Thesis Proposal

Siddhartha Jain

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Ziv Bar-Joseph, Chair
Jaime Carbonell
Eric Xing
Naftali Kaminski

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy.

Keywords: genomics, computational biology, signaling networks, network inference, regulatory networks

Abstract

Cells need to be able to sustain themselves, divide, and adapt to new stimuli. Proteins are key agents in regulating these processes. In all cases, the cell behavior is regulated by signaling pathways and proteins called transcription factors which regulate what and how much of a protein should be manufactured. Anytime a new stimulus arises, it can activate multiple signaling pathways by interacting with proteins on the cell surface (if it is an external stimulus) or proteins within the cell (if it is a virus for example). Disruption in signaling pathways can lead to a myriad of diseases including cancer. Knowledge of which signaling pathways play a role in which condition, is thus key to comprehending how cells develop, react to environmental stimulus, and are able to carry out their normal functions.

Recently, there has also been considerable excitement over the role epigenetics – modification of the DNA structure that doesn't involve changing the sequence may play. This has been buoyed by the tremendous amount of epigenetic data that is starting to be generated. Epigenetics has been heavily implicated in transcriptional regulation. How epigenetic changes are regulated and how they affect transcriptional regulation are still open questions however.

In this thesis we present a suite of computational techniques and tool and deal with various aspects of the problem of inferring signaling and regulatory networks given gene expression and other data on a condition. In many cases, the amount of biological data available for a condition can be very small compared to the number of variables. We will present an algorithm which uses multi-task learning to learn signaling networks from many related conditions. There are also very few tools that attempt to take temporal dynamics into account when inferring signaling networks. We will present a new algorithm which attempts to do so and significantly improves on the state of the art. Finally, we propose to work on integrating epigenetic data into the inference of signaling and regulatory networks.

Contents

Chapter 1

Introduction

1.1 Background and motivation

Transcription is the process by which RNA molecules are creating based on the information stored in the DNA. A DNA molecule is divided into genes, both coding (those subsequently converted into proteins) and non-coding, microRNAs, tRNAs, and many other elements. The processs of transcribing a gene is called gene expression. The process of gene expression is highly complex. To start with, one or more proteins called transcription factors (TFs) bind to so called enhancer sequences which help regulate gene expression. These TFs recruit a series of TFs called general transcription factors (GTFs). The GTFs recruit an enzyme called RNA polymerase II and induce it to bind to the gene promoter (upstream of the actual gene sequence) forming the pre-initiation complex (PIC). After that transcription commences. Just transcribed RNA (termed pre-mRNA) is then processed and converted to messenger RNA (mRNA). These mRNAs are read by ribosomal proteins and converted into proteins which then perform various functions in the cell including regulation of transcription, cell signaling, responding to stimuli, inducing transcriptional patterns to generate more proteins to defend against pathogens, etc. Knowledge of what signaling proteins and TFs are involved in the response to any pathogen is vitally important in understanding how to disrupt the pathways that pathogen might be using to hijack the cellular machinery and self-propagate (for example by targeting proteins aiding viral reproduction or cancer propagation via drugs).

Many previous attempts to detect genes that play a functional role in a phenotype (such as the propagation of a viral infection) rely on gene expression knockdowns or knockouts. There remain several problems with such an approach. While a gene knockdown or knockout may have little effect on a phenotype (such as cell division) under normal conditions, it could have very different effects under chemical or environmental stress conditions [40]. In addition, even gene knockdown studies meant to test gene relevance to phenotype under similar or even virtually identical conditions can drastically differ in their results. For example, three well known knockdown studies for detecting genes related to HIV-1 had a pairwise overlap of $< 7\%$ in the genes they detected [14]. Various explanations are suggested, including experimental noise, differences in timing of sampling and differences in filtering criteria used to selected hits. In fact, the authors of one of the screens performed a duplicate screen to estimate experimental variance

and found that only 50% of the top 300 hits would be obtained under identical experimental conditions [14, 41]. Such results suggest that to experimentally estimate functional relevance, one would have to do genome-wide knockdowns or knockouts anytime the experimental conditions even slightly change requiring a staggering amount of experimental effort. Compounding the problem are changes like epigenetic modifications which could drastically change the results from one cell type to another or from one condtion to another.

Even if one had the resources to be able to do that, a more troubling problem is that sophisticated backup mechanisms exist in regulatory networks that can obscure the true role of transcription factors (TFs). One would expect the expression of genes directly bound by a TF to be affected by the knockdown of that TF. In [43], 269 TFs in yeast were knocked down one at a time. The differentially expressed genes so obtained were compared to the protein-DNA binding data from [39]. Surprisingly, they found that only 3% of bound genes were affected by the knockdown. A large part of the explanation is the existence of redundant TFs which can obscure the role the TFs in general may play [32]. Another way to put it is that TFs (and perhaps signaling proteins in general) can act in concert. If we had the ability to perform knockdowns of every combination of genes, then we would be able to solve this problem but that would quickly lead to combinatorial explosion and is thus infeasible.

A third problem which so far has received less attention in literature is *when* do signaling pathways and TFs get triggered in terms of timing relative to each other. For example, if we have a time series gene expression dataset, then we want to understand the different signaling pathways and TFs that trigger differential gene expression at the different time points. This is tough to detect experimentally. Gene knockdowns via siRNA or shRNA usually require upto 48-72 hours to result in a substantial knockdown of the gene expression in a majority of the cells [24, 84]. Thus any signaling events happening on a timescale smaller than that are not possible to differential between temporally. However the temporal annotation can turn out to be relevant biologically. For example, the Src kinase LCK is involved in HIV-1 viral assembly. We know that the viral assembly phase of HIV-1 occurs starting about 16 hours after the cell is infected with the virus. Thus, if we are able to detect LCK as being relevant at that time point, we could subject LCK to more rigorous testing to see if there is a link between the late phase activities of HIV-1 infection and LCK (as we show later, our temporal annotation algorithm is indeed able to detect LCK as a late phase signaling protein). While there has been work on inferring which TFs are active at which time points [11, 25], there has been no work, as far as we are aware, on temporal annotation of signaling pathways.

Given that experimental techniques are not sufficient, we need to turn to computational methods to aid us. High throughput data measuring various aspects of several biological systems is rapidly accumulating. These include RNA-Seq studies [64], profiling of microRNAs [88], ChIP-Seq, epigenetics studies [30], information about protein interactions within a cell [72] and information on interactions between host proteins and pathogen / environmental factors [65]. Such datasets provide extensive information about the sets of genes that are activated, their regulation and their interactions both within a cell and between cellular proteins and the environment or pathogen. However, integrating these datasets to reconstruct a unified view of the networks and pathways that are activated in order to identify potential interventions that may lead to a desired response remains a major challenge.

1.2 Thesis goals

In this thesis, we propose to address three aspects of the above problem :-

1. **Using multitask learning to reduce overfitting**. The number of samples available for a particular condition is usually very limited in comparison to the number of possible biological variables when reconstructing signaling and regulatory networks. We use *multi-task learning* to alleviate this problem. We develop the tool, Multi-Task Signaling and Dynamic Regulatory Events Miner (MT-SDREM), which uses multi-task learning to reconstruct response pathways and temporal regulatory networks.
2. **Constructing temporal pathways which explain the differential gene expression**. While several methods have been proposed to reconstruct signaling networks, there has been no work, as far as we are aware, that tells you when particular signaling pathways were activated – i.e. gives a temporal annotation to the signaling proteins of the reconstructed networks. We develop an *Integer Programming* formulation to solve this problem.
3. **Incorporating epigenetic data into signaling and regulatory network inference**. There is a large body of literature on how to infer signaling and regulatory networks for a given condition. However an important aspect that all of the above methods don't consider is the role epigenetic modifications play in regulating gene expression. Given our focus on trying to infer signaling pathways and active TFs for various conditions, we propose modeling how epigenetic modifications can affect TF-DNA interactions and thus affect gene regulation.

1.3 High-throughput data used in this thesis

Many high-throughput experimental methods have been developed to study various aspects of transcriptional regulation either directly or indirectly. Below we provide short descriptions of data used.

1.3.1 RNA sequencing

RNA-Seq is a recently developed approach to transcriptome profiling that uses deep-sequencing technologies to reveal the presence and quantity of RNA in a biological sample at a given moment in time. All our gene expression data comes from RNA-seq. See Figure 1.1 for an overview of how a typical RNA-seq experiment is conducted. In [20], a detailed review of the RNA-seq pipeline and a survey of best practices for RNA-seq data analysis is provided.

1.3.2 Chip-Chip and Chip-Seq

Chromatin immunoprecipitation (ChIP) followed by microarray (ChIP-chip) [13] or sequencing (ChIP-seq) [66] has been developed to study genome-wide TF binding in vivo. The *in vivo* protein-DNA interactions are first cross linked by formaldehyde, and then these cross linked chromatin is sheared into fragments. The TF of interest is immunoprecipitated with specific antibody, and then the cross linking is reversed to release the bound DNA fragments. The location

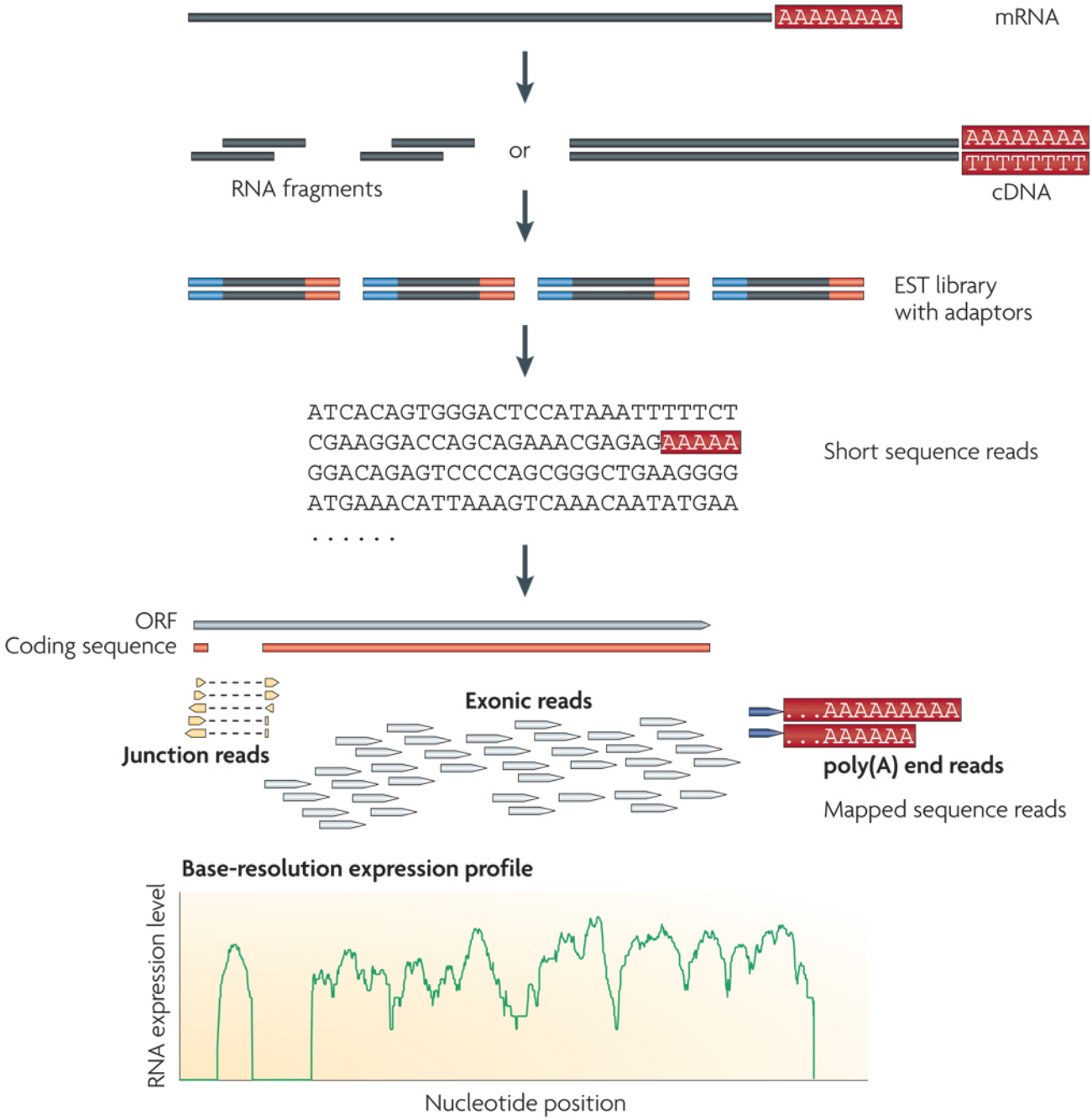

Figure 1.1: Briefly, long RNAs are first converted into a library of cDNA fragments through either RNA fragmentation or DNA fragmentation. Sequencing adaptors (blue) are subsequently added to each cDNA fragment and a short sequence is obtained from each cDNA using high-throughput sequencing technology. The resulting sequence reads are aligned with the reference genome or transcriptome , and classified as three types: exonic reads, junction reads and poly(A) end-reads. These three types are used to generate a base-resolution expression profile for each gene, as illustrated at the bottom (TopHat and Cufflinks are a popular tool combination to do this [86]). The reads are typically converted to RPKM/FPKM/TPM units which are a measure of the number of transcripts in the cell. Figure is taken from [91]

