNAs compared to control, across a wide array of knockdown conditions. This makes finding our molecular interaction network more akin to other biological problems, such as finding mutational pathways to escape in HIV. Escape mutants are viral strains that have mutated to evade human immune pressure or drugs that target specific viral proteins. Often HIV acquires a combination of mutations to evade the attack while maintaining viral fitness. The order of mutations leading to escape matters greatly, as certain combinations of mutants are very unfavorable. Thus understanding which mutants are highly correlated can be helpful to inferring pathways to escape [24].

The problem of determining the prevalence of single and double mutants and reconstructing mutational pathways to escape has been approached via a maximum entropy or least biased model known as the Potts model in studies by Ferguson et al, Barton et al, and Mann et al [30, 24, 31]. Potts models, also known as Markov random fields, are a family of undirected categorical graph models. They have been used in diverse biological problems such as predicting structural contacts from protein sequence data, describing neural activity patterns, and finding mutational pathways...
to escape in HIV. The Potts model is intended to disentangle direct from indirect interactions given variable correlations. In addition, The Potts model constructs a sparse interaction network sufficient to reproduce observed data and yields sparser models when sampling is poor to provide more robust inference [24, 25]. These properties make Potts models a potentially useful approach to reconstructing direct connections in a biological network.

In the Barton study, they created the ACE (Adaptive Cluster Expansion) package to solve this problem. In prior experiments the ACE package has been shown to capture interaction strength better than faster Gaussian and pseudo-likelihood models [25]. Given the similarities of the escape mutant problem to ours and the potential usefulness of Potts models for our problem, we explored using the ACE method to reconstruct information about our simulated networks.

**Our Study**

Motivated by a T cell leukemia experiment, we sought to simulate their experiments on known example systems in order to test models for deriving molecular pathways from their shRNA knockdown results. We will simulate a shRNA knockdown experiment on a small known network based on the PI3K network, calculate correlations between the different molecules. We’ll then apply the ACE package to these correlations and assess how well ACE reconstructs the network.

**Methods**

*shRNA knockdown data*

The data motivating our study comes from a shRNA knockdown experiment done by Jeroen Roose’s lab at UCSF. The experiment focused on two T cell leukemia cell lines, Jurkat and MOLT-
3. Jurkat is a leukemic cell line with PI3K overexpression, whereas MOLT-3 lacks PI3K overexpression and thus served as one form of a control. Both cell lines were treated with a wide library of 55,000 barcoded shRNAs, with 26 shRNAs targeting each gene on average. Each cell was transfected with one shRNA packaged in a lentivirus, and thus in theory only one gene was shut down per cell. The cells were split into a control and treatment group, where a dosage of PI3K inhibitor not strong enough to kill the cells was added to the treatment group. The cells were grown for 3 weeks, after which the shRNA counts in the treated and control groups were recorded.

The log-fold change between the shRNA counts in the treated and untreated groups were considered an indicator of how the combined shRNA and PI3K inhibitor affected cell growth. Cells where an essential gene was targeted by shRNA should drop out at the beginning, leaving the study to focus on those cells where the shRNAs alone didn’t kill the cell. The remaining shRNAs could have little effect on the log fold change, demonstrating they likely played little role in the PI3K signaling pathway. On the other hand, some shRNAs could be significantly enriched, indicating that knockdown of shRNA compensated for the effect of PI3K inhibition, or depleted, indicating that the gene targeted by the shRNA was important for cell growth and therefore may be a potential therapeutic target. As a result, we have data on which shRNAs have a negative or positive effect on cell growth in combination with the PI3K inhibitor.

As some shRNAs likely suppress their respective gene with differing strengths and there’s a risk of off target effects, there is a problem of systematic bias that should be considered in analysis. Most shRNA studies focus on 2-5 shRNAs per gene, which is considered to help control for off-target effects, so it may be possible that the inclusion of on average 26 shRNAs per gene will be
sufficient to ignore such a bias [9]. On the other hand, the screens found that for many genes there are numerous shRNAs with neutral or opposing effects on the log fold change, as shown in Figure 2. Therefore it’s possible the large number of shRNAs used per cell increase the risk of including ineffective shRNAs that bias the sample. In addition, the in vitro analysis might also neglect certain intercellular interactions that may be significant in vivo.

