Inferring Temporal Signaling Pathways and Regulatory Mechanisms from High-Throughput Data

Siddhartha Jain

CMU-CS-17-124
October 19, 2017

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Ziv Bar-Joseph, Chair
Jaime Carbonell
Eric Xing
Naftali Kaminski (Yale University)

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

This research was sponsored by the National Science Foundation under grant numbers IIS-0905390, DBI-0965316, CCF-1101668, IIP-1127832, and DBI-1356505, the McDonnell Foundation under grant number 22020366, the National Institutes of Health under grand numbers 1R01GM085022 and R01GM122096, Mt. Sinai under grant number 025578554609, and Yale University under grant number M16A12324A10472. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of any sponsoring institution, the U.S. government or any other entity.

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 that are focused on modeling the dynamic regulation of biological processes. These methods address the various aspects of the problem mentioned above focusing on the reconstruction of dynamic signaling and regulatory networks. In many cases, the amount of biological data available for a specific condition can be very small compared to the number of variables. We 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. The thesis presents a new algorithm that utilizes and extends Integer Programming methods for inferring such dynamic regulation. Finally, we present a new strategy to integrate epigenetic data with other temporal datasets using deep neural networks. We use this new method to reconstruct dynamic disease progression networks in Idiopathic Pulmonary Fibrosis (IPF).

Acknowledgments

This thesis would not be possible without the incredible guidance and mentorship I received from my advisor Ziv Bar-Joseph. I am very grateful to him for having taken a chance on me back when I had neither much knowledge of computational biology nor a deep background in machine learning. Ziv's focus on having a research plan in place when starting out a project and an understanding of how the end result may help a biologist has been immensely valuable in getting concrete results at the end of a project.

I would like to thank my committee members, Jaime Carbonell, Naftali Kaminski, and Eric Xing for being part of my committee and providing useful feedback on the thesis. I would also like to thank my amazing co-authors Joel Arrais, Velpandi Ayyavoo, Anthony Gitter, Chieh Lin, and Jay Venkatachari without whom this PhD would not have been possible.

As a member of Ziv's Systems Biology Group, I have greatly benefitted with conversations and consistently useful advice from Ercument Cicek, Jun Ding, Jason Ernst, Xin He, Hai-Son Le, Saket Navlakha, Sabrina Rashid, Matthew Ruffalo, Marcel Schulz, Aaron Wise, Shan Zhong, and Guy Zinman. Having Amir Alavi, and Michael Kleyman around in the last few years of my PhD made for great random conversations on research as well as good times outside of work.

I would also like to thank CMU professors Jian Ma, Andreas Pfenning, Ariel Procaccia, and Tuomas Sandholm for their support and useful research and career advice over the years.

I have been very fortunate to share great times, tasty food, and conversations about life and research with friends in and around the wider CMU community including (listed alphabetically) Sruti Bhagavatula, Noam Brown, Alejandro Carbonara, Stephanie Cathryn, Zack Coker, John Dickerson, Aram Ebtekar, Sanjana Gupta, An Ka, Meghana Kshirsagar, Ashique Khudhabuksh, Christian Kroer, David Kurokawa, Kelvin Liu, Janine Love, Abe Othman, Jakub Pachoki, Nick Resch, Ram Raghunathan, Steven Rich, Shayak Sen, Sahil Singla, Petar Stojanov, Jocelyn Sunseri, Dougal Sutherland, Ameya Velingker, John Wright, Jay Yoon, Erik Zawadzki, Veronica Zheng, and Goran Zuzic.

Lastly, I'm deeply indebted to my parents Veena and Jitender Jain for their unconditional love and continual support throughout my undergraduate and graduate studies.

Contents

List of Figures

List of Tables

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 [117]. 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 [46]. 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 [46, 118]. 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 [120], 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 [111]. 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 [92]. 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 [73, 255]. 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 [34, 78], 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 [184], profiling of microRNAs [270], ChIP-Seq, epigenetics studies [90], information about protein interactions within a cell [206] and information on interactions between host proteins and pathogen / environmental factors [189]. 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 DNA methylation 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 do not consider is the role DNA methylation plays in regulating gene expression. Given our focus on trying to infer signaling pathways and active TFs for various conditions, we try and model how DNA methylation can affect TF-DNA interactions and thus affect gene regulation.

1.3 Diseases studied in this thesis

While this thesis presents general methods that can be globally applied, to illustrate their usefulness, we applied them to analyze the following diseases.

1.3.1 Influenza (Flu) infection

Influenza, commonly known as ”the flu”, is an infectious disease caused by an influenza virus. The most common symptoms of flu are fever, runny nose, sore throat, coughing, headache, muscle pains, and feeling tired. The flu virus has several subtypes, the most common of which is Influenza A. That itself has multiple subtypes. H1N1 is the most common and the one that most people usually get. The prognosis for it is usually 1-3 weeks. H3N2 is another common strain of flu and kills about 36,000 people in the United States each year. H5N1, also known as avian or bird flu, is one of the deadliest strains of flu. It kills tens of millions of birds worldwide every year [160]. While so far, H5N1 has rarely infected humans, when it does do so, it can be very deadly.

1.3.2 Human immunodeficiency virus (HIV) infection

HIV is a lentivirus that causes HIV infection and over time, acquired immunodeficiency syndrome (AIDS) [280]. HIV attacks the immune system infecting vital cells like helper T cells (CD4+ cells), macrophages, and dendritic cells. Following infection, it is typical not not notice any symptoms. This period can go for a long time, sometimes upto several years. As the infection progresses, it interferes more and more with the immune system. This stage is what is referred to as AIDS.

HIV is spread primarily by unprotected sex, contaminated blood transfusions, hypodermic needles, and from mother to child during pregnancy, delivery, or breastfeeding [221]. In 2016, about 37 million people were living with HIV and it caused 1 million deaths [4].

Without treatment the average survival time after infection is 9-11 years [266]. With treatment, life expectancy can be 10-40 years [274].

1.3.3 Idiopathic pulmonary fibrosis (IPF)

Idiopathic Pulmonary Disease (IPF) is the most common of the interstitial lung diseases and the most severe with median survival ranging from 3-5 years [102]. It is described as a chronic, progressive fibrosing interstitial pneumonia of unknown etiology that occurs more commonly in older male subjects with smoking being the major risk factor. IPF belongs to a large group of more than 200 lung diseases known as interstitial lung diseases charaterized by involvement of the lung interstitium [261]. IPF is estimated to occur in 14.0 and 42.7 per 100,000 persons in the United States [213] depending on how IPF is being defined. It is more common in men than in women and usually diagnosed in people over the age of 50 [214]. The median survival time can be between 2 to 5 years after diagnosis [214]. The 5-year survival for IPF ranges between 20 and 40% [135] – a mortality rate that is higher than diseases like colon cancer, bladder cancer, myeloma, etc [135].

1.4 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. As we discuss in the next section, a key challenge is how to integrate them in order to reconstruct a complete model of cellular activity under various conditions. Here we provide short descriptions of data used.

1.4.1 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 [60], a detailed review of the RNA-seq pipeline and a survey of best practices for RNA-seq data analysis is provided. This data is used throughout the thesis, in particular in Chapters §2, §3, and §4.

1.4.2 Chip-Chip and Chip-Seq

Chromatin immunoprecipitation (ChIP) followed by microarray (ChIP-chip) [44] or sequencing (ChIP-seq) [198] 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 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) [301]. 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 [89] for 348 transcription factors to get our human TF-DNA interaction network as in [230] comprising of 59K TF-DNA interactions. This data is also used throughout the thesis.

1.4.3 Epigenetic modifications

Epigenetic modifications, such as DNA methylation and histone modification, alter DNA accessibility and chromatin structure, thereby regulating patterns of gene expression. These processes are crucial to normal development and differentiation of distinct cell lineages in the adult organism. They can be modified by exogenous influences, and, as such, can contribute to or be the result of environmental alterations of phenotype or pathophenotype. Importantly, epigenetic programming has a crucial role in the regulation of pluripotency genes, which become inactivated during differentiation

DNA methylation is the covalent attachment of a methyl group to the C5 position of cytosine residues in CpG dinucleotide sequences (referred to as CpG throughout this review) [31]. Recent findings suggest that in undifferentiated stem cells, cytosines, other than those in CpG, can be methylated, as well [164], and that methylation of non-CpG cytosines is crucial for gene regulation in embryonic stem cells in particular. CpG methylation is, however, an important mechanism to ensure the repression of transcription of repeat elements and transposons, and also plays a crucial role in imprinting and X-chromosome inactivation [219]. Transcriptional gene silencing by CpG methylation also restricts the expression of some tissue-specific genes during development and differentiation by repressing them in non-expressing cells.

Histones are highly alkaline proteins found in eukaryotic cell nuclei that package and order the DNA into structural units called nucleosomes [190, 293]. They are the chief protein components of chromatin, acting as spools around which DNA winds, and playing a role in gene regulation. Without histones, the unwound DNA in chromosomes would be very long (a length to width

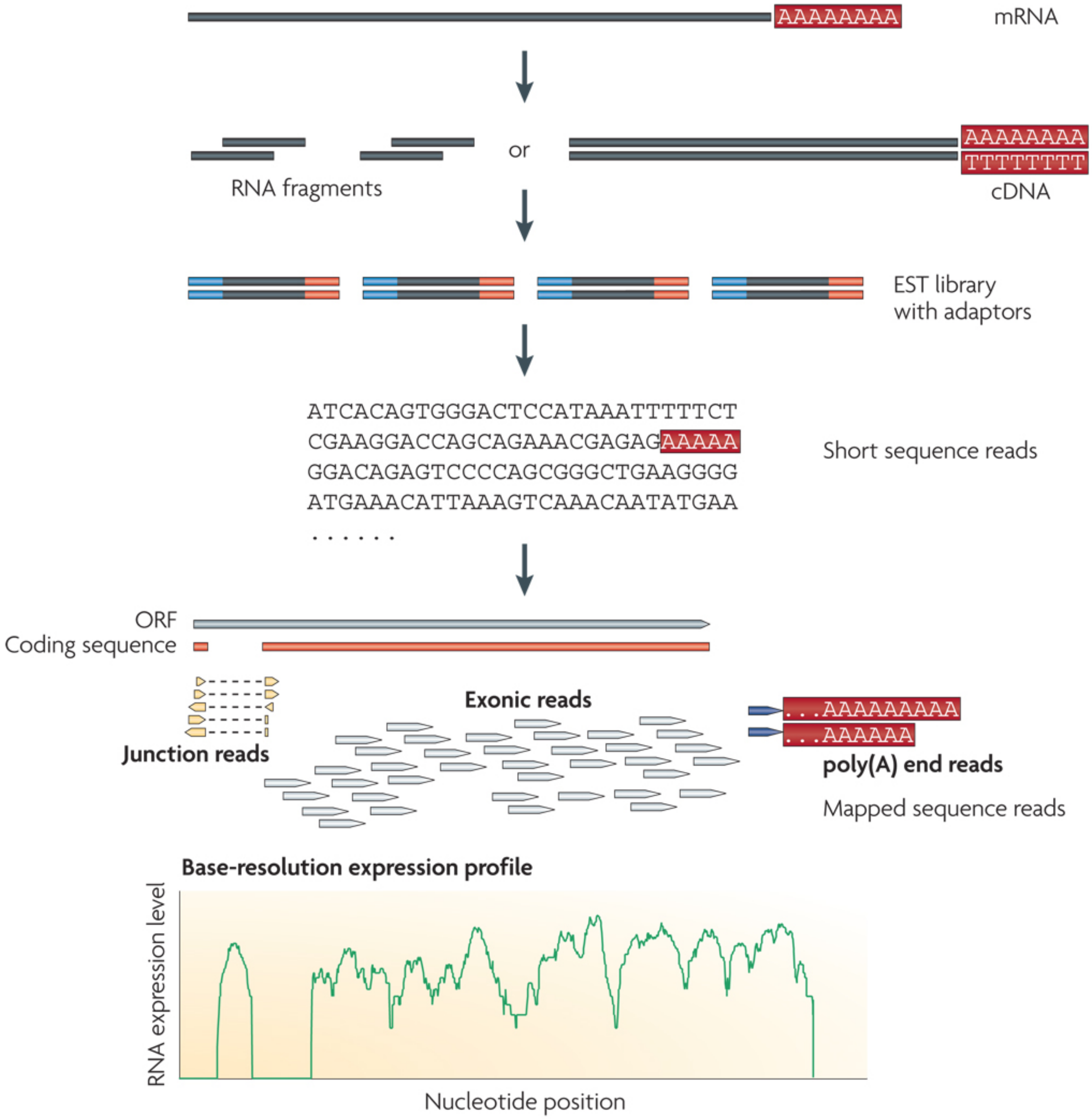

Figure 1.1: **Graphic showing a typical RNA-seq experiment.** Briefly, long RNAs are first converted into a library of cDNA fragments through either RNA fragmentation or DNA fragmentation. Sequencing adaptors (blue) are subsequently added to each cDNA fragment and a short sequence is obtained from each 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 [260]). 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 [278]

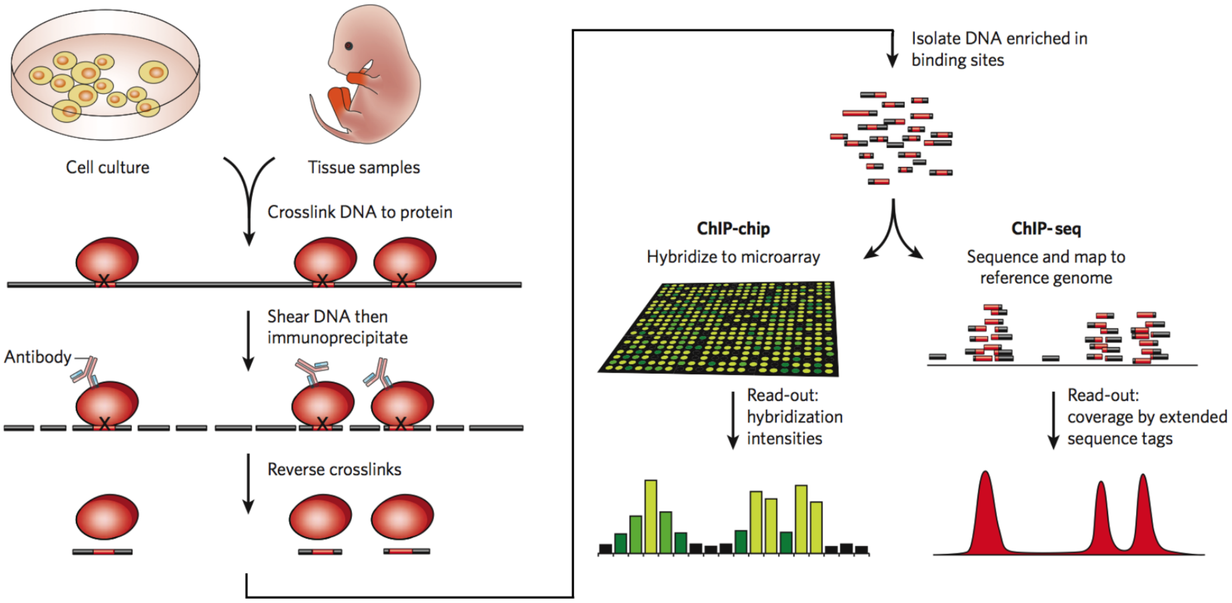
Cell culture
Tissue samples
Crosslink DNA to protein
Shear DNA then immunoprecipitate
Antibody
Reverse crosslinks
Isolate DNA enriched in binding sites
ChIP-chip
Hybridize to microarray
ChIP-seq
Sequence and map to reference genome
Read-out: hybridization intensities
Read-out: coverage by extended sequence tags

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 [301]

ratio of more than 10 million to 1 in human DNA). Histones can be post-translationally modified to restructure chromatin in many ways, including phosphorylation, ubiquitination, acetylation, and methylation [145, 175].

Other epigenetic modifications include (but are not limited to), chromatin structure which can vary from celltype to celltype and is measured via a technology called chromatin conformation capture [70]; methylation of messenger RNAs [69]; Prions which are infectious forms of proteins [207]; etc.

Measuring DNA methylation

We make extensive use of DNA methylation data in Chapter §4. Here we describe two popular protocols to measure DNA methylation levels.

Reduced Representation Bisulphite Sequencing (RRBS) Reduced representation bisulfite sequencing (RRBS) is an efficient and high-throughput technique used to analyze the genome-wide methylation profiles on a single nucleotide level. This technique combines restriction enzymes and bisulfite sequencing in order to enrich for the areas of the genome that have a high CpG content. Due to the high cost and depth of sequencing needed to analyze methylation status in the entire genome, Meissner et al. developed this technique in 2005 [179] in order to reduce the amount of nucleotides needed to be sequenced to 1% of the genome. The fragments that comprise the reduced genome still include the majority of promoters, as well as regions such as repeated sequences that are difficult to profile using conventional bisulfite sequencing approaches [6, 104]. This data is used in Chapter §4.

Whole Genome Bisulphite Sequencing (WGBS) Whole Genome Bisulphite Sequencing is another technology used to determine the DNA methylation status of single cytosines. DNA is treated with the chemical sodium bisulfite. This compound converts unmethylated cytosines to uracils. The cytosines that have not been converted into uracil are the methylated ones. Then the resulting treated DNA is sequenced. Unmethylated cytosines appear as thymines. Comparison with the sequence of the untreated DNA tells you which cytosines are methylated. This data is used in Chapter §4.

1.4.4 Chromosome conformation capture

The human DNA, if arranged in a perfectly straight line would stretch out to 3 meters [199]. As the typical cell is on teh order of micrometers, DNA needs to be folded into a complex 3D shape to be able to fit inside the cell. This folding pattern is called the chromatin structure and it has a huge influence in cell biology and gene expression [9, 243].

Chromosome conformation capture is a set of molecular biology techniques used to try and detect the chromatin structure. They consist of the following :-

- 3C (one-vs-one) :- 3C tries to detect the interactions between just a single pair of genomic loci. This technique has the highest resolution out of all often reaching resolutions of 1-8 kilobases (kb) on average (i.e. it is able to detect 1-8kb sized genome segments which interact with each other) [244].

- 4C (one-vs-all) :- As the description implies, 4C tries to capture interactions between one locus vs all other genomic loci. It typically reaches resolutions of 3-4 kb [99]
- 5C (many-vs-many) :- This version detects interactions between all restriction fragments within a given region, with this region's size typically no greater than a megabase [70, 108]. However it suffers from low coverage.
- Hi-C (all-vs-all) :- Hi-C uses high-throughput sequencing to find the nucleotide sequence of fragments are interacting with each other [108, 162]. All possible pairwise interactions between different genomic loci are tested. However the resolution for this variant is poor with it being as low as 40kb.

This data is used in Chapter §4.

1.4.5 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. [202] 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 [245] and HPRD [206] databases which collate interactions for the above such experimental sources. An interaction could have been detect in multiple independent experiments. This data is used in Chapters §2, and §3. We processed the interactions as follows to obtain a weight set of edges between the proteins.

Edge score

As mentioned above, the protein interaction network is collected from BioGRID and HPRD. The PPI (protein-protein interaction) and PTM (Post-translational modification) scores are calculated based on experimental methodology and number of independent detections as in [93]. More specifically, for an edge e_{ij} between proteins i, j, the score is

$$\mathbb{P}(e_{ij} = 1 - \Pi_{k \in I_{ij}}(1 - c(k))$$

where I_{ij} is the set of all distinct instances of i, j interacting in the PPI or PTM data based on experiment type (yeast 2-hybrid, coimmunoprecipitation, etc.) and $c(k)$ is the confidence in the class of experiments to which k belongs. The values for $c(k)$ for the BioGRID interactions are taken from [93]. HPRD included the more generic types of interaction evidence 'in vivo' and 'in vitro', both of which were given a confidence of 0.6.

Protein signaling pathway score

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 computed as specified above. It can be interpreted as the probability of the existing of each pathway assuming that the edge existence probabilies($\mathbb{P}(e)$) are independent.

1.4.6 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 [189].

This data is used in Chapters §2, and §3.

1.4.7 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 [38]. 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.

This data is used in Chapters §2, and §3.

1.4.8 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 [20]. 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.

This data is used throughout the thesis.

1.5 Computational techniques used in this thesis

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

1.5.1 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

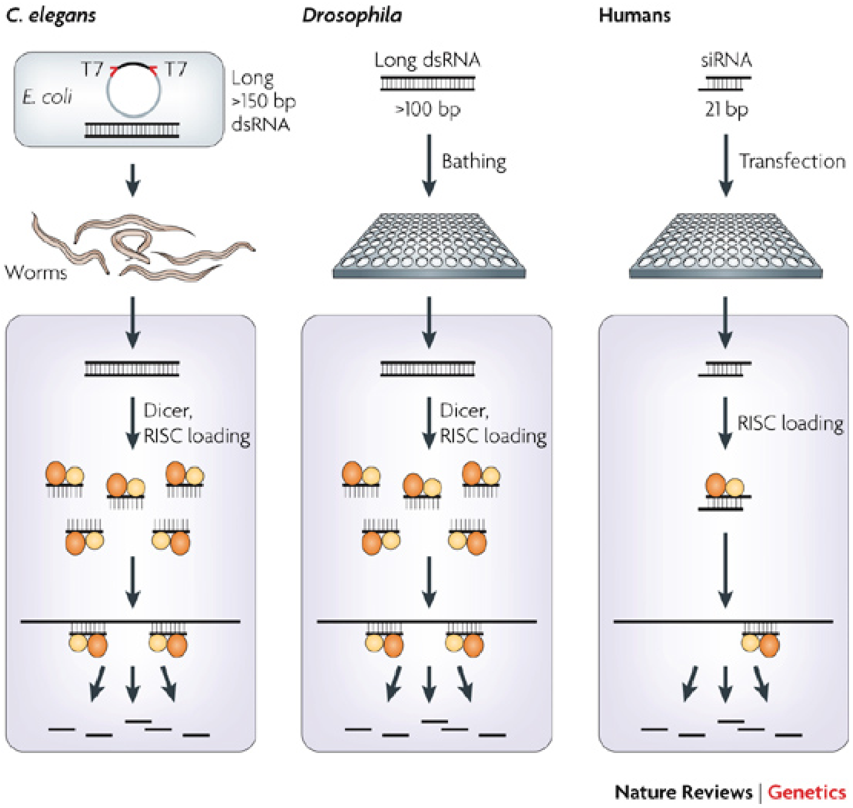

Figure 1.3: **Overview of RNAi screening approaches used in different organisms.** Long double-stranded (ds) RNAs are introduced into a cell or and are intracellularly diced into small-interfering RNAs (siRNAs). This leads to highly efficient knockdown because many different siRNAs are generated from 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 [38]

parameters to fit is very large relative to the number of samples. In addition, extensive data 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 [281], genome-wide association studies [136, 137], protein structure [126], and pairwise protein-protein interaction prediction [147, 208].

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 [80]

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

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.

1.5.2 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 [285] 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) [285]. This is what is 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 do not 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. [157]. As we shall see later in this chapter, due to the size of our problem, we are forced to resort to the latter techniques.

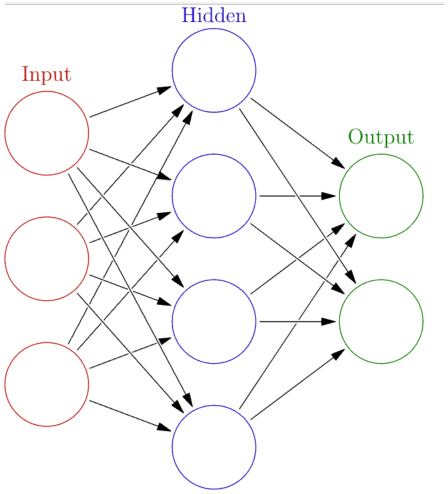

Figure 1.4: **Example neural network.** Image taken from [5]

1.5.3 One-hot encoding

One-hot encoding transforms categorial features into unit vectors which work better with classification and regression algorithms. Thus if the set of categories is C and the size of C is N, then each element in C is assigned a number from 1 to N. The one hot encoding for a feature x then is a vector with a 1 at the position k where k is the number assigned to the category represented by x, and 0s at all other positions. In particular, for the purpose of this thesis, the one-hot encoding of a DNA base is a vector of size 4 with position 1 in the vector having value 1 if the base is A, position 2 if the base is C, position 3 for base G, and position 4 for base T with all other positions having value 0.

1.5.4 Neural networks

An (artificial) neural network is a network of simple elements called neurons, which receive input, change their internal state (activation) according to that input, and produce output depending on the input and activation. The network forms by connecting the output of certain neurons to the input of other neurons forming a directed, weighted graph. The weights as well as the functions that compute the activation can be modified by a process called learning which is governed by a learning rule. Typically the neurons are organized in layers with the output of one layer being fed as input to the next layer as shown in Figure 1.4.

The heart of a neural network is the neuron. An example neuron is shown in Figure 1.5. It is essentially a function that elementwise multiplies its associated weights (which are the parameters to be learned during training) with the inputs, sums the resulting vector and then passes it through an activation function which is typically non-linear. The figure shows the sigmoid activation function though rectified linear units, exponential linear units, etc. are more commonly used activation functions these days.

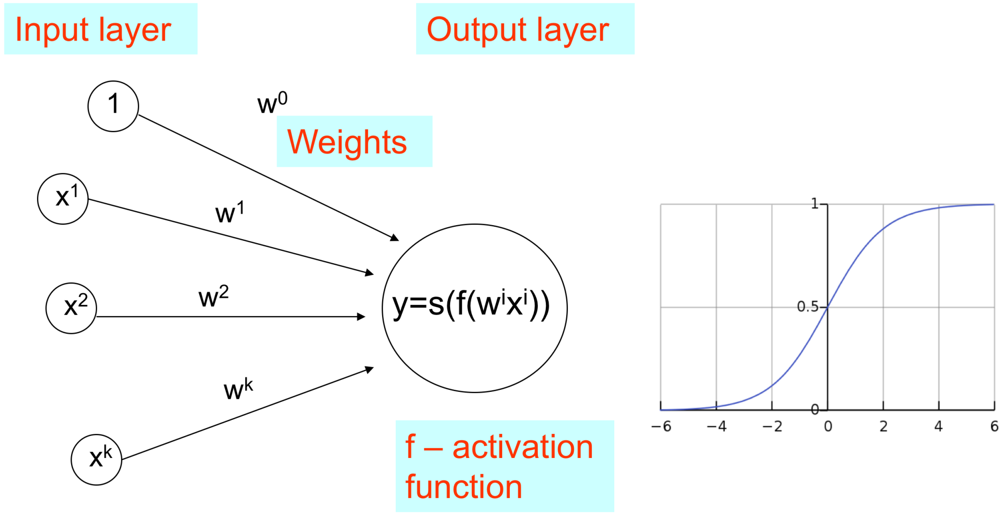

Figure 1.5: **An example neuron.** The plot on the right shows a sigmoid activation function.

Types of neural networks

As mentioned previously, then neurons in a neural network are typically organized into layers. There exist a wide variety of connectivity patterns between layers however, some of which are described below.

- **Dense layers**

  The most common type of layers, every neuron is connected to every input node.

- **Convolutional layers**

  In convolutional layers, each neuron is connected to only a small subset of *contiguous* input nodes (for example, adjacent pixels in an image). The next neuron adjacent to the previous one is then connected to a subset of input nodes that is contiguous as before but also *overlaps* with the set of input nodes for the previous neuron. An example of a convolutional layer is given in Figure 1.6. The set of input nodes is called the kernel or receptive field. The reason to prefer convolutional layers over dense ones is that they massively reduce the number of parameters that need to be learned from $m \cdot n$ where m is the number of input nodes and n the number of neurons in the layer to just the number of nodes in the receptive field. They have been very successfully used in computer vision and other areas.

- **Pooling layers**

  Used for reducing the dimensionality, they are similar to convolutional layers with the key difference that the receptive fields of adjacent neurons do not overlap. Thus adjacent neurons do not share any inputs. An example of a pooling layer is given in Figure 1.6.

- **Recurrent neural networks**

  Described above are neural networks which work on fixed size input. However if the input is not fixed size or if you want to generate time series predictions from your network, then

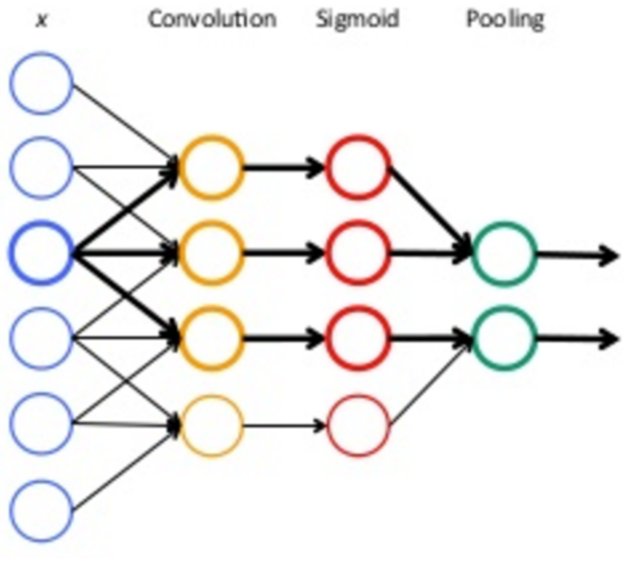

Figure 1.6: **Example convolutional and pooling layers.**

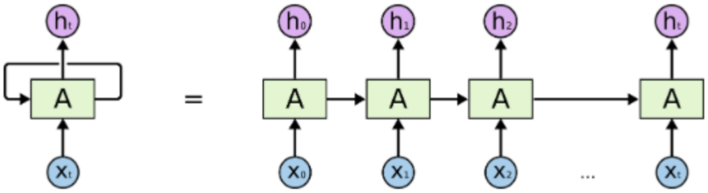

Figure 1.7: **Example recurrent neural network.**

a recurrent neural network is appropriate. Essentially in recurrent neural networks, the connections between layers from a directed cycle. So the output of a layer is fed back into the input. An example of a recurrent neural network is given in Figure 1.7. The most common recurrent neural network in vogue right now is LSTM (long short-term memory). There is also a variant of it called Bi-LSTM which has been successfully used for image analysis and natural language processing.