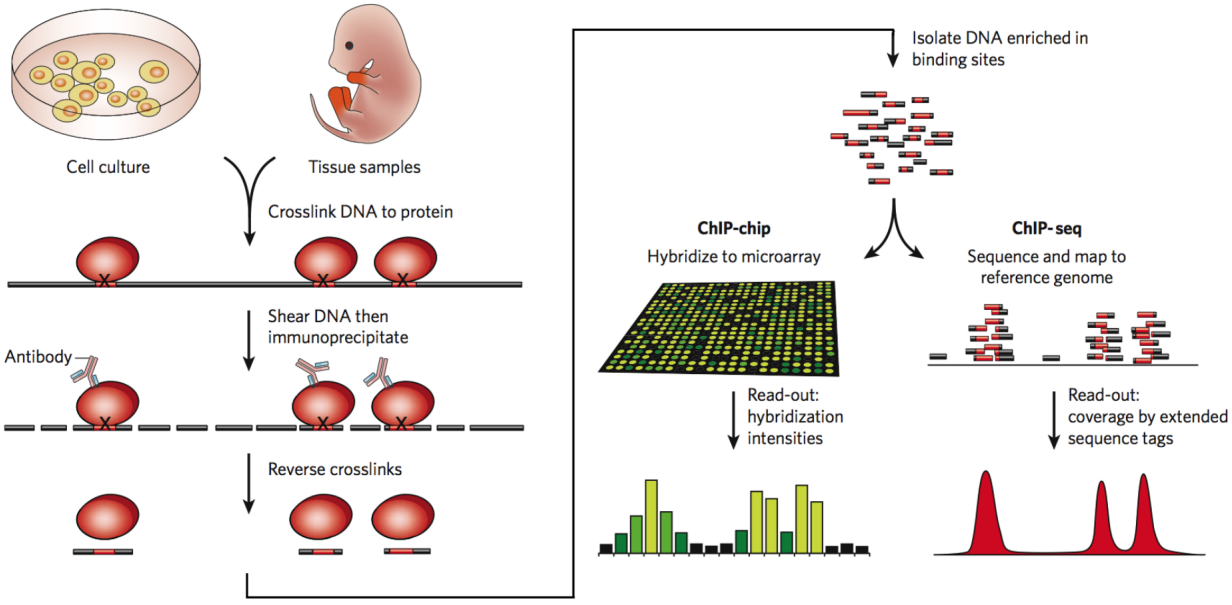

Figure 1.2: An overview of a typical Chip-chip or Chip-seq experiment. It shows cells to examine being taken from a culture or tissue sample. Proteins attached to DNA are then cross linked to the DNA (usually formaldehyde is used). Then the chromatin is sheared and the protein of interest precipitated out using antibodies. The cross links are reversed, the DNA extracted and then sequenced using a microarray or next generation sequencing. The sequenced DNA is mapped to a reference genome to figure out the genome sites to which the antigen protein binds. Image has been taken from [101]

of these DNA fragments bound by the TF is then determined by either hybridization to specific microarray containing promoter regions from the genome (ChIPchip), or by direct sequencing and aligning to the reference genome computationally (ChIP-seq) [101]. In the end we obtain hybridization intensities (in case of ChIPchip) and tag densities (in case of ChIPseq) for the whole geneome. Peak calling software can be run to identify true binding sites. Chip-Seq can also be used to detect mehthylation patterns and histone marks by using the appropriate antibodies. An overview of the experimental method is given in Figure 1.2.

We process Chip-Seq data from ENCODE [29] for 348 transcription factors to get our human TF-DNA interaction network as in [76] comprising of 59K TF-DNA interactions.

1.3.3 Protein-protein interactions

Several experimental techniques of varying levels of accuracy exist to detect protein-protein interactions including Yeast-2-hybrid, Immunoprecipitation, Co-crystallization, etc. [68] gives a nice overview of the various experimental techniques to detect protein-protein interactions. For our human protein-protein interaction network, we used the BIOGRID [80] and HPRD [72] databases which collate interactions for the above such experimental sources. An interaction

could have been detect in multiple independent experiments. We processed the interactions as in [46] to obtain a weight set of edges between the proteins.

1.3.4 Virus-Host interactions

These are interactions between viral proteins and host cell proteins (that the virus has invaded). These interactions are detected using the same techniques as general protein-protein interactions. We obtained interactions between viral proteins and host cellular proteins from HPRD as well as VirHostNet [65].

1.3.5 RNAi screens

RNAi is an endogenous cellular process by which messenger RNAs are targeted for degradation by double-stranded (ds) RNA of identical sequence, leading to gene silencing. These can be small interfering RNA (siRNA) or small hairpin RNA (shRNA). Initially used to knock down the function of individual genes of interest, the technology was harnessed in several organisms on a global scale with the production of RNAi libraries to silence most of the genes in their genomes, allowing genome-wide loss-of-function screening [12]. For example, They are often used to check whether a gene is causally related to a phenotype of interest (e.g. viral load) but knocking down the gene and then measuring the phenotype. We use genome-wide RNAi screen for HIV and Flu (H1N1 and H5N1) as a means to validate our predictions. An overview of the RNAi process is in Figure 1.3.

1.3.6 Gene ontology

Gene ontology (GO) attempts to annotate genes with their biological context – specifically which cellular components they are usually present in, what molecular functions they perform, and what biological processes they are involved in [8]. Checking for enrichment of GO categories among a group of genes is a useful and quick way to get an idea of the biological meaning of one's results. We use this technique as another method of validating our findings.

1.4 Computational techniques used in this thesis

Below we give a very brief overview of two of the main computational techniques we have used in this thesis.

1.4.1 Multi-task learning

Multi-task learning is an approach to machine learning that learns a group of related problems together, using a partly shared representation. This allows one to effectively increase the amount of data available per parameter and reduce overfitting. This is especially important when reconstructing biological response networks from high-throughput data because the number of parameters to fit is very large relative to the number of samples. In addition, extensive data

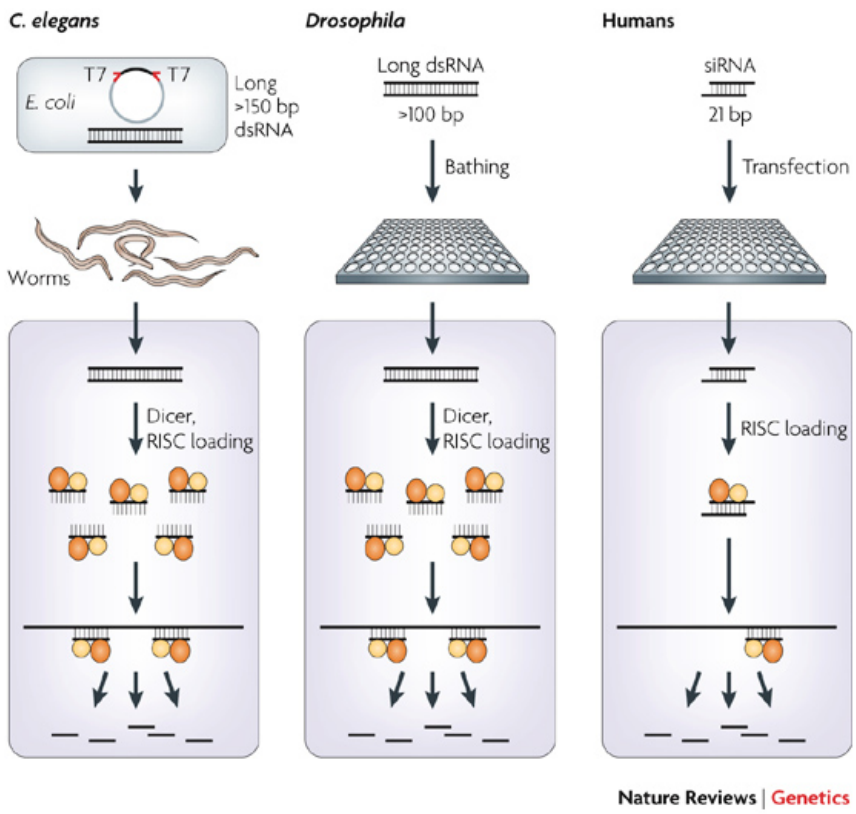

Figure 1.3: Overview of RNAi screening approaches used in different organisms. Long double-stranded (ds) RNAs are introduced into a cell or and are intracellularly diced into small-interfering RNAs (siRNAs). This leads to highly efficient knockdown because many different siRNAs are generated from each dsRNA. Introduction of siRNAs into human (or vertebrate) cells requires transfection. RNAi screens in human cells usually require multiple independent siRNAs, either in individual wells or delivered as pools. Other methods for human cells include viral transduction of hairpin expression constructs or endoribonuclease-derived siRNAs (esiRNAs), essentially pool of extracellular diced long dsRNAs. RISC, RNA-induced silencing complex; T7, bacteriophage T7 promoter. Image taken from [12]

from a well-characterized condition may be able to compensate for sparse data in a similar, less-understood condition. Multi-task learning has been applied to many problems in the biological domain including classification [92], genome-wide association studies [51, 52], protein structure [45], and pairwise protein-protein interaction prediction [53, 73].

As a primer on the general multitask framework, we discuss a common formulation of the multitask learning problem.

The objective function commonly used for multi-task learning combines two related goals: First, similar to standard machine learning applications (for example, classification) it tries to minimize the loss (i.e. error) for each task while at the same time regularizing the parameters used by each task to avoid overfitting. Second, it further regularizes the parameters *across tasks* so that the final parameters are similar. A typical objective function is the following [26]

$$\underset{\mathbf{w_1},\ldots,\mathbf{w_C}}{\operatorname{argmin}} \left[\left\{ \sum_{i=1}^{C} L(\mathbf{y_i}, f(\mathbf{w}_i^T \mathbf{x_i})) + \lambda_1 \cdot ||w_i||_p \right\} + \left\{ \lambda_2 \cdot \sum_{i=1}^{C} \sum_{j=i+1}^{C} ||w_i - w_j||_p \right\} \right]$$

where C is the number of tasks, L is the loss function, f is a function of the dot product of the task-specific weight vector and the data for the task, and p is the L_p norm for the regularization. The left, red part, T_1 is the *task-specific* part of the objective function while the right, blue part, T_2 is the regularization *across* tasks.

1.4.2 Integer programming

Integer programming is a mathematical optimization technique in which one has a linear objective function to minimize or maximize, a set of linear inequality constraints, and a subset of the variables are restricted to only integer values.

An integer program in canonical form is expressed as

$$\begin{aligned} &\max \mathbf{c^T x} \\ &\text{subject to } \mathbf{Ax} \leq \mathbf{b} \\ &\mathbf{x} \geq \mathbf{0} \end{aligned}$$

This is in general an NP-hard problem [95] and thus unlikely to have an efficient solution in all cases. However, over the past decades, there has be a tremendous amount of progress in making this problem tractable for many practical cases. A typical strategy involves using a branch and bound algorithm in combination with sophisticated branching heuristics and solving linear programs to upper bound the optimal solution in case of a maximization problem (or lower bound for minimization problem) [95]. This is what's known as a complete algorithm – as in such an algorithm will eventually find the optimal solution and provide a proof of its optimality. However, often times, we don't need to find the absolute optimal solution. The advantage of settling for a solution that is close to the optimal (but not actually so) is that we can apply much faster algorithms that scale much better, for example simulated annealing, tabu search, large neighborhood search, etc. [56]. As we shall see later in this chapter, due to the size of our problem, we are forced to resort to the latter techniques.

1.5 Structure of this proposal document

In Chapter §2, we look at the multitask aspect of the problem mentioned in point (1) and present our algorithm MT-SDREM. We also discuss an ongoing project to apply MT-SDREM to time series gene expression data from Arabidopsis Thaliana. In Chapter §3, we present TimePath which can be used to temporally annotate signaling pathways. We also discuss plans to apply the algorithm to the Arabidopsis data as well as expression data from IPF lung disease samples. Finally in Chapter §4, we discuss future plans to incoporate epigenetic data into our models.

Chapter 2

MT-SDREM

MT-SDREM extends the Signaling and Dynamic Regulatory Events Miner (SDREM) which has so far only been applied to reconstruct response networks for a single condition at a time [34]. Prior to discussing the multi-task learning procedures we first briefly discuss the SDREM method. SDREM is an iterative procedure that combines regulatory and signaling network reconstruction to model response pathways. For the regulatory part, SDREM uses time series gene expression data with protein-DNA interaction data to identify bifurcation events in a time series (places where the expression of previously co-expressed set of genes diverges – see Figure 2.1), and the transcription factors (TFs) controlling these split events. While some TFs are transcriptionally activated, others are only activated post-translationally via signaling networks. To explain these TFs, the second part of SDREM links sources (host proteins that directly interact with the virus / treatment) to the TFs determined to regulate the regulatory network. This part of SDREM uses protein-protein interaction (PPI) and protein modification data to infer such pathways – while imposing the constraint that the *direction* of PPI in the inferred pathways is consistent. These two parts (regulatory and signaling reconstruction) iterate a fixed number of times until the final network is obtained. See [34] for complete details.

Like its single-condition predecessor [34], MT-SDREM iterates between finding pathways that connect the upstream proteins that directly interact with an external stimulus (called source proteins) and the downstream transcription factors (TFs) that regulate the response and learning dynamic regulatory networks activated by these TFs. The learning process involves the simultaneous reconstruction of several such networks. While a different network is learned for each condition, the joint learning framework allows sharing and/or constraining parameters across the different networks which helps overcome the overfitting problem that is often an issue when reconstructing biological networks.

We demonstrate how MT-SDREM can be used to gain insights into a clinically-relevant problem: characterizing the human response to viral infection. In particular, we explore the differences between mild, seasonal strains of the influenza A virus, which are typically H1N1 or H3N2 strains [28], and lethal, pandemic strains such as the H1N1 1918 Spanish flu and highly pathogenic avian H5N1 strains.