To get a sense of the network, correlation data is also required. Presently there is correlation data for 10 genes but not for the entire library of genes. Thus, initial analysis would focus on this subset of genes and then would be applied to the larger dataset when it is produced.

![Figure 2](image.png)

Figure 2. The log fold change in shRNA counts across the PI3K inhibitor treated and control groups for shRNAs targeting the RAC1 gene in the Jurkat line of the UCSF study. Red triangles indicate statistically significant results, black triangles indicate results that are not statistically significant, and grey triangles represent results where the shRNA counts were too few to make a conclusion. Most shRNAs treated cells display a downregulated phenotype, but some shRNAs did not result in statistically significant downregulation, and one shRNA had the opposite effect. These effects are likely due to differing effectiveness and off-target effects across the different shRNAs, and must be accounted for in our analysis.
Model and Simulations

We designed a model system of molecular interactions on which to simulate the shRNA knockdown experiment. Our model was inspired by the known PI3K signaling pathway with some nodes removed for simplicity as shown in Figure 3. In particular, Ras was removed and PI3K is represented as directly activating Raf, and the PIP2 PIP3 activation and deactivation modulated by PI3K and PTEN, respectively, was instead represented as PI3K activating and PTEN inhibiting the downstream targets of PIP3. The model features two pathways regulated by PI3K, as well as a type 1 coherent feedforward loop, thus including complexity in the network structure.

Figure 3. The network used in our simulations, based on the known PI3K network. Graph drawn in Cytoscape.

The system was represented as a set of molecules where each molecule was in one of two states: active or inactive. Active molecules could in turn act upon downstream molecules in the network.
Activation and deactivation interactions were modeled, but binding of molecules and production and degradation of molecules were not modeled. For simplicity, molecules were assumed to be evenly distributed across the compartment, had same copy number, and activation and inactivation rate constants were the same across interactions. Basal activation or inactivation rates were also added for several component molecules. The full simulation script can be seen in the appendix.

The shRNA knockdown experiments record growth in the form of the inhibited shRNA counts recovered from each respective experiment as opposed to a high-throughput assay of mRNA or protein production per cell. As a result, we added an approximate metric for cellular growth and proliferation for our simulation. We added proxy molecules in our model that would be activated or inhibited by the furthest downstream genes known to promote growth.

To model multiple pathways that can have an impact on cellular growth and proliferation, we recorded two outputs, out1 and out2, activated by mTORC1 and Erk 1/2, respectively (Figure 3). mTORC1 is known to upregulate factors that contribute to cell growth and protein synthesis, while Erk 1/2 is known to promote cell proliferation and survival [26].

Simulations were run using the Stochastic Simulation Compiler (SSC), a program created by the Chakraborty lab that performs stochastic simulations of biological reaction networks based on the Gillespie Stochastic Simulation Algorithm (SSA) with algorithmic improvements for speedup, allowing it to scale better for larger systems [27]. SSA is a Monte-Carlo method based on the chemical master equation, which given a series of equations with rate constants and initial reactant counts, repeatedly asks the question of when will the next reaction fire and which reaction will fire.
next. Stochastic algorithms such as SSC and SSA are better suited than deterministic formulations of chemical kinetics for representing the stochastic nature of biological systems at low copy numbers [28].

We performed two types of simulations in SSC: one where each note was shRNA inhibited in conjunction with PI3K, and another where pairs of molecules were inhibited along with PI3K. Having pairwise knockdowns would allow us to have better insight on interactions between molecules. We also ran a control simulation where there was no PI3K or shRNA inhibition.