Training neural networks

Neural networks are usually trained using the backpropagation algorithm [114] which is essentially gradient descent. In particular, these days they are trained using stochastic gradient descent [37] which uses only a small number of training examples to estimate the gradient greatly speeding up training. A large amount of effort has also been devoted to proper initialization of the parameters and the step size one should use for gradient descent yielding significant speedups [96, 139, 296]. The advent of GPUs to train neural networks in particular has yielded 10-20x speedups in terms of training time [146, 194]. In this thesis, we use the nVidia 1080i GPU to train and test our neural networks.

1.6 Prior approaches to reconstructing regulatory and signaling networks

The relative ease of high-throughput data collection enables profiling a system of interest in many ways with complementary assays, at different times, and under various perturbations to compare and contrast the outcomes. The resulting computational challenge is to develop analysis strategies that maximally leverage these related experiments to improve our ability to reconstruct biologically accurate models.

Even when applied to study the same condition, different types of high-throughput data (e.g., functional genetic screens and gene expression) often times implicate largely disjoint groups of genes or proteins because each experiment highlights different facets of the biological processes and networks involved [290]. Consequently, there has been extensive research to develop techniques for integrating one or more types of *condition-specific* high-throughput data with *general purpose* physical interaction networks, such as protein-protein interactions (PPIs), to reconstruct signaling and regulatory networks.

Below we give brief surveys of such methods that have been used to reconstruct static and dynamic signaling and regulatory networks.

1.6.1 Methods to reconstruct static networks

There has been a huge amount of work in trying to reconstruct static regulatory networks. We provide a very brief overview of these here since our main focus in this thesis is on the modeling of time series based networks. In Gardner et al. [86], they use linear models (without assuming any prior knowledge of the network structure) to study the SOS response system in *E. coli*. In Tyson et al. [265] and Rao et al. [216], they go further and develop non-linear kinetic and stochastic models to understand the behavior of regulatory networks. Pe'er et al. [201] use Bayesian networks to infer a variety of metabolic, signaling, and regulatory pathways for *S. cervesiae*. Bayesian networks are also used by Friedman et al. in [83]. Shmulevich et al. [238] use a boolean network approach. Bonneau et al. [35] use a biophysical based model although their model can be applied to both static and time course data. ResponseNet [153] uses a linear programming model to connect differentially expressed genes with potential source genes that may be regulating them. Tuncbag et al. [264] extend that model to a multitask setting where several different conditions are modeled simulataneously. Finally, in SDREM [95], they reconstruct static signaling pathways that could be regulating gene expression (even though the transcriptional regulatory network is dynamic).

1.6.2 Methods to reconstruct time series networks

Static networks however, do not provide temporal information making it hard to determine the various stages associated with the system being studied (for example, waves of expression changes [50]) or the optimal time to apply an intervention. Consider the HIV-1 infection. While the development of highly active antiretroviral therapy has made it possible to delay the progression of HIV infection, the persistence of the virus, rapid development of resistance and inability to completely eliminate the virus still pose major challenges for effective HIV-1 management [239]. HIV-1 infects a host cell by a sequential process involving several temporal events. These start with binding of the viral envelope protein to the host cell receptor followed by reverse transcription and integration of proviral DNA (early infection stage). Next, viral proteins are produced facilitating viral replication (intermediate stage). Finally, new viruses are released (late stage). While several studies have experimentally quantified the large scale changes and host-pathogen interactions for HIV-1 infection [223], to date no models exist to fully link these high throughput temporal datasets with the underlying dynamic networks that lead to the observed responses.

A small number of methods have been proposed for reconstructing dynamic interaction networks from high throughput data. These methods utilize the (relatively small number of) time series datasets to determine temporal information for the (mostly) static interaction datasets either directly (by projecting the time series data on the known interaction networks [68]) or indirectly (by looking at targets of transcription factors (TFs) and associating temporal information for the interactions based on these targets [78, 230]. Since gene expression is the primary source of time series data these methods use, they have primarily focused on the reconstruction of regulatory networks [23]. Signaling networks proved to be more challenging since much of the activity in these networks is post transcriptional [82] and often faster than regulatory networks which made it hard to use time series gene expression data to obtain temporal information about the activity of these networks.

Several other methods have been developed and evaluated for reconstructing regulatory networks using gene expression data [112, 171, 173, 257]. These methods utilize expression levels to determine regulatory interactions based on various statistical techniques including correlation, mutual information, regression etc. Particularly interesting is Kolar et al. [142], in which they use markov random fields to estimate *time-varying* networks. While such methods can be successfully applied in some cases, they are less appropriate for modeling immune response dynamics since they cannot model post-transcriptional events (including the effects of virus-host and protein-protein interactions) which, as we show, play a major role in such responses.

To address these issues, two new methods have been proposed recently to jointly reconstruct dynamic signaling and regulatory networks by integrating static and time series data. SDREM [93] relies on a method for orienting protein interaction networks which are then combined with TFs and the networks they regulate using a separate input-output hidden markov model (IOHMM). While SDREM has been successfully applied to study yeast and human response networks [91, 94, 127] it does not provide temporal information about the pathways it finds. In 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, SDREM 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 [200] 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.

1.6.3 Network models applied to disease

There are several instances of literature of network models being successfully applied to disease. Zhai et al. [299] applied the WGCNA algorithm [154] to flu data to construct network modules from gene co-expression analysis and uncovered several relevant gene networks (Figure 1.8). Schulz et al. [230] applied the DREM model [78] to study lung development in mice and idiopathic pulmonary fibrosis (IPF). They found several miRNAs involved in the regulation of IPF that they validated using proliferation assays. Novershtern et al. [192] applied a Bayesian model termed Physical Module Networks to study the response of primary human epithelial cells to the H1N1 flu virus and found several relevant pathways. Yosef et al. [291] used novel computational methods to study Th17 cell differentiation and identified novel drug targets for controlling it. In Alvarez et al. [13], they use a network-based inference algorithm for protein activity to characterize somatic mutations in cancer. Fu et al. [84] identify a microRNA-mRNA regulatory network in colorectal cancer using bioinformatics analysis. Brichta et al. [41] were able to use ARACNe [173] to identify key factors that determine neuronal survival or death in degenerative disorders. Finally, in Hajingabo et al. [106], they use network models to identify functional effects of genetic alterations. A more comprehensive review can be found in [77].

1.7 Structure of the thesis

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 our attempts at incorporating DNA methylation data into our models and we finally conclude in Chapter §5.

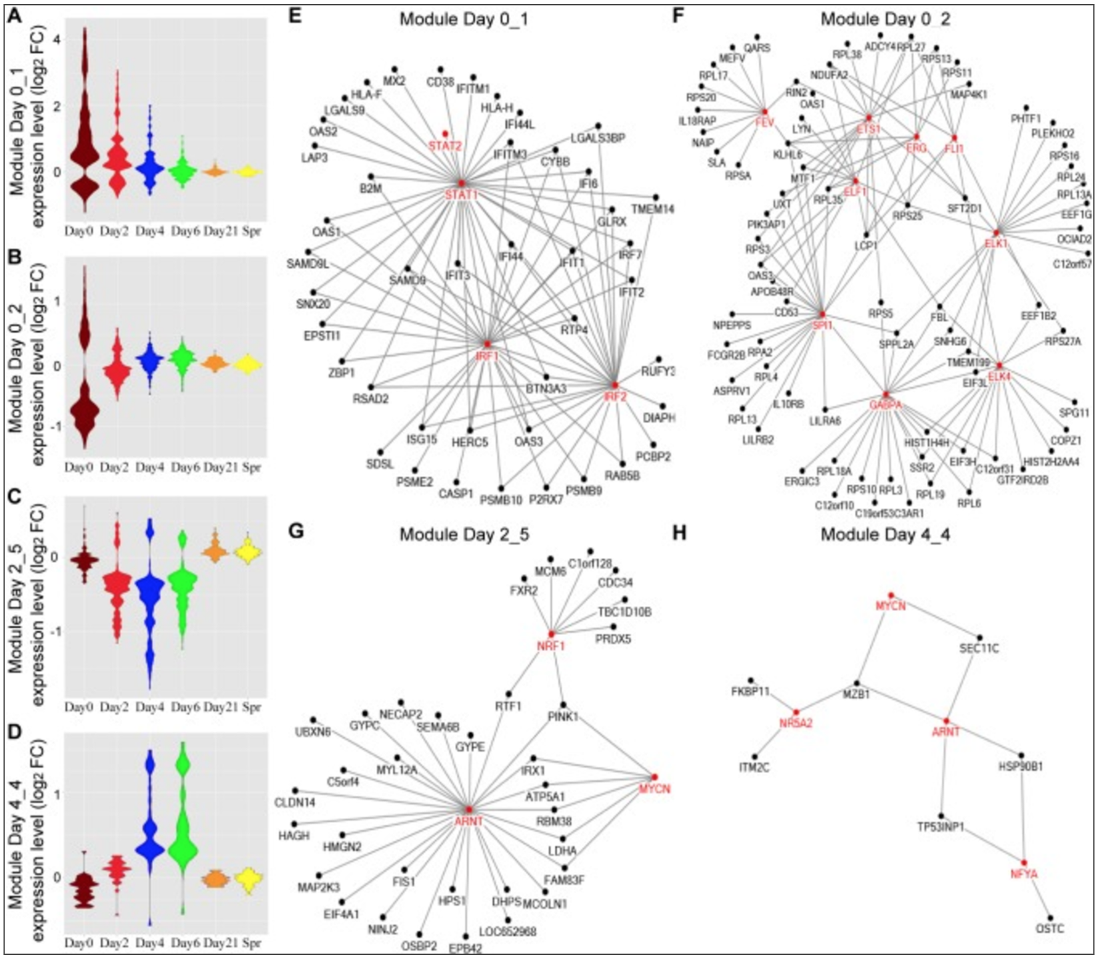

Figure 1.8: **TF networks within the WGCNA modules over the course of influenza illness.**. **(A-D)** Groups, or modules, of co-regulated DEGs were identified by WGCNA. Representative Gene Ontology (GO) categories for each module were identified by functional enrichment analysis and shown in Table 6. Module expression patterns across different time points were represented by violin plots of log2 fold-change in gene expression relative to baseline. **(E-H)** Pscan was used to scan the promoter regions of all genes in each module and identify the over-represented transcription factor binding sites (TFBS). The predicted transcription factors, which marked in red and their target genes (z-score ¿ 2) were connected by edges in the networks. Image taken from [299].

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 [94]. Like its single-condition predecessor [94], 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.

MT-SDREM [127] uses multi-task learning to reconstruct response pathways and temporal regulatory networks. It 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 [94], 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 [88], and lethal, pandemic strains such as the H1N1 1918 Spanish flu and highly pathogenic avian H5N1 strains.

As MT-SDREM builds on SDREM which builds on DREM, in the next couple of sections, we briefly describe both methods.

2.1 DREM

DREM uses protein-DNA binding interactions and time series gene expression data to reconstruct dynamic regulatory networks by identifying bifurcation events, places in the time series where a set of genes that were previously co-expressed diverges. DREM annotates these split events with TFs that are predicted to regulate genes in the outgoing upward and/or downward paths allowing us to associate temporal information (the timing of the splits) with the static protein-DNA interaction data. An input-output hidden Markov model (IOHMM) [26], which unlike traditional HMMs also includes additional observed (in our case static) input data that can influence transition probabilities, is the underlying probabilistic graphical model. In DREM, protein-DNA interactions serve as the static input data that influence transitions between hidden states. An L1-regularized logistic regression classifier is trained at all expression profile bifurcations to assign transition probabilities to genes based on the set of TFs that bind them. DREM searches the state space of possible splits in gene expression profiles to predict a compact set of diverging regulatory paths and the TFs that control them. It was successfully applied to reconstruct networks in a large number of species including yeast [78], Escherichia coli [79], fly [222], and human [103].

2.2 SDREM

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 2.2), 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 [94] for complete details.

2.3 MT-SDREM

MT-SDREM simultaneously investigates and infers regulatory networks and signaling pathways for several biologically related conditions. The relation could be for example, in terms of overlap in terms of biological processes governing the conditions, or similarities in gene expression profiles, or similarity in phenotypes, or a combination of all those. 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.

2.3.1 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.

2.3.2 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.

2.3.3 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:

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

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 [93].

2.3.4 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 [91, 94] 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 [93] 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.

2.4 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 $s(n)$.

2.5 Optimizing the MT-SDREM objective

Using α we maximize the following objective:

$$\rho(B^1, ..., B^K) \propto \sum_{c \in C} \sum_{t \in T_c} \sum_{p \in P_t^c} n_t^{\alpha} \cdot I(p) \cdot h_p \cdot s_{tc}$$

Here n_t is the number of conditions the TF t is predicted to regulate, where T_c is the list of TFs predicted to regulate the time series for condition c, P_t^c is the set of paths that start from a source of condition c 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, s_{tc} is the score of the TF t for condition c, and $I(p)$ 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).

For $\alpha \geq 1$ the objective above would prefer selecting joint TFs to equally explanatory TFs that are not shared. Thus α represents a trade-off between fitting individual networks (specifically, $\alpha = 1$ means that we are back to our condition independent network learning) and learning a single joint network (very high values of α will lead to the selection of the same TFs for *all* networks). Note that the n_t^{α} factor implements the ρ function (for regularizing between tasks). The procedure to select an appropriate value for α is described later.

We use a greedy algorithm to optimize the objective. We randomly select a direction for every edge that has conflicting direction, i.e. it is present in opposite directions in two different pathways. We then do a local search to arrive at a local minimum. We flip the directions of the conflicting edges, always choosing the flip that increases our objective by the highest amount until we cannot find any flip that would still improve the objective. This approach is similar to SDREM's approach which has been shown to work well, both on real and on simulated data [93].

After we optimize the above objective, we obtain a single oriented network for each condition. We then use that network to obtain new priors for TFs for DREM. First we compute the weights for the TF for each condition using the equation

$$w_t^c = \sum_{p \in P_t^c} I(p) \cdot h_p \cdot s_t$$

where t is the TF and P_t^c is the set of selected paths for condition c that end in TF t. To normalize these scores, we further run the above orientation procedure L number of times, each time with an *additional* set of randomly selected TFs which are *not* predicted to regulate *any* of the conditions. We use the random score to adjust the score for the predicted TF [93].

2.6 Detailed description of the algorithm

We constructed a probability for each protein-protein interaction (ppi) using the formula $1 - \Pi_{i=1}^{n}(1 - p_i)$ where p_i is our confidence that the ith experimental evidence for the ppi is a true positive. The network so constructed has in the HGNC symbol naming scheme, 228,159 edges and 16,671 proteins with maximum degree 9,368 and average degree 27.372.

1. First, we pre-process the data. Let $S = \cup_{c \in C} S_c$ where S_c is the set of sources for condition c. We exhaustively search for all simple (non-cylic) paths from all sources in S to all TFs in our protein interaction network. The weight of a path is defined as the multiplication of the probabilities of the edges in the path (actual calculation done by summation in log space). We select and keep the top k paths by weight. Denote this set of paths P and let the weight of path p be h_p.
2. We run DREM (Dynamic Regulatory Events Miner) on the time series data for every condition individually with default prior $\pi_{tg} = 0.5$ for all TF-gene interactions. DREM is based on an input output Hidden Markov model and annotates the various time points of the time series with TFs that are supposed to be regulating the genes at those time points.
3. We extract a list of TFs from the results output by DREM and assign them a score in the same manner as done in SDREM. Let T_c be the set of TFs for condition c and let s_{tc} be the score of TF t in condition c. Let n_t be the number of conditions TF t is present in. Let $T = \cup_c T_c$ and T_{all} be the set of all possible TFs.
   In addition, the score for a TF is increased by multiplication with the n_t^α factor discussed in the previous section. As mentioned before, this factor is how we implement the ρ function of the objective.
4. We create the TF sets $T^i = T \cup T_r$ where T_r^i is a randomly selected set of TFs from $T_{all} \backslash T$ such that $|T_r| = |T|$. We create L such sets. In addition we also create the set $T_r^{L+1} = \emptyset$ and thus $T^{L+1} = T$.
5. For every TF $t \in T^i, 1 \leq i \leq L+1$, we then compute $f_{ti} = \sum_c \sum_p h_p \cdot s_{tc}$ where the path p ends in TF t and p has edges that are of the same orientation as those reached in optimization problem i.
6. We then create a list M consisting of all the f_{ti} so computed and sort that list according to the f_{ti}.
7. Then for a TF $t \in T$, if $f_{t,L+1}$ is in the 80th percentile of list M, we increase its prior via the formula $\pi_{tg}^{new} = (\pi_{tg} + 1)/2$ for all genes g. In addition, if $f_{t,L+1}$ is greater than the node threshold parameter, the prior is also increased similarly.
   If neither of the two conditions hold, we decrease the prior via the formula $\pi_{tg}^{new} = \max(0.01, \pi_{tg}/2)$
8. We run steps 2-7 for 10 iterations which is the default number of iterations of SDREM.

2.7 Learning parameters for the multi-task objective

For handling parameters for the SDREM component, we refer the reader to [93]. We use the default provided parameters for SDREM inside of MT-SDREM. MT-SDREM adds the α parameter to the set of parameters. α encodes our prior on how much the given conditions are related.

One way to choose α is to perform cross validation on say the number of RNAi hits one obtains in the top k ranked genes. Another, which is what we use here, is to look for an α which is in a stable region – i.e. perturbing it would not change the results in terms of the TFs that we extract. We found that $\alpha = 6$ achieves such stability and we thus use that for all our experiments. Later on in this chapter, we also explore using a *pairwise* similarity parameter between each pair of tasks.

2.8 Constructing the joint signaling network

To construct the joint signaling network in Figure 1 of the main text, we took the top 200 source, intermediate, and target (TFs predicted to regulated the condition's gene expression) proteins from each of the 3 conditions. We then looked at the set of paths from a source protein to an intermediate protein to a target protein and computing the condition-specific path flow for each protein (see Materials and Methods for details on how to compute it given a set of paths). We also computed the path flow for each edge between the selected proteins. Edge path flow was computed in a similar manner to node path flow by summing path scores for all paths containing that edge. We then only selected nodes which had a flow of at least 1000 and edges which had a flow of at least 200.

2.9 Results on Influenza data

MT-SDREM simultaneously infers signaling and dynamic regulatory networks for multiple related conditions. It extends the SDREM tool [91, 94] which discovers signaling pathways by orienting edges in protein interaction networks. To demonstrate the performance of such multitask network learning we looked at data from 3 different flu viruses: H1N1, H3N2, and H5N1.

For each of these viruses we obtained time series gene expression measurements of cells infected with the virus. For H1N1 the data is from [233] and contains 10 time points. For H5N1, we obtained data from [158] with 5 time points, and the H3N2 data from [124] had 6 time points. In addition, for each of these viruses we obtained a set of sources (host proteins interacting with the virus proteins) from mass spec experiments. Data for H1N1 is from [189] and literature [233, 251] and contains 200 human proteins that were experimentally determined to interact with H1N1 proteins. Data for H3N2 is from [189] and consists of 153 host proteins and source data for H5N1 is from [189] and literature [54, 123, 165, 235, 251, 276] and consists of 41 sources.

2.9.1 MT-SDREM reconstructed networks

Joint signaling network

Figure 2.1 presents the joint signaling network learned for the three conditions (Methods). The top layer (nodes colored in red) are sources for at least one of the conditions. The bottom layer (nodes colored green) are TFs identified in at least one of the conditions, and the middle layer (blue nodes) are signaling proteins linking the sources and TFs in the networks. We colored each node with multiple colors depending on the condition for which it was identified as a top network protein (Methods). The lightest shade for each color represent nodes from the H1N1 reconstructed network, the darkest is from the H5N1 network and the middle shade is for the H3N2 network.

While sources (red shades) are provided as inputs, all other nodes were automatically identified by MT-SDREM. Several of the proteins identified in multiple networks, both as intermediate and as TFs are well known immune response regulators. For example, we identify a pathway from UBE2I (a source for both H1N1 and H3N2) to SUMO1 (signaling protein identified for all strains). SUMO activates E1 and transfers it to conjugating enzyme E2. Then UBE2I interacts with and transfers SUMO to a target viral protein. Indeed, it has been recently shown that SUMO interacts with the key flu protein, NS1, via UBE2I [288]. In addition, TRAF6, part of the TRAF (TNF receptor associated factor) family of proteins, is identified as an important protein for H5N1. Pro-inflammatory cytokines including TNF-α are known to be hyper-induced in H5N1 infected macrophages [49].

We also identify several TFs as common amongst the 3 conditions. SMAD4 is present in all 3 conditions. The SMAD family of TFs is part of the TGFβ pathway which is responsible for regulating macrophage activation and proliferation of T cells [168]. STAT1 and JUN, both key immune response regulators, are also identified in all 3 conditions. We also identify NR3C1 which produces the GR protein that is known to inhibit T and B cells as well as suppressing phagocyte function [51] (this could be a viral strategy to reduce the effects of immune response). Interestingly, we identify the AKT1 gene in all 3 conditions, part of the PI3K/AKT pathway, which has recently been shown to be activated by the influenza A virus's NS1 protein [75]. We also identify the PPARG TF which has been linked to immune response by regulation of immune and inflammation related genes [250]. Other TFs belonging to the AP-1 TF complex are also identified for various conditions – ATF2 for H1N1 and H5N1, and FOSL2 for H1N1 and H3N2. NFKB1 and RELA, both part of the NF-κB complex are identified for H1N1 and H5N1 respectively.

Regulatory networks

In addition to the signaling parts of the networks, MT-SDREM also reconstructs dynamic regulatory networks for each of the different flu strains. We show the regulatory network inferred for H1N1 in Figure 2.2. For space reasons not all TFs presented in Figure 2.1 are shown for the model in Figure 2.2, though all TFs that are associated with H1N1 are used by the model. Full list of TF assignments to paths in the regulatory networks is available on the Supporting Website [1]. Corresponding networks for H3N2 and H5N1 are in Figures 2.3 and 2.4. Several of the TFs identified as controlling the first splits in both the H1N1 and H3N2 networks belong to the IRF family of TFs, known to regulate interferons, which play an important role in viral immune response [168]. TFs belonging to the FOS, ATF, and JUN families appear in both the H1N1 and H5N1 networks. These TFs are part of the AP-1 TF complex (which is known to regulate gene expression in response to a variety of stimuli including cytokines, and viral infections [116]). We also identify the SMAD family of TFs to play a part in all 3 networks. The STAT family of TFs

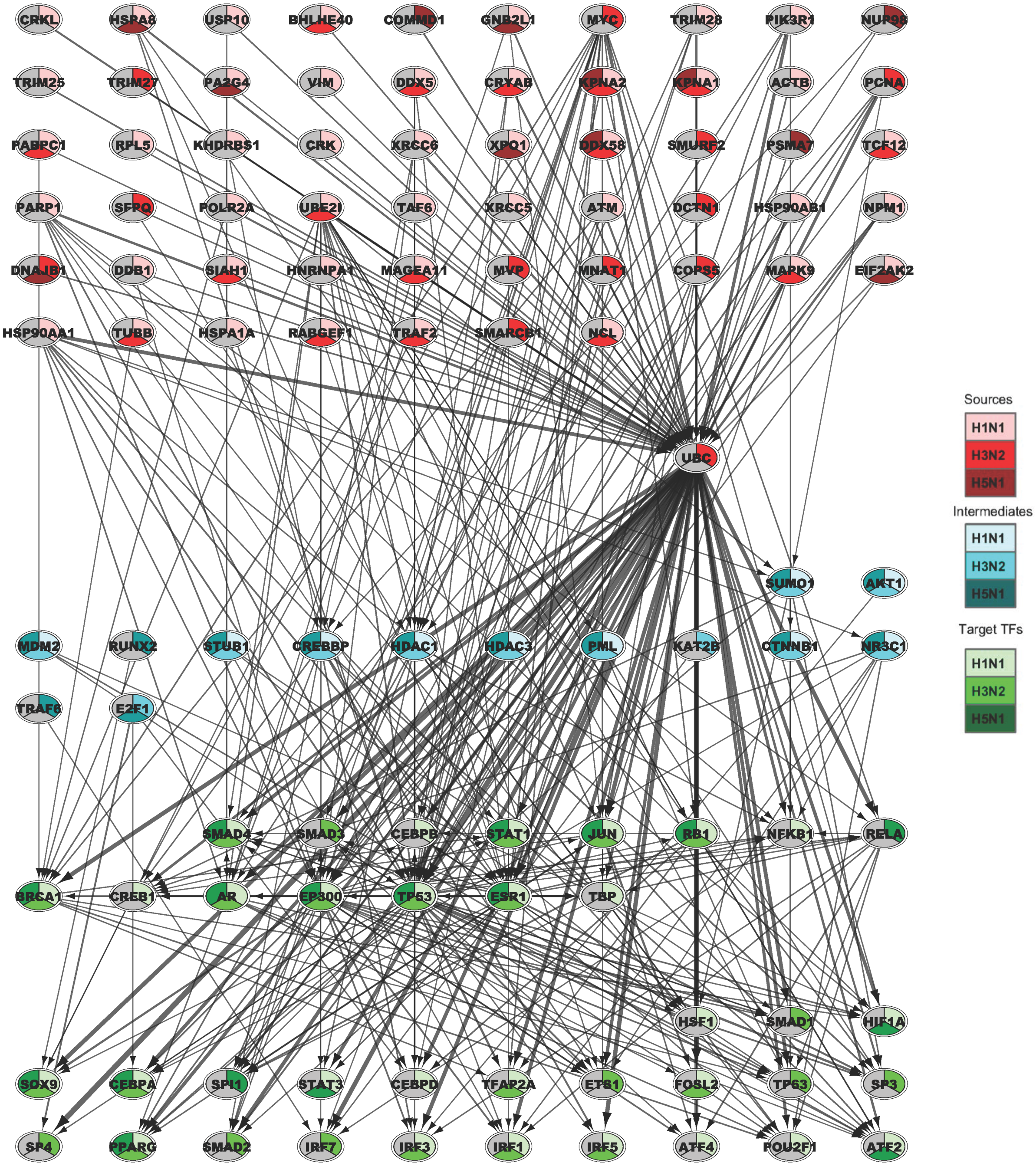

Figure 2.1: **Joint signaling network inferred by MT-SDREM for the three flu viruses.** Top: Sources. Middle: Signaling (intermediate) proteins. Bottom: TFs. Nodes are colored according to the role the protein is determined to play in the pathway (red - source, blue - signaling, green -TF). Each node is also denoted with the set of strains it was predicted for (color shades). For example, JUN is a TF predicted for all three strains whereas TCF12 is identified as a source for H1N1 and H3N2 but not for H5N1.

is found to play a role in all 3 conditions. This family of TFs is part of the JAK-STAT signaling pathway. This is a class of pathways responsible for activating transcription in response to extracellular signals from messengers such as interferons, interleukins, growth factors, etc. [7, 272].

In addition to analyzing the TFs identified we performed an enrichment analysis using the Gene Ontology (GO) terms associated with each path in the reconstructed regulatory networks.

All p-values that we give below are after correcting for multiple hypothesis testing.

In the H1N1 regulatory network, the gene cluster corresponding to the path labeled **A** is predicted to be regulated by STAT1, part of the JAK-STAT signaling pathway, IRF1 and IRF2. This path is enriched for 'defense response to virus', 'immune response', 'type I interferon signaling pathway', and 'cytokine-mediated signaling' categories (p-value of <0.001 for both). We also find enrichment for similar categories in paths labeled **B**, **C** of the H1N1 network and the paths labeled **D-H** of the H3N2 network (Figure S1). In addition, we also find enrichment for 'toll-like receptor signaling pathway' in path **F**, and 'T cell activation' and 'lymphocyte activation' in path **H** (p-value of < 0.001). Path **D** is also predicted to be regulated by several members of the IRF family.

We find enrichment for the more general categories of 'defense response' and 'immune response' in the path labeled **I** of the H5N1 network (Figure S2, p-value of < 0.001). Notably, in all 3 conditions, the genes in the relevant paths are being upregulated indicating a response to all three pathogens that has shared features.

The complete list of GO categories for all the labeled paths can be found on the Supporting Website [1].

Strain specific proteins

In addition to looking for common response, we used MT-SDREM to identify strain-specific factors and proteins. These represent potential targets for individual strains and may explain why some are more virulent than others. Table 2.1 presents the set of unique proteins identified for each strain (defined as those appearing in the top 100 proteins set for that strain, but not in the top 100 of the other two). While many of the proteins on the list are not well characterized in the three conditions making it hard to validate the results, some are known and the results agree with very recent experimental data. For example, IRF7 which was only identified by MT-SDREM for H3N2 was recently tested for H5N1 and shown to be significantly lower in H5N1 response when compared to less virulent strains [271]. Similarly, as mentioned above, the regulatory networks for H1N1 and H3N2 contain several IRFs as key regulators while the networks reconstructed for H5N1 does not pointing to a potential target for improving prognosis from this infection.