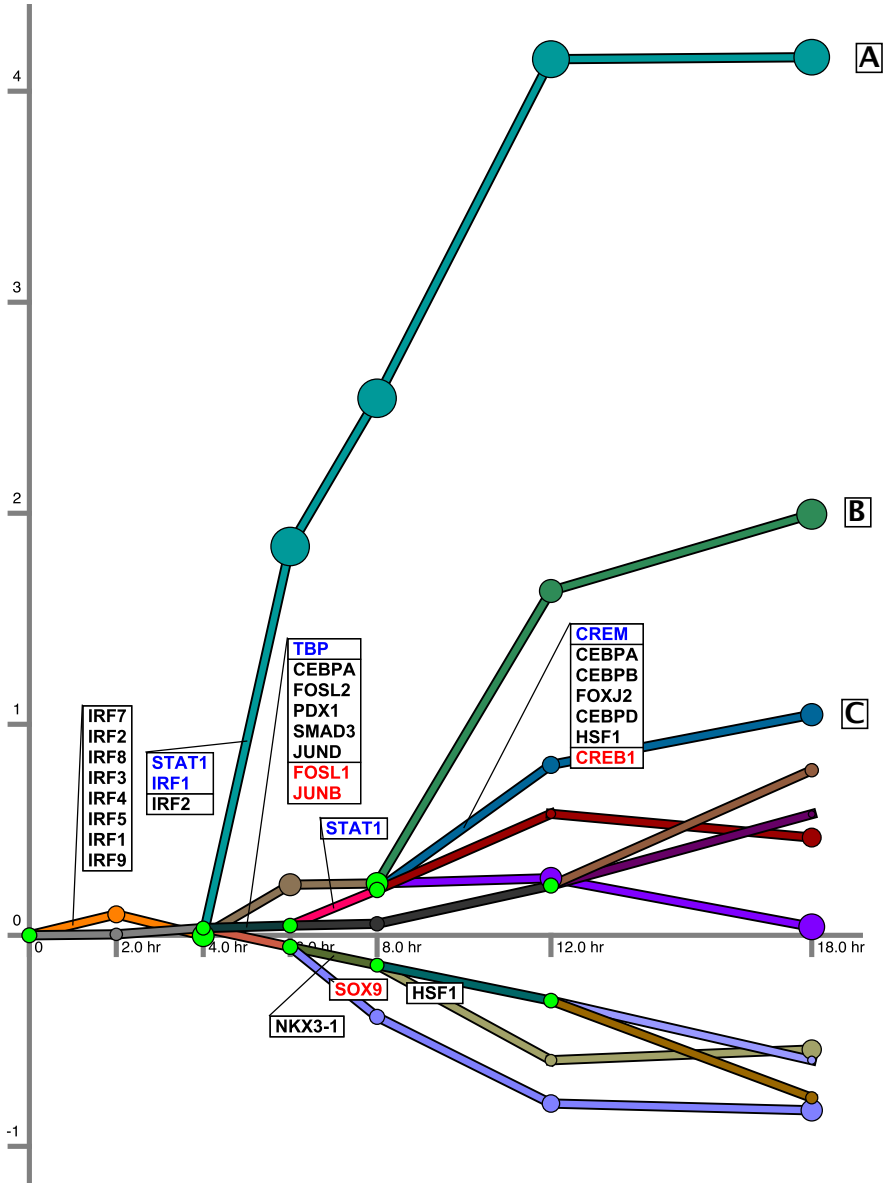

Figure 2.1: H1N1 Regulatory network. Each path represents a set of genes with a similar expression profile. Split nodes are colored green and are annotated with the TFs that are predicted to regulate genes in the paths going out of the split at the time point associated with the split. The blue TFs are up-regulated at that split time point while the red TFs are down-regulated. The black TFs are not differentially expressed at the split point. Note that several of the TFs included in this latter group are likely post-transcripitionally regulated.

2.1 Completed research

Here we present the Multi-Task Signaling and Dynamic Regulatory Events Miner (MT-SDREM) [46], which uses multi-task learning to reconstruct response pathways and temporal regulatory networks. MT-SDREM is equipped to capitalize on the many dimensions in complex systems biology datasets by integrating different types of experimental data in each condition, explaining the time-dependent elements of a response (as observed in gene expression data), and constraining the inferred networks to be similar for related conditions or perturbations. Like its single-condition predecessor [34], MT-SDREM iterates between finding pathways that connect the upstream proteins that directly interact with an external stimulus (called source proteins) and the downstream transcription factors (TFs) that regulate the response and learning dynamic regulatory networks activated by these TFs. The learning process involves the simultaneous reconstruction of several such networks. While a different network is learned for each condition, the joint learning framework allows sharing and/or constraining parameters across the different networks which helps overcome the overfitting problem that is often an issue when reconstructing biological networks.

We demonstrate how MT-SDREM can be used to gain insights into a clinically-relevant problem: characterizing the human response to viral infection. In particular, we explore the differences between mild, seasonal strains of the influenza A virus, which are typically H1N1 or H3N2 strains [28], and lethal, pandemic strains such as the H1N1 1918 Spanish flu and highly pathogenic avian H5N1 strains.

2.1.1 MT-SDREM

MT-SDREM simultaneously investigates and infers regulatory networks and signaling pathways for several biologically related conditions. For this, it uses both condition-specific gene expression and interaction data and general interaction data. We first discuss the input data that the method utilizes and then present the modeling and learning frameworks.

Input Data

We use C to denote the set of conditions that are jointly modeled by MT-SDREM. Below we list the datasets used by MT-SDREM.

1. *Condition-specific: Time series gene expression data* for each of the conditions that are modeled by MT-SDREM.
2. *Condition-specific: Sources* S_c - the set of sources or host proteins which are known experimentally to interact with the pathogen / treatment applied when studying condition c.
3. *Condition-specific (optional): Screen hits* A list of proteins for each condition whose removal is known to phenotypically impact the response of the cells in that condition.
4. *General and / or condition-specific: TF-gene binding data*: A list of potential TF-gene interactions with an optional probabilistic prior / likelihood for the interaction. If data is available for the specific condition / cell type being studied these can be used, otherwise

general data can be used as well. We denote by $\pi_{t,g}$ the interaction prior for TF t binding with gene g.

5. *General: Protein interaction network*: A list of protein-protein interactions which may be directed or undirected. The method can also use information regarding the confidence in each interaction. We denote such confidence in edge e by π_e and by E the set of all edges.

Application of multi-task learning to the inference of signaling and regulatory networks

One way to infer networks for each condition would be to run SDREM individually on the expression data for different infections to infer regulatory and signaling cascades for each of these conditions. However, several shared attributes can be jointly learned for these conditions and given the scarcity of data compared to the number of variables (very few time points for each expression experiment with thousands of genes in each model) such an approach can improve the accuracy of the reconstructed networks for each condition. Specifically, the direction of (the originally undirected) PPIs is likely to be similar for all conditions since several pathways are likely used by multiple conditions. Similarly, TFs that are active in response to one virus are more likely to be active in response to other viruses as well. MT-SDREM defines an optimization function that captures these expected similarities while still allowing for a condition-specific response component.

Multi-task objective for MT-SDREM

Recall that in the introduction, we called T_1 the task-specific of a multitask objection function, and T_2, the part of the objective that enforces regularization across tasks. In MT-SDREM, the loss minimizing part, T_1, is achieved by the regulatory network learning procedure which learns parameters for a IOHMM that uses a logistic regression classifier to compute transition probabilities. The logistic regression classifier is regularized using Lasso to reduce the number of active TFs inferred for each split. Thus in terms of the multi-task objective, $\mathbf{y}_i$ corresponds to the prediction regarding a gene trajectory at any split and $\mathbf{x}_i$ is the *TF-gene binding information*. $\mathbf{w}_i$ is the set of logistic regression weights learned for each split. Note that the TF-gene binding information $\mathbf{x}_i$ is *not* specific to each split but is the same for the entire times series.

In addition to expression data, we use signaling network information to infer TFs that are reachable from the infection sources. Such TFs are more likely to explain how the infecting agents affects gene expression and so their weights are increased in our framework. To find such TFs we need to orient the undirected edges and determine a weight for the paths leading to these TFs from sources. These two procedures (edge orientation and TF re-weighting) are shared across tasks and both affect the TF priors used by the logistic regression function. Thus for MT-SDREM, the objective function is:

$$\underset{\mathbf{w_1},...,\mathbf{w_C}}{\operatorname{argmin}} \left\{ L(\mathbf{y_i}, f(\phi(\mathbf{w_i}, \mathbf{B^i})^\mathbf{T}(\mathbf{x_i}))) + \lambda_1 \cdot ||w_i||_p \right\} - \rho(B^1, ..., B^C)$$

where B is the weight matrix learned for TFs for all tasks in the signaling network and B^i are the weights determined for task i. ρ is the similarity function used to constrain parameters across

tasks which is described below (hence the negative sign in front of it as we are minimizing the objective but we want to maximize the similarity).

An important difference between the standard multi-task learning framework and our method is that while we regularize the within task parameters (w_i's), the between task parameters (B^i's) are not explicitly regularized. The reason is that the B^is are already constrained by the input protein interaction network and so are inherently bounded.

Given B^i, the above equation can be optimized by fitting parameters to the IOHMM and logistic regression function as was previously done in [33]. See Supplement Methods for details.

Between task regularization

Next we discuss how we use the signaling network to determine the values for B, the TF weights used to reconstruct the regulatory networks. While the main goal of the regulatory network reconstruction method is to explain the temporal gene expression trajectories using the dynamic activation of TFs, the main objective when reconstructing the signaling network is to explain how these TFs are activated by the infecting viruses. For this, we attempt to link sources (protein interacting with the virus) and targets (TFs controlling virus-specific expression response) using paths in the network. The orientation is determined by specifying edge directionality to optimize the following equation:

$$\max \sum_{t \in T} \sum_{p \in P_t} I(p) \cdot h_p \cdot s_t$$

where T is the list of TFs predicted to regulate the time series for a specific condition, P_t is the set of paths that start from a source of this condition and end in TF t, h_p is the weight of the path which is defined as the multiplication of the probabilities of the edges in the path, and s_t is the score of the TF t obtained from the regulatory network reconstruction. $I(p)$ is an indicator function indicating whether path p is satisfied or not (a path is satisfied if all the edges in the path are oriented in a direction that links the source to the target) and thus optimizing the above equations requires the assignment of directionality to the PPI edges (see [31, 34] for details). Note that a Breadth First Search or a Depth First Search are not enough to solve this since we assume PPI edges may be undirected. Thus, certain paths can contradict each other in terms of the specific edge direction making this a non trivial optimization problem (in fact, it is NP complete – see [33] for details and algorithm for solving this problem).

If we have multiple conditions we can simply run this function independently for each of them leading to the following set of optimization problems:

$$\max \sum_{t \in T_c} \sum_{p \in P_t^c} I(p) \cdot h_p \cdot s_{tc} \;\; \forall c \in C$$

Here c goes over each of the conditions and the function is optimized independently for that condition. However, such independent optimization may lead to contradictory directionality assignments. In addition, it does not utilize shared properties between the conditions. Instead, we would like to -

1. Constrain the model to use shared parameters – thus the direction of the edges in the signaling networks is constrained to be the same in all models.
2. Favor pathways which end in TFs that are used in more than one condition.

To achieve the first goal above we attempt to maximize the objectives for each condition using a shared, directed, network. For this we modify the search procedure by assigning edge direction to maximize the sum of the objectives across all networks.

The second requirement is more involved since it requires us to change node scores based on TF usage across the conditions. To obtain more shared TFs we add an additional term to the objective function. We introduce a new, global, parameter, α which is used to increase the weight assigned to shared TFs.

2.1.2 Ranking proteins in reconstructed networks

Following the multi-task learning procedure we arrive at directed, weighted networks for each of the conditions being studied. To further select the key proteins from each of these networks we rank the proteins based on the "path flow" going through a node. The path flow f through a node n is defined as follows –

$$f(n) = \sum_{p \in P} I(p) \cdot h_p$$

where P is the set of paths containing node n.

To combine the rankings from each condition into a single ranking, we compute the total flow through all the nodes

$$F_i = \sum_{n \in N} f_i(n)$$

where N is the set of genes and i is the condition and then we computed the % flow $\hat{f}_i(n) = \frac{f_i(n)}{F_i}$ through a node. To get the combined score for a gene across conditions, we sum up the condition-specific % flows to get $s(n) = \sum_{i=1}^{C} \hat{f}_i(n)$ where C is the number of conditions. Then we rank the genes in descending order of the final score $s(n)$.

2.1.3 Results on Influenza data

RNAi screen hits

Using the screen hit data for H1N1 and H5N1 we compared the performance of MT-SDREM, I-SDREM and Endeavour [2, 85]. Endeavour is a gene prioritization algorithm which uses a set of seed genes (the sources) to rank genes based on several types of evidence including gene expression, interaction networks derived from various sources, text mining, sequence similarity, and functional annotations. It combines the individual rankings to create a global ranking for all genes. For the MT-SDREM and I-SDREM results we ranked proteins based on the total number of paths weighted by their score going through them. See Supplementary Methods for details. For Endeavour, we configured it to use only BioGRID and HPRD as data sources as those are the only sources we use to construct our PPI network. The expression data is not used by Endeavour. We gave the source proteins as the seed genes to Endeavour. We further compared these three

methods with a baseline method that is condition-independent: ranking nodes by their weighted degree in the PPI network. The results are presented in Figure 2.2. For H1N1, the top 100 genes in the Endeavour ranking include only 20 screen hits (p-value is 4.9E-7). For I-SDREM the number increases to 35 (p-value 2.0E-19) whereas MT-SDREM obtains the highest number of protein in the overlap 39 (p-value 1.7E-23). The baseline comparison where we rank by degree has an overlap of 30 genes (p-value 9.4E-15). For H5N1, the top 100 genes for Endeavour and for ranking by degree include only 5 screen hits (p-value 1.2E-6) whereas both I-SDREM and MT-SDREM have an overlap of 9 screen hits (p-value 1.7E-13).

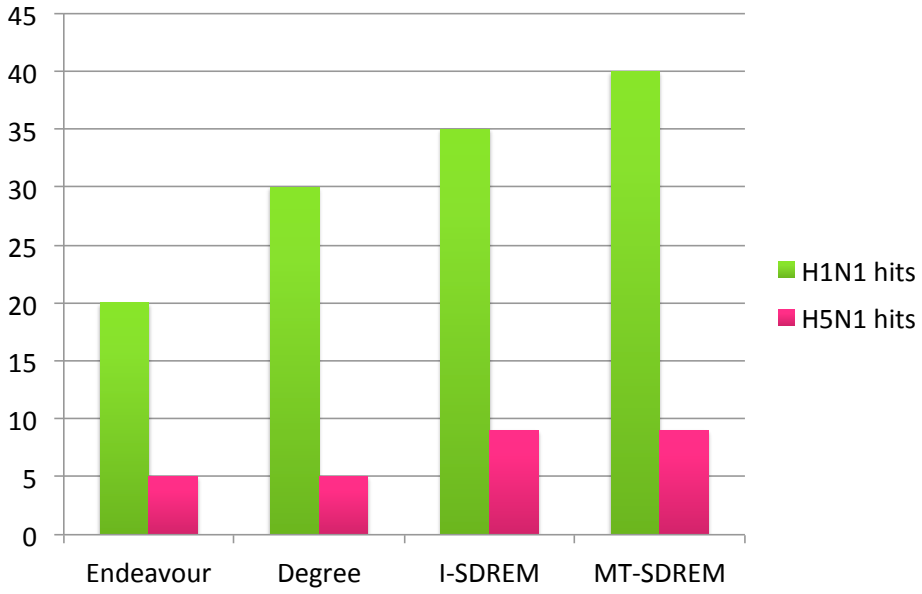

Figure 2.2: Screen hits overlap for top 100 ranked genes for both H1N1 and H5N1. 925 H1N1 and 32 H5N1 screen hit proteins were present in our network.