For both single and double inhibitions, we ran several sets of simulations, sampling multiple combinations of PI3K inhibition and shRNA inhibition rates (Table S1). Each simulation was run for 5000 time steps and each simulation condition was repeated for 20 trials. The various combinations of PI3K and shRNA inhibition rates allowed us to investigate the impact of varying levels of inhibition on network reconstruction. In addition, the varying PI3K levels could also act as modelling multiple cell lines where PI3K activation varies.

For all our simulations, we recorded output counts at each time step. To assess the accuracy of our simulations, we also recorded counts of molecules in the network at each time step, although such information was not gleaned from the in vitro experiments.

**Correlation Matrix**

To begin our analysis of the simulation data, we created correlation matrices to capture the relations between pairwise and single knockdowns. The correlation matrices were $m \times m$, where $m$
is the number of molecules involved in the pairwise knockdowns. In the case of our simulation, \( m = 9 \). For each correlation matrix \( C \), we calculated each index \( C_{ij} \) as a correlation function between knockdowns of molecules \( i \) and \( j \) across trials as shown in Equation 1. This correlation function is additive as opposed to multiplicative as we are looking at whether the effect of inhibition is additive as opposed to a probability measure.

\[
C_{ij} = \frac{1}{N} \sum_{t=1}^{N} \frac{O_{ij}^t - (O_{i}^t + O_{j}^t)}{2 \cdot O_{con}^t} \tag{1}
\]

\( O_i \) represents the output with molecule \( i \) under shRNA knockdown, \( O_{ij} \), the output with molecules \( i \) and \( j \) under shRNA knockdown. \( O_{ii} \) was approximated as \( 2 \cdot O_i \). The control, output with no shRNA or PI3K repression, \( O_{con} \), was used as a normalization to account for the fact that the output from the in vitro knockdown experiments is the relative change in expression compared to control. The superscript \( t \) represents the trial from which the output was collected, while \( N \) is the number of trials (in our case, 20).

**ACE and Potts Models**

To reconstruct information about the simulated network, we used the ACE method, which infers Potts models based on correlation data from biological and artificial systems. A Potts model is a system of \( N \) variables, in which each variable \( x_i \) can take on \( q \) categories. In our case, the variables are the molecules in our network, each which can take on one of two states: active or inactive.
ACE takes in both pairwise correlation values, which are the $O_{ij}$’s in our correlation matrix, as well as probability priors for the different values each variable can take on, which we calculated as $O_i / O_{con}$, representing the relative amount each molecule is active. ACE returns Potts parameters, $P_i$, for each variable, and $P_{ij}$ for each pair of variables. We interpreted the $P_{ij}$ values as edge weights for the reconstructed network.

**Results and Discussion**

**SSC Simulation Results**

Several patterns were apparent across all our simulation outputs. Due to the stochastic nature of the simulation, molecule levels oscillated across a narrow range. Non-output molecules levelled off after very few timesteps, while output values levelled off after 1000-2000 timesteps, largely due to the slower rate constants for output activation (Figure 4). To model the system at steady state, data for subsequent analysis was taken after the 2000 timesteps. Further, larger inhibition values led to a greater difference in molecule and output values across conditions.
Figure 4. SSC simulation output for single knockdowns with different levels of PI3K and shRNA inhibition. Results were averaged across 20 trials. “Normal” corresponds to the control case with no inhibition. a) out1 b) out2

Looking at output values across different simulation conditions revealed some interesting patterns (Figure 4). For the single molecule knockdowns for out1, we found that as shRNA inhibition levels increase with no PI3K inhibition, only output levels for Rheb and mTORC1 knockdowns decrease, with output levels far lower in the mTORC1 case. With shRNA inhibition, as PI3K inhibition increases, output levels decrease across all conditions except for the TSC inhibition case, which remains level with the control. At higher levels of PI3K and shRNA inhibition these patterns are
more extreme, and the output level under AKT inhibition appears to start falling below the cluster of other output levels. mTORC1 is the closest molecule in the pathway to the output, followed by Rheb. Hence, it appears that the output in our simulation is most sensitive to the nodes more directly associated with it. This is further evidenced by the fact that AKT, as the next closest molecule to upregulate output, appears to have the next greatest effect on decreasing output after mTORC1 and Rheb, if only slight. As TSC is an inhibitor of Rheb, inhibiting TSC would increase output, indicating why output levels under TSC inhibition and PI3K inhibition remain close to control. It is possible that varying rate constants and adding other paths affecting the output could disrupt these patterns.