Several proteins that are only predicted for H5N1 response are known to have important roles in H5N1 infection. Knockdown of DDX39B, also known as UAP56, decreased H5N1 viral titre nearly 10 fold in infected cells [22]. MAPK8 (JNK) was strongly induced in H5N1 (and H3N2) infection, but not H1N1 infection [88]. NUP98 recruits the H5N1 protein NS2 to the nucleoli, and disrupting this interaction impedes viral propagation [54]. Mice with wild type MX1 were protected against infection by a highly lethal H5N1 strain relative to mice with defective MX1 [263]. H5N1-derived NS1 stimulates the ERK pathway, increasing cell viability and promoting infection [180]. Through interactions with viral NS1 and another host factor, IVNS1ABP (NS1-BP) can counteract this NS1-induced ERK phosphorylation [180].

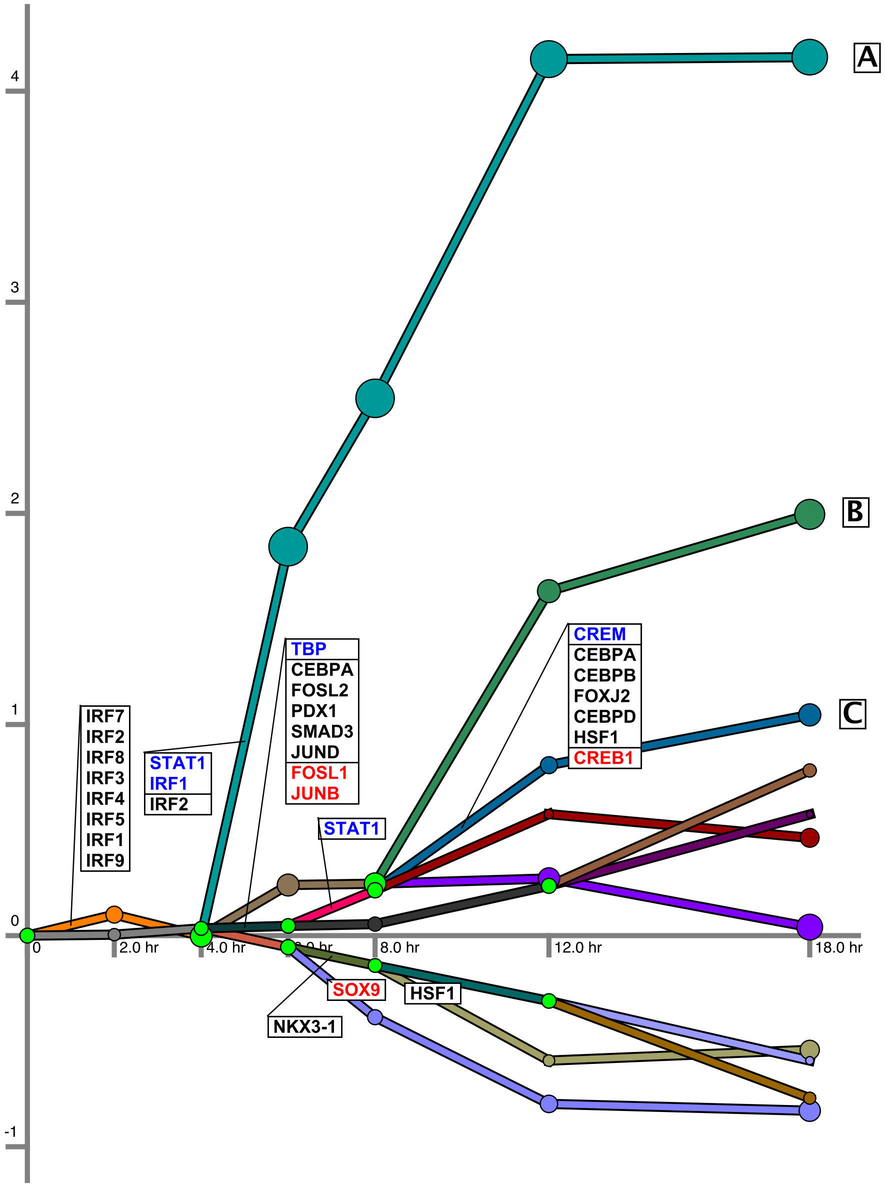

Figure 2.2: **H1N1 Regulatory network.** Each path represents a set of genes with a similar expression profile. Split nodes are colored green and are annotated with the TFs that are predicted to regulate genes in the paths going out of the split at the time point associated with the split. The blue TFs 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.

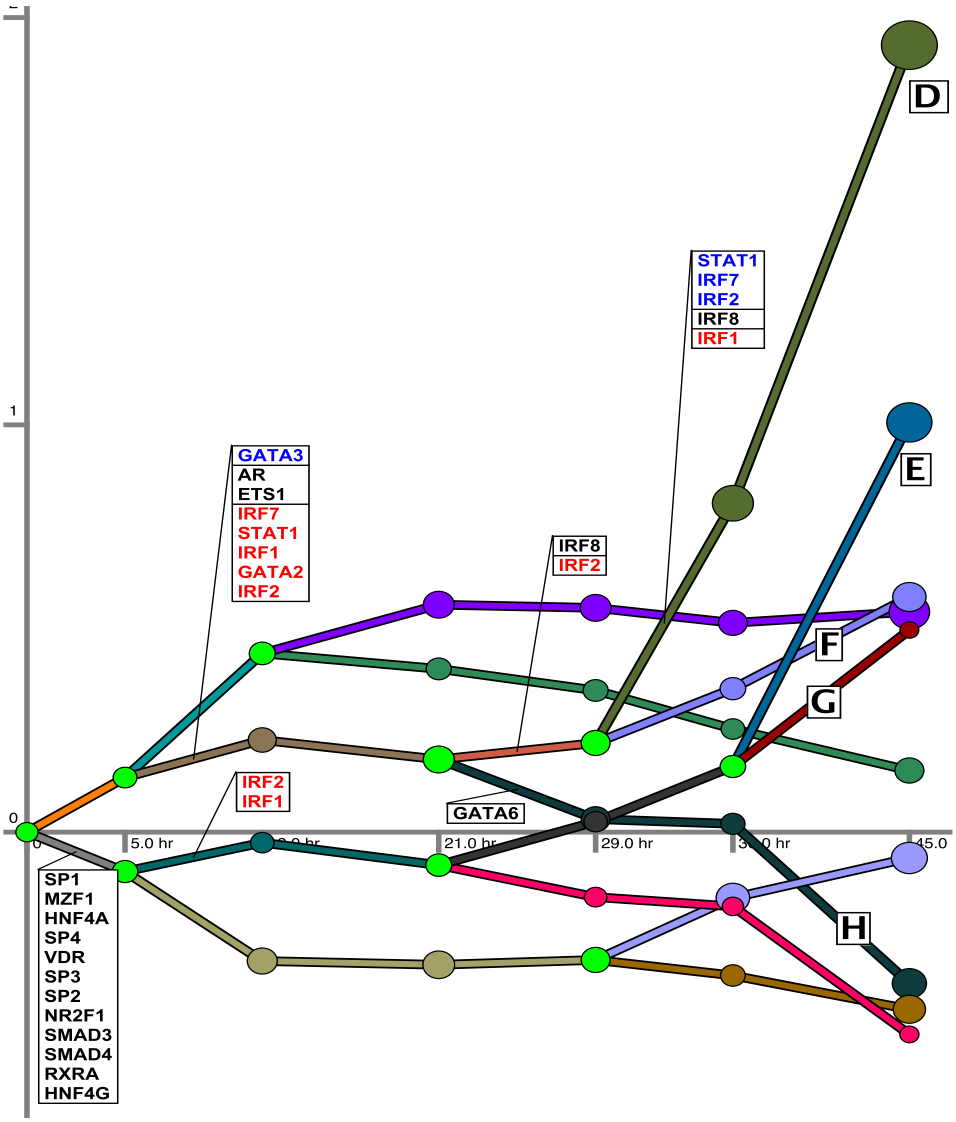

Figure 2.3: **The H3N2 regulatory network is presented.** The paths represent the different gene sets which are coexpressed. The TF lists are the TFs predicted to regulate the path that they are connected to.

2.10 Comparison of MT-SDREM with prior work

To test the advantages of multi-task learning we compared MT-SDREM with previous methods that can be used to analyze expression and interaction data. Since we are not aware of prior methods that utilize multi-task learning in biological network reconstruction we first looked at the differences between applying MT-SDREM and applying SDREM separately to each of the three flu datasets. We have also compared MT-SDREM's results to a baseline *joint* ranking of differentially expressed (DE) genes from different experiments in a single analysis. This

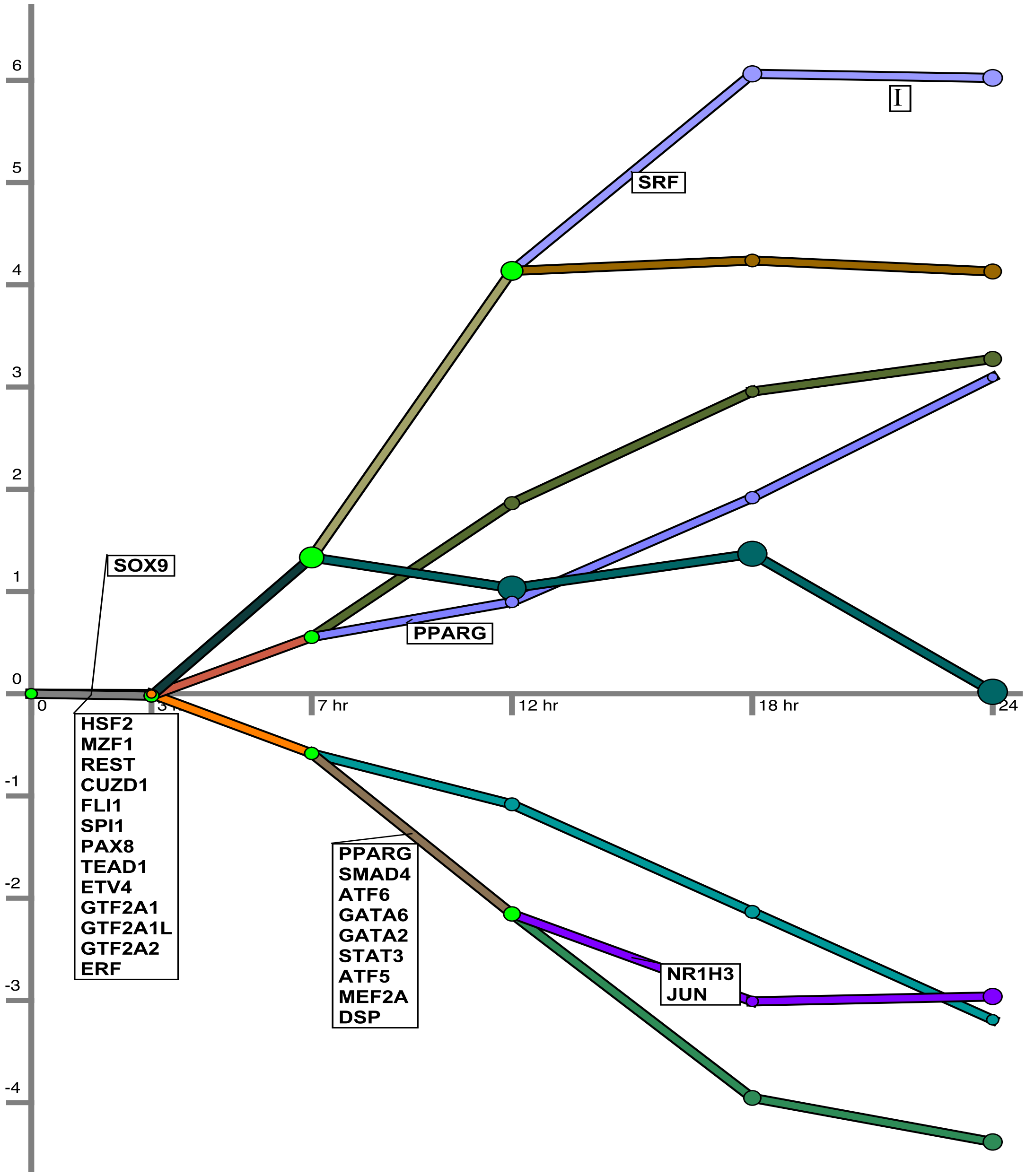

Figure 2.4: **The H5N1 regulatory network is presented.** The paths represent the different gene sets which are coexpressed. The TF lists are the TFs predicted to regulate the path that they are connected to.

approach is similar to several previous studies that perform follow up analysis using such joint sets [11].

Since the 'ground truth' (complete underlying networks for each condition) is obviously unknown, we used three different types of complementary information for these comparisons. First, we examined the set of TFs identified by each of these methods and determined their relevance to the condition being studied. Next, we used the Gene Ontology (GO) to test the differences in the identified functional categories between the different analysis methods. Expression experiments and RNA interference (RNAi) screens have revealed a multitude of host pathways and processes that are involved in viral host response including MAPK signaling, apoptosis, trafficking, mRNA export, splicing, and proteolysis [39, 144, 233]. A statistical meta-analysis implicates nearly 3000 host genes [110] in these pathways. Although many processes as a whole are relevant to influenza response, not all genes participating in those processes necessarily are important. Therefore we focused our TF and GO evaluation on immune processes, which were shown to compose a critical component of the host response that kills infected cells, protects uninfected cells, combats viral components, and promotes inflammation [253].

Finally, we used a set of RNAi experiments that were performed for H1N1 and H5N1 to test the ability of these different methods to identify key disease related proteins. In these experiments proteins are knocked down one at a time and the impact on viral load is measured. A protein affecting viral load is likely participating in the host response and so methods that can identify such proteins more accurately are in better agreement with the observed response. The RNAi data for H1N1 was obtained from [36, 39, 132, 144, 233] resulting in a total of 980 screen hits, 925 of which were present in our initial interaction network (which contained 16671 genes, Methods). 32 screen hits for H5N1 were obtained from [36], all of which are present in our interaction network.

2.10.1 Comparison of identified TFs

In Table 2.2 we present the overall and pairwise overlap of the inferred TFs for the 3 conditions (extracted by same mechanism as in SDREM [91, 94]) for MT-SDREM and compare it to when SDREM is run independently on the 3 conditions (I-SDREM). Note that the pairwise intersections shown are *in addition* to the overall intersection between all of the 3 conditions.

The shared TFs identified by MT-SDREM among all 3 conditions that are missed by I-SDREM include several that are known to be immune response related. In particular, CEBPA is known to be responsible for regulating a large variety of cell functions including immune and inflammatory response [204]. MT-SDREM also identifies SMAD4 in all three conditions. SMAD family proteins are part of the TGFβ pathway as mentioned above. MT-SDREM also identifies RB1 which has been implicated in viral immune response [182], JUN which is part of the AP-1 TF complex, and PPARG an important TF regulating immune response mentioned above. In contrast, I-SDREM does not identify any TF in the intersection that MT-SDREM does not.

In addition, we also find several immune response related TFs in the pairwise overlaps for MT-SDREM that we do not see for I-SDREM. For the overlap between H1N1 and H3N2, MT-SDREM identifies IRF1/3/5 which are known to regulate interferons and thus important for immune response. For the overlap between H1N1 and H5N1, MT-SDREM finds the the STAT3 gene which is part of the JAK-STAT signaling pathway and ATF2, part of the AP-1 TF complex.

For the pairwise intersection of H1N2 and H3N2, I-SDREM identifies NR3C1 as a TF while MT-SDREM only selects it as an intermediate (signaling) protein. It also identifies another member of the SMAD family (SMAD3 whereas MT-SDREM identifies SMAD4). For H3N2 and H5N1 it identifies AHR whose activation inhibits inflammation [161] and RELA in the intersection of H1N1 and H5N1, which as part of the NF-κB complex.

We also compared MT-SDREM to the popular TF prediction tool oPossum [151]. Our primary goal when comparing MT-SDREM with oPossum is to highlight the fact that using network information in the multi-task learning framework is useful. The input to oPossum is a list of genes identified by the experiment(s) and using this list it attempts to find overrepresented TF-binding sites. To select a common gene list from all three experiments we ranked the genes for each condition according to their differential expression and then merged the 3 rankings using the Kemeny-Young method [292]. Similar to the number of genes used by MT-SDREM we used the top 3000 in the joint ranking as input to oPossum. In Table 2.3 we present the comparison. Note that since we used oPossum as the tool for the comparison of MT-SDREM with other methods for integrating data from several conditions, the results shown for Table 2.3 are different from the intersection results of Table 2.2. Here, for the MT-SDREM rankings we used the *sum* of % path flow going through each gene across the 3 networks to rank TFs (Methods). The oPossum TFs are ranked according to their Z-score.

While oPossum is able to identify a few relevant TFs, for most of the TFs identified by oPossum, we could not find significant roles in immune response regulation for them. In contrast, several of the shared MT-SDREM TFs that are not identified by oPossum are known to play major roles in immune response as discussed above. These include STAT1/3, JUN/ATF2, CEBPA/B which regulate a large number of immune response genes, RB1 which has been implicated in viral immune response networks [182], PPARG, and SMAD. MT-SDREM also uniquely identifies IRF1 which plays a major role in viral immune response by regulating interferons. oPossum was able to identify only two relevant TFs that were not found by MT-SDREM. These are ZEB1 which regulates the IL2 interleukin, part of the immune response system and AHR, part of the ANTR-AHR complex.

We also tried to compare MT-SDREM with the Inferelator method [34] but following email discussions with the authors of that method determined that such comparison is not feasible since Inferelator requires expression data for a large number of conditions while we only had time series response for three types of infections.

2.10.2 RNAi screen hits

Using the screen hit data for H1N1 and H5N1 we compared the performance of MT-SDREM, I-SDREM and Endeavour [8, 259]. 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. 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 2.5. 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).

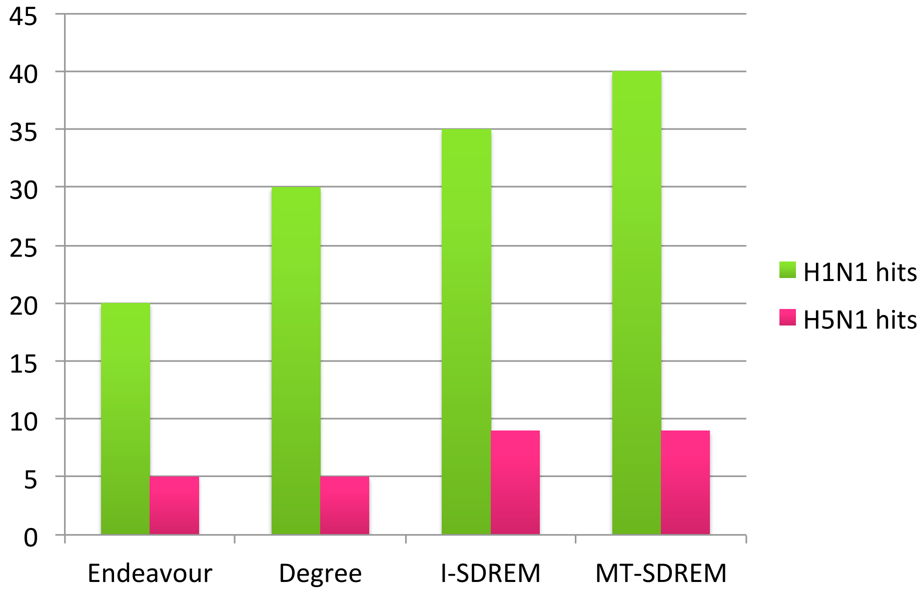

Figure 2.5: **Screen hits overlap for top 100 ranked genes for both H1N1 and H5N1.** 925 H1N1 and 32 H5N1 screen hit proteins were present in our network.

2.10.3 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 [28, 29] 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 2.4. 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 2.5. Note that we find significant enrichment for a 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.

2.11 Learning task-relatedness parameters

So far we have only had a single task relatedness parameter α for each pair of tasks. One interesting avenue to explore would be different α parameters for each pair of tasks. As we had 3 tasks (the 3 flu viruses), the parameters we had to set were $\alpha_{12}, \alpha_{23}, \alpha_{13}$. We computed the overlap of differentially expressed genes between the differentially expressed genes for each pair of the 3 strains of flu. The overlap was $862/11569 = 0.074$ genes for H1N1 and H3N2, $458/11463 = 0.04$ for H3N2 and H5N1, and $645/12549 = 0.051$ for H1N1 and H5N1. As $\alpha = 6$ worked well for us in past experiments, we normalized the overlap fraction values such that the maximum value (in this case 0.074 was 6) and used those as our α values. Thus we got $\alpha_{12} = 6, \alpha_{23} = 3.2, \alpha_{13} = 4.1$. Rerunning MT-SDREM with these parameters, we looked at the enrichment of RNAi screen hits for H1N1 and H5N1 in the top 100 proteins in our new ranking. We found 39 of the RNAi screen hits for H1N1 and 9 for H5N1. This is in contrast to 40 for H1N1 and 9 for H5N1 when using a single α value of 6 for all three conditions.

While we did not improve on our existing results, the above evaluation does suggest that using overlap of differentially expressed genes as a measure of relatedness between the different tasks may be a good way to select the α parameters.

2.12 Conclusion

In this chapter we presented a novel method to jointly reconstruct the signaling and regulatory networks of multiple related conditions in a multitask fashion. We gave extensive anecdotal and statistical evidence that our method could recover known biology of 3 different types of flu viruses – H1N1, H3N2, and H5N1. We also gave statistical evidence based on enrichment of RNAi screen hits as well as GO analysis that our method outperformed existing methods like oPossum, Endeavour. Importantly we also showed that doing a joint reconstruction of the networks of the three viruses improved the network quality compared to doing it individually.

H1N1	**H3N2**	**H5N1**
CREB1	PCNA	CASP8
PARP1	COPS5	ERBB3
XRCC6	IRF7	COMMD1
POLR2A	SMAD1	PSMA7
CEBPD	ETS1	SPI1
TRIM28	SP3	NUP98
ATM	SMARCB1	HNRNPF
ACTB	TP63	KPNA6
XRCC5	SFPQ	EIF4G1
HSF1	BCR	PCBP1
EEF1A1	SMURF2	DDX39B
ATF4	TRIM27	STAU1
KHDRBS1	ANXA1	IPO5
HNRNPA1	DCTN1	PABPN1
HSP90AB1	CHAF1A	TLR3
DDB1	DVL3	HSPA4
POU2F1	KAT2B	GTF3C3
CRKL	DDX3X	MX1
CRK	RPS3A	GLUL
RPL5	RABGEF1	CCND1
RUVBL2	AIMP2	NQO2
RPL11	SP4	CDKN1B
CDC42	HNRNPH1	TLR8
MCM7	PSMD8	MAPK8
DDX17	MAGEA11	NOMO2
VIM	MLH1	CDKN1A
EWSR1	GSK3B	SIRT1
RPS7	NCOR1	RUNX2
	RBM14	FOS
	PIN1	IVNS1ABP
	NMI	SNAPC4
		ERBB2
		NCOA3
		TRAF6
		TP73
		CASP3
		PRKDC
		CDK1

Table 2.1: **Strain-specific protein list**

Table 2.2: **TF comparison for I-SDREM and MT-SDREM**

H1N1 & H3N2 & H5N1		**H1N1 & H3N2**		**H3N2 & H5N1**		**H1N1 & H5N1**	
I	**MT**	**I**	**MT**	**I**	**MT**	**I**	**MT**
AR	AR	IRF1	IRF1	AHR◇		EP300	ATF2†
BRCA1	BRCA1	IRF3	IRF3	JUN		RELA◇	HIF1A†
ESR1	ESR1	FOSL2	FOSL2	PPARG		TP53	STAT3†
STAT1	STAT1	CEPBA	IRF5†	RB1			
	CEBPA†	NR3C1◇	TFAP2A†	SMAD4			
	EP300†	SMAD3◇		SOX9			
	JUN†						
	PPARG†						
	RB1†						
	SMAD4†						
	SOX9†						
	TP53†						

TFs predicted to regulate two or all three response networks. Each set of conditions is divided to two columns with the first column containing TFs at the intersection of the SDREM output for the conditions and the second the MT-SDREM results for these conditions. TFs identified by MT-SDREM but not SDREM have a † next to them and vice versa have a ◇ next to them. Note that TFs listed for the pairwise overlap are **in addition** to the ones listed for the overall overlap. Thus JUN in the I-SDREM column of H3N2 & H5N1 is not highlighted since it was identified by MT-SDREM for all three conditions.

Table 2.3: **TF comparison for oPossum and MT-SDREM**

oPossum	**MT-SDREM**
MZF1_1-4	EP300
SP1	TP53
ZNF354C	BRCA1
MZF1_5-13	JUN$\star$
NFYA	ESR1
ZEB1$\star$	AR
MIZF	RB1$\star$
ROAZ	SMAD4$\star$
GABPA	STAT1$\star$
TEAD1	CEBPA$\star$
TLX1-NFIC	PPARG$\star$
SPIB	STAT3$\star$
Hand1-Tcfe2a	SMAD3$\star$
ARNT-AHR$\star$	HIF1A
ELF5	RELA$\star$
MYC-MAX	MYC
TP53	ATF2$\star$
ELK1	CEBPB$\star$
REL$\star$	SOX9
AR	IRF1$\star$

oPossum and MT-SDREM comparison. Immune response related TFs have a $\star$ next to them. oPossum TFs are ranked according to their Z-score. MT-SDREM TFs are ranked according to the path flow measure as described in the text

Table 2.4: **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 [28]. 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

Table 2.5: **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 [28]. Only categories with MT-SDREM adjusted p-value of $\leq$ 0.001 and DE genes p-value of $\geq$ 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

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 [200] 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 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.

3.1.1 Cell culture, HIV infection, and Reagents

Sup-T1 cell lines were obtained through the NIH AIDS Research and Reference Reagent Program, Division of AIDS, NIAID, NIH (A Sup-T1 from Dr. James Hoxie [242] and were maintained in RPMI containing 10% FBS, 1% l-glutamine and 1% penicillinstreptomycin (GIBCO). HIV-wt-EGFP reporter virus was obtained by transfecting HEK293 T cells (2×10^6 per plate) with 10 μg of HIV-1 vpr(+)/EGFP proviral construct by Polyjet following manufacturers protocol. Forty-eight hours post transfection, the supernatants were collected, filtered through a 0.4-μm filter to remove cellular debris, and centrifuged at 22,000 rpm for 1 h. The virus pellets were resuspended in PBS and stored in aliquots at 80 C for subsequent assays. Multiplicity of infection (MOI) for virus was calculated by TZM blue assay using the HIV-1 reporter cell line cMAGI (AIDS Research and Reference Reagent Program [RRRP], National Institutes of Health [NIH]). The Sup-T1 cells were infected at a MOI of 0.3 either in the presence or absence of specific inhibitor at indicated time points. Forty hours post infection, the cells were washed and fixed with 1% paraformaldehyde and the samples were analyzed using Fortessa (BD Biosciences) with 10,000 gated events acquired for each sample, and the results were analyzed using FlowJo software (Tree Star, Inc., OR). The infected cells were detected by the expression of reporter virus EGFP. Azidothymidine (AZT) obtained from Sigma-Aldrich was used as positive control. IKK2 inhibitor V, Dasatinib, and Dinaciclib were obtained from CalBiochem. SP600125 and WP 1066 were obtained from Abcam Biochemicals and Enzo, respectively. SNS-032, Regorafenib, Carfilzomib, and Veliparib, Olaparib were obtained from selleckchem.com. SAHA and 5-Azacytidine were obtained from Sigma-Aldrich. The viability of cells was estimated by Trypan blue staining. We conducted the experiments 3 times with duplicate wells for each experiment.

3.1.2 Data description

The overall goal of TimePath is to determine the dynamics of both the signaling and the regulatory events that take place as part of a cellular response process. For this, TimePath integrates time series gene expression data, static protein interaction data (both within and across species) and protein-DNA interaction data. We constructed a weighted, partially directed, protein interaction network using several databases including BIOGRID [245], HPRD [206] and have also used Post-translational Modification Annotations from the HPRD. Protein-DNA interactions are based on data from [230]. Sources (host proteins that interact with the HIV-1 proteins) were obtained from VirHostNet [189]. Time series gene expression and miRNA expression data following HIV-1 infection in Sup-T1 was obtained from [181]. Differentially expressed genes were computed using DESeq [16] and ranked using the p-value generated by the package for the differential expession with smallest p-value first.

3.1.3 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 (Introduction). 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) [289].
   (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 [19]. We thus use a heuristic to compute the set of paths.

3.1.4 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 is 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.

3.1.5 Detailed path search algorithm description

Algorithm 1 `search-paths`$(n, q, S, T_1, \dots, T_k, E)$

```
$P \leftarrow \{\}$
$Q \leftarrow S$ with priority 0
while $|P| < n$ do
    $p \leftarrow$ pop($Q$) where $p$ has the lowest priority
    end-gene $\leftarrow$ end($p$)
    end-gene-nbrs $\leftarrow \{(t, score(e)) : e = (s, t) \in E \wedge$ end-gene $= s\}$
    if end-gene$\in T_i$ then
        if $i = 0$ or $\exists g \in p \setminus \{$end-gene$\}, g \in T_j, j < i$ then
            $P \leftarrow P \cup \{p\}$
        end if
    end if
    for $(g, sc) \in$end-gene-nbrs do
        insert $[p\ g]$ into $Q$ with priority $priority(p) + (-log(sc))$
        if $|Q| > q$ then
            remove highest priority element from $Q$
        end if
    end for
end while
return $P$
```

The algorithm takes as input the following arguments –

- n which is the final number of paths to output. In our case $n = 10$ million.
- q which is the maximum number of elements the breadth first queue is allowed to hold.
- S which is the set of sources.
- $T_1, \dots, T_k$ which is the set of targets for phase 1 upto the phase k, the total number of phases.

- Q which is the set of edges.