GO enrichment comparisons

To compare the GO enrichment of shared genes / proteins we examined the top 500 genes in the combined MT-SDREM network (ranked using the same sum of % of path flow going through genes across the 3 networks as we did for the oPossum comparison) with the top 500 genes from the combined ranking of the differentially expressed (DE) genes from each condition (combined using the Kemeny-Young method as explained before). We used FuncAssociate [9, 10] to compute standard GO enrichment for the genes. We found **3** categories, only 2 of which were immune response related for which the p-value for DE genes was ≤ 0.001 but which were not present in the MT-SDREM list or if present, their p-value was < 0.01. The categories are listed in Table 2.1. However, for the vice versa comparison, we found a large number of categories for which the MT-SDREM p-value was ≤ 0.001 but which were either not enriched for in the DE genes list (most common outcome) or if present, their p-value was ≤ 0.01. A subset of the immune response related categories are listed in Table 2.2. Note that we find significant enrichment for a

Table 2.1: **GO categories enriched in DE genes that are not enriched as significantly in MT-SDREM** GO comparison between the joint DE gene list and the joint MT-SDREM for the top 500 genes. The enrichment was performed using the FuncAssociate tool [9]. Only categories with DE genes adjusted p-value of ≤ 0.001 and MT-SDREM genes p-value of ≥ 0.01 are presented. If a p-value for MT-SDREM is NA, that means that that category was not enriched for in the MT-SDREM list. **All** immune response related categories are presented.

GO Category	**DE p-value** $\leq$	**MT-SDREM p-value**	**GO Category Description**
GO:0045071	0.001	NA	negative regulation of viral genome replication
GO:0048525	0.001	0.019	negative regulation of viral process

very varied set of immune response processes including T cell activation, cytokine production, activation of immune response, etc. as well as categories related to viral genome expression and positive regulation of viral process. The DE genes list is only enriched for negative regulation of viral process and viral genome replication.

2.2 Proposed research

2.2.1 Elucidating plant hormonal signaling networks

Plant hormones regulate myriad processes involved in almost all growth, development and environmental responses. These hormones make competing demands for cellular resources and may have conflicting or complementary objectives. Correct regulation of such complex processes can only be achieved by a high degree of interaction (cross-regulation) between distinct hormonal signaling pathways. Extensive evidence demonstrates that cross-regulation exists between these signaling pathways. For example, many studies have highlighted the extensive cross-regulation of all other hormone signaling pathways by the primary ET response transcription factor (TF), EIN3, along distinct temporal profiles [16]. SA is a crucial component of plant basal and induced response defenses, whilst BR is classically thought to be involved in multiple growth processes, including cell division and elongation [37, 89, 103]. SL is a more recently characterized hormone, involved primarily in branching but also drought responses, whose signaling pathway is less well described than others [87] . Primarily, JA is considered a defense hormone [69]. However, the diverse roles of JA and ET signaling demonstrate that plant hormones are multifunctional, exhibiting extensive cross-regulation of growth and defense processes [19, 69, 70, 77].

In collaboration with a group at Salk Institute, we are working on applying MT-SDREM to time series gene expression data and Chip-seq data from Arabidopsis Thaliana when it is stimulated by a variety of drugs. The data was collected using RNA-seq and examines the response of Arabidopsis to five different hormone treatments – Methyle jasmonate (JA), Ethylene (ET), Brassinosteroid (BR), Salicylic acid (SA), and Strigolactone (SL). Expression for each treatment was document at six time points. The hormone-host protein interaction data was culled from a literature search and the protein-protein interaction data was taken from [23]. We have already run MT-SDREM on the data and are in the process of starting experiments to validate our predictions. We also have histone modification data from the same group (though only for a subset

Table 2.2: **GO categories enriched in MT-SDREM that are not enriched as significantly in Differentially Expressed (DE) genes** GO comparison between the Differentially Expressed gene list and MT-SDREM gene list for top 500 genes. The enrichment was performed using the FuncAssociate tool [9]. Only categories with MT-SDREM adjusted p-value of ≤ 0.001 and DE genes p-value of ≥ 0.01 are presented. If a p-value for DE genes is NA, that means that that category was not enriched for in the DE genes list. Only **select** immune response related categories are presented.

GO Category	**MT-SDREM p-value** $\leq$	**DE genes p-value**	**GO Category Description**
GO:0002218	0.001	NA	activation of innate immune response
GO:0002684	0.001	NA	positive regulation of immune system process
GO:0002429	0.001	NA	immune response-activating cell surface receptor signaling pathway
GO:0046328	0.001	NA	regulation of JNK cascade
GO:0001816	0.001	NA	cytokine production
GO:0001959	0.001	NA	regulation of cytokine-mediated signaling pathway
GO:0042113	0.001	NA	B cell activation
GO:0042110	0.001	NA	T cell activation
GO:0043923	0.001	NA	positive regulation by host of viral transcription
GO:0019080	0.001	NA	viral genome expression
GO:0048524	0.001	NA	positive regulation of viral process
GO:0007259	0.001	NA	JAK-STAT cascade
GO:0002573	0.001	NA	myeloid leukocyte differentiation

of the conditions and time points) and are looking into how to incorporate it into our models. We will be talking further about this aspect in §4.

Chapter 3

TimePath

While MT-SDREM is very good at inferring signaling networks, it does not provide temporal information about the pathways it finds. In MT-SDREM, all pathways from source proteins (protein interacting with the environment / pathogen) to TFs are assumed to be activated concurrently which does not explain expression waves and response phases. Further, it does not optimize a single target function but rather two, separate, functions for different models (one for the IOHMM and the other for the combinatorial orientation algorithm) making it hard to determine optimal parameters for the networks. TimeXnet [67] is another method for reconstructing such networks. It uses linear programming to formulate a max-flow problem imposing a constraint that the flow through expressed genes has to be greater than 0 so that they are accounted for in the networks identified. TimeXnet has been applied to study immune response in mice. However, TimeXnet does not directly consider the (often post-transcriptionally activated) source of the resulting response which may lead to missing important pathways. In addition, TimeXnet does not explain why some genes are activated early while others are only activated at a later stage.

Here we present TimePath, a new method for reconstructing fully dynamic signaling and regulatory networks. TimePath uses a single Integer Programming (IP) based optimization function to jointly construct the networks. Before delving further into the details of our method, we give a brief overview of Integer programming.

3.1 Completed research

3.1.1 Methods

We initially select a large set of pathways that are rooted in source proteins and end in differentiall expressed (DE) genes. This allows us to include sources that are only post-transcriptionally and / or post-translationally activated. Pathways for later DE genes are required to contain DE genes or miRNAs from earlier phases to explain their delayed response. Next, we use the IP to select a small subset of pathways that, together, explain the full set of DE genes. These selected pathways are analyzed to determine phase specific proteins and miRNAs and select those that are key to the response observed.

We applied TimePath to reconstruct dynamic models for HIV-1 immune response. As we show, the method accurately reconstructed the response networks identifying several known and novel pathways. We have performed experiments based on novel predictions made by TimePath several of which validated the ability of TimePath to determine a specific time for targeting a protein in order to reduce viral loads.

Candidate pathways

To reconstruct the dynamic set of signaling pathways that are activated we first divide the time series gene expression data into K phases. Initial response is likely driven by host proteins that interact directly with virus proteins. However, later changes in expression data (for example, expression changes that only occur 10 hours after infection) are likely driven by genes or TFs that have been activated as part of an earlier expression response. In general we assume that expression changes in phase i can be partially explained by activation / repression of a gene(s) in phase $i-1$. To guarantee that our reconstructed pathways satisfy this we impose the constraint that any pathway that explains differential gene expression for a gene in phase $i > 1$ has to include at least one gene that was differentially expressed (DE) in phase $i-1$.

Based on these assumptions we initially select a subset of pathways that can be used to explain the DE genes as follows:

1. We divide the time series into k phases each consisting of T/k time points where T is the total number of points. We use $k = 3$ for this paper.
2. We extract the top N_1 DE genes for each phase (we use $N_1 = 200$).
3. We then search for the highest scoring N_2 acyclic paths from the source proteins (host proteins interacting with the virus of drug) to the targets (DE genes) for each phase (we use N_2 = 10 million here). We use the edge weights to compute a score for each path (Supplementary methods). We also guarantee that the following constraints are satisfied for each pathway:-
   (a) The last edge in the path has to be a protein-dna interaction (i.e. we need a TF to activate / repress the gene) [96].
   (b) A path to a phase $i > 1$ target has to contain a node that is a target for phase $i-1$.

In general, searching for the top $N2$ acyclic paths in a graph is a #P-complete problem which is not considered to be solvable efficiently [7]. We thus use a heuristic to compute the set of paths. See Supporting Methods for a detailed description of the above process.

Integer program to select subset of pathways

Given a set of top paths for each target, our next goal is to combine them to identify the actual pathways that are activated as part of the response. Consider 2 targets g_1 and g_2 in phase k that are known to be bound by the same TF A. If we believe that A explains the activation of g_1 in that phase it increases our belief that A is also the TF activating g_2. More generally, our goal is to select a subset of these pathways that, together, would minimize the number of intermediate signaling and regulatory proteins that are used across all pathways while at the same time maximize the number of targets that can be explained.

To accomplish this we define a new Integer Programming (IP) problem which includes 3 sets of binary variables (bv)

1. bv for a path to indicate whether it is selected or not
2. bv for a target to indicate whether there is at least one path ending at it
3. bv for protein to determine whether it is part of a path selected.

Using these variables we maximize the following objective

$$\max \sum_{p \in P} w(p) \cdot b_p^P + \lambda_1 \sum_{k=1}^{K} \sum_{g \in T_k} f_g - \lambda_2 \sum_{g \in G} b_g^G \tag{3.1}$$

with the constraints

$$\forall p \in P, \forall g \in p, b_g^G \geq b_p^P \tag{3.2}$$

$$\forall p \in P, \sum_{g \in P} b_g^G \leq |p| - 1 + b_p^P \tag{3.3}$$

$$\forall g \in G, \forall p \in P(:, g), f_g \geq b_p^P \tag{3.4}$$

$$\forall g \in G, \sum_{p \in P(:,g)} b_p^P \geq f_g \tag{3.5}$$

where

- K is the number of phases.
- T_k is the targets for phase k.
- P is the set of all paths
- G is the set of all genes
- $P(:, g)$ is the set of paths ending at gene g.
- $w(p)$ is the weight of path p. The score of each a pathway p is defined as $\Pi_{e \in E_p} \mathbb{P}(e)$ where E_p is the set of edges in pathway p and $\mathbb{P}(e)$ is the edge score.
- b_p^P is whether path p is selected or not.
- f_g is whether gene g has even one selected path ending at it.
- b_g^G is whether gene g is selected.
- λ_{1-2} are the weights for balancing the minimization requirements in terms of intermediate nodes and the maximization requirements in terms of the number of targets. They are the parameters that decide in the end, how large of a network in terms of number of genes and edges will be chosen.

Note that setting $b_g^G = 0$ for a specific gene immediately implies that b_p^P for a path containing that gene is 0 and similarly that f_g is 0 for that gene and so these variables are not independent as the constraints above imply. We set $b_p^P = 1$ if and only if all the genes in the path are selected

as enforced by constraints 1-2. f_g is 1 if and only if there's at least one path with $b_p^P = 1$ ending at the gene g as enforced by constraint 3.

Since this is a problem with linear constraints, a linear objective and since the b_g variables are binary, this is an IP and not an Linear Program (LP). The IP we are dealing with however is too large for standard IP solvers and we thus solve it using a greedy approach followed by a tabu search heuristic to escape local minimum. Briefly, we start with all the nodes selected. Then at each step, we search for a node whose addition or removal from network would increase the objective the most (this is accomplished by flipping the b_n variable for that gene). Paths that contain a gene that is not in the current network are removed (i.e. their corresponding b_p variable is 0). Once we find such a node, we add or remove it and keep going until we can find no node whose addition or removal will improve the objective. We randomly select nodes if there are ties between them. Thus the results can differ from one run to another – however the actual genes selected by the network change little according to our experimental results. See Supplementary Results for details.

Ranking genes

After solving the IP we obtain a subset of the pathways that, combined, explain the observed expression response over time. While we attempt to minimize the number of proteins in these networks, we still end up with hundreds of proteins in the set of selected pathways. To identify key proteins for follow up analysis, we rank genes for each phase based on the "path flow" going through them. The path flow f through a node n for phase i is defined as follows –

$$f(n) = \sum_{p \in P} I(p) \cdot w(p)$$

where P is the set of paths ending at a target in phase i and containing node n. $I(p)$ is 1 when the path p is selected and 0 otherwise. We further refine the phase specific genes for later phases to remove those already identified by earlier phases. See Supporting Results for details.

3.1.2 Results

TimePath analysis of HIV data

We used TimePath to examine cell response to HIV infection. Time series expression data for HIV-1 was obtained from Mohammadi et al [63] which profiled genes using SAGEseq every 2 hours for 24 hours after transfection with HIV-1 in Sup-T1 cell line. Expression data was Normalized using DESeq [6]. In addition to HIV expression data we obtained interaction data for HIV-1 proteins and host (human) proteins from VirHostNet [65]. Of the 235 proteins in VirHostNet, 231 are present in our protein-protein interaction (ppi) network and were used as potential sources.

TimePath also uses general protein-protein interactions from BIOGRID [80] and HPRD [72], Post-translational Modification Annotations from HPRD and Protein-DNA interaction data [76] (Methods).