Similar patterns arise for the single molecule knockdowns for out2. As shRNA inhibition levels increase with no or constant PI3K inhibition, output levels decrease under Raf and Erk inhibition, decreasing far more in the latter case. With PI3K inhibition, the output levels decrease to a greater degree under Raf and Erk inhibition than they do for the corresponding level of shRNA inhibition without PI3K inhibition.

With no shRNA inhibition and PI3K inhibition output levels are lower for out1 than the control case. Such a pattern does not occur for the out2 case.
Looking at the pairwise knockdowns, we find that more complex patterns form in the output. For out1, the pairwise knockdown of mTORC1 and Rheb consistently produced the lowest output. As shRNA inhibition increases several clusters formed, with mTORC1 pairwise knockdowns in general having lower output, followed by Rheb knockdown outputs. As shRNA inhibition increased knockdowns between AKT and mTORC1 and PDK1 and mTORC1 started to have lower output levels. The other exception to the patterns were knockdowns between mTORC1 or Rheb...
and TSC, each of which had a higher output level than the respective mTORC1 or Rheb knockdown. The TSC knockdown pairs that did not involve mTORC1 or Rheb were level with control.

With PI3K inhibition and no shRNA inhibition, the output is lower than the control. With a combination of shRNA and PI3K inhibition the patterns seen in the shRNA knockdowns without PI3K inhibition were more extreme and the output levels were lower for pairwise knockdowns that did not involve TSC.

In the case of out2, the pairwise knockdown of Raf and Erk results in the lowest output. All other pairwise knockdowns with Erk formed a band with the next lowest output, followed by a band of all other pairwise knockdowns with Raf. The remainder of knockdowns had output levels close to the control. This pattern is more pronounced as shRNA increases. With PI3K inhibition and no shRNA inhibition there is no noticeable difference from control, but PI3K inhibition in conjunction with shRNA knockdown enhances the aforementioned patterns.

Comparing the single and pairwise knockdowns at the same strength of PI3K and shRNA inhibition, we found that most pairwise knockdowns have an output corresponding to the output of the molecule that had the largest impact in the single knockdown. The only exception came for pairwise knockdowns of molecules that both had strong effects in the single knockdown (Figure 6).
For out1, all TSC knockdowns except for those with mTORC1 or Rheb had about the same output as the single knockdown of TSC. TSC knockdowns with mTORC1 or Rheb had an output level in between the output of a single TSC knockdown and the respective mTORC1 or Rheb single knockdown. Knockdowns between mTORC1 and Rheb produced lower output than the single knockdown for either mTORC1 or Rheb. Knockdowns between mTORC1 or Rheb and PDK1 or AKT produced output also slightly below the output for the respective mTORC1 or Rheb single knockdown.
For out2, pairwise knockdowns with Raf or Erk had about the same output as the respective Raf or Erk single knockdown case. The only exception was the case in which both Raf and Erk were knocked down, which was much lower. This implies that in this case knockdowns in the other branch did not have a significant impact on output.

For the simulations, mTORC1, Rheb, and TSC, and to a lesser extent AKT and PDK1 knockdowns had the greatest impact on out1. In the case of out2, only Erk and Raf knockdowns had significant impact on output. These results show that in the case of this network molecules with a positive net effect on output that are closer to the output have a greater impact on output. TSC, the only molecule in the simulated network with a net negative effect on out1, also has a significant effect on out1 when knocked down, particularly in combination PI3K inhibition. Given different inhibition conditions led to variation in the outputs, there is potential for the output values for network reconstruction.