In line 1, we initialize the set of paths P to an empty set and fill the priority queue Q with the sources all with the same priority 0. P contains the set of final paths we have selected and Q contains the candidate paths.

Then until we have filled the set of paths P with the target number of paths n, we do the following. In line 4, we pop off the path p with the lowest priority (and thus the highest scoring path) from Q. In line 5, we extract the current last gene `end-gene` in p and in line 6, we extract the set of neighbor genes of `end-gene` using the set of edges E. Then if the `end-gene` is a target, that means we already have a path from a source to a target and we may be able to add path p to the final set of paths P. However we still have to ensure the constraint that any pathway to a target in phase i has to have a protein that was a target in phase $i - 1$ unless $i = 0$. This condition is checked for in line 8 and if satisfied, we add p to P.

In line 12, then we iterate over all the neighbor genes of `end-gene`. We extend path p with the every neighbor gene g in line 13. The score of the new path is the score of p and the negative logarithm of the edge score between `end-gene` and g. Then if the queue exceeds the maximum size allowed, the highest priority element (and thus the lowest scoring path) is removed in line 15.

Finally we return the final set of paths in line 19.

3.1.6 Detailed IP algorithm description

Given the 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 $g1$ and $g2$ in phase k that are known to be bound by the same TF A. If we believe that A explains the activation of $g1$ in that phase it increases our belief that A is also the TF activating $g2$. 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.

Specifically, we try to balance three different criteria –

1. Maximize the total path weight of the selected paths (i.e. select the pathways we are most confident in).
2. Maximize the number of targets – i.e. the number of targets with at least one selected path ending at them.
3. Minimize the number of selected nodes in the network. A node is selected if there is at least one selected path that has uses node.

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 objective 3.6 :-

$$\max \sum_{p\in P} w(p)\cdot b_p^P + \lambda_1 \sum_{k=1}^{K} \sum_{g\in T_k} f_g - \lambda_2 \sum_{g\in G} b_g^G \tag{3.6}$$

with the constraints

$$\forall p \in P, \forall g \in p, b_g^G \geq b_p^P \tag{3.7}$$

$$\forall p \in P, \sum_{g\in P} b_g^G \leq |p| - 1 + b_p^P \tag{3.8}$$

$$\forall g \in G, \forall p \in P(:,g), f_g \geq b_p^P \tag{3.9}$$

$$\forall g \in G, \sum_{p\in P(:,g)} b_p^P \geq f_g \tag{3.10}$$

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.
- 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. each part of the objective and are the parameters that decide in the end, how large of a network in terms of number of genes and edges will be chosen.

Since standard solver cannot be used due to the scale of this problem, we solve this IP using a greedy approach. We run a local search to converge to a local minimum and then augment this with the tabu search heuristic to escape the local minimum. In the end we return the best solution found.

The algorithm is described in detail in Figures 2– 6.

We augment the local search described above with a metaheuristic called tabu search [97] to escape out of local minima. Briefly, we maintain a list called the tabu list of the past L solutions and ensure that the current solution is not the same as any of the previous L solutions. This helps us escape local minima in the following way – suppose we are at a local minimum with b_1^G being the binary vector of whether a gene is turned on or off. We then flip a random gene to escape the local minimum and arrive at the solution b_2^G. Now the solution b_1^G is part of the tabu list which prevents us from using it in future rollbacks allowing us to escape the current minimum.

Algorithm 2 filter-paths$(\mathbf{b}^P, \mathbf{b}^G, P, \mathbf{w}, \lambda_1, \lambda_2, \tau, L, N)$

```
$l \leftarrow []$
$\mathbf{b}^P \leftarrow [0]^{|P|}$
$\mathbf{b}^G \leftarrow [0]^{|G|}$
$\boldsymbol{\delta} \leftarrow$ compute-initial-delta
$O \leftarrow$ compute-initial-objective
$i \leftarrow 1$
while $\max(\boldsymbol{\delta}) > \tau \vee i \leq N$ do
    $i \leftarrow i + 1$
    $g \leftarrow \arg\max_i \delta_i s.t. f(g, b^G) \notin l$
    $O \leftarrow O + \delta_i$
    $T \leftarrow$ compute-target-genes-table$(P_g)$
    add-target-penalty$(\boldsymbol{\delta}, \mathbf{f}^G, \mathbf{f}, T, -\lambda_2)$
    if $b_g^G = 1$ then
        $b_g^G \leftarrow 0$
        $\delta_g \leftarrow \delta_g - 2\lambda_1$
        update-delta-on-deactivation($\boldsymbol{\delta}, g$)
        for $p \in P_g$ do
            $b_p^P \leftarrow 0$
        end for
    else
        $b_g^G \leftarrow 1$
        $\delta_g \leftarrow \delta_g + 2\lambda_1$
        update-delta-on-activation($\boldsymbol{\delta}, g$)
    end if
    add-target-penalty$(\boldsymbol{\delta}, \mathbf{f}^G, \mathbf{f}, T, \lambda_2)$
    if $|l| = L$ then
        remove first element of $l$
    end if
    append current $b^G$ to $l$
end while
```

Algorithm 3 `compute-target-genes-table`(P_g)

```
$T \leftarrow \{\}$
for $p \in P_g$ do
    $t \leftarrow \texttt{target}(p)$
    if $t \notin T$ then
        $T \leftarrow T \cup \{t : \{\}\}$
    end if
    for $g' \in p$ do
        $T[t] \leftarrow T[t] \cup \{g'\}$
    end for
end for
return $T$
```

Algorithm 4 `add-target-penalty`($\boldsymbol{\delta}, g, \mathbf{f}^G, \mathbf{f}, T, \lambda_2$)

```
for $t \in T$ do
    for $g' \in T[t]$ do
        if $f_t \geq 1$ then
            if $f_t + \Delta f_t^g < 1$ then
                $\delta_g \leftarrow \delta_g - \lambda_2$
            end if
        else
            if $f_t + \Delta f_t^g \geq 1$ then
                $\delta_g \leftarrow \delta_g + \lambda_2$
            end if
        end if
    end for
end for
```

Algorithm 5 `update-delta-on-activation`$(\boldsymbol{\delta}, g, \mathbf{f}^G, \mathbf{f})$

```
for $p \in P_g$ do
    $r \leftarrow |p| - \sum_{g' \in p} b^G_{g'}$
    $t \leftarrow \texttt{target}(p)$
    if $r = 0$ then
        for $g' \in p$ do
            $\delta_{g'} \leftarrow \delta_{g'} - w_p$
            $\Delta f_t^{g'} \leftarrow \Delta f_t^{g'} - w_p$
            $f_t \leftarrow f_t + w_p$
        end for
        $\Delta f_t^{g} \leftarrow \Delta f_t^{g} - w_p$
    else if $r = 1$ then
        $\delta_g \leftarrow \delta_g + w_p$
        $\Delta f_t^{g} \leftarrow \Delta f_t^{g} + w_p$
    end if
end for
```

Algorithm 6 `update-delta-on-deactivation`$(\boldsymbol{\delta}, g, \mathbf{f}^G, \mathbf{f})$

```
for $p \in P_g$ do
    $r \leftarrow |p| - \sum_{g' \in p} b^G_{g'}$
    $g'' \leftarrow$ the only other inactive gene if $r = 2$
    $t \leftarrow \texttt{target}(p)$
    if $r = 1$ then
        for $g' \in p$ do
            $\delta_{g'} \leftarrow \delta_{g'} - w_p$
            $\Delta f_t^{g'} \leftarrow \Delta f_t^{g'} + w_p$
            $f_t \leftarrow f_t - w_p$
        end for
        $\Delta f_t^{g} \leftarrow \Delta f_t^{g} + w_p$
    else if $r = 2$ then
        $\delta_{g''} \leftarrow \delta_{g''} - w_p$
        $f_t^{g''} \leftarrow f_t^{g''} - w_p$
    end if
end for
```

Algorithm 2 is the main procedure. We first initialize the binary variable vectors $\mathbf{b}^P$ and $\mathbf{b}^G$. We also initialize a $\boldsymbol{\delta}$ vector whose value is the change in the objective that would result if we turned a particular gene on or off. The vector $\mathbf{f}$ is the expected number of paths going to a target t (i.e. the sum of the path scores of all active paths ending in target t). The vector $\Delta\mathbf{f}^G$ is the change in $\mathbf{f}^T$ that would be caused by flipping a gene.

In the main loop starting at line 5, we first get the gene g that would cause the biggest increase in the objective. We then compute all the target gene pairs T that are present in the paths that contain gene g. We then update the $\boldsymbol{\delta}$ vector in lines 13 and 20 and fix the target penalty in lines 9 and 22. We iterate until convergence.

Run time and scalability

For our experiments, the search for paths from the source proteins to the possible Tfs, takes 104s. The search from those paths to the differentially expressed target genes takes 146s. The previous two procedures should scale linearly with the number of source + targets (assuming the number of paths per source and target is kept the same). We already have 235 sources and 600 targets and as the number of sources rarely exceed 300, and the number of targets (DE genes) is unlikely to exceed 2000, the algorithm should not take more than 4-5 times longer in the worst cases.

The IP filtering algorithm takes 115 s. It scales linearly with the number of starting genes × number of iterations (typically about 1000 iterations). In our experiment, we started with 1374 genes. As the number of genes in the genome is 20000 which is 15 times larger, we do not expect the algorithm to take more than 1725s in the worst case assuming the number of iterations remain the same.

3.1.7 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.

3.1.8 Selecting phase-specific genes

To further identify *phase specific* genes we use the following procedures. For phase 1 we selected the top K genes. Then for each phase $i > 1$, we select the top K genes such that the gene was not in the top K for any previous phase $j < i$, and the minimum fold change in rank from any previous phase to phase i was atleast δ. In other words we trying to find nodes / proteins that are first used in that phase. We used the value of $K = 50, \delta = 2$ for this thesis.

3.2 Results

3.2.1 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 [181] 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 [16]. In addition to HIV expression data we obtained interaction data for HIV-1 proteins and host (human) proteins from VirHostNet [189]. Of the 235 proteins in VirHostNet, 231 are present in our protein-protein interaction (ppi) network and were used as potential sources.

As metnioned before, TimePath also uses general protein-protein interactions from BIOGRID [245] and HPRD [206], Post-translational Modification Annotations from HPRD and Protein-DNA interaction data [230] (Methods).

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).

3.2.2 Pathways and proteins identified for HIV response

The resulting dynamic network is presented in Figure 3.1.

3.2.3 Relation of the phase genes to HIV

During the initial phase following HIV-1 infection, which corresponds to early events starting from virus entry to integration (0-8 hours), the reconstructed network is enriched for transcription factors associated with DNA modification and cell cycle regulation. Transcription factors YY1 and MYC, which have repressive effect on HIV-1 LTR transcription are repressed in this early phase of infection, whereas other transcription factors such as EP300, NFKB1, STAT1, MAPK1, and TBP which are enhancers of HIV-1 LTR activity are increased. Genes such as TP53, RELA, and NR3C1 which could potentially upregulate HIV-1 LTR transcription are repressed. Genes associated with DNA modification (acetylation) HDAC1, HDAC2, KAT2B;(methylation) DNMT1, and cell cycle regulators - CTNNB1, CSNK2A1, CDK2, E2F1

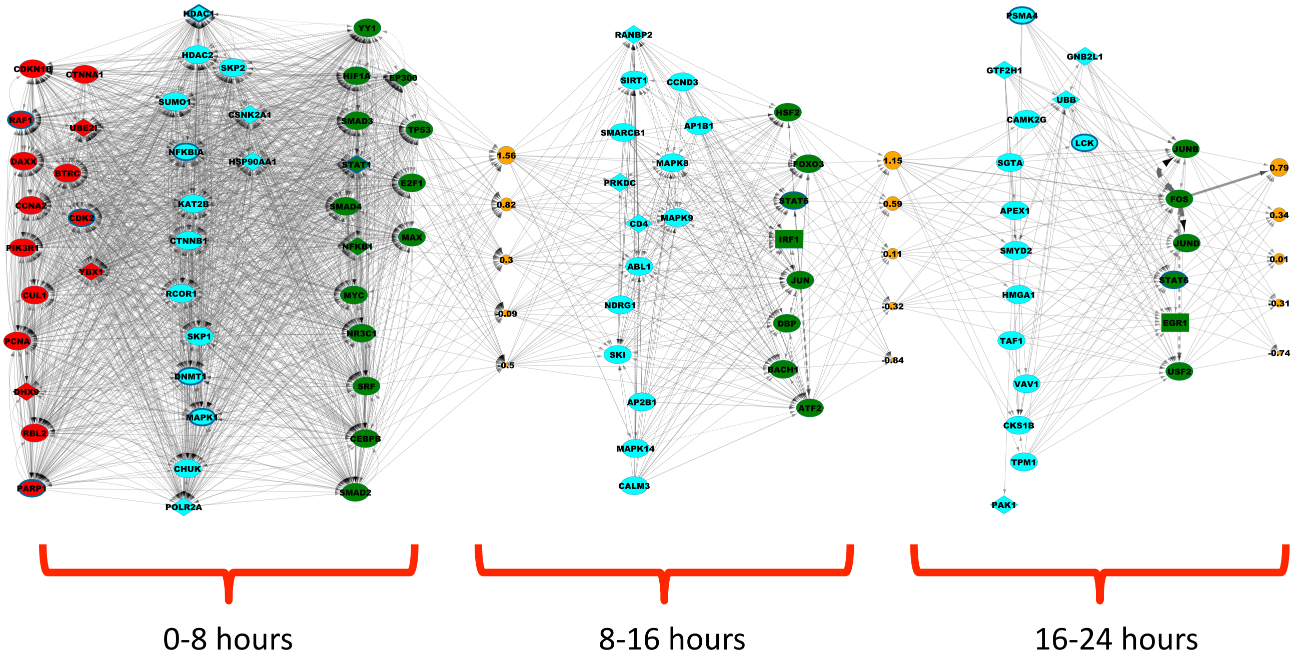

Figure 3.1: **Dynamic signaling and regulatory network for HIV-1 immune response**. The red nodes are the host proteins that interact with the HIV-1 proteins (selected sources). Blue nodes are intermediate signaling proteins and green nodes are the TFs that are predicted to directly up/down-regulate 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.

are also modulated in this early stage of infection. These changes observed in the genes related to DNA modification enzymes and the genes that regulate cell cycle support the ability of HIV-1 to infect non-dividing cells and indicate that very early in the infection process the virus dysregulates the cellular machinery to favor its transcription.

While many of the genes listed above have been known to play a role in HIV expression response, less was known about genes assigned by TimePath to later phases in the process. We have thus focused on the later two stages (8-24h) and have characterized the roles of top network genes in these two phases on virus replication and immune response (Table 3.1).

During Phase two of HIV infection (8-16 hours), when there is sequential synthesis of early, intermediate and late viral transcripts and proteins, we observe a significant decrease in the expression of genes that have a critical role in antigen presentation and host defense regulation. Adapter proteins, such as AP1 and AP2, that are essential for sorting of MHC molecules, and CD28 costimulatory molecules involved in T cell-DC interaction showed reduced expression. There is a significant 500-1000 fold decrease in the expression of these proteins in the infected T cell line. Furthermore, expression of CD4 transcripts is also reduced. This reduction in CD4 expression along with the reduced expression of PTPN7 could have adverse affect on signal transduction in T helper cells and on the induction of immune response. Since CD4 is also the primary receptor of HIV, the decreased expression of CD4 transcripts could also potentially contribute to super infection interference. It is also interesting to note that Actin, which is essential for trafficking of incoming virions is also downregulated, suggesting that HIV also modulates additional host cellular proteins to prevent reinfection of the infected target cells. During this phase, CALM3 and NDRG1 that can promote apoptosis are reduced which may help the survival of infected cells. Additional changes observed in ATM, IRF1, PIN1, SIRT1, NBN, KPNB1, SMARCB1 and XRCC5 could have a role in suppression of virus replication and could be related to host defense response. XRCC5 encoded Ku80 protein could be a of part of the host defense, as Ku80 are involved in double strand DNA repair mechanism that is caused by HIV during integration of proviral DNA during infection cycle. There is a decreased expression of KPNB1 and SMARC1, which is required for efficient integration of HIV-1, while the expression of NBN, which has a key role in post integration repair is increased. HIV-1 viral proteins Env, Rev, Tat, Vif, Vpr and Integrase interact with Ataxia-telangiectasia mutated (ATM) kinase in diverse role to promote virus replication and DNA repair pathways. Decrease in ATM can reduce virus replication and also decrease the survival of infected cells as a consequence of impaired genome stability.

Changes observed in genes such as AP1B1, AP2B1, CALM3, CCND3, CD4, ACTB, NDRG1 and others listed in Table 3.1, can be explained by changes in regulatory genes modulated in phase 1. Surprisingly, out of 16 such genes that are differentially regulated in phase 2 changes in 8 genes facilitate virus replication, while 1 gene SMARCB1 may suppress virus production and the role of the five other genes (SKI, DBP, NCOR2, STUB1, PRRC2A) are not well studied in the context of HIV infection. Differential regulation of STAT5B can either inhibit immune response or contribute to suppress HIV-1 LTR activity.

A number of the changes observed in genes in Phase 3 could have an regulatory effect on HIV-1 replication, similar to that is observed in phase 2. For instance, increase in expression of GTF2H1 helps in elongation of RNA transcripts and aid in Tat and Vpr dependent enhancement of HIV LTR activity. Similarly, increase in TAF1 is also associated with increased transcriptional activity of HIV-1 LTR. GNB2L1 is found in Staufen-RNP complexes along with other viral proteins such as Env, Gag, Tat and Nef and is considered to be involved in phosphorylation of Nef. Decreased expression of GNB2L1 can result in reduced sequestration of viral proteins in Staufen that might help to optimize Nef phosphorylation at the late stage of virus infection. DDIT3 is known to increase HIV transcription but also induce apoptosis, decrease in DDIT3 expression at later stage of virus infection may contribute to resistance to apoptosis in infected cells. Other genes that are differentially regulated in phase 3 independent of changes in earlier phase may be activated as part of the host defense response. For example, PAK1 is activated by HIV-1 viral protein Nef and is shown to have a role in HIV pathogenesis and decrease in PAK1 expression can minimize HIV-1 induced pathogenesis. Also a reduction in RAD23A, which is essential for virus replication also adversely affect virus production, though protective effect of reduced RAD23A in Vpr-mediated apoptosis cannot be ruled out.

Overall the results indicate that the virus upon entry (during the initial phase) either immediately or in a delayed manner exploits the signaling pathways and intracellular protein interactions to facilitate its replication and/or evade the innate immune defense. This occurs prior to the induction of cellular host cellular immune mechanism that seems to be progressively enriched at later stages though to a limited extent.

We also performed a literature search to assess if and how the genes we uncovered for each phase were related to HIV. The results are presented in Tables 3.2– 3.4.

Gene	Fold change (log2)	Phase	Predicted outcome	Overall impact of expression change on virus replication
AP1B1	-9.04	2 (8 - 16 hours)	Inhibiton of antigen presentation; dysregulation of immune response	+
AP2B1	-8.20		Dysregulation of surface expression of immune molecules	+
CALM3	-10.89		Prevent apoptosis of infected cells	+
CCND3	-10.99		cell cycle arrest; reduced transactivation of HIV-1 LTR	+/-
CD4	-10.71		prevent superinfection; dysregulation of immune response; altered signal transduction; membrane targeting of Env	+

NDRG1	-8.08		inhibit p53 mediated caspase activation and induction of apoptosis	+
SKI	-7.51		Not known	
SMARCB1	-7.54		reduced HIV integration; reduce gag processing and release	-
DBP	-9.08		Not known	
STAT5B	-6.91		inhibit or activate HIV-1 LTR based on STAT5B isoform; impair immune response	+/-
ACTB	-12.05		reduced HIV movement , prevent superinfection	+
NCOR2	-7.66			
PTPN7	-10.02		impair T cell signal transduction	+
STUB1	-8.46			
CDC34	-9.05		Cell cycle arrest	+
PRRC2A	-10.35			
GNB2L1	-7.99	3 (16 - 24 hours)	inhibits virus incorporation in staufen; Optimizes nef function	+
GTF2H1	1.93		increased elongation of viral transcripts, transactivation of HIV-1 LTR	+
PSMA4	2.58		Multiple role , proteasomal subunit	+/-
SGTA	-5.37		suppress Vpu mediated Gag release	-
TAF1	1.37		Transactivation of HIV-1 LTR	+
TPM1	1.20			
UBB	-2.97		Multiple role	+/-
VAV1	-3.75		Inhibition of T cell transduction	+
DDIT3	-2.59		inhibition of HIV transactivation; inhibition of apoptosis	+/-

Table 3.1: **Analysis of predicted most likely genes playing a role in viral replication**. Note that these are subset of genes predicted as being important for the respective phases

Table 3.2: **Phase 1 proteins and their relation to HIV**

Protein Name	Relation to HIV
EP300	Acetylation of the HIV-1 Tat protein by p300 is important for its transcriptional activity [196]
TP53	HIV-1 viral infectivity factor interacts with TP53 to induce G2 cell cycle arrest and positively regulate viral replication [125]
HDAC1	NF-κB p50 promotes HIV latency through HDAC recruitment and repression of transcriptional initiation [282]
RELA	NF-κB p50 promotes HIV latency through HDAC recruitment and repression of transcriptional initiation [282] and also regulates HIV-transcription
HDAC2	NF-κB p50 promotes HIV latency through HDAC recruitment and repression of transcriptional initiation [282]
CEBPB	Regulated HIV-1 gene expression through CDK9 association [169]
CTNNB1	Plays a role in HIV transcription repression in multiple cell types including astrocytes via the beta-catenin/Wnt pathway [188]
NFKBIA	NF-κB p50 promotes HIV latency through HDAC recruitment and repression of transcriptional initiation [282] and also regulates HIV-transcription
NFKB1	NF-κB p50 promotes HIV latency through HDAC recruitment and repression of transcriptional initiation [282] and also regulates HIV-transcription
DAXX	Daxx interacts with HIV-1 integrase and inhibits lentiviral gene expression [122]
YY1	Human transcription factor YY1 represses human immunodeficiency virus type 1 transcription and virion production [174]
SMAD3	MH2 domain of Smad3 reduces HIV-1 Tat-induction of cytokine secretion [76]
E2F1	Downregulates HIV transcriptional activity [150]
NR3C1	The HIV-1 virion-associated protein vpr is a coactivator of the human glucocorticoid receptor [140]
MYC	Expression of the c-myc proto-oncogene is essential for HIV-1 infection in activated T cells [249], HIV-1 Tat transactivation requires c-Myc [40], c-Myc and Sp1 contribute to proviral latency by recruiting histone deacetylase 1 to the human immunodeficiency virus type 1 promoter [129]
STAT1	HIV-1 Nef activates STAT1 in human monocytes/macrophages through the release of soluble factors [81], STAT1 signaling modulates HIV-1induced inflammatory responses and leukocyte transmigration across the blood-brain barrier [52]
RAF1	Raf-1 activates HIV-1 LTR expression [43]
CDK2	CDK2 involved in HIV-1 transcription [14, 15]
SKP2	Ubiquitylation of Cdk9 by Skp2 facilitates optimal Tat transactivation [24]
SRF	While not directly related to HIV, it is associated with a variety of early response genes like FOS and JUN [48] which are targets (differentiall expressed genes) in the later phases of the HIV time series
MAPK1	Expression of Nef (HIV-1 protein) in podocytes induced significant MAPK1,2 phosphorylation [113]

KAT2B	Tat (HIV-1 protein) stimulates HIV-1 transcription and its activity is dependent on PCAF (alias of KAT2B) [185]
PARP1	Poly (ADP-ribose) polymerase-1 (PARP1) is required for efficient HIV-1 integration [105]
CDKN1B	The CDKN family has been suggested to inhibit HIV-1 transcription [276]
YBX1	Interaction of YB-1 with human immunodeficiency virus type 1 Tat and TAR RNA modulates viral promoter activity [18]
CCNA2	Phosphorylation of SAMHD1 by Cyclin A2/CDK1 Regulates Its Restriction Activity toward HIV-1 [62]
TBP	HIV-1 Tat stimulates transcription complex assembly through recruitment of TBP in the absence of TAFs [215]
MAX	Max: a helix-loop-helix zipper protein that forms a sequence-specific DNA-binding complex with Myc [32] and thus may be relevant to Myc's role in HIV-1 Tat's activation/repression as documented in the entry for Myc
CSNK2A1	Biochemical characterization of HIV-1 Rev as a potent activator of casein kinase II in vitro [195]
PIK3R1	Interaction between Nef and phosphatidylinositol-3-kinase leads to activation of p21-activated kinase and increased production of HIV [163]
DHX9	RNA helicase A (DHX9) modulates translation of HIV-1 and infectivity of progeny virions [33]

Table 3.3: **Phase 2 proteins and their relation to HIV**

Protein Name	**Relation to HIV**
JUN	HIV-Tat protein activates c-Jun N-terminal kinase and activator protein-1 [149]. In addition, AP-1 formed by JUN/FOS has binding sites on HIV LTR and can regulate viral transcription
CD4	Receptor for the HIV virus; In later stages, teh viral proteins Nef, Vpr, Env dysregulate (reduce) surface expression of CD4 to prevent superinfection (new infection of the same cell) [66]
ACTB	Actin is known to interact with the viral protein Gag
ATM	Suppression of HIV-1 infection by a small molecule inhibitor of the ATM kinase [155]
SIRT1	SIRT1 regulates HIV transcription via Tat deacetylation [197]
RANBP2	HIV integration targeting: a pathway involving Transportin-3 and the nuclear pore protein RanBP2 [193]
PRKDC	HIV-1 Tat depression PRKDC expression [248]
PIN1	Concerted action of cellular JNK and Pin1 restricts HIV-1 genome integration to activated CD4+ T lymphocytes [170]. Also interacts with anti-HIV factor APOBEC3 to reduce its antiviral activity, hence it has replication promoting effect in the presence of A3 proteins [279].
NCOR2	SNP in NCOR2 associated with HIV-1 transmission [56]
NBN	Evidence that it is involved in HIV-1 post integration repair [242]
XRCC5	Ku80 (protein from XRCC5) depletion associated with delay in HIV-1 viral replication [128]
KPNB1	Evidence that it is required for viral integration [143]

Table 3.4: **Phase 3 proteins and their relation to HIV**

Protein Name	Relation to HIV
FOS	AP-1 formed by JUN/FOS has binding sites on HIV LTR and can regulate viral transcription
PSMA4	Proteasome inhibition interferes with relase and maturation of HIV [229]
SGTA	Binds to Vpu and Gag and its overexpression can reduce efficiency of HIV-1 particle release [72]
GNB2L1 (RACK1)	Rack1 Binds HIV-1 Nef and Can Act as a Nef–Protein Kinase C Adaptor [85]
VAV1	Human immunodeficiency virus type 1 Nef recruits the guanine exchange factor Vav1 via an unexpected interface into plasma membrane microdomains for association with p21-activated kinase 2 activity [218]
USF2	USF/c-Myc enhances the promoter activity of CXCR4, a coreceptor for HIV-1 entry [183]
EGR1	HIV-1 Tat inhibits NGF-Induced Egr-1 transcriptional activity and consequent p35 expression in neural cells [67]
CKS1B	Inhibition of human immunodeficiency virus type 1 transcription by chemical cyclin-dependent kinase inhibitors [275]
MXI1	Suggested to inhibit Myc function [298] which is implicated in HIV-1 infection as stated in the entry for Myc in Table 3.2
HMGA1	High-mobility-group protein I can modulate binding of transcription factors to the U5 region of the human immunodeficiency virus type 1 proviral promoter [115]
PAK1	A PAK related kinase is activated by HIV Nef protein [224]
RAD23A	Homolog binds to Vpr and helps in Vpr dependent replication of HIV-1 in non-proliferating cells and primary macrophages.
LCK	The Src kinase Lck facilitates as-

Table 3.5: **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

3.2.4 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 [121], relatively few are focused on analyzing dynamic response networks. These include SDREM [91, 93], which combines a HMM method for modeling dynamic regulatory networks with a combinatorial algorithm for signaling network reconstruction and TimeXnet [200] 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 since several methods for analyzing gene expression data still focus on such DE genes [217].

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 [46]. We have filtered the combined list to select a subset of the hits that are supported by at least two lines of evidence resulting in 389 supported hits, 364 of which were present in our initial network.

The results are in Table 3.5. 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

Table 3.6: **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

the TimeXnet output ranking for that method. The RNAi overlap is presented in Tables 3.5. 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 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 [28] 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 [3] and includes ”toll-like receptor signaling pathway”, an important component of innate immune response [168], ”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 3.7 using the FuncAssociate [28] tool. We compared the % of significantly enriched GO categories that were immune response related. 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 3.7).

Results for Reactome are presented in Table 3.5. 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.

Table 3.7: **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 [28].

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 3.8: **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}

We have also analyzed the usefulness of the various stages of TimePath. As can be seen in Table 3.6, 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 3.8, we obtain almost 3 times as many edges in the overlap compared to the network without the time constraint with correspondingly better p-value.