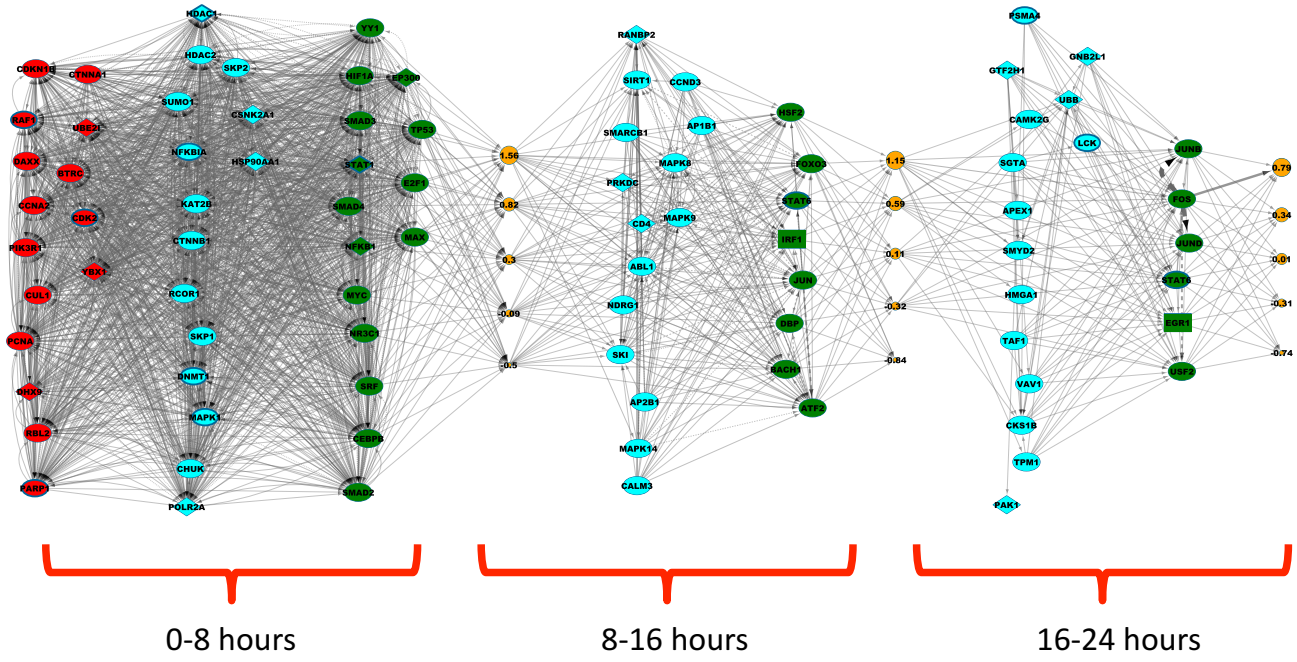

Figure 3.1: **Dynamic signaling and regulatory network for HIV-1 immune response**. The red nodes are the host proteins that interact with the HIV-1 proteins (selected sources). Blue nodes are intermediate signaling proteins and green nodes are the TFs that are predicted to directly up/down-regulate the differential expression of target genes (targets not shown in figure, but the average levels of the regulated targets for each TF is presented by the yellow nodes while the size of each of the yellow nodes indicates how many genes belong to the cluster represented by the node). The figure displays the top predicted nodes for each of the three phases and also demonstrates is directly linked to the sources via the signaling proteins and DE genes in earlier phases. Diamond shaped nodes were identified as supported RNAi screen hits (text) and rectangular nodes are targets for the phase they are in. Nodes with bold blue border represent proteins we experimentally tested. Note that some intermediate proteins may also be TFs. The functional role in the network figure is based on the location of the protein in the selected paths based on the IP.

To identify pathways for specific response phases we divided the time series expression into 3 phases (every 8 hours) and extracted 200 targets (DE genes) for each phase (Methods). We next used the static interaction data to identify a large number of potential pathways connecting sources and targets constraining potential pathways for *later targets* to contain a gene that is DE at an earlier phase. A subset of these pathways that, together, explain the observed response to HIV infection are then selected by the IP method. Pathways retained by the IP for this data included a total of 607 genes of which 319 are targets. We next ranked proteins in these pathways based on their importance to each phase (Methods).

Pathways and proteins identified for HIV response

The resulting dynamic network is presented in Figure 3.1.

Table 3.1: **Overlap between RNAi screen hits and top 100 genes for the different dynamic network reconstruction methods and between edge list from Reactome (1265 edges in network) and the edges extracted by the different methods. Comparison with a baseline ranking of the differentially expression (DE) genes is also presented.**

Method	Overlap with screen hits	p-value	Overlap with Reactome edges	p-value
TimePath	23	1.7×10^{-17}	101/3203	7.9×10^{-44}
SDREM	21	3.2×10^{-16}	74/3203	3.9×10^{-24}
TimeXnet	16	4.9×10^{-10}	54/2585	3.9×10^{-16}
DE ranking	5	0.23	NA	NA

Statistical validation of the reconstructed network and comparison with other methods

To more globally assess the ability of TimePath to accurately identify pathways and proteins, and to compare its performance with prior methods that were developed to reconstruct dynamic signaling and regulatory networks we used several complementary datasets to test the reconstructed pathways.

While several methods have been proposed for reconstructing biological networks [44], relatively few are focused on analyzing dynamic response networks. These include SDREM [31, 33], which combines a HMM method for modeling dynamic regulatory networks with a combinatorial algorithm for signaling network reconstruction and TimeXnet [67] which uses a linear programming (LP) formulation to find important genes. Note that neither of these methods uses miRNA expression data and so we constrained our comparison to TimePath models that do not utilize such data.

In addition to comparing TimePath with prior methods that construct both signaling and regulatory networks, we have also compared the top ranked genes from TimePath to the top DE genes in the dataset (Supporting Methods) since several methods for analyzing gene expression data still focus on such DE genes [74].

RNAi screen hits

First, we looked at RNAi screen experiments which test the impact of gene knockdown on HIV viral load. Three such experiments were conducted though a meta-analysis of the results determined that only 3 proteins were detected by all studies [14]. We have filtered the combined list to select a subset of the hits that are supported by at least two lines of evidence (Supplementary Results) resulting in 389 supported hits, 364 of which were present in our initial network.

The results are in Table 3.1. We find that the pathways obtained by TimePath are significantly enriched for screen hits (p-value of 1.7×10^{-17}). This significant overlap also holds separately for each the subset of proteins identified for the three phases. We next compared these results to results from the other two network reconstruction methods and to the top DE genes. For this comparison we ranked the genes using path flow for TimePath and SDREM (Methods) and used the TimeXnet output ranking for that method. The RNAi overlap is presented in Tables 3.1. As can be seen, rankings for all network reconstruction methods greatly outperforms the DE genes rankings highlighting the importance of post-transcriptional and post-translational events in the

Table 3.2: **Overlap with HIV screen hits at various stages of the algorithm. "Pre-algorithm" is the initial overlap for all genes in the network. "Unexpressed genes filtered" is when we remove all genes from our interaction network that are unexpressed. "After pathway search" is that stage that uses all genes included in the initial top scoring set of pathways. "After IP" is the final stage after the IP (and thus the whole algorithm) has run. As can be seen, the IP step seems to further improve the resulting set of genes indicating that the selection process indeed identifies HIV response pathways.**

Stage	Overlap	Overlap %
Pre-algorithm	364/16671	2.1
Unexpressed genes filtered	246/6604	3.7
After pathway search	144/1374	10.4
After IP	85/607	14.0

response process. Further, both TimePath and SDREM significantly outperform TimeXnet in this analysis with almost a quarter of the top ranked genes supported by screen hits.

Analysis using GO and Reactome

To further analyze the pathways identified by TimePath we looked at the agreement between them and two complementary databases: The Gene Ontology (GO) and the set of HIV curated pathways in Reactome. GO analysis was performed on the top 100 genes (nodes) identified based on the path flow metric (Methods) using FuncAssociate [9] while Reactome analysis was performed using the set of pathway edges. The results indicate that the pathways obtained by TimePath agree very well with known pathways involved in HIV response. The full list of enriched GO categories (corrected p-value ≤ 0.001) is presented on the Supporting Website and includes "toll-like receptor signaling pathway", an important component of innate immune response [58], "positive regulation of defense response", "innate immune response-activating signal transduction", etc. We also find that TimePath achieves a higher number and a higher percentage of significantly enriched immune related categories compared to SDREM and TimeXnet 3.3 using the FuncAssociate [9] tool. We compared the % of significantly enriched GO categories that were immune response related (Supporting Methods). TimePath again has a both a slightly higher number and a higher percentage of significantly enriched immune related categories compared to SDREM and TimeXnet (Table 3.3).

Results for Reactome are presented in Table 3.1. As can be seen, we achieve a significant overlap between edges in the selected pathways and those present in the HIV Reactome pathways. Comparison with the other methods clearly demonstrates the advantages of TimePath which is able to identify a much larger number of correct interactions than the other two network reconstruction methods. Note that Reactome comparison is not available for the DE gene list since it does not contain interactions.

We have also analyzed the usefulness of the various stages of TimePath. As can be seen in Table 3.2, each step in the TimePath method further improves the overlap with the screen hits. Initially, only 3.7% of the expressed genes are screen hits. The initial pathway extraction step increases the overlap to 10% while the overlap following IP increases to 14%.

Finally, we investigated the impact of the constraint imposed on later paths in our network to include a DE gene from an earlier phase. As we show in Table 3.4, we obtain almost 3

Table 3.3: **GO comparison**. We give the % of immune-related categories as well as the absolute number of immune related categories and total categories enriched for in parenthesis. The p-value cutoff for all categories was 0.05. The GO enrichment was performed on the top 100 genes as ranked by path flow (Methods) using the FuncAssociate tool [9].

Method	% of immune-related categories	p-value
TimePath	11.16 (72/645)	2.074×10^{-5}
SDREM	8.04 (71/883)	0.077
TimeXnet	10.44 (66/632)	3.18×10^{-4}

Table 3.4: **Validation for the time constraint**

Method	Overlap	p-value
TimePath	101/3203	7.9×10^{-44}
TimePath without time constraint	37/3203	3.6×10^{-5}

times as many edges in the overlap compared to the network without the time constraint with correspondingly better p-value.

Experimental results

To experimentally test the temporal predictions of TimePath we selected top ranking phase proteins for which we could obtain commercial inhibitors and examined the impact of blocking these proteins at various time points in the response (Figure 3.2). Note that the RNAi knockdown screens discussed above were performed on a different cell type (Hela/TZM-bl and 293T) and so, while they are useful for statistical validation, they may not completely reflect pathways activated in Sup-T1 cells. More importantly, these screens do not provide information about the dynamics of the response while our experiments are aimed at testing not just the predictions regarding top ranked proteins but also their phase specific assignment. We performed experiments in which we varied the time of applying the inhibitors w.r.t the infection time. For each of the proteins tested, inhibitors were applied 2 hours prior to infection (phase 1), 4 hours (phase 2) and 14 hours (phase 3) post infection. amount of infection was determined at 40 hours post infection for all experiments. We concurrently measured cell viability to test the toxicity of the inhibitor.

The results are presented in Figure 3.2. As can be seen, for 5 of the inhibitors we tested (targeting 11 of the 22 proteins tested) we observed a significant impact on viral load as predicted by TimePath. Note that the screen results indicate that less than 1.5% of all proteins lead to decreased viral load, and so such a high validation rate is a strong indication for the accuracy of TimePath. Importantly, several of the time specific predictions were validated in these experiments. We expected that inhibiting proteins that are ranked at the top for all phases or for phase 3, at any time, would lead to reduction in viral load since even early inhibition prevents them from being activated at a later stage. We indeed see this effect for the STAT inhibition (ranked in the top 30 for all phases) and for PSMA4 (ranked at the top only for phase 3). In contrast, for proteins ranked high in phase 1 and lower at the next phases we expected to see a much greater impact for the early treatment vs. later ones since their impact may have already been exerted by the time of the later treatments. This is exactly what we see for two of these proteins. For

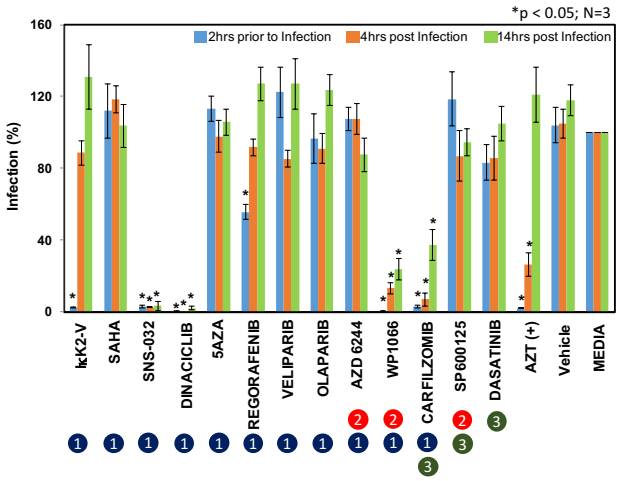

Figure 3.2: **Experimental validations.** Relative infection after treatment with inhibitors. Significant changes in infection are highlighted with a *. The inhibitor names are given on the X axis and the target proteins of the inhibitors are given in parenthesis.

both NFKB1 (ranked 14 in the first phase but dropping to 50 in the 2nd) and for Raf1 (dropping from 28 to 66) we see significant response when treated early but a much lower impact on viral load when treated at later stages strongly supporting TimePath's predictions. Published studies suggest that NF-kB has a major role in HIV-1 transcription due to it is binding sites in HIV-1 LTR and TAR-RNA [55, 82, 83, 93, 94]. Results from our analyses predicted a role for NF-kB during the early phase (phase 1) and blocking this TF inhibited virus replication only in pretreatment (2 hours) and did not affect virus replication when treated at the later stages and this effect is independent of cellular toxicity. Similarly, another protein Raf1, predicted as early phase response to HIV-1 also exhibited similar phase dependent inhibition. Though Raf1 is known to interact with HIV-1 Nef and perturb T cell signaling and activation pathway [42], the mechanisms by which Raf1 exerts its effects is unclear. It is possible to predict that blocking Raf1 might have an effect on the function of HIV-1 early protein Nef, thus altering T cell signaling and virus infection. Another phase 1 protein, CDK2 (dropping from 29 to 59) also showed strong impact when treated at the early time point but unlike the other phase 1 predictions, later treatments continued to have a significant impact on viral loads. CDK is known to play a role in HIV-1 transcription by the viral transactivator, Tat [22], thus there is a direct correlation predicted by TimePath. However, blocking CDK using inhibitors blocked both at the early and late phase suggest that these inhibitors might have direct and indirect effect on virus replication.