**Correlation Matrix Calculations**

We applied our correlation function to our simulation data for each set of PI3K and shRNA inhibition conditions. We found that our correlation function always returns negative values for pairwise correlations from our data, indicating that our simulation always leads to a situation where the output for a double knockdown is less than the sum of the output for the respective single knockdowns. This obscures a clear difference between activating and inhibitory interactions that we were seeking. However, it does indicate a high degree of cooperativity. Furthermore, the fact
that the values within a matrix vary indicate that there may be some interesting insights to be garnered from the data.

**Eigenvalue Analysis**

As a preliminary analysis of our correlation matrices, we made some linear algebra-based inferences by calculating the eigenvalues and eigenvectors of our correlation matrices. For all our correlation matrices, two of the eigenvalues were negative, one of which was consistently the eigenvalue with the largest magnitude. This is due to how our correlation matrices always consist of negative values with zeros along the diagonal. The presence of large negative eigenvalues indicates that our system is nonlinear and that there is a corrective factor in the network pushing the system back towards the control system.

Looking at the two eigenvalues of largest magnitude across PI3K and shRNA inhibition conditions for correlations based on both outputs, we found that the greatest eigenvalue generally increases as shRNA inhibition increases and decreases as PI3K inhibition increases. The eigenvalue with second greatest magnitude is generally positive, and increases as PI3K and shRNA inhibition increases. For correlation matrices for out1 as shRNA grows particularly high with P high, the other negative eigenvalue is second greatest in magnitude.

As there are negative eigenvalues, one cannot calculate the amount of variance in the system each eigenvector captures. However, the eigenvectors associated with the eigenvalues of greatest magnitude still represent the main modes of the system. As a result we plotted the eigenvectors associated with the two eigenvalues with greatest magnitude. The results can be seen in Figure 7.
Figure 7. Plot of the eigenvectors of the correlation matrices associated with the eigenvalues with the greatest magnitude by the second greatest magnitude. Points labelled by associated molecule. a) out1 b) out2

While the orientation of the points shifts across conditions, when looking at locations of points relative to each other, a couple of patterns are evident. At low levels of inhibition, Raf is far away from the other points, although as shRNA inhibition increases it draws closer. At high levels of shRNA inhibition in which the second eigenvalue is negative the points shift to a diagonal. For out1 TSC is on one end of the diagonal and Rheb and mTORC1 on the other end, and PTEN, Erk,
and PDK1 are clustered. For out2 mTORC1 and Rheb are on one end of the diagonal and Raf and Erk on the other. Given TSC, Rheb, and mTORC1’s impact on out1 and Raf and Erk’s impact on out2, the order along the diagonal for high levels of inhibition may correspond to the amount of impact inhibition of each molecule has on the output.

With no shRNA inhibition no diagonal forms, Raf remains far from the other nodes, although the order of the nodes on the Y axis does shift. The addition of PI3K inhibition does result in the eigenvectors forming a diagonal with lower levels of shRNA inhibition.

With a greater number of nodes in play, this method may be able to reveal more interesting information about the system.

**ACE and Network Reconstruction**

We then investigated whether our output metrics from the single and pair-knockdown simulations could be used to reconstruct the original network via ACE. We applied ACE to our correlation matrices to calculate the Potts parameters. We plotted the resultant networks in Cytoscape, a program for visualizing and analyzing networks [29]. We weighted edges by the magnitude of the Potts parameter between nodes. Some examples of our results are shown in Figure 8.
Figure 8. Potts variables from our simulations plotted as edge weights in a network across different PI3K and shRNA inhibition levels. Graphs generated in Cytoscape. a) out1  b) out2
Instead of assigning larger Potts parameters to direct connections in the network and low parameters to molecules that don’t directly interact in the network as we hoped, ACE seemed to concentrate edge weights around certain nodes. This indicates that for this small network, the method is capturing effective pairwise interactions that include long-range cooperative effects. As a result, instead of defining a cutoff and selecting edge weights above a cutoff and comparing the resultant graph to our original graph, we focused our analysis on the edge weights associated with each node to gain insight into patterns from our ACE output. As Potts parameters were almost entirely negative, more negative edge weights were thicker. Colors were added to elucidate patterns, with blue corresponding to more negative edge weights. As shRNA inhibition increases the range of Potts variables increases, which might help with statistical comparison of Potts variables. However, to better compare patterns across simulations for this study we set edge weights on a set by set basis as opposed to using the same weights across inhibition conditions.