3.2.5 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

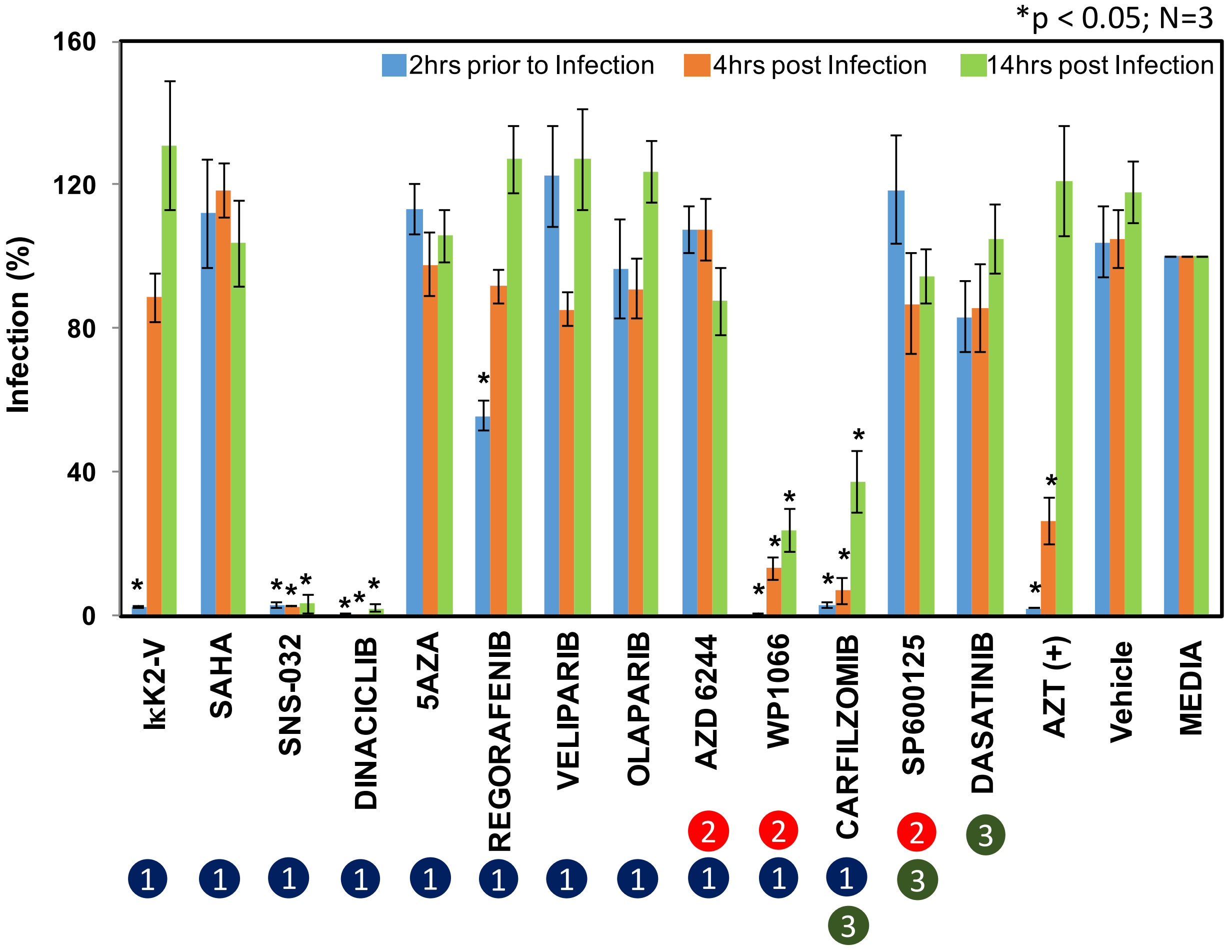

Figure 3.2: **Experimental validations.** Relative infection after treatment with inhibitors. Significant changes in infection are highlighted with a *. The inhibitor names are given on the X axis and the target proteins of the inhibitors are given in parenthesis.

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 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 [152, 252, 254, 283, 284]. 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 [119], 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 [64], 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 [295] and cullin ligases are recruited by HIV-1 viral proteins to overcome host viral restriction factors, HIV-1 Vif degrades APOBEC proteins [98] and HIV-1 Vpr induces degradation of UNG and SMUG uracil-DNA glycosylases [228]. 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 [107]. 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.

3.3 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 [269].

HIV-1 associated neurocognitive disorder (HAND) is one of the major co-morbidities of HIV-1 infection. HAND includes a spectrum of clinical manifestations associated with cognitive and behavioral impairments, based on increasing severity is classified as asymptomatic neurocognitive disorder (ANI), mild neurocognitive disorder (MND) and HIV-associated dementia (HAD) [101]. These clinical manifestations are the consequence of progressive loss of neurocognitive function [225]. Nearly half of the HIV-1 infected population is known to have some degree of HAND [58, 191] and understanding how HIV-1 contributes to neuronal dysfunction remains a priority.

3.3.1 Methods

Study Population

Frozen PBMCs were obtained from participants of the Multicenter AIDS Cohort Study (MACS), as per the protocol [25, 71]. The study population comprised of HIV-1 seronegative controls (N=36), well-characterized HIV-1 seropositive individuals who did not have any clinical neurocognitive symptoms on standard clinical neurological testing (N=16) and those who were identified as MND (N=8) or HAD (N=16), based on well-established clinical evaluation. All the subjects were men of unknown ethnicity.

mRNA profiling and data analysis

Total RNA was isolated from PBMCs using the MirVANA kit (Applied Biosystems), as suggested by the manufacturer and was profiled with HT-12 V4 array bead chips (Illumina, San Diego, CA, USA) as described previously [71, 268]. Genome Studio was used to analyze the data and identify the differentially regulated gene transcripts. Rank invariant method and no background subtraction was included to normalize the data. Additionally, the missing samples were excluded. A detection cut-off of p ¡0.01 was used. For calculating differential expression, the Illumina custom model was included along with multiple testing corrections using Benjamini and Hochberg False Discovery Rate. q¡0.05 was considered as the cut-off to identify significantly regulated gene transcripts.

3.3.2 Results

We next explored the contribution of HIV-1 viral proteins using TimePath analysis. Results (Figure 3.3) identified CCND3, CDK4, CCND1, ESR1 and RB1 as the top 5 regulators of the transcriptome changes observed in MND (Table 3.9). It can also be noted that HIV-1 Env is ranked higher than the other viral proteins at rank 26, with Gag-pol at 33 and Rev at 37. Similarly analyses of the HAD stage, with the restriction to include the cellular networks associated

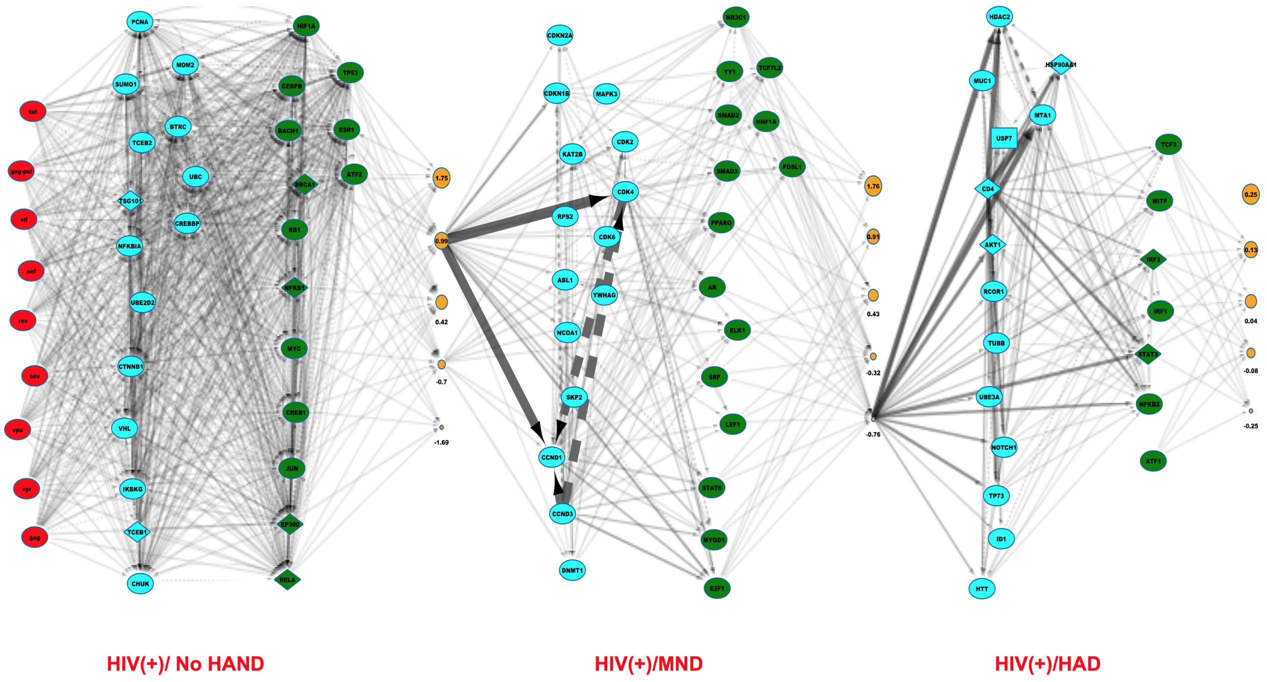

Figure 3.3: **Network recovered by TimePath for HIV-related dementia.** The red nodes represent the HIV-1 proteins (sources). Blue nodes are intermediate signaling proteins and green nodes are the TFs that are predicted to directly up/down-regulate the differential expression of target genes (targets 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 proteins for each of the three stages and also demonstrated is the relation to the HIV-1 proteins via the signaling proteins and differentially expressed genes in earlier phases. Note that some intermediate proteins may also be TFs.

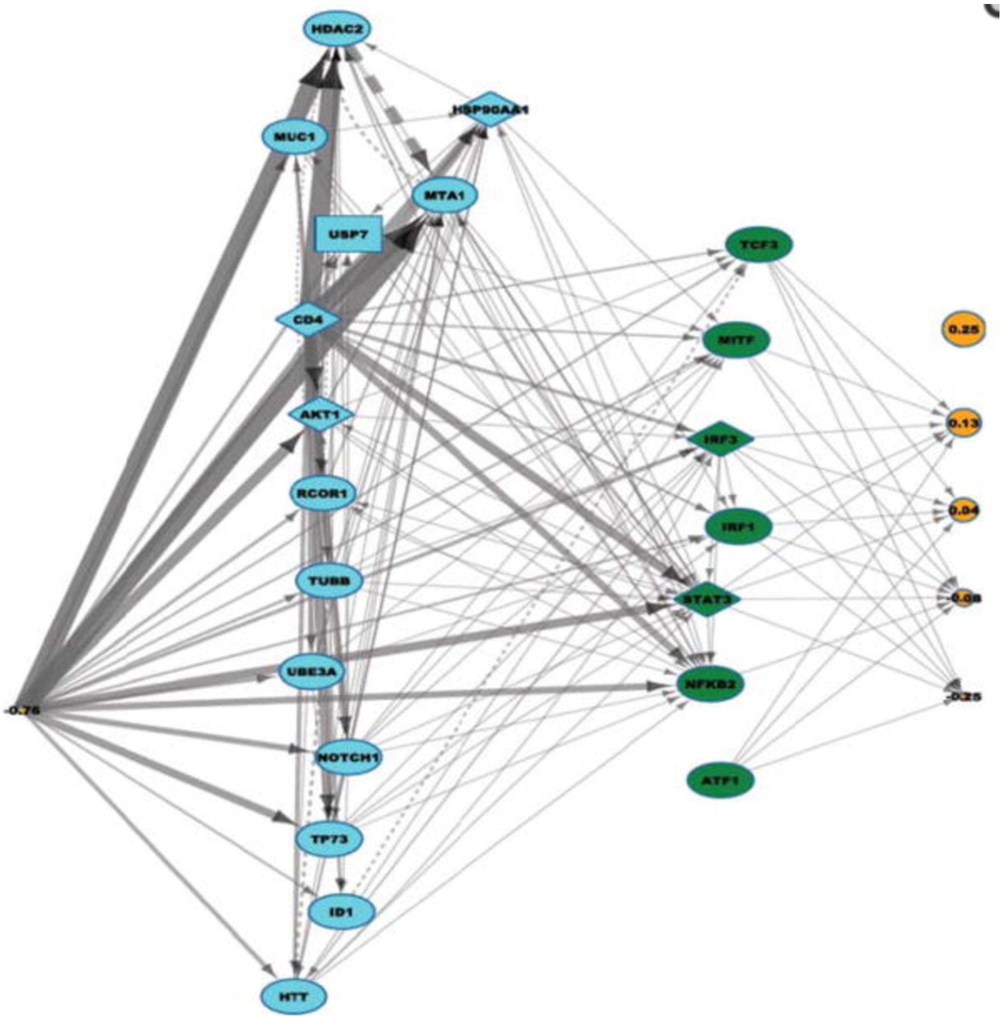

Figure 3.4: **Magnified view of section of the network recovered by TimePath.** Blue nodes are intermediate signaling proteins and green nodes are the TFs that are predicted to directly up/down-regulate the differential expression of target genes (targets not shown in figure, but the average levels of 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)

with HIV-1 seropositive group and MND, shows that the viral proteins are ranked relatively high (between ranks 2039), suggesting that the viral proteins and/or virus infection may play a major role in progression of disease from MND and HAD. Other proteins that ranked high include the host protein CD4, which is the main receptor of HIV-1 virus along with transcription factors including TP53, EP300, RELA, RB1, and ESR1, which are known to regulate virus replication, further strengthening the association of virus replication/infection with HAD (Figure 3.4). Additionally specific HIV-1 viral proteins were identified to regulate pathways: TRAFCD40RNF31, CREBBPSREBF1MYH9, CEBPB/SUMO1HSF1HSPH1 (Table 3.10), which have been previously identified to regulate monocyte/macrophage chemotaxis, inflammation and regulation of intracellular signaling, these were identified during HAD. Interestingly, other significant pathways (Table 3.10) regulated by HIV-1 viral proteins, especially those regulating NRGN and CIRBP were identified in patients who did not have HAND symptoms while the rest of the other significant pathways were enriched in HAD (Table 3.10), suggesting that some of the early molecular events associated with neurological pathogenesis caused due to HIV-1 viral proteins are observed in PBMC in the absence of any HAND symptoms. The HIV-1 proteins regulating these pathways in HAD were due to Nef, Vpu and Env, while the changes in NRGN and CIRBP in HIV seropositive subjects with no HAND can be attributed to Tat, Vpr, Vpu, Vif, Nef and Gag-Pol.

	NO HAND	**MND**	**HAD**
Vif	4	109	20
Gag	6	110	23
Nef	12	111	26
Tat	13	68	27
Gag-pol	10	33	28
Vpu	9	102	29
Vpr	11	113	32
Rev	14	37	34
Env	15	26	39
CD4	NP	NP	1
UBC	1	6	2
EP300	5	10	3
TP53	2	7	4
RELA	16	41	5
RB1	34	5	6
ESR1	21	4	7
HDAC1	36	24	8
HIF1A	3	19	9
CTNNB1	19	22	10
PCNA	18	30	11
MDM2	8	28	12
BRCA1	20	13	13
CEBPB	50	17	14

CREBBP	7	77	15
CCND1	59	3	16
JUN	27	15	17
SUMO1	17	47	18
CHUK	26	NP	22
CDKN1A	37	12	24
NFKBIA	22	53	25
HDAC2	71	NP	30
MTA1	87	NP	31
NR3C1	88	36	33
SMAD3	91	38	35
KAT2B	102	18	36
MYOD1	115	14	37
NFKB1	33	81	38
BTRC	23	NP	40
PPARG	103	29	42
TSG101	25	86	43
AKT1	NP	NP	44
EGFR	39	95	45
VHL	24	NP	47
HSF1	148	43	49
IKBKG	31	NP	50
NCOR2	129	48	51
BRCA2	46	NP	55
SRC	35	NP	56
UBE2D2	28	NP	57
NOTCH1	138	NP	59
MYC	43	23	60
MDM4	32	89	61
TCEB1	29	NP	62
TCEB2	30	NP	63
ATF2	45	55	64
CDK2	54	16	68
E2F1	57	11	69
RANGAP1	40	NP	71
CBL	42	NP	72
TRAF2	41	72	73
RUNX1	139	66	74
NCOA3	124	52	78
CREB1	38	NP	81
PIAS1	177	88	83

IRF3	NP	NP	86
VCP	49	NP	88
ELK1	152	46	90
RANBP2	47	NP	92
RUNX2	153	67	95
PAK1	172	50	98
USP7	NP	NP	101
ABL1	175	32	104
SRF	168	79	106
RAC1	192	71	107
NCOA1	76	21	109
STAT6	77	20	110
NR1I2	183	87	116
FN1	48	45	124
NCK1	119	51	126
ATF1	NP	NP	128
SKP2	98	25	143
CDK4	189	2	144
CDKN2A	109	39	154
CDK6	108	40	155
CDKN1B	132	9	157
BACH1	44	63	190
CCND3	191	1	191
CDC5L	198	74	197
MAPK3	194	27	199
YWHAG	NP	8	NP
PTGES3	NP	31	NP
RPS2	NP	44	NP
YWHAZ	NP	54	NP
SUMO3	NP	57	NP
SMARCA5	NP	59	NP
ERGIC3	NP	60	NP
APEX1	NP	65	NP
SUMF2	NP	90	NP
TARS	NP	98	NP
EIF3L	NP	108	NP
EIF3D	NP	114	NP

Table 3.9: **Rank of HIV-1 viral proteins and host proteins in cellular networks associated with different stages of HAND.** Each phase consisted of one time point starting with the HIV positive subjects without HAND. Top 200 DE genes were extracted for each phase relative to the previous phase. These were the DE genes were included as targets in TimePath. All the HIV-1 viral proteins and cellular proteins were included as sources in the analyses. Ranking corresponds to their relative role in the changes observed in transcriptome for each phase. Higher the rank (lower the value) denotes greater role in the transcriptome regulation.

Pathway	Viral proteins	Comment
CREBBP → CREB1 → CIRBP	Tat,Gag-pol,Vif,Vpr,Vpu	Associated with Huntingtons Disease,and disorders of basal ganglia
TBP → RXRA → NRGN	Tat	Associated with Alzheimers,Huntingtons Disease,and disorders of basal ganglia
AR → NCOR1 → PPARA → NRGN	Tat,Nef,Vpu,Gag-pol,Vif,Vpr	Associated with Alzheimers,Huntingtons Disease,and disorders of basal ganglia
TRAF → CD40 → RNF31	Nef,Vpu,Env	Regulate NFKB pathway in macrophages
CREBBP → RELA → SEC24A	Nef,Vpu,Env	Associated with Huntingtons Disease,and disorders of basal ganglia
CREBBP → SREBF1 → MYH9	Nef,Vpu,Env	Role in chemotaxis of monocytes
MYC → MAX → BRD2	Nef,Vpu,Env	Associated with Neuromuscular disease
RB1 → E2F1 → IFNAR1	Nef,Vpu,Env	Role in regulation of Inflammation,virus infection
EP300 → YY1/CEBPB → PREPL	Nef,Vpu,Env	Associated with central nervous system cancers,and Huntingtons Disease
SMAD2 → SMAD4 → NDUFS8	Nef,Vpu,Env	Associated with Leukoencephalopathy
EP300 → ELK1 → NDUFS3	Nef,Vpu,Env	Associated with Leukoencephalopathy,Huntingtons Disease and disorders of basal ganglia
CEBPB/SUMO1 → HSF1 → HSPH1	Nef,Vpu,Env	Role in macrophage differentiation
EP300 → NFATC1 → S1PR1	Nef,Vpu,Env	Associated with disorders of basal ganglia

Table 3.10: **List of pathways identified by TimePath analysis that are associated with HAND pathogenesis.**

3.4 Conclusion

In this chapter, we presented an integer programming based method to jointly reconstruct the signaling and regulatory network for time series gene expression data. We applied our method to expression data obtained from blood cells when infected with the HIV-1 virus. We showed that we outperformed existing methods (specifically TimeXnet [200] and SDREM [93]) in terms of enrichment of RNAi screen hits and enrichment of protein-protein interactions present in 'gold standard' Reactome pathways pertaining to HIV. Furthermore, we introduced a novel constraint in the reconstruction of our network and provided statistical evidence that it greatly improved the quality of the reconstructed network. We also performed followup experiments to verify the temporal predictions of our reconstructed network. Finally we applied our method to HIV-related dementia and showed that it was able to recover a lot of known biology as well as generate ideas for future exploration.

Effect of sampling rate on TimePath For the HIV-1 infection model in this chapter, the sampling rate for measuring gene expression was every 2 hours for 24 hours. This was carefully chosen so as to not miss out on any important biological events [181]. This raises the question however, as to how the predictions of the algorithm might change if the sampling rate changes. In particular, a lower sampling rate could cause the method to miss some regulatory events. However there are a couple of mitigating factors :-

1. Even if a differential gene expression event at a particular time point is missed because that time point is not sampled, differential expression changes can take 30 minutes–2 hours and last for several hours. So one could still be able to observe differential gene expression for that gene at a later time point.
2. The largest amount of differential expression usually occurs during the early part of the time series. Thus sampling densely around that time and less densely at later time points is a strategy that can be used if there are not enough resources to do a dense sampling throughout the time period being looked at (and indeed, several experiments do follow this design).

Furthermore, we want to note that missing regulatory events due to undersampling is a problem that would affect any method that tries to do inference on time series gene expression data and not just ours. In addition, see Kleyman et al. [141] for a more detailed discussion of how sampling rates can affect downstream analysis.

Chapter 4

Incorporating epigenetic data for network inference and application to Idiopathic Pulmonary Fibrosis

As explored in this thesis, 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 do not consider is the role epigenetic modifications play in regulating gene expression. As described in the introduction, 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 like cancer [236], Alzheimer's [57], etc. A comprehensive review is available in [74].

The primary means via which epigenetic modifications cause phenotypic change is by altering gene expression by various mechanisms [205]. Enhancers are genomic elements $50 - 1500$ bp long, situated anywhere from 1 bp to 1Mbp from the transcription start site (TSS) of a gene that can regulate the expression of that gene [131, 237]. 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 [148, 176]. 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] [277, 297]. For example, the histone

[1] Active enhancers are those aiding in ongoing transcription, Poised enhancers are those that are not but are just

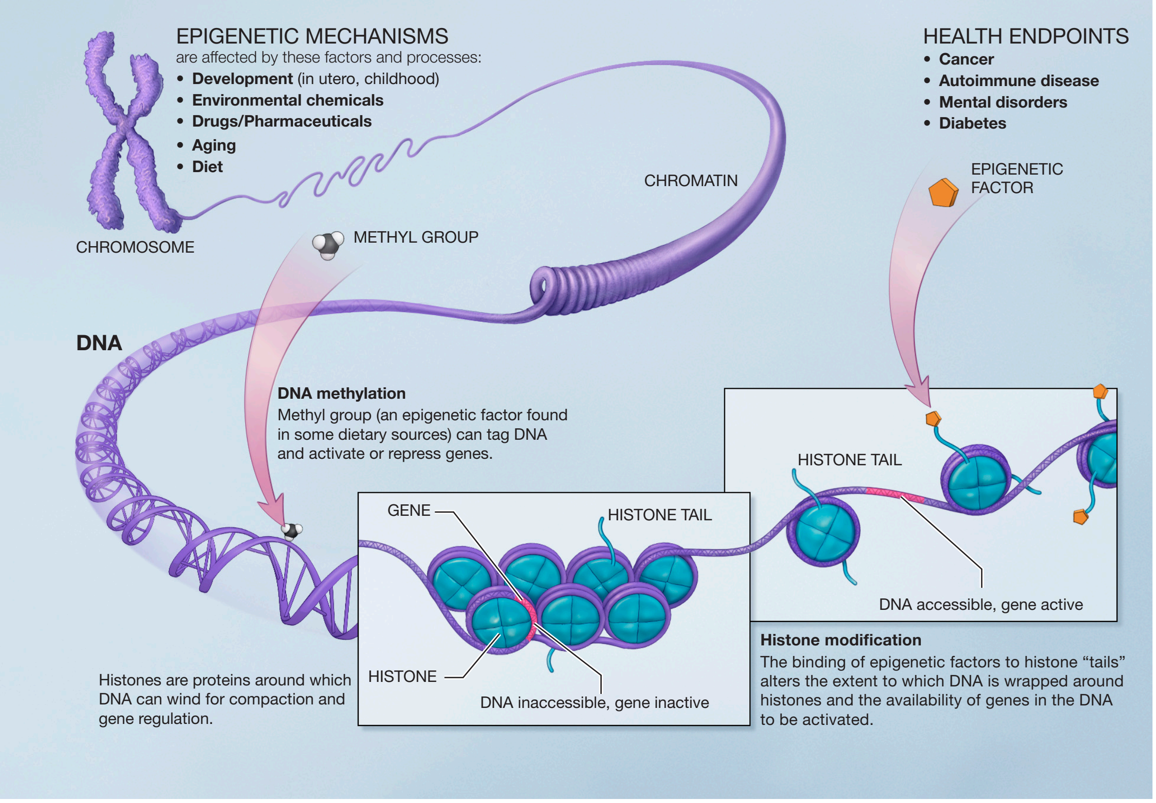
EPIGENETIC MECHANISMS
are affected by these factors and processes:
• Development (in utero, childhood)
• Environmental chemicals
• Drugs/Pharmaceuticals
• Aging
• Diet
HEALTH ENDPOINTS
• Cancer
• Autoimmune disease
• Mental disorders
• Diabetes
CHROMOSOME
CHROMATIN
EPIGENETIC FACTOR
METHYL GROUP
DNA
DNA methylation
Methyl group (an epigenetic factor found in some dietary sources) can tag DNA and activate or repress genes.
HISTONE TAIL
GENE
HISTONE TAIL
DNA accessible, gene active
Histone modification
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.
Histones are proteins around which DNA can wind for compaction and gene regulation.
HISTONE
DNA inaccessible, gene inactive

Figure 4.1: **Illustration of various epigenetic modifications.** 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 [2]

modification H3K27ac has been shown to be associated with active enhancers [47, 61].

Recently, there has started to be an increasing interest in the role epigenetics plays in cell 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 [130, 241] For example, some TFs like CREB bind less well to methylated DNA. There also exist transcriptional repressors like MeCP1 and MeCP2 that recognize methylated DNA and bind to it thus inhibiting transcription [133]. Thus it would be of great interest to have models that are able to incorporate epigenetic information when inferring signaling and regulatory networks.

4.1 Prior work

There have been some attempts to examine the influence of epigenetics on gene expression. Li et al. [159] use epigenetic and other genome features to predict differential gene expression between lung cancer and control patients. Yu et al. [294] use a bayesian network model to try and infer causal links between epigenetic modifications within ±1kb of the TSS. Cheng et al. [55] develop support vector machine and support vector regression models to quantify the effect of epigenetic modifications on gene expression. They bin the DNA region ±4kb of the TSS into 100 bp 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 [53, 300]. Both of the latter methods use the correlation between epigenetic profiles of genes as a prior when inferring gene regulatory networks. In Gong et al. [100], 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. In Singh et al. [240], they use a deep neural network to predict gene expression using histone modification information from segments of the genome close to the gene as features. However this approach does not take into account DNA methylation.

There have also been attempts at using sequence and epigenetic features to predict binding of transcription factors (TFs) to DNA. In [12, 134, 210, 302], deep learning methods are used to predict TF binding, and histone modifications as a function of the DNA sequence. DeepBind [12] 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 [302] 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 step away from being active

There have also been attempts to incorporate epigenetic information for predicting TF binding. In [167], they found several histone modifications that were predictive of TF binding even excluding sequence information. FactorNet [211] is a deep learning model that uses the DNA sequence and DNAse-seq signal as input to try and predict TF binding. Recently, the ENCODE-DREAM challenge also took place with the task being TF binding prediction as a function of sequence and DNAse-seq (https://www.synapse.org/ENCODE). However as using DNAse-seq to improve binding prediction, seemed like a well explored direction, we thought it more fruitful to focus on other types of epigenetic data like histone marks or DNA methylation.

There are several problems however with trying to use histone marks. You need an antibody for each type of histone mark. The resolution of the histone modification is on the order of several hundred base pairs – whereas methylation can be on a single nucleotide level. Thus, we focused on DNA methylation as the epigenetic modification of interest.

4.2 Initial attempts at incorporating methylation

4.2.1 Data used

All experiments in this section were conducted using the ENCODE database consisting of binding information for 180 TFs (measured via Chip-Seq) across 90 celltypes (methylation measured via RRBS – restricted reduced bisulphite-sequencing) though not all TFs had binding information for all celltypes.

We preprocessed the above data in the following way. We first segmented the entire human genome into 200 bp non-overlapping segments. We then considered a TF bound to a segment (for a particular celltype) if some Chip-Seq peak for that TF-celltype pair overlapped half or more of that segment.

4.2.2 Trying to improve binding prediction

Baseline models

For each TF and (200 bp) segment in the genome, we computed the pearson correlation between the binding and the methylation levels across all the celltypes for which we had binding information for that TF. We then took the average of this correlation across all the segments of the genome. We compared this average correlation with the average correlation we obtained if we shuffled the methylation levels across the different methylated sites for a celltype. For several TFs (ex CTCF), we found a significant improvement in the correlation for the original vs the shuffled version (Table 4.1).

This seemed to indicate that methylation could be predictive of TF binding. An important point to note is that the correlation was computed for the *same* genome segments. Thus this experiment tried to measure the correlation between binding and methylation *after* accounting for the confounding factor of genome sequence. However we then trained a logistic regression classifier (again for each TF) with 2 features – the (binary) binding indicator for the segment of the genome in question from an arbitrarily selected celltype and the methylation level for that segment for the celltype whose binding we were trying to predict. However this did not yield any improvement over using no methylation at all or over shuffling the methylation across the genome 4.1.

Neural network models

One possible reason that methylation was not yielding a better predictor was that the hypothesis space that we were using (logistic regression with only 2 features) was too small. Given the success of neural networks in predicting binding from sequence (and DNAse-Seq), we turned to them to see if we could improve binding prediction by adding in methylation features.