PSMA41 is part of the proteasomal complex and so inhibiting this protein with Carfilzomib not only blocks the proteasomal pathway, but could also alter additional cellular processes such as sumoylation, ubiquitination and Cul1 activity. These results are further supported by the early time points predictions that identified SUMO1, UBE2I and CUL1 in Phase 1. Sumoylation of HIV-1 integrase is essential for efficient viral replication [98] and cullin ligases are recruited by

HIV-1 viral proteins to overcome host viral restriction factors, HIV-1 Vif degrades APOBEC proteins [35] and HIV-1 Vpr induces degradation of UNG and SMUG uracil-DNA glycosylases [75]. Also HIV-1 Vpr is known to interact with damaged DNA binding protein 1 (DDB1) to induce G2/M arrest which contributes to efficient viral replication [38]. Indeed, many of the factors predicted for the early stage response (Phase 1: 0-8 hours) are related to DNA modification and chromatin remodeling (HDAC1, HDAC2, DNMT1, KAT2B) and cell cycle (CTNNB1, CSNK2A1, CDK2, E2F1). Also there is an enrichment of transcription factors (P53, RELA, NFKB1, NR3C1, Stat1, MYC, RAF1, TBP, YY1), which have binding sites on HIV-1 LTR. These factors may have a role in integration of proviral DNA and regulation of HIV-1 transcription.

3.2 Proposed research

3.2.1 Application to HIV related dementia

In collaboration with a group at the University of Pittsburgh, we applied TimePath to HIV data from patients with HIV and increasingly severe forms of dementia (under submission). The goal is to explore what exact role does HIV play in the progression of dementia in these patients.

3.2.2 Application to plant hormonal signaling

We're also exploring the application of TimePath to time series gene expression data under various hormonal stimulations from Arabidopsis Thaliana (collaboration with group at Salk Institute). The data for Arabidopsis is the same as that described in § §2.2.1. Our initial plan is to run TimePath on the given data with each phase covering two time points (total of 6 time points for each experiment).

3.2.3 Application to IPF lung disease

We are also applying TimePath to gene expression data from IPF lung disesase at various stages of progression (collaboration with group at Yale University). The IPF expression data is also RNA-seq based and is taken from 15 different patients (10 of whom have the disease) at different stages of disease progression. We also have epigenetics data for the patients. Our goal is to identify the signaling pathways and regulatory TFs active at the different stages of the disease. The approach we are taking is to infer a consensus gene expression profile as a function of how far the disease has progressed for all genes that are differentially expressed in the disease. We have used B-splines with 5 control points and patient specific gaussian noise to model the gene expression profile. Our plan is to sample a time series expression profile for each gene related to the disease at fixed disease progression points and use that as input for TimePath to obtain the signaling pathways. We also have plans to incorporate the epigenetic data which we will talk about in Chapter §4.

Chapter 4

Incorporating epigenetic data for network inference

4.1 Proposed research

As explored in this proposal, there is a large body of literature on how to infer signaling and regulatory networks for a given condition. However an important aspect that all of the above methods don't consider is the role epigenetic modifications play in regulating gene expression. Epigenetic modifications are changes to the DNA structure or associated chromatin proteins but not involve changes to the DNA sequence itself. An illustrative figure is given in Figure 4.1. They can take two forms – DNA methylation or histone protein modifications. They can be caused by DNA damage, change in the environment, etc. They are key players in the differentiation of a stem cell into different cell types and misregulation of epigenetics has been implicated in a wide variety of diseases.

The primary means via which epigenetic modifications cause phenotypic change is by altering gene expression by various mechanisms. Enhancers are genomic elements $50 - 1500$bp long, situated anywhere from 1bp to 1Mbp from the transcription start site (TSS) of a gene that can regulate the expression of that gene [49, 78]. DNA methylation of enhancer regions can impede the binding of transcription factors to that region. Methylated DNA can also be bound by methyl-CpG-binding domain (MBD) proteins which can recruit chromatic remodeling proteins to change the chromatin structure to make it much more compact (and thus hard for TFs to bind to). The role of *intra*-genic methylation is less understood but is suspected to be important for the regulation of transcript elongation, expression of intragenic coding and non-coding transcripts, alternative splicing, and enhancer activation [54, 60]. Histone modifications can similarly cause changes to chromatin structure which can increase or decrease the ability of an enhancer to be bound or a gene to be expressed. In fact, histone modifications have also been shown to be predictive of active and poised enhancers[1] [90, 99]. For example, the histone modification H3K27ac has been shown to be associated with active enhancers [15, 21].

Recently, there has started to be an increasing interest in the role epigenetics plays in cell

[1] Active enhancers are those aiding in ongoing transcription, Poised enhancers are those that are not but are just one step away from being active

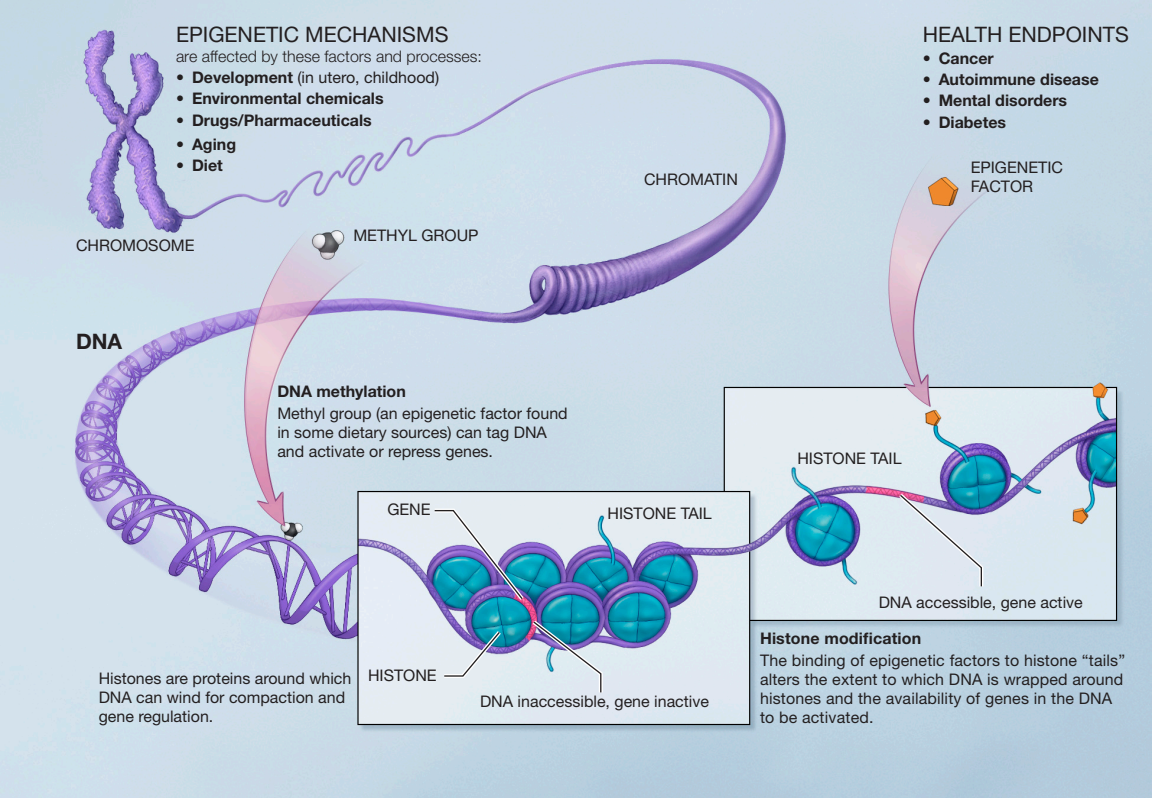

Figure 4.1: Epigenetic mechanisms are affected by several factors and processes including development in utero and in childhood, environmental chemicals, drugs and pharmaceuticals, aging, and diet. DNA methylation is what occurs when methyl groups, an epigenetic factor found in some dietary sources, can tag DNA and activate or repress genes. Histones are proteins around which DNA can wind for compaction and gene regulation. Histone modification occurs when the binding of epigenetic factors to histone "tails"; alters the extent to which DNA is wrapped around histones and the availability of genes in the DNA to be activated. All of these factors and processes can have an effect on people's health and influence their health possibly resulting in cancer, autoimmune disease, mental disorders, or diabetes among other illnesses. Image taken from [1]

biology. A large amount of epigenetic data is now regularly generated, thanks to next generation sequencing methods. And the number of ways in which epigenetic modifications can affect transcription are numerous [47, 79]. Thus it would be of great interest to have models that are able to incorporate epigenetic information when inferring signaling and regulatory networks.

There have been some attempts to examine the influence of epigenetics on gene expression. In [57], they use epigenetic and other genome features to predict differential gene expression between lung cancer and control patients. [97] uses a bayesian network model to try and infer causal links between epigenetic modifications within $\pm$1kb of the TSS. [18] develops support vector machine and support vector regression models to quantify the effect of epigenetic modifications on gene expression. They bin the DNA region $\pm$4kb of the TSS into 100bp sized bins and feed the aggregate chromatin features in each bin as features for the SVM and SVR. Other methods have tried to integrate epigenetic priors into gene regulatory network inference [17, 100]. Both of the latter methods use the correlation between epigenetic profiles of genes as a prior when inferring gene regulatory networks. In [36], they develop a two-stage model. First, for a given cell line, they infer a gene sequence specific score of it being bound by *any* TFs using position-weight matrices (PWM), histone modifications and expression of nearby genes as features and experimental binding data for 17 TFs for that cell line as the training data. They then use that score as a prior to whether a TF binding location is actually bound when inferring regulatory networks (they use a dynamic bayesian network for the second part). However this approach is not applicable when no such TF binding data is available for a cell line. It is also not applicable for epigenetic changes that are specific to a condition rather than a cell line.

Our focus is broader. Given an expression dataset, we want to figure out which signaling pathways and transcription factors are active for that condition. There are three questions in particular that we're interested in :-

1. **How do epigenetic modifications affect TF-DNA interaction strength and expression in general**. While several methods have been developed to predict TF-DNA binding energy (or sites) [3, 5, 27, 71, 81], to our knowledge, none of these methods take epigenetic modifications into consideration. As such modifications can vary from cell type to cell type and even from one condition to another within the same cell type, it is important to incorporate them to get an accurate picture.
2. **Can the pattern of epigenetic marks in enhancer regions be used to infer which TFs have bound to the enhancer regions**. We already have strong evidence in literature that epigenetic marks can distinguish between enhancers and non-enhancers and even poised and active enhancers. The pattern of those epigenetic marks (for example, which particular histones H3K27ac is present at) and cross-referencing that with the potential TF-DNA interaction map (derived from PWMs or Chip-Seq) could help us infer active TFs in the nucleus. Relatedly, histone modification levels within a few thousand base pairs of the TSS, have been shown to be to be highly predictive of gene expression [50, 61, 62] which is additional evidence that histone modifications would supplement gene expression as additional training data, when inferring which TFs are causing gene expression.
3. **How can we integrate the above two models into our existing signaling and regulatory network inference models**. One way to integrate models for the strength of TF-DNA interactions would be to put priors on the TF-DNA interaction strength in our models.

Integrating (2), assuming there is some signal in the pattern, would be more complicated as it would involve appropriately weighting that signal, and the signal from gene expression.

4.1.1 Effect of epigenetics on TF-DNA interaction strength and expression

There are several ways in which a TF can affect gene expression. It could bind to an enhancer and recruit an activator or repressor protein (or act as one itself) [4]. Then depending on the cellular context it could have varying effects on the gene expression. Or it could mediate changes in the chromatin structure. This change could either facilitate a direct effect by the TF on the expression of the gene or an indirect effect by increasing the local concentration of RNAP II [49]. Given the myriad number of ways a TF can affect the expression of a gene, estimating the impact of the epigenetic profile at the TF binding site on the expression directly would be an immense challenge. Thus our goal is to focus on estimating the TF-DNA interaction strength itself and use that as a prior for our regulatory inference models.

There is considerable evidence that tag density in Chip-Seq datasets is highly correlated with binding affinity [48, 59] for most TFs. Given that, our aim is to try and predict Chip-Seq profiles using the nucleotide and epigenetic sequence. There are a lot of methods that try and predict TF binding or binding affinity based on the raw nucleotide sequence. Two in particular stand out as methods that could act as a starting point – DeepBind [5] and DeepSea [102]. DeepBind takes in raw sequence data to try and predict the ChipSeq, SELEX, and CHIP/CLIP profiles. It shows excellent correlation with experimental data (~ 0.8). DeepSea was designed to predict effects of changing the nucleotide sequence (down to the single nucleotide level) on both TF binding and on the epigenetic code. Both have code available online and should be a good platform to build off of.

One challenge here would be that some of the histone marks might be a consequence of TF binding rather than a cause. So in effect the binding affinity predictor could predict a strong binding of a TF *because* of a particular epigenetic mark when the actual causation is the other way around. However, as our plan is to use the affinity predictions as priors to the regulatory network inference, this causation reversal may not matter as the final prior of that particular TF-DNA interaction occuring would still be correct.

4.1.2 Can the pattern of epigenetic marks in enhancer regions be used to infer which TFs have bound to the enhancer regions

As we just mentioned at the end of the last subsection, TF binding can also change histone marks at or near the site the TF bound to. A very interesting question is whether one could use these histone marks to try and infer active TFs. There are two key challenges however.

1. The Chip-Seq signals of histone marks are very broad and can range from several hundred to thousand base pairs. Thus they may turn out to be too noisy to be able to have a strong signal as to which TFs are active and bound to the genome.
2. While there are a large number of possible histone modifications, most studies choose to examine only a handful of them. Thus there is a big missing data problem here that we will have to deal with.

There is no obvious way we can think of currently to handle (1) beyond developing a model and seeing how it performs. Our current plan for is to start by using the TF-DNA binding data obtained from Chip-Seq or PWM scan on the whole genome to classify TFs as active and non-active which is then validated by checking how good the inferred set of TFs are at predicting differentially expressed genes (via another model that uses the TF-DNA interactions in the vicinity of the TSS of the gene as features). To handle (2), one possibility would be to simply train a different model for each histone mark and combine the predictions for each model into one. We are still exploring ideas however.