With no PI3K inhibition, as the amount of shRNA inhibition increased, the edge weights for mTORC1, followed by Rheb, were the weakest. In the case of out2, the edge weights associated with Erk, followed by Raf, were weakest. In the case of out1, edge weights associated with TSC were the greatest, while in out2 the edge weights associated with mTORC1 were greatest. Further, in out1 the mTORC1-Rheb edge weight was weakest and in out2 the Raf-Erk edge weight was weakest. This corresponds to the output levels from simulations (Figure 10).
With PI3K inhibition, in the case of out1 the TSC edge weights were more pronounced (Figure 9). For out2 the edge weights from Raf and Erk were stronger compared to the no PI3K inhibition case, suggesting high PI3K inhibition might mask some patterns.

When there was no shRNA inhibition, the edge weights from TSC were relatively higher than edge weights from the other nodes (Figure 11). This pattern was far more extreme in the out1 case than the out2 case. This is likely due to, as mentioned in the analysis of the simulation output, the fact that TSC has a net inhibitory role on out1.

Figure 9. Comparison of Potts model output for out1 (a) and out2 (b) at high inhibition (PI3K inhibition 20, shRNA inhibition 100). Plots generated in Cytoscape.
Figure 10. Comparison of Potts model output for out1 (a) and out2 (b) with no PI3K inhibition (PI3K inhibition 0, shRNA inhibition 100). Plots generated in Cytoscape.

Figure 11. Comparison of Potts model output for out1 (a) and out2 (b) at with no shRNA inhibition (PI3K inhibition 20, shRNA inhibition 0). Plots generated in Cytoscape.
ACE failed to reconstruct the simulated network. Instead, edge weights between two nodes appeared to be correlated with the output level from their pairwise knockdown. This indicates that ACE is capturing and perhaps amplifying which pairwise knockdowns have greatest effect on the output via cooperative effects rather than resolving direct connections.

**Conclusion**

Understanding the pathways involved in leukocytes is key for both enhancing disease treatment and treating autoimmune diseases such as leukemia. We explored this problem in the context of PI3K signaling in leukemia cells through a shRNA knockdown screen across a broad library of genes. Graphical model analysis has the potential to construct a putative network of the interactions in the leukemic PI3K network, which will provide interactions and genes of therapeutic interest. Further, a major question in cancer therapies is whether it is better to thoroughly target one pathway or hit as many distinct and synergistic pathways as possible. Understanding the interactions and pathways involved in various cancer types will aid in answering such a question as well as illuminate strong targets.

To approach this problem, we tested methods to reconstruct information about the PI3K network from shRNA knockdown experiments. We created a toy model of a PI3K pathway and used an SSC simulation to simulate cell growth under different single and double-molecule inhibition conditions for the pathway. From this data we generated a correlation coefficient matrix and used ACE analysis to generate Potts parameters, which in turn were used as edge weights to reconstruct the network. Our reconstructed networks did not capture the direct interactions in our simulated network. Instead the ACE method reflected the impact of single and pairwise knockdowns on
output, including potentially cooperative effects. As the output in our simulation captures to some degree distance from the output and positive and negative net impact of molecules on the output, ACE also captures these metrics. Further, we found that relative inhibition strengths in our simulation affect ACE output.