Briefly, neural networks are a biologically inspired machine learning model. They are typically organized in layers. Each layer takes as input a vector, multiplies it with a weight matrix to obtain another vector and then applies a non-linear function (like sigmoid or hyperbolic tangent or rectified linear function) to this vector to obtain the output vector. The first layer of the neural network accepts the input features as its input vector. The weight matrix can be sometimes constrained to enforce a sparsity pattern or by enforcing that subsets of elements in the weight matrix be the same. This is done to reduce the number of parameters the neural network may have to learn [156]. The reader is referred to the Introduction for a more detailed introduction to neural networks.

We first took the DanQ network [210] (Figure 4.4) that tries to predict binding from just sequence data as it had the state of the art results at the time. The first layers of the DanQ model are designed to scan sequences for motif sites through convolution filtering. The convolution step of the DanQ model contains one convolution layer and one max pooling layer to learn motifs. The max pooling layer is followed by a Bi-LSTM layer. The rationale for including a recurrent layer after the max pooling layer is that motifs can follow a regulatory grammar governed by physical constraints that dictate the in vivo spatial arrangements and frequencies of combinations of motifs, a feature associated with tissue-specific functional elements such as enhancers [209, 212]. Following the Bi-LSTM layer, the last two layers of the DanQ model are a dense layer of rectified linear units and a multi-task sigmoid output.

The task is to predict the binding for any arbitrary 200 bp genome segment. DanQ takes as input the one-hot encoded version of the 200 bp DNA sequence (see introduction for a description of one-hot encoding) as well as flanking sequences on both sides of length 400 bp. This gives an input matrix of size 4×1000. We then take a 1000 element vector per strand where each element indicates the methylation level for that base for a celltype. We then concatenate these two vectors to the input matrix. The rest of the network is suitably extended to incorporate the larger input matrix. Unfortunately the more complex model also failed to extract any predictive value from the methylation data (as an example, the accuracy on CTCF for neural networks was 0.84 for both scenarios and the accuracy for MAX was 0.65 for both as well). We also tried variants of this model including concatenating the methylation vectors with layers much closer to the output, incorporating a much larger genome context for methylation (up to 25000 bp compared to 1000 bp for the actual sequence), and trying to regress the chip-seq profile directly instead of just trying to do a binary classification of whether a TF is binding or not, but the results did not change. In addition to this, we also tested different parameter values for the model. Specifically, we tested different numbers of convolutional layers (between 1 and 3), different number of convolution filters (between 10 and 30), and different convolution filter sizes (between 13 and 26) but they yielded no further improvements to the model.

TF	Avg. r^2 of meth/binding	Avg accuracy of logistic regression	Avg. r^2 of shuffled meth/binding	Avg accuracy of logistic regression with shuffled meth
ATF3	-0.0477	0.7734	0.0388	0.7687
BCL3	-0.1737	0.9101	-0.0127	0.9143
BHLHE40_(NB100-1800)	-0.2416	0.5720	0.0362	0.5931
CEBPB_(SC-150)	-0.2659	0.6686	0.0499	0.6996
CHD1_(A301-218A)	-0.1923	0.8550	-0.0255	0.8476
CHD2_(AB68301)	-0.1450	0.6739	0.1806	0.6597
C-JUN	-0.3137	0.5749	0.1577	0.5934
C-MYC	-0.0948	0.9261	0.0333	0.9267
COREST_(SC-30189)	-0.0374	0.5574	0.2556	0.5710
CTCF_(SC-15914)	-0.0906	0.5622	0.0064	0.5670
CTCF_(SC-5916)	-0.1418	0.6335	-0.0307	0.6422
CTCF	-0.0626	0.9073	-0.0161	0.9082
E2F4	-0.0664	0.7830	0.0747	0.7673
EGR-1	-0.2176	0.7306	-0.0032	0.7307
ELF1_(SC-631)	-0.2909	0.6708	-0.0861	0.6695
ELK1_(1277-1)	-0.1711	0.6191	-0.0226	0.6240
ETS1	0.0971	0.8212	0.1074	0.8227
EZH2_(39875)	-0.2218	0.8380	-0.0366	0.8327
GABP	-0.0407	0.6822	-0.0361	0.6852
GTF2F1_(AB28179)	-0.1587	0.7524	0.2032	0.7404
HDAC2_(SC-6296)	-0.0876	0.8144	-0.0167	0.8163
JUND	-0.1450	0.7292	0.0502	0.7269
MAFK_(AB50322)	-0.1533	0.6738	0.1110	0.6614
MAX	-0.1249	0.8255	0.0953	0.8265
MAZ_(AB85725)	-0.2330	0.6689	0.0379	0.6425
MXI1_(AF4185)	-0.1226	0.6682	-0.0390	0.6745
NF-YA	0.1362	0.8076	0.4722	0.7870
NF-YB	-0.3515	0.5642	-0.3253	0.5670
NRF1	-0.2087	0.7760	-0.0379	0.7820
NRSF	-0.0882	0.6927	-0.0442	0.6827
P300	-0.1708	0.7691	0.0175	0.7729
POL2-4H8	-0.1077	0.8572	0.0312	0.8547
POL2(PHOSPHOS2)	-0.1520	0.6782	0.1171	0.6677
POL2	-0.0942	0.8478	0.0112	0.8484
PU.1	-0.0445	0.6306	0.0182	0.6225

RAD21	-0.1276	0.7603	-0.0267	0.7581
RFX5_(200-401-194)	-0.0421	0.5978	0.1316	0.5970
SIN3AK-20	-0.0157	0.6445	-0.0429	0.5793
SIX5	0.0183	0.7469	0.0739	0.7504
SMC3_(AB9263)	-0.1951	0.5085	0.1679	0.5161
SP1	-0.2443	0.7410	0.0169	0.7442
SP2_(SC-643)	-0.2642	0.6944	0.0037	0.7205
SRF	-0.2243	0.7439	0.0134	0.7553
STAT1	0.0471	0.9204	0.3728	0.9284
TAF1	-0.0760	0.7104	0.0065	0.7254
TBP	-0.1687	0.6481	0.0086	0.6803
TEAD4_(SC-101184)	-0.2637	0.7362	0.0519	0.7328
TR4	0.0684	0.5248	-0.0788	0.5167
USF-1	-0.1695	0.6997	-0.0164	0.6977
USF2	-0.2152	0.6975	0.0414	0.6723
YY1_(SC-281)	-0.0822	0.6680	-0.0170	0.6757
YY1	-0.0593	0.7947	-0.0808	0.7848
ZBTB33	-0.0473	0.8683	0.1157	0.8635
ZNF143_(16618-1-AP)	-0.2264	0.7493	-0.1372	0.7517

Table 4.1: **Results for correlation and logistic regression analysis of methylation and TF binding.** The first column is the average correlation across all possible binding sites between the the binary variable of whether the TF binds to that site and the methylation level for that celltype. The second column is the logistic regression accuracy when using the binding indicator for one pre-selected celltype and the methylation level as features. The third and fourth columns are similar except with the methylation value shuffled across the entire genome

Hi-C analysis

During the course of our experiments, we had used Hi-C data to connect distal enhancer regions of the genome with the genes they could potentially regulate. We used Hi-C data published by Schmitt et al. [227] which was collected and integrated for over 21 primary human tissues and cell types. Fit-Hi-C [21] was used to identify significant chromatin interactions. The resolution of the data was 40kb. We then defned enhancer regions for a gene G as those (40kb) segments of the genome that interacted with segments containing the transcription start site (TSS) of gene G. We only considered interactions that were statistically significant (p-value < 0.05) after correction for multiple hypothesis testing.

We found that the methylation in the enhancer regions for a gene showed good correlation with the expression of the gene. On average, the correlation for a gene's expression with methylation in the enhancer region was 0.15. If we shuffled the methylation values across the entire genome, the correlation with drop down to 0.003 and if we randomized it within only each chromosome, the correlation was 0.135 which was still lower than what we got without shuffling – indicating that there was some enhancer's methylation specific signal for gene expression. As far as we know, this is the first time anyone has looked at the correlation of methylation in distal enhancer regions and gene expression. Given this, we hypothesized that for a given segment of the genome, using the expression of genes it could potentially regulate may be useful for predicting binding.

While we indeed found the expression of the enhancer associated genes to be predictive, doing dimensionality reduction on the expression matrix (with each row of the matrix being for a different celltype) and taking the first 8 components as features was equally predictive. This result indicated that there was not any value that the enhancer gene's expression in particular had in predicting TF binding and thus knowledge of the enhancers (via Hi-C data) was not useful for this task.

4.3 Incorporating methylation via application to idiopathic pulmonary fibrosis

While we could not find any predictive value for binding itself as a function of the methylation data in the ENCODE database, we nevertheless wondered if we could test the effect methylation had on gene regulation via more indirect means. To that end, we turned to the lung disease idiopathic pulmonary fibrosis (IPF) for which we had detailed expression and methylation data.

In the next section we will first describe how we applied DREM (described earlier in the thesis) to infer regulatory models for IPF (without taking methylation into account). After that we will perform a direct comparison of the regulatory models inferred with and without the effects of methylation incorporated.

4.3.1 Using DREM to model IPF

Idiopathic Pulmonary Disease (IPF) is the most common of the interstitial lung diseases and the most severe with median survival ranging from 3-5 years [102]. It is described as a chronic, progressive fibrosing interstitial pneumonia of unknown etiology that occurs more commonly in older male subjects with smoking being the major risk factor. Diagnosis requires a multidisciplinary consensus scoring including radiological and histologic patterns of usual interstitial pneumonia (UIP). IPF was initially considered a chronic inflammatory disease due to the correlation of inflammatory infiltrates and the fibrosis present within these lungs [63, 226] and treated with steroids. More recently, IPF has become thought of as being a disease of repetitive injury of the alveolar epithelial cells with aberrant wound healing resulting in fibrosis. This shift has been further supported by accumulating evidence that steroids have a detrimental effect on the patient and that recently developed antifibrotic drugs have good efficacy [138, 220].

Due to the complexity of the diagnosis, that symptoms may occur 1-2 years before diagnosis, and that patients often present with advanced fibrotic disease, the pathological progression of IPF from its early stage has not been well studied. IPF presents with a heterogeneous distribution of the fibrotic lesions, where more advanced scarring and honeycombing generally occurs in the basal and subpleural regions of the lobule and with increasingly normal tissues presenting towards the centre of the lobule and the apex of the lung. Due to the progressive nature of IPF, the distribution of pathological remodelling has been described as having a temporal heterogeneity leading from chronic, to acute, then to normal tissues ranging from the peripheral regions towards the centre of the lobule, respectively [286]. As such, while obtaining lung tissue samples from patients with early stage IPF may be difficult, the regional heterogeneity characteristic of this disease may provide clues towards elucidating key mechanisms in disease progression. As lung transplantation remains the main treatment option for these patients, use of explanted lungs combined with methodical sampling and detailed characterization of the sampled tissue would allow for IPF tissue with varying degrees of pathological remodelling to be compared. In this study, we applied this methodology to determine the changes that occur during the progression of IPF from mildly affected to severely remodelled areas. As the development of fibrotic tissues is a balance of accumulation of extracellular matrix and the proteases that degrade these proteins, we sought to initially examine the changes in the expression of these genes across the different IPF stages. We also sought to determine changes that are specific to the early/mild stages of IPF to provide information on the processes that may initiate the disease process.

Methods

Patient Data and Sampling Patients with IPF undergoing lung transplantation for their disease at UZ Hospital in Leuven, Belgium were selected. Donor lungs that were not suitable for transplantation due to a number of factors, including trauma or tumours in non-lung organs, were collected to be used as controls. All lungs were collected following local hospital ethical committee approval (ML6385). In total 10 IPF lungs and 6 donor lungs were collected for this study. See Table 1 for patient demographics and lung function information.

Following explantation, lungs were inflated and held at 20 cm H2O pressure inflation while frozen over liquid nitrogen vapour according to previously established protocols [177]. The frozen lung was then imaged using a high resolution CT scanner and subsequently cut for sampling. The lungs were then cut into 2 cm thick slices along the transverse plane with a 1.4 cm diameter coring drill used to systematically sample the slice. Of samples collected, 2 cores were randomly selected from each of upper, mid, and lower lung regions for a total of 6 samples per lung and 96 samples in total.

Imaging and Histology MicroCT scans were done on frozen lung samples using a Bruker Skyscan 1172 (Bruker, Belgium) with cooling stage according to previously established protocols (Verleden 2016). Briefly microCT scans were set at 40kV, 240mA, and 0.5 rotation step at -30C. Temperature was maintained throughout the scan by having sample placed within a Styrofoam cylinder with dry ice on top. Scans were reconstructed using NRecon software (Bruker, Belgium) and images analysed using CTAn software (Bruker, Belgium).

An ROI was placed on intact portions of the sample to exclude regions that were damaged from sampling. A threshold was manually set to segment tissue from air and the software was used to measure tissue % and surface area/ volume (surface density, SD) from each core. Terminal bronchioles (TB) were identified and manually counted within each core then divided by the ROI volume to determine terminal bronchiole density (TB/mL).

A portion of each core was vacuum embedded in 50:50 O.C.T. (Sakura) and PBS (REF: Daisuke) then sectioned using a cryotome for histology. and to collect samples for gene expression. A haematoxylin and eosin (H&E) stain was used for Ashcroft scoring (ref) of the extent of fibrosis within the sample. Sections were also stained with picrosirius red for total collagen, and antibodies were used to detect Collagen I, Collagen III, and Elastin. Surface density was found to have greater correlation with Ashcroft scores and was used for subsequent analyses.

RNA-sequencing and data analysis RNA was extracted using miRNeasy micro kit from each core. cDNA libraries were prepared from 20 ng of total RNA using Ion Ampli-Seq-transcriptome human gene expression kit and sequenced using Ion Torrent (ThermoFisher).

For sequencing analysis, we used a two-stage mapping strategy to map the raw reads to the human genome (UCSC hg19). Cufflinks was used to calculate the Fragments per Kilobase of transcript per Million mapped reads (FPKM) values as the estimated gene expression levels. The 25,276 genes were filtered to remove low and non-expressing genes with expression below 0.01 FPKM in more than 90% of samples resulting in 21,837 genes remaining. Samples were subsequently divided into three tertiles based on surface density with high SD group representing mild fibrosis (IPF1), middle SD group for moderate fibrosis (IPF2), and low SD group for severe fibrosis (IPF3) to examine differential gene expression in each region. Two linear mixed-effects models were used to identify gene expression profiles with extent of fibrosis in each of these tertiles:

$$Gene_{ij} = \beta_0 + \alpha_{SLICE} + \alpha_j + \epsilon_{ij} \tag{4.1}$$

$$Gene_{ij} = \beta_0 + DISEASE_{ij} + \alpha_{SLICE} + \alpha_j + \epsilon_{ij} \tag{4.2}$$

where $i = 1, \dots, 6, j = 1, \dots, 16$.

$Gene_{ij}$ is the FPKM expression value for sample i in patient j for a single gene. SLICE is a random effect controlling for the location the sample was collected and for differences in lung size between control and IPF lungs. α_j represents the random effect for each subject with β_0 representing the intercept and ϵ_{ij} being the random error. Model 4.2 includes the fixed effects of DISEASE; an ANOVA was used to compare the two models to determine differentially expressed genes with a p-value < 0.05 considered significant following use of false discovery rate (FDR) correction. All statistical analyses were conducted using R statistical software (v.3.3.1) and the *lme4* package.

Genes and regulators used We only used genes which were considered to be differentially expressed per the analysis in the previous section. In addtion, we only considered as potential regulators, TFs which were differentially expressed with p-value < 0.05 and microRNAs that were differentially expressed with p-value < 0.1.

For the TF-gene network, we used the human TF-DNA interaction network we described in the Introduction. The miRNA-gene network was obtained from the TargetScan database [10]. Later in the chapter, we will describe how we incorporated methylation into the computation of the TF-gene network.

4.3.2 Results for DREM modeling of IPF

The clusters of genes (pathways) identified by the model are presented in Figure 4.2 and the associated regulators of those clusters are in Table 4.2. Note that while pathways A, B, and M merge from the same split node (IPF_1), M in fact consists of mainly downregulated genes whereas A, B consist of mainly upregulated genes. We also present in Figure 4.3, the Sankey diagram showing the various gene clusters and subclusters within those clusters enriched for different biological processes.

DREM identified distinct pathways that differed in their regulation and temporal pattern (Figure 4.2). The pathway associated with IPF_1 node (labeled as **1** in the figure) was involved in extracellular matrix organization. This is particularly interesting as IPF involves extensive accumulation of extracellular matrix [258]. Indeed this pathway also shows enrichment for genes associated with the IPF disease (Table 4.4). The one associated with IPF_2 node was associated with cell adhesion, the one with IPF_3 with cation channel activity, and the one with IPF_4 with GTPase mediated signal transduction (Table 4.4).

Genes known to be associated with fibrosis were increased across all stages of disease. Impressively, among genes known to be characteristic of IPF, COL1A1 (collagen), MMP14 (matrix metalloproteinase 14), CTSK (cathepsin K), ITGB6 (integrin subunit beta 6) were maximally induced in minimal disease with no further increase, whereas others such as MMP7 and MUC5B (mucin 5b) were induced in minimal disease and continued increase to late stage disease. Blood vessel formation and defense response against microbial infections were decreased at all stages of IPF but innate immune pathways where increased in early stage and adaptive immune response at mid- to late-stage IPF, potentially reflecting specific roles. Among known IPF regulators MIR-29C, MIR-30 and TFs HMGA2, LEF1 and GLI1 regulated early disease; whereas, LET-7, MIR-199 and TFs SMAD3, STAT3 and POU2AF1 regulated later phases (Table 4.2). Preliminary analysis suggested that POU2AF1 KO mice are relatively protected from bleomcyin induced fibrosis.

In normal tissue, alveolar type I and type II cells make up the parenchyma and endothelial cells make up the capillary bed. In the early stages of IPF, one starts losing type II alveolar cells leading to parenchymal tissue collapse and the development of bronchiolized honeycomb tissues and fibrosis. This results in an increased number of fibroblasts, bronchial epithelium (honeycombing), and loss of alveolar epithelium and endothelium. This can be seen in pathways G and H whose regulators are enriched for cation channel activity and cation transmembrane transporter activity respectively (cation channels are a major component of epithelium [231]). Thus this pathway likely represents the formation of honeycomb tissues. The loss of alveolar tissue leads to multiple downstream effects. While there is uncertainty as to what those downstream effects are, the most accepted ones are mesenchymal cell proliferation and excess extracellular matrix (ECM) accumulation [256]. This can be seen in pathways A, B, and D which are enriched for regulators governing extraceullar matrix organization.

As mentioned before, we observed loss of blood vessels in our data. Indeed, Pathway L is enriched for blood vessel development regulation matching that observation. IPF also involves the infiltration of inflammatory cells, in particular B/T cells [172]. You can see that in pathways M, C, D, and E which are enriched for innate immune cells, and lymphocytes.

node IPF_1		
NR6A1	MIR-29C	
node IPF_2		
ERG	GFI1B	PKNOX1
TFE3		
LEF1	KCNH6	EFNA2
SMAD3	KLF12	IRF8
RARB	RARG	HNF4A
ESR1	HOXB8	HOXB7
PBX3	RUNX2	PGR
TP63		
node IPF_3		
MIR-30D	MIR-30A	MIR-30B
MIR-30E	MIR-455	MIR-338
MIR-218	MIR-26B	MIR-26A
MIR-377	MIR-34C	MIR-34A
MIR-127	MIR-376A	MIR-376C
MIR-874	MIR-187	MIR-506
node IPF_4		
MIR-200A	MIR-181A	MIR-181C
MIR-135A	MIR-7G	MIR-7D
MIR-205	MIR-543	MIR-299
MIR-144	MIR-219-1	MIR-181B+MIR-181D
MIR-199B	MIR-199A+MIR-199B	MIR-199A
MIR-27B	MIR-21	LET-7I
MIR-LET-7A	MIR-155	
node A		
NFE2	LET-7G	LET-7D
LET-7A	PKNOX1	
LET-7I		
node B		
MIR-29C	NR6A1	

<table>
<tr><td>POU2AF1</td><td></td><td></td></tr>
<tr><td colspan="3" align="center"><b>node C</b></td></tr>
<tr><td>STAT3</td><td>STAT2</td><td></td></tr>
<tr><td colspan="3" align="center"><b>node D</b></td></tr>
<tr><td>MEF2A</td><td>ZNF219</td><td>FOXD1</td></tr>
<tr><td>FOXF1</td><td>MIR-130A</td><td>PPARA</td></tr>
<tr><td>ATF6</td><td>FOXD3</td><td>HNF1B</td></tr>
<tr><td>POU2AF1</td><td>HMGA2</td><td>OTX1</td></tr>
<tr><td>MIR-183</td><td>FOXC1</td><td>NKX6-1</td></tr>
<tr><td>FOXA1</td><td>ESR2</td><td>EFNA2</td></tr>
<tr><td colspan="3" align="center"><b>node E</b></td></tr>
<tr><td>STAT3</td><td>STAT5B</td><td>GFI1B</td></tr>
<tr><td>MIR-29C</td><td>TFE3</td><td>CUX1</td></tr>
<tr><td>MIR-126</td><td>MIR-141</td><td></td></tr>
<tr><td>E2F3</td><td>TFDP1</td><td>MITF</td></tr>
<tr><td>MIR-182</td><td>TFAP2C</td><td>BHLHE41</td></tr>
<tr><td>DEC1</td><td>MIR-382</td><td>MIR-376A</td></tr>
<tr><td>MIR-376C</td><td>STAT2</td><td>LEF1</td></tr>
<tr><td colspan="3" align="center"><b>node F</b></td></tr>
<tr><td>PPARA</td><td>SP3</td><td>SRY</td></tr>
<tr><td>ESR2</td><td>STAT2</td><td>TCF3</td></tr>
<tr><td>E2F3</td><td>ESR1</td><td></td></tr>
<tr><td colspan="3" align="center"><b>node G</b></td></tr>
<tr><td>MIR-30D</td><td>MIR-30A</td><td>MIR-30B</td></tr>
<tr><td>MIR-30E</td><td>LET-7G</td><td>LET-7D</td></tr>
<tr><td>LET-7A</td><td>MIR-181A</td><td>MIR-181B+MIR-181D</td></tr>
<tr><td>MIR-411</td><td>MIR-185</td><td>MIR-376A</td></tr>
<tr><td>MIR-376C</td><td>MIR-382</td><td>LET-7I</td></tr>
<tr><td>MIR-132</td><td></td><td></td></tr>
<tr><td colspan="3" align="center"><b>node H</b></td></tr>
<tr><td>LET-7G</td><td>LET-7D</td><td>LET-7A</td></tr>
<tr><td>MIR-495</td><td>MIR-34C</td><td>MIR-34A</td></tr>
</table>

LET-7I		
node I		
MIR-203	MIR-181A	MIR-181B+MIR-181D
MIR-181C		
MIR-299	MIR-543	MIR-187
MIR-506	MIR-154	MIR-323A
MIR-99A	MIR-219-1	
node J		
GATA6	GATA1	MIR-135A
CEBPB	SRY	MIR-218
FOXD1	MIR-29C	
PLAU	TFAP2A	BHLHE41
DEC1	FOXC1	MIR-495
node K		
MIR-208A	MIR-335	
MIR-377	MIR-96	MIR-378D
MIR-125B	MIR-127	MIR-183
node L		
CEBPB	NR6A1	TAL1
MIR-203		
node M		
STAT3	STAT2	MIR-185

Table 4.2: **Regulators inferred by DREM for the different split nodes in the model** If the regulator is in red, that means its expression is downregulated. If it is in blue, its expression is upregulated.

Molecular function	
IPF_1 node	
glycosaminoglycan binding sulfur compound binding fibronectin binding	heparin binding extracellular matrix structural constituent
IPF_2 node	
IPF_3 node	

cation channel activity substrate-specific transporter activity ion transmembrane transporter activity	gated channel activity metal ion transmembrane transporter activity
IPF_4 node	
GTPase binding enzyme binding Ras GTPase binding	small GTPase binding sodium channel regulator activity
Biological process	
IPF_1 node	
extracellular matrix organization cell adhesion multicellular organismal catabolic process	extracellular structure organization biological adhesion
IPF_2 node	
cell adhesion locomotion cell-cell adhesion	biological adhesion regulation of cell differentiation
IPF_3 node	
behavior neuromuscular process potassium ion transport	startle response ion transport
IPF_4 node	
organelle localization dendrite morphogenesis regulation of cellular component biogenesis	vesicle-mediated transport small GTPase mediated signal transduction
Cellular Component	
IPF_1 node	
extracellular space proteinaceous extracellular matrix collagen trimer	extracellular matrix extracellular matrix component
IPF_2 node	
proteinaceous extracellular matrix basement membrane lysosomal lumen	extracellular matrix extracellular matrix component
IPF_3 node	
NMDA selective glutamate receptor complex ion channel complex transmembrane transporter complex	neuron part chromosomal region
IPF_4 node	
endosome apical junction complex Golgi apparatus	bicellular tight junction occluding junction
Human phenotype	

IPF_1 node	
Cigarette-paper scars Osteoarthritis Premature birth following premature rupture of fetal membranes	Molluscoid pseudotumors Hyperextensibility of the knee
IPF_2 node	
IPF_3 node	
IPF_4 node	
Disease	
IPF_1 node	
Degenerative polyarthritis Idiopathic Pulmonary Fibrosis Adenocarcinoma	Pulmonary Fibrosis Hamman-Rich syndrome
IPF_2 node	
Pancreatic carcinoma Liver Cirrhosis, Experimental Malignant tumor of colon	Mammary Neoplasms Malignant neoplasm of pancreas
IPF_3 node	
Mental Retardation Seizures Dull intelligence	Intellectual Disability Schizophrenia
IPF_4 node	
Small for gestational age (disorder)	Low Birth Weights

Table 4.3: **Gene enrichment categories for each split** Only the up to the top 5 categories were displayed.

IPF_1 node
Genes down-regulated in the luminal B subtype of breast cancer Invasiveness signature resulting from cancer cell/microenvironment interaction Genes up-regulated in invasive ductal carcinoma (IDC) relative to ductal carcinoma in situ (DCIS, non-invasive) Genes up-regulated in papillary thyroid carcinoma (PTC) compared to normal tissue
IPF_2 node
Human Breast genes Human Embryonic StemCell genes Genes up-regulated in uterus upon knockout of BMP2 Genes forming the macrophage-enriched metabolic network (MEMN) claimed to have a causal relationship with the metabolic syndrome traits Human Sarcoma genes
IPF_3 node
Human Mesenchymal Stem Cells

The 'Cervical Cancer Proliferation Cluster' (CCPC): genes whose expression in cervical carcinoma positively correlates with that of the HPV E6 and E7 oncogenes; they are also differentially expressed according to disease outcome
Genes up-regulated in B lymphocytes: control versus stimulated by anti-IgM for 12h
Genes up-regulated during later stage of differentiation of Oli-Neu cells (oligodendroglial precursor) in response to PD174265
Human Sarcoma genes
IPF_4 node
Genes down-regulated in erythroid progenitor cells from fetal livers of E13.5 embryos with KLF1 knockout compared to those from the wild type embryos
Genes down-regulated in ME-A cells (breast cancer) undergoing apoptosis in response to doxorubicin
Genes down-regulated in fibroblasts expressing mutant forms of ERCC3 after UV irradiation
Human immune genes
Mouse lung genes

Table 4.4: **Coexpressed genes.**

4.3.3 Incorporating DNA methylation into the model

Our general approach for this task was to train the DanQ [210] neural network model on ENCODE data and generate celltype independent predictions (using just the DNA sequence as input) to generate the TF-DNA interaction networks. As the transcription factors in ENCODE differed substantially from the ones used in the previous section and several of them were complexes making it unclear how to filter for differential expression, we decided to not do any filtering on the TFs. The set of microRNAs that were considered as potential regulators as well as the set of genes remained the same as before.

Methylation data analysis

The methylation data was collected using Illumina's Infinium EPIC array [203] which is a whole genome bisulphite sequencing (WGBS) technology. GenomeStudio's Methylation module [30] was used to analyze the data and detect the level of methylation.