4.1.3 Integrating the above two models into our existing signaling and regulatory network inference models

We have already mentioned one possible way of integrating TF-DNA binding affinity predictions by using them as priors for TF-DNA interactions. For integrating the signal from epigenetic marks, one simple way would be to (pursuing thoughts similar to the end of the last subsection), simply have another model which predicts active TFs based on differential expression data and combine the predictions from both models to get the final set of active TFs.

4.1.4 Data available

We already have histone modification data from Arabidopsis Thaliana for the JA hormone treatment described in § §2.2.1. We also expect to obtain methylation data for IPF lung disease project descrbed in § §3.2.3.

Chapter 5

Conclusion and timeline

A cell is a highly sophisticated piece of biological machinery with a staggeringly complex program running it. This complexity can in turn lead to very large variability in cell behavior – even between situations where you would expect no difference. Sophisticated mathematical models thus become essential to taming this vast complexity and making reliable and accurate predictons. The large amount of biological data being generated today presents us with a unique opportunity to use computational techniques to generate such mathematical models.

In this thesis, we have attempted to deal with some aspects of a significant component of cell biology – namely which signaling pathways and transcription factors (TFs) are active for and related to a particular condition. We have talked about why it is so hard for experimental methods to be able give us a complete picture of what is happening and how computational techniques may aid us in completing that picture.

In particular, we have presented our solutions to three problems (1) learning from limited data by using data from related conditions using multi-task learning (MT-SDREM) (2) Temporal annotation of signaling pathways and TFs (TimePath) (3) proposed incoporating epigenetic modifications into our models in order to better infer the signaling and regulatory networks.

We have presented successful application of MT-SDREM to inferring important genes and pathways related to Flu infection and of TimePath to inferring temporally annotated pathways for HIV infection. We have also proposed applications of both methods to plant hormone signaling data and of TimePath to IPF lung disease data. Finally, we have proposed application of the new techniques we develop to incorporate epigenetic modifications on the plant and IPF data.

In regards to the timeline, given that I will be doing an internship from May 16 to August 9, I aim to have a new method for incorporating epigenetic data ready by winter of 2017 so it can be tested and written up. I aim to graduate in June of 2017.

Bibliography

[1] Nih website. http://commonfund.nih.gov/epigenomics, 2015. 4.1

[2] Stein Aerts, Diether Lambrechts, Sunit Maity, Peter Van Loo, Bert Coessens, Frederik De Smet, Leon-Charles Tranchevent, Bart De Moor, Peter Marynen, Bassem Hassan, et al. Gene prioritization through genomic data fusion. *Nature biotechnology*, 24(5):537–544, 2006. 2.1.3

[3] Phaedra Agius, Aaron Arvey, William Chang, William Stafford Noble, and Christina Leslie. High resolution models of transcription factor-dna affinities improve in vitro and in vivo binding predictions. *PLoS Comput Biol*, 6(9):e1000916, 2010. 1

[4] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the Cell*. Garland Science, 5 edition, 2007. 4.1.1

[5] Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature biotechnology*, 2015. 1, 4.1.1

[6] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome biol*, 11(10):R106, 2010. 3.1.2

[7] Sanjeev Arora and Boaz Barak. *Computational complexity: a modern approach*. Cambridge University Press, 2009. 3.1.1

[8] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000. 1.3.6

[9] Gabriel F Berriz, Oliver D King, Barbara Bryant, Chris Sander, and Frederick P Roth. Characterizing gene sets with funcassociate. *Bioinformatics*, 19(18):2502–2504, 2003. 2.1.3, 2.1, 2.2, 3.1.2, 3.3

[10] Gabriel F Berriz, John E Beaver, Can Cenik, Murat Tasan, and Frederick P Roth. Next generation software for functional trend analysis. *Bioinformatics*, 25(22):3043–3044, 2009. 2.1.3

[11] Richard Bonneau, David J Reiss, Paul Shannon, Marc Facciotti, Leroy Hood, Nitin S Baliga, and Vesteinn Thorsson. The inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome biology*, 7(5):R36, 2006. 1.1

[12] Michael Boutros and Julie Ahringer. The art and design of genetic screens: Rna interference. *Nature Reviews Genetics*, 9(7):554–566, 2008. 1.3.5, 1.3

[13] Michael J Buck and Jason D Lieb. Chip-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, 83 (3):349–360, 2004. 1.3.2

[14] Frederic D Bushman, Nirav Malani, Jason Fernandes, Iván D'Orso, Gerard Cagney, Tracy L Diamond, Honglin Zhou, Daria J Hazuda, Amy S Espeseth, Renate König, et al. Host cell factors in hiv replication: meta-analysis of genome-wide studies. *PLoS pathogens*, 5(5):e1000437, 2009. 1.1, 3.1.2

[15] Eliezer Calo and Joanna Wysocka. Modification of enhancer chromatin: what, how, and why? *Molecular cell*, 49(5):825–837, 2013. 4.1

[16] Katherine Noelani Chang, Shan Zhong, Matthew T Weirauch, Gary Hon, Mattia Pelizzola, Hai Li, Shao-shan Carol Huang, Robert J Schmitz, Mark A Urich, Dwight Kuo, et al. Temporal transcriptional response to ethylene gas drives growth hormone cross-regulation in arabidopsis. *Elife*, 2, 2013. 2.2.1

[17] Haifen Chen, DAK Maduranga, Piyushkumar A Mundra, and Jie Zheng. Integrating epigenetic prior in dynamic bayesian network for gene regulatory network inference. In *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2013 IEEE Symposium on*, pages 76–82. IEEE, 2013. 4.1

[18] Chao Cheng, Koon-Kiu Yan, Kevin Y Yip, Joel Rozowsky, Roger Alexander, Chong Shou, Mark Gerstein, et al. A statistical framework for modeling gene expression using chromatin features and application to modencode datasets. *Genome Biol*, 12(2):R15, 2011. 4.1

[19] José-Manuel Chico, Gemma Fernández-Barbero, Andrea Chini, Patricia Fernández-Calvo, Mónica Díez-Díaz, and Roberto Solano. Repression of jasmonate-dependent defenses by shade involves differential regulation of protein stability of myc transcription factors and their jaz repressors in arabidopsis. *The Plant Cell*, 26(5):1967–1980, 2014. 2.2.1

[20] Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, Daniel Gaffney, Laura L Elo, Xuegong Zhang, et al. A survey of best practices for rna-seq data analysis. 2016. 1.3.1

[21] Menno P Creyghton, Albert W Cheng, G Grant Welstead, Tristan Kooistra, Bryce W Carey, Eveline J Steine, Jacob Hanna, Michael A Lodato, Garrett M Frampton, Phillip A Sharp, et al. Histone h3k27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences*, 107(50):21931–21936, 2010. 4.1

[22] Thomas P Cujec, Hiroshi Okamoto, Koh Fujinaga, Jon Meyer, Holly Chamberlin, David O Morgan, and B Matija Peterlin. The hiv transactivator tat binds to the cdk-activating kinase and activates the phosphorylation of the carboxy-terminal domain of rna polymerase ii. *Genes & Development*, 11(20):2645–2657, 1997. 3.1.2

[23] Matija Dreze, Anne-Ruxandra Carvunis, Benoit Charloteaux, Mary Galli, Samuel J Pevzner, Murat Tasan, Yong-Yeol Ahn, Padmavathi Balumuri, Albert-László Barabási, Vanessa Bautista, et al. Evidence for network evolution in an arabidopsis interactome map. *Science*, 333(6042):601–607, 2011. 2.2.1

[24] Christophe J Echeverri and Norbert Perrimon. High-throughput rnai screening in cultured cells: a user's guide. *Nature Reviews Genetics*, 7(5):373–384, 2006. 1.1

[25] Jason Ernst, Oded Vainas, Christopher T Harbison, Itamar Simon, and Ziv Bar-Joseph. Reconstructing dynamic regulatory maps. *Molecular Systems Biology*, 3(1), 2007. 1.1

[26] Theodoros Evgeniou, Charles A Micchelli, Massimiliano Pontil, and John Shawe-Taylor. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6 (4), 2005. 1.4.1

[27] Barrett C Foat, Alexandre V Morozov, and Harmen J Bussemaker. Statistical mechanical modeling of genome-wide transcription factor occupancy data by matrixreduce. *Bioinformatics*, 22(14):e141–e149, 2006. 1

[28] Janina Geiler, Martin Michaelis, Patchima Sithisarn, and Jindrich Cinatl Jr. Comparison of pro-inflammatory cytokine expression and cellular signal transduction in human macrophages infected with different influenza a viruses. *Medical microbiology and immunology*, 200(1):53–60, 2011. 2, 2.1

[29] Mark B Gerstein, Anshul Kundaje, Manoj Hariharan, Stephen G Landt, Koon-Kiu Yan, Chao Cheng, Xinmeng Jasmine Mu, Ekta Khurana, Joel Rozowsky, Roger Alexander, et al. Architecture of the human regulatory network derived from encode data. *Nature*, 489(7414):91–100, 2012. 1.3.2

[30] Casey A Gifford, Michael J Ziller, Hongcang Gu, Cole Trapnell, Julie Donaghey, Alexander Tsankov, Alex K Shalek, David R Kelley, Alexander A Shishkin, Robbyn Issner, et al. Transcriptional and epigenetic dynamics during specification of human embryonic stem cells. *Cell*, 153(5):1149–1163, 2013. 1.1

[31] Anthony Gitter and Ziv Bar-Joseph. Identifying proteins controlling key disease signaling pathways. *Bioinformatics*, 29(13):i227–i236, 2013. 2.1.1, 3.1.2

[32] Anthony Gitter, Zehava Siegfried, Michael Klutstein, Oriol Fornes, Baldo Oliva, Itamar Simon, and Ziv Bar-Joseph. Backup in gene regulatory networks explains differences between binding and knockout results. *Molecular systems biology*, 5(1):276, 2009. 1.1

[33] Anthony Gitter, Judith Klein-Seetharaman, Anupam Gupta, and Ziv Bar-Joseph. Discovering pathways by orienting edges in protein interaction networks. *Nucleic acids research*, 39(4):e22–e22, 2011. 2.1.1, 2.1.1, 3.1.2

[34] Anthony Gitter, Miri Carmi, Naama Barkai, and Ziv Bar-Joseph. Linking the signaling cascades and dynamic regulatory networks controlling stress responses. *Genome research*, 23(2):365–376, 2013. 2, 2, 2.1, 2.1.1

[35] Ritu Goila-Gaur and Klaus Strebel. Hiv-1 vif, apobec, and intrinsic immunity. *Retrovirology*, 5(51):10–1186, 2008. 3.1.2

[36] Wuming Gong, Naoko Koyano-Nakagawa, Tongbin Li, and Daniel J Garry. Inferring

dynamic gene regulatory networks in cardiac differentiation through the integration of multi-dimensional data. *BMC bioinformatics*, 16(1):1, 2015. 4.1

[37] Michael D Grove, Gayland F Spencer, William K Rohwedder, Nagabhushanam Mandava, Joseph F Worley, J David Warthen, George L Steffens, Judith L Flippen-Anderson, and J Carter Cook. Brassinolide, a plant growth-promoting steroid isolated from brassica napus pollen. 1979. 2.2.1

[38] Yoshiyuki Hakata, Masaaki Miyazawa, and Nathaniel R Landau. Interactions with dcaf1 and ddb1 in the crl4 e3 ubiquitin ligase are required for vpr-mediated g2 arrest. *Virology journal*, 11(1):1–11, 2014. 3.1.2

[39] Christopher T Harbison, D Benjamin Gordon, Tong Ihn Lee, Nicola J Rinaldi, Kenzie D Macisaac, Timothy W Danford, Nancy M Hannett, Jean-Bosco Tagne, David B Reynolds, Jane Yoo, et al. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431 (7004):99–104, 2004. 1.1

[40] Maureen E Hillenmeyer, Eula Fung, Jan Wildenhain, Sarah E Pierce, Shawn Hoon, William Lee, Michael Proctor, Robert P St Onge, Mike Tyers, Daphne Koller, et al. The chemical genomic portrait of yeast: uncovering a phenotype for all genes. *Science*, 320 (5874):362–365, 2008. 1.1

[41] Alec J Hirsch. The use of rnai-based screens to identify host proteins involved in viral replication. *Future microbiology*, 5(2):303–311, 2010. 1.1

[42] David R Hodge, K Joyce Dunn, Gou Kui Pei, Mrinal K Chakrabarty, Gisela Heidecker, James A Lautenberger, and Kenneth P Samuel. Binding of c-raf1 kinase to a conserved acidic sequence within the carboxyl-terminal region of the hiv-1 nef protein. *Journal of Biological Chemistry*, 273(25):15727–15733, 1998. 3.1.2

[43] Zhanzhi Hu, Patrick J Killion, and Vishwanath R Iyer. Genetic reconstruction of a functional transcriptional regulatory network. *Nature genetics*, 39(5):683–687, 2007. 1.1

[44] Carol Shao-shan Huang and Ernest Fraenkel. Integrating proteomic, transcriptional, and interactome data reveals hidden components of signaling and regulatory networks. *Science signaling*, 2(81):ra40, 2009. 3.1.2

[45] Laurent Jacob and Jean-Philippe Vert. Efficient peptide–mhc-i binding prediction for alleles with few known binders. *Bioinformatics*, 24(3):358–366, 2008. 1.4.1

[46] Siddhartha Jain, Anthony Gitter, and Ziv Bar-Joseph. Multitask learning of signaling and regulatory networks with application to studying human response to flu. *PLoS computational biology*, 10(12):e1003943, 2014. 1.3.3, 2.1

[47] Roy Joseph, Yuriy L Orlov, Mikael Huss, Wenjie Sun, Say Li Kong, Leena Ukil, You Fu Pan, Guoliang Li, Michael Lim, Jane S Thomsen, et al. Integrative model of genomic factors for determining binding site selection by estrogen receptor-α. *Molecular systems biology*, 6(1):456, 2010. 4.1

[48] Raja Jothi, Suresh Cuddapah, Artem Barski, Kairong Cui, and Keji Zhao. Genome-wide identification of in vivo protein–dna binding sites from chip-seq data. *Nucleic acids research*, 36(16):5221–5231, 2008. 4.1.1