As a result there is potential for our method to reveal knockdowns of interest and possibly be used to capture network structure. It is possible interactions with output dominate too much and even with pairwise knockdowns the method does not reveal enough about molecular interactions with each other. However, our simulation only investigated one small network structure with constant activation and deactivation rates. Replicating this process with different and more realistic large networks will be essential for fully assessing our method’s effect.

**Future Steps**

Moving forward, the first step would be improving our methodology to more accurately capture networks from knockdown data. As our correlation function only returned negative values on the simulation data, we will need to revise our correlation function and/or our simulation to differentiate between positive and negative interactions between molecules. It is also likely that we will need to revise our output metric to better approximate the cellular growth measured by the log-fold change in shRNA counts from in vitro experiments. For multiple pathways that may affect cell growth, we may need to make a new metric that’s the weighted average of outputs from various pathways, or else devise a single output node to which multiple pathways contribute.
To better understand what ACE picks up from simulation data we could systematically simulating various network motifs and determine the ACE output akin to the Blair study [17]. In particular, structures of interest to study would include nodes upstream of PI3K, analyzing individual paths and branching paths, and molecules not linked to the network, as well as motifs such as feedforward and feedback loops. This could allow us to get a better sense of exactly how our methodology performs on different network structures, which could guide revising our methods. Based on this analysis we might determine that ACE is unsuited for the problem or devise a method to translate ACE output into the original network.

Direct Coupling Analysis (DCA) is a potential method that could be applied to further disentangle indirect from direct correlations. DCA is a statistical inference framework based on maximum entropy to determine the strength of direct relationships between two nodes in a biological sequence and disentangles direct contributions to correlations from indirect ones. It has been used in many biological problems such as multiple sequence alignment and predicting contacts in protein folding [21].

Other routes for improvement include considering methods that generate several likely networks since multiple networks could explain the data equally well [17]. Further, as biological networks significantly differ from random networks [15], we could use knowledge of common biological network structure and motifs as heuristics to enhance algorithm performance. Devising a method to infer directionality in the network would also be of interest.
If our revised method performs accurately on reconstructing simulation data, the next step would be to test its performance on in vitro data from a known network before proceeding to applying it to data from unknown networks. Testing performance on in vitro data from a known network is key as variations in in vitro networks compared to the simulated network may require further revisions of the simulation and the methodology.

Another potential concern moving forward is scaling to larger networks. ACE can be applied to systems as large as several hundred variables [25]. Given the search for molecules of interest could extend to larger libraries, it may be necessary to devise methods for filtering out variables less likely to play a role in the system of interest or sampling different combinations of variables and applying ACE analysis. In addition, further testing of simulations of larger systems will need to be performed to determine how performance changes with scaling.

While ACE failed to reconstruct the direct network connections in this study, it produced several patterns of interest indicating its potential for revealing insights into the network from knockdown experiment data. Therefore modifications to our procedure and a more in depth understanding of how ACE performs on different network structures and motifs could improve its performance and allow us to infer more information about the PI3K network from shRNA knockdown studies.
Acknowledgements

I would like to express my gratitude to my advisor Professor Arup Chakraborty for his wisdom, feedback, and motivation. I’d like to thank our collaborators Jeroen Roose and Marsilius Mues at UCSF for their motivating experiment. I’d also like to thank my fellow labmates in the Chakraborty Lab for their support and advice: Assaf Amitai, John Barton, Hanrong Chen, Kevin Kaczorowski, Myung Sun Kang, Allen Lin, Joy Louveau, Dariusz Murakowski, Krishna Shrinivas, and Alexey Shvets. Thanks are also due to Renée Saindon at the Chakraborty lab for her encouragement and support.