TF-gene network construction

We used the DanQ [210] network that tries to predict the binding for each TF-celltype combination for a candidate 200 bp genome sequence. The only change we made was that instead of predicting the binding scoe for each TF-celltype combination, we predicited the same binding score for a TF for all celltypes. The architecture is shown in Figure 4.4. As described before, it took as input, the one hot encoded version of the DNA sequence. It then applied a convolutional layer with 300 filters and receptive field of size 26. Each filter is activated by a subset of sequences of size 26 and thus represents a different "motif". We then applied a max pooling layer of size 13 thus downsampling the output of the convolutional layer by 13. We then applied

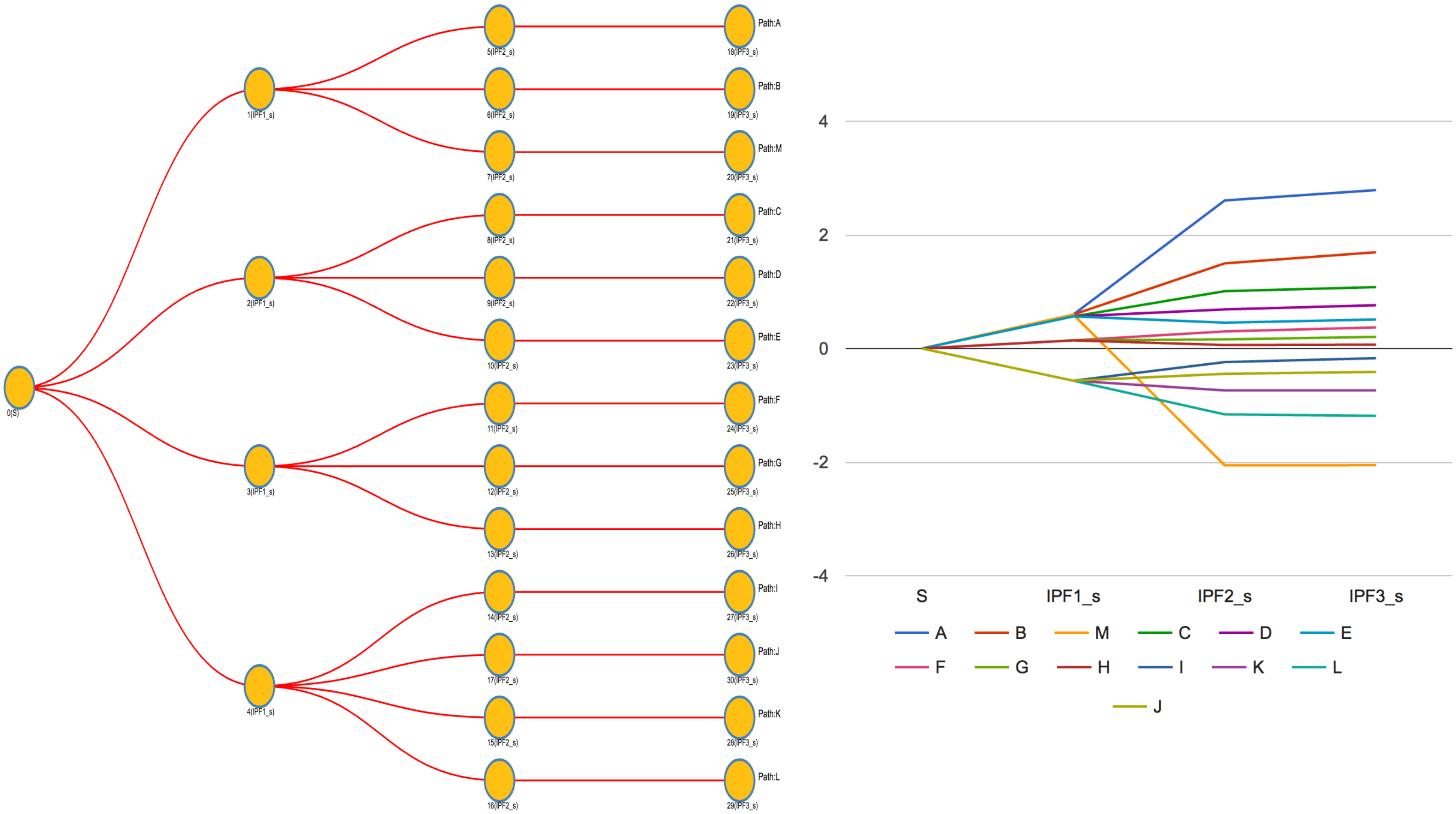

Figure 4.2: **Visualization of the pathways identified by the DREM model for IPF.**

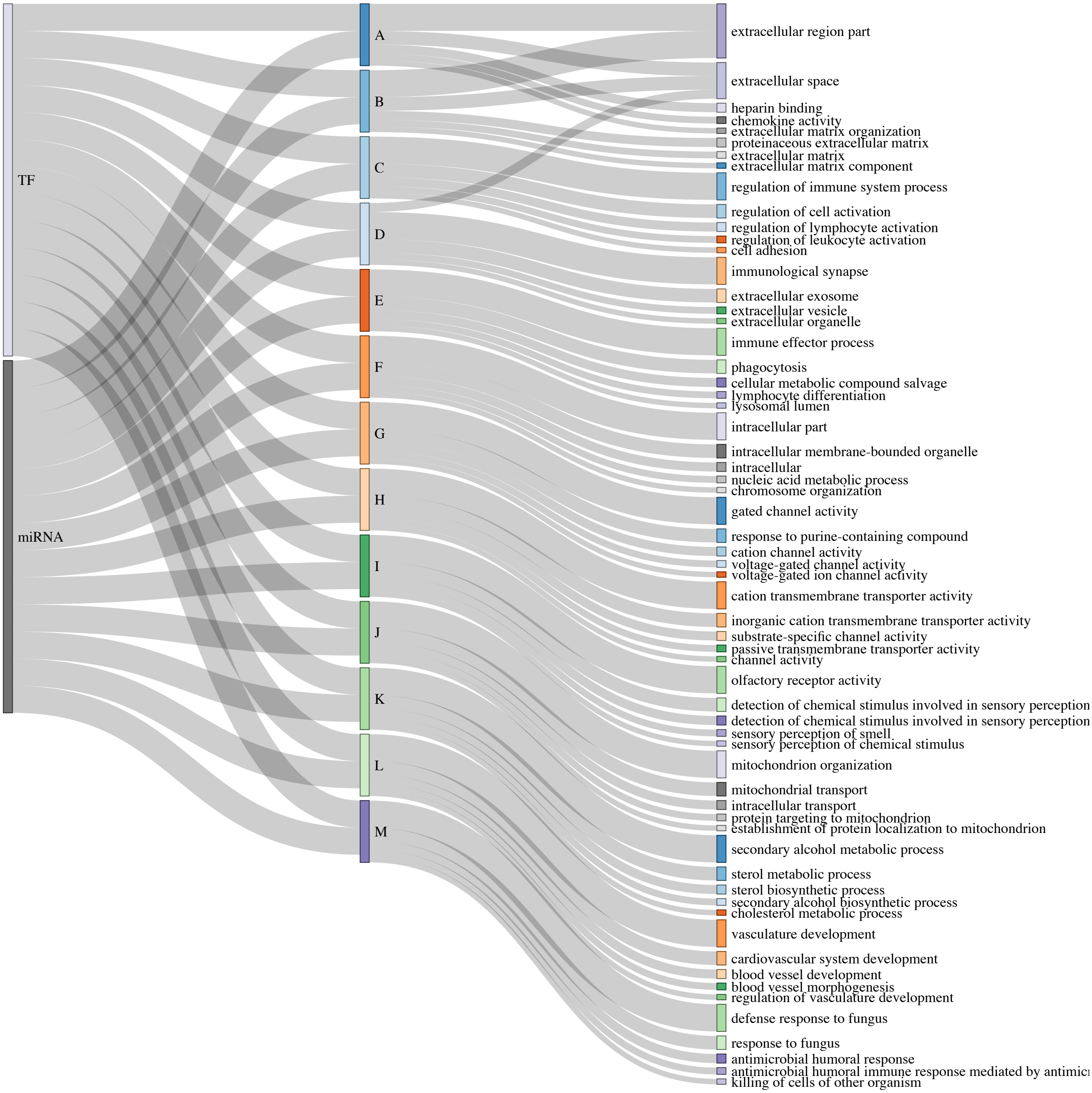

Figure 4.3: **Sankey diagram for the DREM IPF model.** This is a Sankey diagram. Sankey diagrams are a specific type of flow diagram, in which the width of the arrows is shown proportionally to the flow quantity. In this illustration, the width is proportional to the number of genes associated with the flow. A visualization with the paths labeled (A through M) is shown in Figure 4.2. On the right hand side, the biological processes associated with each flow are shown.

a Bi-LSTM layer for the same rationale as the original DanQ network – i.e. motifs can follow a regulatory grammar governed by physical constraints that dictate the in vivo spatial arrangements and frequencies of combinations of motifs, a feature associated with tissue-specific functional elements such as enhancers [209, 212]. Finally we applied a dense layer with output size 925 and then a sigmoid layer to predict the binding scores (between 0 and 1) for each TF. The sigmoid layer was of size 180 (same as the number of TFs) which differed from the sigmoid layer size of the DanQ network (909) as we were trying to predict the scores for each TF regardless of celltype rather than each TF-celltype combination.

The network parameters were initialized using the Glorot initialization scheme [96]. Apart from the sigmoid layer at the very end, all of the other activation functions were rectified linear units [186]. The step size scheme used was Adam [139].

We used this network to generate binding predictions (for 200 bp non-overlapping segments) for the promoter section of all genes in the genome (where the promoter region was defined as ± 10 KB of the transcription start site (TSS) of the gene. The prediction scores were then summed for every TF-gene pair across the promoter sequence as follows :-

$$\frac{\sum_{w \in W} NN_t(g_w)}{\sum_{t \in T} \sum_{w \in W} NN_t(g_w)} \forall t \in T$$

where W is the set of non-overlapping 200bp genome windows that cover the promoter region, NN_t is the binding score the neural network outputs for TF t, T is the set of all TFs, and g_w is the genome sequence associated with the window w.

To incorporate methylation information, we assumed that methylation was in general, inhibitory with respect to binding [178, 187], and thus for each segment for which we predicted the binding score, we set the score to 0 if the methylation level for that segment was above a certain threshold.

We then took the two TF-gene networks we so generated and ran the regulatory network inference component of our method (DREM) to generate models for both networks.

Results

The results are presented in Figures 4.5 and 4.6. The results **with** methylation incorporated are in Model A and the ones without are in Model B. In Table 4.5, for each split at the first time phase, we generated the ***unique*** list of regulators that each model predicted as regulating the expression of genes at that split (by unique we mean that we present the list of regulators that only Model A predicts for a split and similarly for Model B).

As shown in the table, Model A is able to capture a large number of B/T-cell regulators that Model B is not, including BCL11A, PRDM1, and others. It is also able to infer immune response regulators like IRF3. Of particular note are the regulators TAF1 and HDAC2 which regulate PEDF (SERPINF1) induced signaling. This is an inhibitor of angiogenesis and matches the decline of blood vessel genes in our data. It also finds the miRNA MIR-21 which is known to mediate the fibrogenic activation of pulmonary fibroblasts and lung fibrosis [166]. There is also considerable evidence that the activation TGF-beta pathway is related to pulmonary fibrosis [42]. Model A detects ZNF217 which is known to be a suppressor of the TGF-beta pathway.

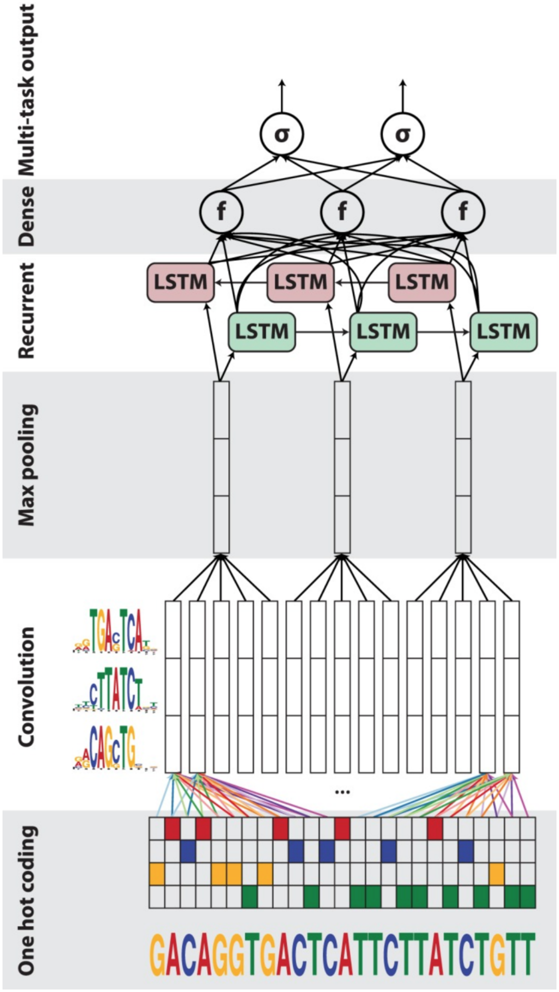

Figure 4.4: **Neural network used by DanQ and in our models**. Neural network for predicting TF binding based on sequence. It shows the one hot encoding of the DNA sequence following by a convolution layer, then a max pooling layer, a recurrent bi-lstm layer, a dense layer, and finally an output layer. Image taken from [210]

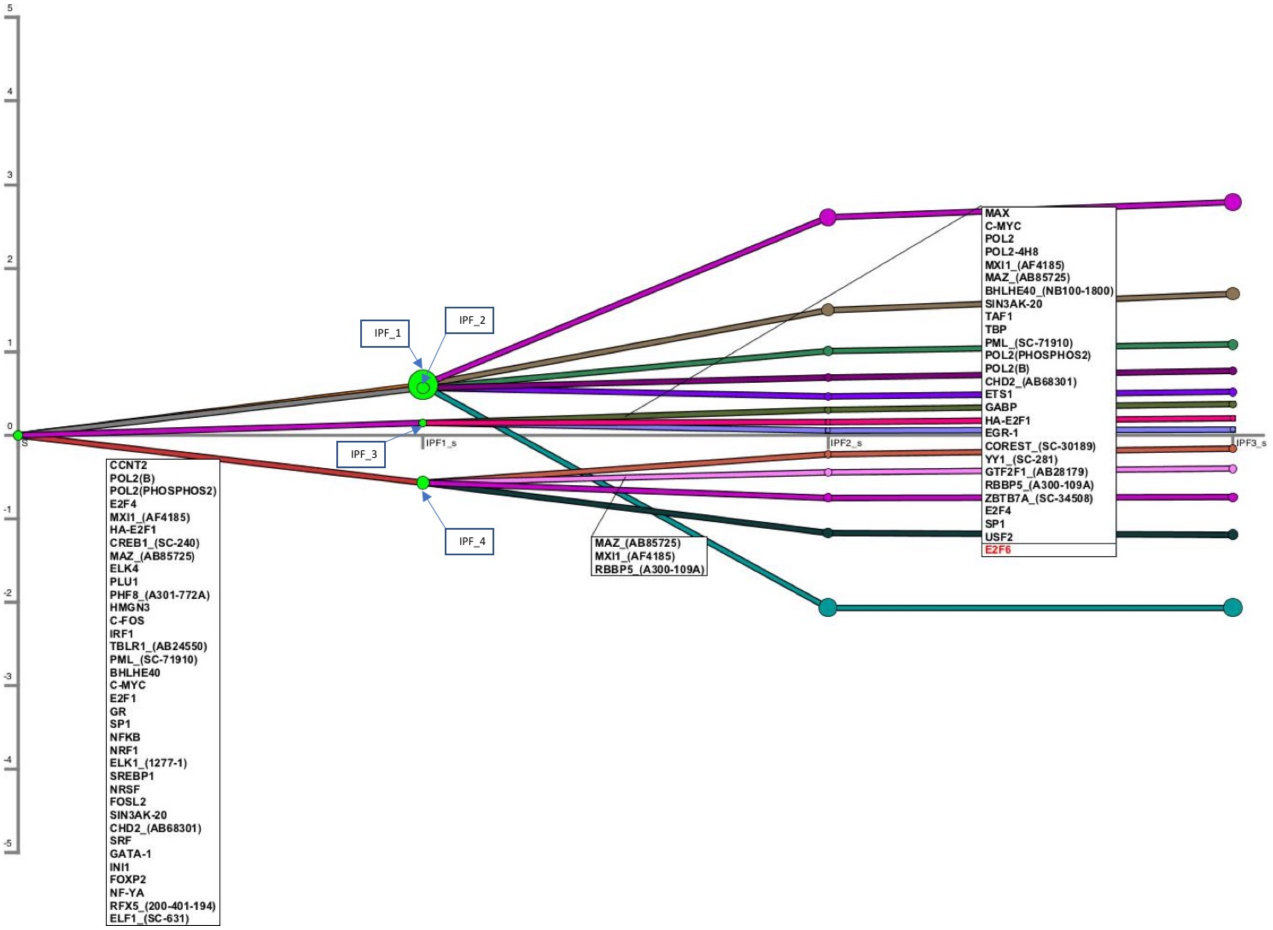

Figure 4.5: **DREM regulatory network model when methylation is incorporated.**

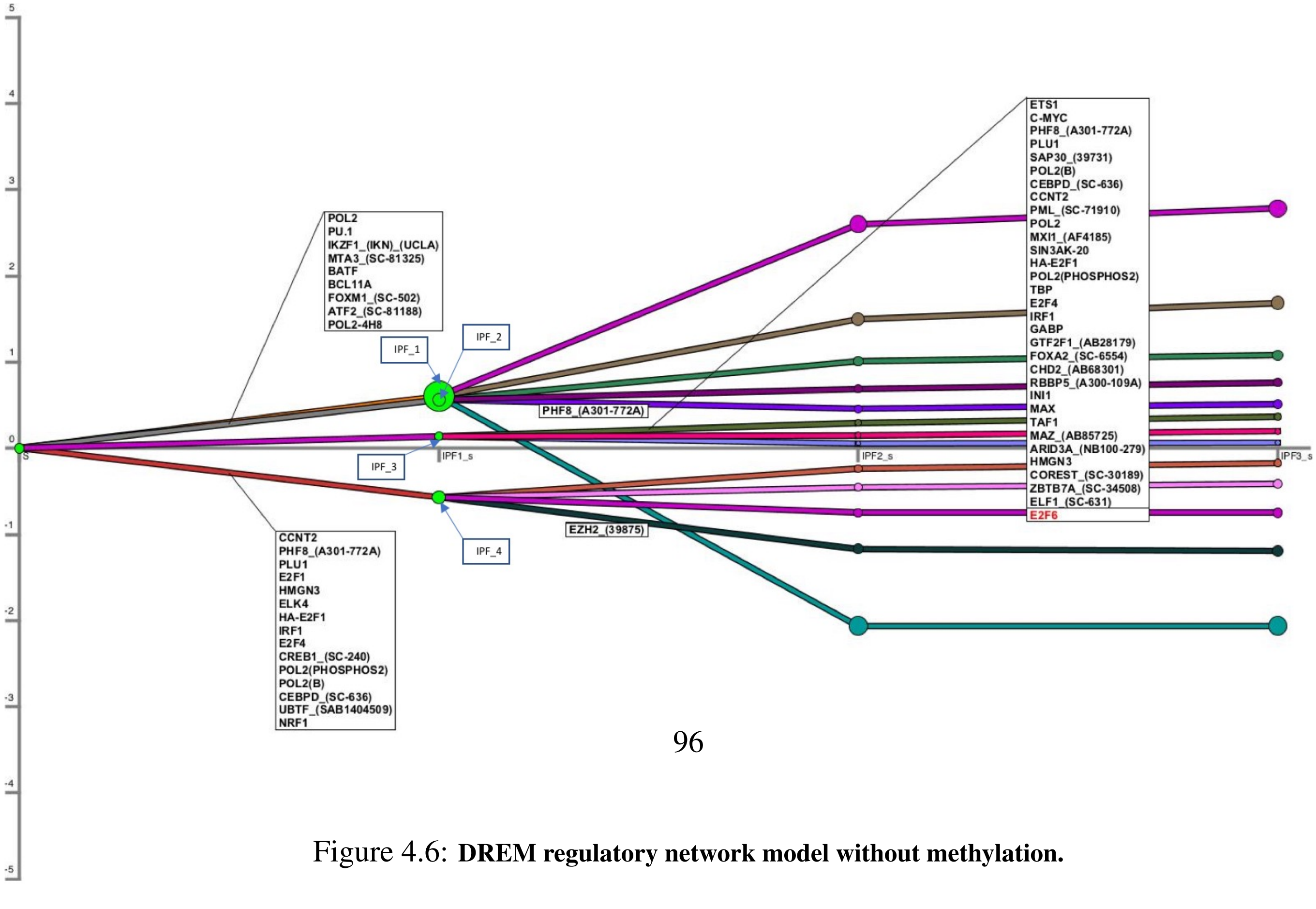

Figure 4.6: **DREM regulatory network model without methylation.**

Regulator	**Function**
IPF_1 node	
Model A	
HSA-MIR-20A-5P+HSA-MIR-20B-5P	DNA damage response
HSA-MIR-423-5P	vascular
HNF4A_(SC-8987)	mesoderm development
BCL11A	Bcell formation
HSA-MIR-21-5P	autophagy
ERRA	estrogen receptor related
SUZ12	senescence
TFIIIC-110	invovled in functional RNA transcription
HSA-MIR-181A-5P	DNA damage response HSC differentiation
HSA-MIR-181B-5P+HSA-MIR-181D	DNA damage response
HSA-MIR-181C-5P	cell differentiation
HSA-MIR-378D	NA
Model B	
EGFP-FOS	TGFbeta pathway signaling
MAFK_(AB50322)	binds FOS platelets and fibrosarcoma
MAFF_(M8194)	cellular stress response
MAFK_(SC-477)	binds FOS platelets and fibrosarcoma
TAL1_(SC-12984)	hemopoietic differentiation
ZNF274_(M01)	transcription repressor
EGFP-JUNB	TGFbeta pathway signaling
ZNF274	transcription repressor
P300	hypoxia related histone modifier
C-JUN	TGFbeta pathway signaling
BDP1	transcription
ZNF217	cell proliferation
IPF_2 node	
Model A	
MAFK_(AB50322)	binds FOS platelets and fibrosarcoma
RAD21	DNA damage response
PRDM1_(9115)	mature B cell beta IFN promoter
ZNF217	cell proliferation
EGFP-FOS	TGFbeta pathway signaling
TAF1	PEDF induced signalling
HDAC2_(A300-705A)	PEDF induced signalling
IRF3	interferon regulatory
HSA-MIR-96-5P	cancer
FOSL1_(SC-183)	cell proliferation differentiation
HSA-MIR-183-5P	cancer
TCF7L2	maintain epithelial stem cell
SUZ12	senescence

TBP	transcription
Model B	
IKZF1_(IKN)_(UCLA)	lymphocyte development
MTA3_(SC-81325)	maintain epithelial architecture
BATF	senescence DNA damage response CD8 Tcell
BCL11A	Bcell formation
FOXM1_(SC-502)	cell proliferation DNA damage reponse
ATF2_(SC-81188)	DNA damage response
EBF1_(SC-137065)	lipid metabolism B-cell
RUNX3_(SC-101553)	TGFbeta induced CDKN1A
IRF4_(SC-6059)	BATF function MHC1 regulator
WHIP	DNA damage response
POU2F2	Ig regulator
P300	hypoxia related histone modifier
TCF12	general transcription factor including B/T cells
ELF1_(SC-631)	lymphoid function
IPF_3 node	
Model A	
HSF1	negative reglator DNA damage repair
ZZZ3	chromatin organization
HSA-MIR-455-5P	NA
MAFK_(SC-477)	binds FOS platelets and fibrosarcoma
FOXM1_(SC-502)	cell proliferation DNA damage reponse
TEAD4_(SC-101184)	Hippo signalling
Model B	
EZH2_(39875)	senescence
JARID1A_(AB26049)	histone demethylase AR reponse gene
HSA-MIR-338-3P	Parkinson disease
HSA-MIR-506-3P	NA
HSA-MIR-218-5P	NA
TFIIIC-110	involved in functional RNA transcription
IPF_4 node	
Model A	
MXI1_(AF4185)	HSC differentiation
MAZ_(AB85725)	Regulates inflammation-induced expression of
serum amyloid A proteins	
C-FOS	TGFbeta pathway signaling
TBLR1_(AB24550)	proteasome
PML_(SC-71910)	senescence DNA damage response antiviral
BHLHE40	chondrocyte differentiation
C-MYC	activate growth related genes
Model B	
CEBPD_(SC-636)	works with C-MYC

UBTF_(SAB1404509)	transcription
NRF1	mitochondrial function
GTF2F1_(AB28179)	PEDF induced signalling
INI1	PEDF induced signalling Regulate chromatin
SREBP1	sterol biosynthesis
SIX5	organogenesis

Table 4.5: **Unique regulators for each split.**

4.4 Conclusion

In this chapter we tried to incorporate DNA methylation data in the reconstruction of regulatory networks. Our initial attempts involved using DNA methylation to improve prediction of TF binding using both simple models like logistic regression as well as more sophisticated, neural network based models. Unfortunately we were unable to get any signal out of DNA methylation that improved TF binding prediction. In the course of our experiments, we noticed that methylation of distal enhancer regions associated with a gene was correlated with the expression of that gene. We wondered if that meant that the gene's expression could improve the TF binding prediction for its enhancer regions better than the first few components of the PCA transformed expression data but unfortunately that was not the case.

Finally we applied our regulatory network inference model to idiopathic pulmonary fibrosis (IPF) which yielded several novel and interesting predictions. We also incorporated methylation into the protein-DNA interaction network which yielded a model notably different from the one that did *not* incorporate methylation and recovered several interesting regulators that the method that ignored methylation did not yield.

An interesting direction to go into would be to examine single cell data for IPF. This would enable us to have a much deeper understanding of what part of the enrichment we observe above is due to an increased population of fibroblasts or loss of epithelium and what part is due to regulatory changes in the other cells. It would also help develop a better understanding of extracellular signaling.

Chapter 5

Conclusion

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.

5.1 Summary of contributions

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) incorporated DNA methylation into our models in order to better infer the signaling and regulatory networks and validated the approach via application to idiopathic pulmonary fibrosis and comparing the model that ignores methylation versus one that does not.

5.1.1 MT-SDREM

We developed MT-SDREM a multi-task learning framework that simultaneously reconstructs signaling and dynamic regulatory networks across related conditions. Given the small number of condition-specific samples that are often available (i.e. time series expression data and host-pathogen interaction data) sharing parameters across related conditions allows the reconstruction of more accurate networks while still retaining the ability to explain condition-specific signaling and regulation.

We applied MT-SDREM to reconstruct networks for 3 related influenza A virus infections – H1N1, H3N2, and H5N1. The resulting signaling and regulatory networks were able to identify several known and novel regulators of immune and viral response. Many of these were shared between all condition including PPARG, FOS, ATF, and JUN. Similarly, we identify key signaling proteins, some shared by all conditions while others are unique to one or two of the conditions. Specifically, we identified the signaling protein SUMO1 as part of pathway from UBE2I for all 3 conditions. This agrees with recent findings that UBE2I interacts with SUMO1 to degrade influenza A's virus, NS1 which is present in all three strains [109]. We also identified the AKT1 gene, part of the PI3K/AKT pathway that is activated by NS1 in all conditions.

MT-SDREM is the first method to jointly reconstruct such dynamic networks. Comparing MT-SDREM with methods that have been suggested to integrate gene expression data or with methods reconstruct such networks independently for each condition highlighted the advantages of multi-task network learning. MT-SDREM outperformed previous methods in identifying a set of TFs controlling immune response, a set of functionally relevant proteins and a set of proteins whose knockdown affects viral loads.

5.1.2 TimePath

Since most of the high throughput data used to reconstruct cellular response networks is static, current models based on these data are often unable to provide specific *temporal* hypotheses regarding the effects of perturbations and drugs on cellular responses. Here we formulated a new Integer Programming (IP) optimization function to connect observed temporal responses (from gene expression data) with the underlying sources, to further identify the pathways and transcription factors that activate them. We then use the pathways and their predicted time to reconstruct the full response network leading to insights regarding the propagation of cellular responses, key proteins controlling the responses and testable hypothesis regarding the effects of perturbing proteins at various time points following infection.

Applying TimePath to model HIV response networks led to the identification of known and novel proteins and miRNAs for the HIV response pathways. The reconstructed network explains the roles of several HIV screen hits, the function of TFs and miRNA controlling expression levels and is enriched for functional categories related to immune and viral responses.

The pathways identified can be divided to those induced by the virus to promote survival/replication and those induced by the host to curtail virus infection and promote cellular survival. Our temporal regulatory model indicates that these can also be divided based on their dynamics.

Follow up experiments using inhibitors confirmed the prediction of TimePath, where 11 of the 22 predicted proteins (that were evaluated in the experiment) were identified to have a role in HIV infection. NFKB and related genes are exclusively essential for virus infection in the initial phase as predicted by TimePath, similarly, RAF1 was also confirmed to have an important role in the initial phase. As predicted by TimePath, these genes may either be required for virus infection during the initial phase, or the changes triggered by these genes in the initial phase can temporally affect downstream events that are essential for virus infection. It is also noted that CDKs, STATs and proteasomal machinery are essential during all phases of HIV infection, and TimePath had predicted a role for these genes starting with phase 1 (CDKs) and/or a combination of phases -

phase 1 and phase 2 (STATs) or phase 3 (proteasomal machinery and related processes). Though TimePath identifies the role for these genes or processes in specific phase, it suggests that the event occurs at the identified phase; however, it does not rule out that the events are continuing over time and have a role in later stages too.

Unlike other methods that attempt to link treatments to disease stages (for example, in cancer which uses pathological analysis to determine tumor grades) TimePath is fully based on the molecular data, thus could be applied to much shorter time scales. This approach enables the programme to obtain a more fine resolution of the disease stage, which cannot be observed by other methods. With higher resolution, it may be possible to use TimePath to tailor appropriate treatment options to treat infected individuals.

5.1.3 Application to IPF and incorporating DNA methylation data

We applied the DREM algorithm to construct a disease progression model for idiopathic pulmonary fibrosis (IPF). We analyzed the results and presented evidence that the model recovers existing biology and found potential new targets like POU2AF1 to explore.

Finally, we wanted to test the usefulness of epigenetic data in the reconstruction of regulatory networks. In particular, we wanted to use DNA methylation to improve the transcription factor (TF)-DNA interaction network that we used as input to our signaling and regulatory network inference models. We presented several different ideas on how to use DNA methylation data to try and get better TF-DNA binding predictions. While we were not been able to directly get better TF-DNA binding predictions using ENCODE data, we attempted to assess the performance by looking directly at the regulatory networks generated for idiopathic pulmonary fibrosis (IPF). We tried two models – one that ignored methylation, and the other that assumed methylation inihibited binding and obtained some evidence that the model that incorporates methylation uncovered some biological aspects of the disease that the model that ignored methylation did not.

5.2 Future directions

In this section, we talk about possible future directions to extend our work.