[49] Stephan Kadauke and Gerd A Blobel. Chromatin loops in gene regulation. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1789(1):17–25, 2009. 4.1, 4.1.1

[50] Rosa Karlić, Ho-Ryun Chung, Julia Lasserre, Kristian Vlahoviček, and Martin Vingron. Histone modification levels are predictive for gene expression. *Proceedings of the National Academy of Sciences*, 107(7):2926–2931, 2010. 2

[51] Seyoung Kim and Eric P Xing. Statistical estimation of correlated genome associations to a quantitative trait network. *PLoS genetics*, 5(8):e1000587, 2009. 1.4.1

[52] Seyoung Kim and Eric P Xing. Tree-guided group lasso for multi-task regression with structured sparsity. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 543–550, 2010. 1.4.1

[53] Meghana Kshirsagar, Jaime Carbonell, and Judith Klein-Seetharaman. Multitask learning for host–pathogen protein interactions. *Bioinformatics*, 29(13):i217–i226, 2013. 1.4.1

[54] Marta Kulis, Ana C Queirós, Renée Beekman, and José I Martín-Subero. Intragenic dna methylation in transcriptional regulation, normal differentiation and cancer. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1829(11):1161–1174, 2013. 4.1

[55] Hakju Kwon, Nadine Pelletier, Carmela DeLuca, Pierre Genin, Sonia Cisternas, Rongtuan Lin, Mark A Wainberg, and John Hiscott. Inducible expression of iκbα repressor mutants interferes with nf-κb activity and hiv-1 replication in jurkat t cells. *Journal of Biological Chemistry*, 273(13):7431–7440, 1998. 3.1.2

[56] Jan K Lenstra. *Local search in combinatorial optimization*. Princeton University Press, 1997. 1.4.2

[57] Jeffery Li, Travers Ching, Sijia Huang, and Lana X Garmire. Using epigenomics data to predict gene expression in lung cancer. *BMC bioinformatics*, 16(Suppl 5):S10, 2015. 4.1

[58] Tak W. Mak and Mary E. Saunders. *The Immune Response: Basic and Clinical Principles*, volume 1. 2006. 3.1.2

[59] Anthony Mathelier and Wyeth W Wasserman. The next generation of transcription factor binding site prediction. *PLoS Comput Biol*, 9(9):e1003214, 2013. 4.1.1

[60] Alika K Maunakea, Raman P Nagarajan, Mikhail Bilenky, Tracy J Ballinger, Cletus DSouza, Shaun D Fouse, Brett E Johnson, Chibo Hong, Cydney Nielsen, Yongjun Zhao, et al. Conserved role of intragenic dna methylation in regulating alternative promoters. *Nature*, 466(7303):253–257, 2010. 4.1

[61] Robert C McLeay, Tom Lesluyes, Gabriel Cuellar Partida, and Timothy L Bailey. Genome-wide in silico prediction of gene expression. *Bioinformatics*, page bts529, 2012. 2

[62] Jane Mellor, Peter Dudek, and David Clynes. A glimpse into the epigenetic landscape of gene regulation. *Current opinion in genetics & development*, 18(2):116–122, 2008. 2

[63] Pejman Mohammadi, Sébastien Desfarges, István Bartha, Beda Joos, Nadine Zangger, Miguel Muñoz, Huldrych F Günthard, Niko Beerenwinkel, Amalio Telenti, and Angela Ciuffi. 24 hours in the life of hiv-1 in a t cell line. *PLoS pathogens*, 9(1):e1003161, 2013. 3.1.2

[64] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature methods*, 5(7): 621–628, 2008. 1.1

[65] Vincent Navratil, Benoît de Chassey, Laurène Meyniel, Stéphane Delmotte, Christian Gautier, Patrice André, Vincent Lotteau, and Chantal Rabourdin-Combe. Virhostnet: a knowledge base for the management and the analysis of proteome-wide virus–host interaction networks. *Nucleic acids research*, 37(suppl 1):D661–D668, 2009. 1.1, 1.3.4, 3.1.2

[66] Peter J Park. Chip–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10(10):669–680, 2009. 1.3.2

[67] Ashwini Patil, Yutaro Kumagai, Kuo-ching Liang, Yutaka Suzuki, and Kenta Nakai. Linking transcriptional changes over time in stimulated dendritic cells to identify gene networks activated during the innate immune response. *PLoS computational biology*, 9(11): e1003323, 2013. 3, 3.1.2

[68] Eric M Phizicky and Stanley Fields. Protein-protein interactions: methods for detection and analysis. *Microbiological reviews*, 59(1):94–123, 1995. 1.3.3

[69] Corné MJ Pieterse, Dieuwertje Van der Does, Christos Zamioudis, Antonio Leon-Reyes, and Saskia CM Van Wees. Hormonal modulation of plant immunity. *Annual review of cell and developmental biology*, 28:489–521, 2012. 2.2.1

[70] Corné MJ Pieterse, Ronald Pierik, and Saskia Van Wees. Different shades of jaz during plant growth and defense. *New Phytologist*, 204(2):261–264, 2014. 2.2.1

[71] Roger Pique-Regi, Jacob F Degner, Athma A Pai, Daniel J Gaffney, Yoav Gilad, and Jonathan K Pritchard. Accurate inference of transcription factor binding from dna sequence and chromatin accessibility data. *Genome research*, 21(3):447–455, 2011. 1

[72] TS Keshava Prasad, Renu Goel, Kumaran Kandasamy, Shivakumar Keerthikumar, Sameer Kumar, Suresh Mathivanan, Deepthi Telikicherla, Rajesh Raju, Beema Shafreen, Abhilash Venugopal, et al. Human protein reference database2009 update. *Nucleic acids research*, 37(suppl 1):D767–D772, 2009. 1.1, 1.3.3, 3.1.2

[73] Yanjun Qi, Judith Klein-Seetharaman, and Ziv Bar-Joseph. Random forest similarity for protein-protein interaction prediction from multiple sources. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 531–542, 2004. 1.4.1

[74] Franck Rapaport, Raya Khanin, Yupu Liang, Mono Pirun, Azra Krek, Paul Zumbo, Christopher E Mason, Nicholas D Socci, and Doron Betel. Comprehensive evaluation of differential gene expression analysis methods for rna-seq data. *Genome Biol*, 14(9): R95, 2013. 3.1.2

[75] Bärbel Schröfelbauer, Qin Yu, Samantha G Zeitlin, and Nathaniel R Landau. Human immunodeficiency virus type 1 vpr induces the degradation of the ung and smug uracil-dna glycosylases. *Journal of virology*, 79(17):10978–10987, 2005. 3.1.2

[76] Marcel H Schulz, William E Devanny, Anthony Gitter, Shan Zhong, Jason Ernst, and Ziv Bar-Joseph. Drem 2.0: Improved reconstruction of dynamic regulatory networks from

time-series expression data. *BMC systems biology*, 6(1):104, 2012. 1.3.2, 3.1.2

[77] Fabian Schweizer, Patricia Fernández-Calvo, Mark Zander, Monica Diez-Diaz, Sandra Fonseca, Gaétan Glauser, Mathew G Lewsey, Joseph R Ecker, Roberto Solano, and Philippe Reymond. Arabidopsis basic helix-loop-helix transcription factors myc2, myc3, and myc4 regulate glucosinolate biosynthesis, insect performance, and feeding behavior. *The Plant Cell*, 25(8):3117–3132, 2013. 2.2.1

[78] Daria Shlyueva, Gerald Stampfel, and Alexander Stark. Transcriptional enhancers: from properties to genome-wide predictions. *Nature Reviews Genetics*, 15(4):272–286, 2014. 4.1

[79] Matthew Slattery, Tianyin Zhou, Lin Yang, Ana Carolina Dantas Machado, Raluca Gordân, and Remo Rohs. Absence of a simple code: how transcription factors read the genome. *Trends in biochemical sciences*, 39(9):381–399, 2014. 4.1

[80] Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. Biogrid: a general repository for interaction datasets. *Nucleic acids research*, 34(suppl 1):D535–D539, 2006. 1.3.3, 3.1.2

[81] Wenjie Sun, Xiaoming Hu, Michael HK Lim, Calista KL Ng, Siew Hua Choo, Diogo S Castro, Daniela Drechsel, François Guillemot, Prasanna R Kolatkar, Ralf Jauch, et al. Thermos: Estimating protein–dna binding energies from in vivo binding profiles. *Nucleic acids research*, 41(11):5555–5568, 2013. 1

[82] Norio Takada, Takaomi Sanda, Hiroshi Okamoto, Jian-Ping Yang, Kaori Asamitsu, Lilen Sarol, Genjiro Kimura, Hiroaki Uranishi, Toshifumi Tetsuka, and Takashi Okamoto. Rela-associated inhibitor blocks transcription of human immunodeficiency virus type 1 by inhibiting nf-κb and sp1 actions. *Journal of virology*, 76(16):8019–8030, 2002. 3.1.2

[83] Mahmud Tareq Hassan Khan, Carlo Mischiati, Arjumand Ather, Tatsuya Ohyama, Kenichi Dedachi, Monica Borgatti, Noriyuki Kurita, and Roberto Gambari. Structure-based analysis of the molecular recognitions between hiv-1 tar-rna and transcription factor nuclear factor-kappab (nfkb). *Current topics in medicinal chemistry*, 12(8):814–827, 2012. 3.1.2

[84] Debra J Taxman, Chris B Moore, Elizabeth H Guthrie, and Max Tze-Han Huang. Short hairpin rna (shrna): design, delivery, and assessment of gene knockdown. *RNA Therapeutics: Function, Design, and Delivery*, pages 139–156, 2010. 1.1

[85] Léon-Charles Tranchevent, Roland Barriot, Shi Yu, Steven Van Vooren, Peter Van Loo, Bert Coessens, Bart De Moor, Stein Aerts, and Yves Moreau. Endeavour update: a web resource for gene prioritization in multiple species. *Nucleic acids research*, 36(suppl 2): W377–W384, 2008. 2.1.3

[86] Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn, and Lior Pachter. Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. *Nature protocols*, 7(3):562–578, 2012. 1.1

[87] Chien Van Ha, Marco Antonio Leyva-González, Yuriko Osakabe, Uyen Thi Tran, Rie

Nishiyama, Yasuko Watanabe, Maho Tanaka, Motoaki Seki, Shinjiro Yamaguchi, Nguyen Van Dong, et al. Positive regulatory role of strigolactone in plant responses to drought and salt stress. *Proceedings of the National Academy of Sciences*, 111(2):851–856, 2014. 2.2.1

[88] Thanasis Vergoulis, Ioannis S Vlachos, Panagiotis Alexiou, George Georgakilas, Manolis Maragkakis, Martin Reczko, Stefanos Gerangelos, Nectarios Koziris, Theodore Dalamagas, and Artemis G Hatzigeorgiou. Tarbase 6.0: capturing the exponential growth of mirna targets with experimental support. *Nucleic acids research*, 40(D1):D222–D229, 2012. 1.1

[89] A Corina Vlot, D'Maris Amick Dempsey, and Daniel F Klessig. Salicylic acid, a multifaceted hormone to combat disease. *Annual review of phytopathology*, 47:177–206, 2009. 2.2.1

[90] Zhibin Wang, Chongzhi Zang, Jeffrey A Rosenfeld, Dustin E Schones, Artem Barski, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Weiqun Peng, Michael Q Zhang, et al. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature genetics*, 40(7):897–903, 2008. 4.1

[91] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009. 1.1

[92] Christian Widmer, Jose Leiva, Yasemin Altun, and Gunnar Rätsch. Leveraging sequence classification by taxonomy-based multitask learning. In *Research in Computational Molecular Biology*, pages 522–534. Springer, 2010. 1.4.1

[93] Samuel A Williams, Hakju Kwon, Lin-Feng Chen, and Warner C Greene. Sustained induction of nf-κb is required for efficient expression of latent human immunodeficiency virus type 1. *Journal of virology*, 81(11):6043–6056, 2007. 3.1.2

[94] Emily S Wires, David Alvarez, Curtis Dobrowolski, Yun Wang, Marisela Morales, Jonathan Karn, and Brandon K Harvey. Methamphetamine activates nuclear factor kappa-light-chain-enhancer of activated b cells (nf-κb) and induces human immunodeficiency virus (hiv) transcription in human microglial cells. *Journal of neurovirology*, 18(5):400–410, 2012. 3.1.2

[95] Laurence A Wolsey and George L Nemhauser. *Integer and combinatorial optimization*. John Wiley & Sons, 2014. 1.4.2

[96] Chen-Hsiang Yeang, Trey Ideker, and Tommi Jaakkola. Physical network models. *Journal of computational biology*, 11(2-3):243–262, 2004. 3a

[97] Hong Yu, Shanshan Zhu, Bing Zhou, Huiling Xue, and Jing-Dong J Han. Inferring causal relationships among different histone modifications and gene expression. *Genome research*, 18(8):1314–1324, 2008. 4.1

[98] Alessia Zamborlini, Audrey Coiffic, Guillaume Beauclair, Olivier Delelis, Joris Paris, Yashuiro Koh, Fabian Magne, Marie-Lou Giron, Joelle Tobaly-Tapiero, Eric Deprez, et al. Impairment of human immunodeficiency virus type-1 integrase sumoylation correlates with an early replication defect. *Journal of Biological Chemistry*, 286(23):21013–21022, 2011. 3.1.2

[99] Gabriel E Zentner, Paul J Tesar, and Peter C Scacheri. Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. *Genome research*, 21(8): 1273–1283, 2011. 4.1

[100] Jie Zheng, Iti Chaturvedi, and Jagath C Rajapakse. Integration of epigenetic data in bayesian network modeling of gene regulatory network. In *Pattern Recognition in Bioinformatics*, pages 87–96. Springer, 2011. 4.1

[101] Shan Zhong. *Computational Study of Transcriptional Regulation-From Sequence To Expression*. PhD thesis, Carnegie Mellon University, 5 2013. 1.2, 1.3.2

[102] Jian Zhou and Olga G Troyanskaya. Predicting effects of noncoding variants with deep learning-based sequence model. *Nature methods*, 12(10):931–934, 2015. 4.1.1

[103] Jia-Ying Zhu, Juthamas Sae-Seaw, and Zhi-Yong Wang. Brassinosteroid signalling. *Development*, 140(8):1615–1620, 2013. 2.2.1