Bibliography


Appendix

Simulation Code

region World box width 1 height 1 depth 1
subvolume edge 1

-- Define initial vals
new PI3K(active="yes") at count1
new PI3K(active="no") at count2
new PTEN(active="yes") at count1
new PTEN(active="no") at count2
new PDK1(active="yes") at count1
new PDK1(active="no") at count2
new AKT(active="yes") at count1
new AKT(active="no") at count2
new mTORC2(active="yes") at count1
new mTORC2(active="no") at count2
new TSC(active="yes") at count1
new TSC(active="no") at count2
new Rheb(active="yes") at count1
new Rheb(active="no") at count2
new mTORC1(active="yes") at count1
new mTORC1(active="no") at count2
new Raf(active="yes") at count1
new Raf(active="no") at count2
new Erk(active="yes") at count1
new Erk(active="no") at count2
new out1(active="no") at count4
new out2(active="no") at count4

-- Auto reactions
rxn x:TSC(active="no") at rate1 -> x.active="yes"
xrn x:Rheb(active="no") at rate1 -> x.active="yes"
xrn x:mTORC1(active="yes") at rate2 -> x.active="no"
xrn x:Raf(active="yes") at rate2 -> x.active="no"
rxn x:Erk(active="yes") at rate2 -> x.active="no"
rxn x:out1(active="yes") at rate6 -> x.active="no"
rxn x:out2(active="yes") at rate6 -> x.active="no"
rxn x:PI3K(active="no") at rate1 -> x.active="yes"
rxn x:PDK1(active="yes") at rate2 -> x.active="no"
rxn x:AKT(active="yes") at rate2 -> x.active="no"
rxn x:mTORC2(active="no") at rate1 -> x.active="yes"
rxn x:PTEN(active="no") at rate1 -> x.active="yes"

-- Phosphorylation reactions
-- Left branch
rxn PI3K(active="yes") x:PDK1(active="no") at rate3 -> x.active="yes"
rxn PI3K(active="yes") x:AKT(active="no") at rate3 -> x.active="yes"
rxn PDK1(active="yes") x:AKT(active="no") at rate3 -> x.active="yes"
rxn mTORC2 x:AKT(active="yes") at rate4 -> x.active="no"
rxn PTEN x:PDK1(active="yes") at rate4 -> x.active="no"
rxn PTEN x:AKT(active="yes") at rate4 -> x.active="no"
rxn AKT(active="yes") x:TSC(active="yes") at rate4 -> x.active="no"
rxn TSC(active="yes") x:Rheb(active="yes") at rate4 -> x.active="no"
rxn Rheb(active="yes") x:mTORC1(active="no") at rate3 -> x.active="yes"
-- Right branch
rxn PI3K(active="yes") x:Raf(active="no") at rate3 -> x.active="yes"
rxn Raf(active="yes") x:Erk(active="no") at rate3 -> x.active="yes"

--Inhibition rxns
rxn x:PI3K(active="yes") at rate7 -> x.active="no"
--Variable names modified for different simulations to be knocked down
rxn x:AKT(active="yes") at rate8 -> x.active="no"
rxn x:Erk(active="yes") at rate8 -> x.active="no"

-- Output rxns
rxn mTORC1(active="yes") x:out1(active="no") at rate5 -> x.active="yes"
rxn Erk(active="yes") x:out2(active="no") at rate5 -> x.active="yes"
Simulation Parameters
#Constants for test simulation
rate3 = 1.00 #activation
rate2 = 1.00 #constitutive deactivation
rate1 = 1.00 #constitutive activation
rate4 = 1.00 #deactivation
rate5 = 0.0001 #output activation
rate6 = 0.001 #output deactivation
rate7 = 1.00 #weak PI3K inhibitor, modified based on simulation
rate8 = 5.00 #shRNA inhibition, modified based on simulation

count1 = 10 #initially active
count2 = 20 #initially inactive
count4 = 500 #output count

Table S1. Inhibitor values tested across simulations.

<table>
<thead>
<tr>
<th>shRNA</th>
<th>PI3K</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>20</td>
<td>10</td>
</tr>
<tr>
<td>50</td>
<td>20</td>
</tr>
<tr>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>