5.2.1 Extensions to MT-SDREM

More granular sharing of parameters

While we have built a multitasking model to jointly model the signaling networks for different but related conditions (Chapter §2), the joint modeling of the regulatory network is still global (via sharing of prior scores for the transcription factors (TFs) predicting to play a role in different conditions) i.e. across all time points. An interesting extension would be to allow for more granular sharing of TF priors such that splits representing the same or similar time would be more likely to share TFs compared to other splits.

Learning priors on regulatory program

In the model presented thus far, MT-SDREM is typically used to model only a small set of related tasks. It could be very useful however, to learn the prior joint probability on the activity of transcription factors. As an example, if we have learnt the prior that that TFs A and B are often active together, then if we infer that A is likely to be active for a particular condition, we could increase the prior probability that B is active in that condition as well. This would effectively allow us to do transfer learning across a large number of conditions, related or unrelated.

To incorporate this into DREM (the regulatory network inference component of MT-SDREM), we would want to have a function $\mathbb{P}$ that outputs the probability of any input TF activity vector. This would form part of the DREM objective, penalizing any TF activity vectors that deviate from the joint prior. It is important to note that we would need to know the normalization constant for $\mathbb{P}$ in order to set the magnitude of the penalty correctly.

Learning how related the tasks are

At the moment, we have a parameter in our model (α, see Chapter §2) which regulates how correlated the TF priors between the different tasks end up being. However one could imagine various schemes to learn this automatically. One possibility could be to compute the set of differentially expressed (DE) genes for the different conditions and then set the parameter as function of how many DE genes are common between the different conditions.

5.2.2 Extensions to TimePath

Multitask extension

For TimePath, we have tried to model only one task at a time. Extending it to a multitask setting (as we extended SDREM to MT-SDREM) would be useful. The simplest extension would be to share priors for the activity of different proteins in the different conditions.

Incorporating post-translational modifications

Post-translational modifications (PTMs) of proteins are the modification of proteins after they are translated. They usually occur on amino acid side chains or at the protein's C or N termini [273]. Protein phosphorylation, which involves the addition of a phosphoryl group to the protein is a common post-translational modification. Other common PTMs are glycosylation, lipidation, etc.

PTMs can cause structural changes to the protein structure and are often necessary to activate signaling pathways. For example, the enzyme GSK-3 is phosphorylated by AKT as part of the insulin signaling pathway [267]. Histone acetylation/deacylation is very important in regulating transcription [87].

Currently, TimePath has no mechanism to incorporate data on PTMs. As mentioned above, however, post-translational modification of a protein can be often responsible for the activation of that protein in a signaling pathway. Thus a simple way to incorporate PTM data might be to look at proteins that are differentially phosphorylated between two different time points, and increase the prior probability of those proteins being active between those two time points in proportion to the how strong the signal for differential phosphorylation is.

5.2.3 Using histone modifications to predict TF binding

Recently there has been work on applying deep learning to histone modification data. Singh et al. [240] use neural networks to try and predict gene expression as a function of histone modifications. DeepBind [12] and others [302] try and predict histone modifications as a function of sequence. Benveniste et al. [27] use *TF binding* to try and predict histone modifications.

In this thesis, we only looked at the possibility of using DNA methylation to improve upon predictions of TF binding. While less widely applicable (as discussed in Chapter §4), it would nevertheless be interesting to see whether histone modification data can be used to predict TF binding. A simple model might be to append histone modification features to the sequence features, suitably extend existing networks, and see if one obtains any improvement. Histone state can change rapidly in response to change in a cell's condition so being able to predict TF binding as a function of that, apart from being interesting from a scientific perspective, could also give a lot of insight as to the active TFs for a particular condition.

5.2.4 Single cell extensions

Most microarray and RNA-Seq studies to date have focused on profiling large populations of cells. While such approaches have led to many important results, they tend to overlook the heterogeneity of the population being profiled [246]. This may be problematic in cases where the population contains a mixture of different cell types with different regulatory programs (for example, in cancer samples [65] or when studying immune response [232] and development [262]). In such cases, expression experiments that profile populations along the differentiation trajectories may not be able to identify the specific regulatory networks that lead to the desired cellular fate.

Recently, new technologies based on RNA-Seq experiments have been developed to profile global gene expression in single cells. By profiling different cells in the population the contribution of different cell types to changes in tissue level expression can be analyzed allowing researchers to address several of the problems mentioned above. However, the single cell based approaches have also raised new computational challenges leading to new methods for the analysis of such data. These include issues related to sample quality, issues related to normalization of single cell data (which is more challenging, especially for lowly expressed genes [234, 287]), and the development of clustering methods to identify the different components within a specific mixture/time point [45].

Thus a potential future direction would be to develop methods to be able to construct a 'time series' from single cell expression data and then perform network reconstruction based on that time series. In particular, it would be useful to be able to *jointly* model the regulatory network reconstruction for single cell data one one and and single cell clustering (where cells are grouped together for each time point and linked up to other groups at adjacent time points). One possible route to accomplish this might be to first cluster the single cells into different clusters, then try and infer regulatory networks. We then iterate between the two steps, allowing cells to switch into a different cluster if the regulatory program for that cluster represents the cell better until we reach a fixed point.

5.2.5 Joint learning of network and interactions

So far we have looked at either inferring the regulatory and signaling network while assuming a fixed TF-gene interaction network or learning the interaction network while ignoring the regulatory and signaling network. An interesting direction would be to *jointly* learn the signaling/regulatory networks as well as the TF-gene interaction network. The simplest way to do a joint learning would be to use the neural network model described previously (or some other model) to predict TF-DNA interaction, aggregate the predictions across the promoter/enhancer of a gene, and thus get the TF-gene interaction prediction. These predictions could then be fed into DREM with the TF-gene interactions being variables through which we backpropagate the error instead of them being fixed. The neural network would be pre-trained on TF-DNA interaction data derived from Chip-Seq and fine tuned in the course of this procedure. As such it would be the reverse of the standard paradigm where supervised fine tuning follows unsupervised training.

The biggest challenge with such a model would be computational. The logistic classifier has to be retrained for every maximization step in the Baum-Welch algorithm which would also entail fine tuning the neural network for every such iteration. One way to resolve this might be to parallelize the training. One task that could be easily parallelized would be the prediction of TF-DNA interactions for the promoter/enhancer sections of each gene/TF combination (and the backpropagation of errors through them).

Bibliography

[1] Supporting website for mt-sdrem. http://sb.cs.cmu.edu/mtsdrem/, 2015. 2.9.1

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

[3] Supporting website for timepath. http://sb.cs.cmu.edu/timepath/, 2015. 3.2.4

[4] Fact sheet latest statistics on the status of the aids epidemic — unaids. www.unaids.org, 2017. 1.3.2

[5] Artificial neural network. https://en.wikipedia.org/wiki/File:Colored_neural_network.svg, 2017. 1.4

[6] Reduced representation bisulfite sequencing. https://en.wikipedia.org/wiki/Reduced_representation_bisulfi 2017. 1.4.3

[7] David S Aaronson and Curt M Horvath. A road map for those who don't know jak-stat. *Science*, 296(5573):1653–1655, 2002. 2.9.1

[8] 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. 2.10.2

[9] Suneet Agarwal and Anjana Rao. Modulation of chromatin structure regulates cytokine gene expression during t cell differentiation. *Immunity*, 9(6):765–775, 1998. 1.4.4

[10] Vikram Agarwal, George W Bell, Jin-Wu Nam, and David P Bartel. Predicting effective microrna target sites in mammalian mrnas. *elife*, 4:e05005, 2015. 4.3.1

[11] Frank W Albert, Mehmet Somel, Miguel Carneiro, Ayinuer Aximu-Petri, Michel Halbwax, Olaf Thalmann, Jose A Blanco-Aguiar, Irina Z Plyusnina, Lyudmila Trut, Rafael Villafuerte, et al. A comparison of brain gene expression levels in domesticated and wild animals. *PLoS genetics*, 8(9):e1002962, 2012. 2.10

[12] 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. 4.1, 5.2.3

[13] Mariano J Alvarez, Yao Shen, Federico M Giorgi, Alexander Lachmann, B Belinda Ding, B Hilda Ye, and Andrea Califano. Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nature genetics*, 48(8):838–847, 2016. 1.6.3

[14] Tatyana Ammosova, Reem Berro, Fatah Kashanchi, and Sergei Nekhai. Rna interference directed to cdk2 inhibits hiv-1 transcription. *Virology*, 341(2):171–178, 2005. 3.2

[15] Tatyana Ammosova, Reem Berro, Marina Jerebtsova, Angela Jackson, Sharroya Charles, Zachary Klase, William Southerland, Victor R Gordeuk, Fatah Kashanchi, and Sergei Nekhai. Phosphorylation of hiv-1 tat by cdk2 in hiv-1 transcription. *Retrovirology*, 3(1): 78, 2006. 3.2

[16] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome biol*, 11(10):R106, 2010. 3.1.2, 3.2.1

[17] JL Andersen, ES Zimmerman, JL DeHart, S Murala, O Ardon, J Blackett, J Chen, and V Planelles. Atr and gadd45α mediate hiv-1 vpr-induced apoptosis. *Cell Death & Differentiation*, 12(4):326–334, 2005. 3.4

[18] Sameer A Ansari, Mahmut Safak, Gary L Gallia, Bassel E Sawaya, Shohreh Amini, and Kamel Khalili. Interaction of yb-1 with human immunodeficiency virus type 1 tat and tar rna modulates viral promoter activity. *Journal of general virology*, 80(10):2629–2638, 1999. 3.2

[19] Sanjeev Arora and Boaz Barak. *Computational complexity: a modern approach*. Cambridge University Press, 2009. 3.1.3

[20] 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. 1.4.8

[21] Ferhat Ay, Timothy L Bailey, and William Stafford Noble. Statistical confidence estimation for hi-c data reveals regulatory chromatin contacts. *Genome research*, 24(6):999–1011, 2014. 4.2.2

[22] Vinod RMT Balasubramaniam, Tham Hong Wai, Bimo Ario Tejo, Abdul Rahman Omar, and Sharifah Syed Hassan. Highly pathogenic avian influenza virus nucleoprotein interacts with trex complex adaptor protein aly/ref. *PloS one*, 8(9):e72429, 2013. 2.9.1

[23] Ziv Bar-Joseph, Anthony Gitter, and Itamar Simon. Studying and modelling dynamic biological processes using time-series gene expression data. *Nature Reviews Genetics*, 13 (8):552–564, 2012. 1.6.2

[24] Matjaz Barboric, Fan Zhang, Mojca Besenicar, Ana Plemenitas, and B Matija Peterlin. Ubiquitylation of cdk9 by skp2 facilitates optimal tat transactivation. *Journal of virology*, 79(17):11135–11141, 2005. 3.2

[25] James T Becker, Lawrence A Kingsley, Samantha Molsberry, Sandra Reynolds, Aaron Aronow, Andrew J Levine, Eileen Martin, Eric N Miller, Cynthia A Munro, Ann Ragin, et al. Cohort profile: Recruitment cohorts in the neuropsychological substudy of the multicenter aids cohort study. *International journal of epidemiology*, 44(5):1506–1516, 2014. 3.3.1

[26] Yoshua Bengio and Paolo Frasconi. An input output hmm architecture. In *Advances in neural information processing systems*, pages 427–434, 1995. 2.1

[27] Dan Benveniste, Hans-Joachim Sonntag, Guido Sanguinetti, and Duncan Sproul. Transcription factor binding predicts histone modifications in human cell lines. *Proceedings of the National Academy of Sciences*, 111(37):13367–13372, 2014. 5.2.3

[28] 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. 2.10.3, 2.4, 2.5, 3.2.4, 3.7

[29] 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. 2.10.3

[30] Marina Bibikova, Bret Barnes, Chan Tsan, Vincent Ho, Brandy Klotzle, Jennie M Le, David Delano, Lu Zhang, Gary P Schroth, Kevin L Gunderson, et al. High density dna methylation array with single cpg site resolution. *Genomics*, 98(4):288–295, 2011. 4.3.3

[31] Adrian P Bird. Cpg-rich islands and the function of dna methylation. *Nature*, 321(6067): 209–213, 1986. 1.4.3

[32] Elizabeth M Blackwood and Robert N Eisenman. Max: a helix-loop-helix zipper protein that forms a sequence-specific dna-binding complex with myc. *Science*, 251(4998):1211–1217, 1991. 3.2

[33] Cheryl Bolinger, Amit Sharma, Deepali Singh, Lianbo Yu, and Kathleen Boris-Lawrie. Rna helicase a modulates translation of hiv-1 and infectivity of progeny virions. *Nucleic acids research*, page gkp1075, 2009. 3.2

[34] 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, 2.10.1

[35] 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. 1.6.1

[36] Eric Bortz, Liset Westera, Jad Maamary, John Steel, Randy A Albrecht, Balaji Manicassamy, Geoffrey Chase, Luis Martínez-Sobrido, Martin Schwemmle, and Adolfo García-Sastre. Host-and strain-specific regulation of influenza virus polymerase activity by interacting cellular proteins. *MBio*, 2(4):e00151–11, 2011. 2.10

[37] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010. 1.5.4

[38] Michael Boutros and Julie Ahringer. The art and design of genetic screens: Rna interference. *Nature Reviews Genetics*, 9(7):554–566, 2008. 1.4.7, 1.3

[39] Abraham L Brass, I Huang, Yair Benita, Sinu P John, Manoj N Krishnan, Eric M Feeley, Bethany J Ryan, Jessica L Weyer, Louise van der Weyden, Erol Fikrig, et al. The ifitm proteins mediate cellular resistance to influenza a h1n1 virus, west nile virus, and dengue virus. *Cell*, 139(7):1243–1254, 2009. 2.10

[40] Vanessa Brès, Tomonori Yoshida, Loni Pickle, and Katherine A Jones. Skip interacts with c-myc and menin to promote hiv-1 tat transactivation. *Molecular cell*, 36(1):75–87, 2009. 3.2

[41] Lars Brichta, William Shin, Vernice Jackson-Lewis, Javier Blesa, Ee-Lynn Yap, Zachary Walker, Jack Zhang, Jean-Pierre Roussarie, Mariano J Alvarez, Andrea Califano, et al. Identification of neurodegenerative factors using translatome-regulatory network analysis. *Nature neuroscience*, 18(9):1325–1333, 2015. 1.6.3

[42] Thomas J Broekelmann, Andrew H Limper, Thomas V Colby, and John A McDonald. Transforming growth factor beta 1 is present at sites of extracellular matrix gene expression in human pulmonary fibrosis. *Proceedings of the National Academy of Sciences*, 88 (15):6642–6646, 1991. 4.3.3

[43] Joseph T Bruder, Gisela Heidecker, Tse-Hua Tan, John C Weske, David Derse, and Ulf R Rapp. Oncogene activation of hiv-ltr-driven expression via the nf-b binding sites. *Nucleic acids research*, 21(22):5229–5234, 1993. 3.2

[44] 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. 1.4.2

[45] Florian Buettner, Kedar N Natarajan, F Paolo Casale, Valentina Proserpio, Antonio Scialdone, Fabian J Theis, Sarah A Teichmann, John C Marioni, and Oliver Stegle. Computational analysis of cell-to-cell heterogeneity in single-cell rna-sequencing data reveals hidden subpopulations of cells. *Nature biotechnology*, 33(2):155–160, 2015. 5.2.4

[46] 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, 3.2.4

[47] Eliezer Calo and Joanna Wysocka. Modification of enhancer chromatin: what, how, and why? *Molecular cell*, 49(5):825–837, 2013. 4

[48] J Chai and AS Tarnawski. Serum response factor: discovery, biochemistry, biological roles and implications for tissue injury healing. *Journal of physiology and pharmacology*, 53(2):147–157, 2002. 3.2

[49] MCW Chan, CY Cheung, WH Chui, SW Tsao, JM Nicholls, YO Chan, RWY Chan, HT Long, LLM Poon, Y Guan, et al. Proinflammatory cytokine responses induced by influenza a (h5n1) viruses in primary human alveolar and bronchial epithelial cells. *Respiratory research*, 6(1):135, 2005. 2.9.1

[50] 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. 1.6.2

[51] WW Chatham and RP Kimberly. Treatment of lupus with corticosteroids. *Lupus*, 10: 140–147, 2001. ISSN 1474-1768. URL `http://www.biomedsearch.com/nih/`

Why-do-viruses-cause-cancer/21102637.html. 2.9.1

[52] Anathbandhu Chaudhuri, Bo Yang, Howard E Gendelman, Yuri Persidsky, and Georgette D Kanmogne. Stat1 signaling modulates hiv-1–induced inflammatory responses and leukocyte transmigration across the blood-brain barrier. *Blood*, 111(4):2062–2072, 2008. 3.2

[53] 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

[54] Jingjing Chen, Shengping Huang, and Ze Chen. Human cellular protein nucleoporin hnup98 interacts with influenza a virus ns2/nuclear export protein and overexpression of its glfg repeat domain can inhibit virus propagation. *Journal of General Virology*, 91(10): 2474–2484, 2010. 2.9, 2.9.1

[55] 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

[56] Leslie W Chinn, Minzhong Tang, Bailey D Kessing, James A Lautenberger, Jennifer L Troyer, Michael J Malasky, Carl McIntosh, Gregory D Kirk, Steven M Wolinsky, Susan P Buchbinder, et al. Genetic associations of variants in genes encoding hiv-dependency factors required for hiv-1 infection. *Journal of Infectious Diseases*, 202(12):1836–1845, 2010. 3.3

[57] Leonidas Chouliaras, Bart PF Rutten, Gunter Kenis, Odette Peerbooms, Pieter Jelle Visser, Frans Verhey, Jim van Os, Harry WM Steinbusch, and Daniel LA van den Hove. Epigenetic regulation in the pathophysiology of alzheimer's disease. *Progress in neurobiology*, 90(4):498–510, 2010. 4

[58] David B Clifford and Beau M Ances. Hiv-associated neurocognitive disorder. *The Lancet infectious diseases*, 13(11):976–986, 2013. 3.3

[59] Y Collette, H Dutartre, A Benziane, F Ramos-Morales, R Benarous, M Harris, and D Olive. Physical and functional interaction of nef with lck hiv-1 nef-induced t-cell signaling defects. *Journal of Biological Chemistry*, 271(11):6333–6341, 1996. 3.4

[60] 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. 1.4.1

[61] 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

[62] Alexandra Cribier, Benjamin Descours, Ana Luiza Chaves Valadão, Nadine Laguette, and

Monsef Benkirane. Phosphorylation of samhd1 by cyclin a2/cdk1 regulates its restriction activity toward hiv-1. *Cell reports*, 3(4):1036–1043, 2013. 3.2

[63] Ronald G Crystal, Peter B Bitterman, Stephen I Rennard, Allan J Hance, and Brendan A Keogh. Interstitial lung diseases of unknown cause: disorders characterized by chronic inflammation of the lower respiratory tract. *New England Journal of Medicine*, 310(4): 235–244, 1984. 4.3.1

[64] 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. 3.2.5

[65] Piero Dalerba, Tomer Kalisky, Debashis Sahoo, Pradeep S Rajendran, Michael E Rothenberg, Anne A Leyrat, Sopheak Sim, Jennifer Okamoto, Darius M Johnston, Dalong Qian, et al. Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nature biotechnology*, 29(12):1120–1127, 2011. 5.2.4

[66] Angus G Dalgleish, Peter CL Beverley, Paul R Clapham, Dorothy H Crawford, Melvyn F Greaves, and Robin A Weiss. The cd4 (t4) antigen is an essential component of the receptor for the aids retrovirus. *Nature*, 1984. 3.3

[67] Nune Darbinian, Armine Darbinyan, Marta Czernik, Francesca Peruzzi, Kamel Khalili, Krzysztof Reiss, Jennifer Gordon, and Shohreh Amini. Hiv-1 tat inhibits ngf-induced egr-1 transcriptional activity and consequent p35 expression in neural cells. *Journal of cellular physiology*, 216(1):128–134, 2008. 3.4

[68] Ulrik de Lichtenberg, Lars Juhl Jensen, Søren Brunak, and Peer Bork. Dynamic complex formation during the yeast cell cycle. *Science*, 307(5710):724–727, 2005. 1.6.2

[69] Yohei Doi and Yoshichika Arakawa. 16s ribosomal rna methylation: emerging resistance mechanism against aminoglycosides. *Clinical Infectious Diseases*, 45(1):88–94, 2007. 1.4.3

[70] Josée Dostie, Todd A Richmond, Ramy A Arnaout, Rebecca R Selzer, William L Lee, Tracey A Honan, Eric D Rubio, Anton Krumm, Justin Lamb, Chad Nusbaum, et al. Chromosome conformation capture carbon copy (5c): a massively parallel solution for mapping interactions between genomic elements. *Genome research*, 16(10):1299–1309, 2006. 1.4.3, 1.4.4

[71] Karolina Duskova, Pruthvi Nagilla, Hai-Son Le, Priyadarshini Iyer, Anbupalam Thalamuthu, Jeremy Martinson, Ziv Bar-Joseph, William Buchanan, Charles Rinaldo, and Velpandi Ayyavoo. Microrna regulation and its effects on cellular transcriptome in human immunodeficiency virus-1 (hiv-1) infected individuals with distinct viral load and cd4 cell counts. *BMC infectious diseases*, 13(1):250, 2013. 3.3.1, 3.3.1

[72] Sujit Dutta and Yee-Joo Tan. Structural and functional characterization of human sgt and its interaction with vpu of the human immunodeficiency virus type 1,. *Biochemistry*, 47 (38):10123–10131, 2008. 3.4

[73] 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

[74] Gerda Egger, Gangning Liang, Ana Aparicio, and Peter A Jones. Epigenetics in human disease and prospects for epigenetic therapy. *Nature*, 429(6990):457, 2004. 4

[75] Christina Ehrhardt, Thorsten Wolff, Stephan Pleschka, Oliver Planz, Wiebke Beermann, Johannes Bode, Mirco Schmolke, and Stephan Ludwig. Influenza A Virus NS1 Protein Activates the PI3K/Akt Pathway To Mediate Antiapoptotic Signaling Responses. *The Journal of Virology*, 81(7):3058–3067, April 2007. URL `http://jvi.asm.org/cgi/content/abstract/81/7/3058?etoc`. 2.9.1

[76] Mazen B Eldeen, Satish L Deshmane, Kenneth Simbiri, Kamel Khalili, Shohreh Amini, and Bassel E Sawaya. Mh2 domain of smad3 reduces hiv-1 tat-induction of cytokine secretion. *Journal of neuroimmunology*, 176(1):174–180, 2006. 3.2

[77] Frank Emmert-Streib, Matthias Dehmer, and Benjamin Haibe-Kains. Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks. *Frontiers in cell and developmental biology*, 2, 2014. 1.6.3

[78] 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, 1.6.2, 1.6.3, 2.1

[79] Jason Ernst, Qasim K Beg, Krin A Kay, Gábor Balázsi, Zoltán N Oltvai, and Ziv Bar-Joseph. A semi-supervised method for predicting transcription factor–gene interactions in escherichia coli. *PLoS computational biology*, 4(3):e1000044, 2008. 2.1

[80] 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. 1.5.1

[81] Maurizio Federico, Zulema Percario, Eleonora Olivetta, Gianna Fiorucci, Claudia Muratori, Alessandro Micheli, Giovanna Romeo, and Elisabetta Affabris. Hiv-1 nef activates stat1 in human monocytes/macrophages through the release of soluble factors. *Blood*, 98 (9):2752–2761, 2001. 3.2

[82] Witold Filipowicz, Suvendra N Bhattacharyya, and Nahum Sonenberg. Mechanisms of post-transcriptional regulation by micrornas: are the answers in sight? *Nature Reviews Genetics*, 9(2):102–114, 2008. 1.6.2

[83] Nir Friedman. Probabilistic models for identifying regulation networks. *Bioinformatics*, 19(suppl_2):ii57–ii57, 2003. 1.6.1

[84] Jihong Fu, Wentao Tang, Peng Du, Guanghui Wang, Wei Chen, Jingming Li, Yunxiang Zhu, Jun Gao, and Long Cui. Identifying microrna-mrna regulatory network in colorectal cancer by a combination of expression profile and bioinformatics analysis. *BMC systems biology*, 6(1):68, 2012. 1.6.3

[85] A Gallina, F Rossi, and G Milanesi. Rack1 binds hiv-1 nef and can act as a nef–protein kinase c adaptor. *Virology*, 283(1):7–18, 2001. 3.4

[86] Timothy S Gardner, Diego Di Bernardo, David Lorenz, and James J Collins. Inferring genetic networks and identifying compound mode of action via expression profiling. *Sci-*

ence, 301(5629):102–105, 2003. 1.6.1

[87] Audrey P Gasch, Paul T Spellman, Camilla M Kao, Orna Carmel-Harel, Michael B Eisen, Gisela Storz, David Botstein, and Patrick O Brown. Genomic expression programs in the response of yeast cells to environmental changes. *Molecular biology of the cell*, 11(12): 4241–4257, 2000. 5.2.2

[88] 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, 2.9.1

[89] 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. 1.4.2

[90] 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

[91] Anthony Gitter and Ziv Bar-Joseph. Identifying proteins controlling key disease signaling pathways. *Bioinformatics*, 29(13):i227–i236, 2013. 1.6.2, 2.3.4, 2.9, 2.10.1, 3.2.4

[92] 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

[93] 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. 1.4.5, 1.6.2, 2.3.3, 2.3.4, 2.5, 2.7, 3.2.4, 3.4

[94] 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. 1.6.2, 2, 2.2, 2.3.4, 2.9, 2.10.1

[95] 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. 1.6.1

[96] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 249–256, 2010. 1.5.4, 4.3.3

[97] Fred Glover and Manuel Laguna. *Tabu search*. Springer, 1999. 3.1.6

[98] Ritu Goila-Gaur and Klaus Strebel. Hiv-1 vif, apobec, and intrinsic immunity. *Retrovirology*, 5(51):10–1186, 2008. 3.2.5

[99] Anita Göndör, Carole Rougier, and Rolf Ohlsson. High-resolution circular chromosome conformation capture assay. *Nature protocols*, 3(2):303, 2008. 1.4.4

[100] 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

[101] Enrique Gonzalez, Brad H Rovin, Luisa Sen, Glen Cooke, Rahul Dhanda, Srinivas Mummidi, Hemant Kulkarni, Michael J Bamshad, Vanessa Telles, Stephanie A Anderson, et al. Hiv-1 infection and aids dementia are influenced by a mutant mcp-1 allele linked to increased monocyte infiltration of tissues and mcp-1 levels. *Proceedings of the National Academy of Sciences*, 99(21):13795–13800, 2002. 3.3

[102] Jonathan Gribbin, Richard B Hubbard, Ivan Le Jeune, Chris JP Smith, Joe West, and Laila J Tata. Incidence and mortality of idiopathic pulmonary fibrosis and sarcoidosis in the uk. *Thorax*, 61(11):980–985, 2006. 1.3.3, 4.3.1

[103] Fei Gu, Hang-Kai Hsu, Pei-Yin Hsu, Jiejun Wu, Yilin Ma, Jeffrey Parvin, Tim HM Huang, and Victor X Jin. Inference of hierarchical regulatory network of estrogen-dependent breast cancer through chip-based data. *BMC systems biology*, 4(1):170, 2010. 2.1

[104] Hongcang Gu, Zachary D Smith, Christoph Bock, Patrick Boyle, Andreas Gnirke, and Alexander Meissner. Preparation of reduced representation bisulfite sequencing libraries for genome-scale dna methylation profiling. *Nature protocols*, 6(4):468, 2011. 1.4.3

[105] Hyo Chol Ha, Krishna Juluri, Yan Zhou, Steve Leung, Monika Hermankova, and Solomon H Snyder. Poly (adp-ribose) polymerase-1 is required for efficient hiv-1 integration. *Proceedings of the National Academy of Sciences*, 98(6):3364–3368, 2001. 3.2

[106] Leon Juvenal Hajingabo, Sarah Daakour, Maud Martin, Reinhard Grausenburger, Renate Panzer-Grümayer, Franck Dequiedt, Nicolas Simonis, and Jean-Claude Twizere. Predicting interactome network perturbations in human cancer: application to gene fusions in acute lymphoblastic leukemia. *Molecular biology of the cell*, 25(24):3973–3985, 2014. 1.6.3

[107] 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. 3.2.5

[108] Ofir Hakim and Tom Misteli. Snapshot: chromosome conformation capture. *Cell*, 148(5): 1068–1068, 2012. 1.4.4

[109] Benjamin G Hale, Richard E Randall, Juan Ortín, and David Jackson. The multifunctional ns1 protein of influenza a viruses. *Journal of General Virology*, 89(10):2359–2376, 2008. 5.1.1

[110] Linhui Hao, Qiuling He, Zhishi Wang, Mark Craven, Michael A Newton, and Paul Ahlquist. Limited agreement of independent rnai screens for virus-required host genes owes more to false-negative than false-positive factors. *PLoS computational biology*, 9 (9):e1003235, 2013. 2.10

[111] 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

[112] Anne-Claire Haury, Fantine Mordelet, Paola Vera-Licona, and Jean-Philippe Vert. Tigress: trustful inference of gene regulation using stability selection. *BMC systems biology*, 6(1): 145, 2012. 1.6.2

[113] John Cijiang He, Mohammad Husain, Masaaki Sunamoto, Vivette D DAgati, Mary E Klotman, Ravi Iyengar, Paul E Klotman, et al. Nef stimulates proliferation of glomerular podocytes through activation of src-dependent stat3 and mapk1, 2 pathways. *The Journal of clinical investigation*, 114(5):643–651, 2004. 3.2

[114] Robert Hecht-Nielsen et al. Theory of the backpropagation neural network. *Neural Networks*, 1(Supplement-1):445–448, 1988. 1.5.4

[115] Angus Henderson, Michael Bunce, Nicole Siddon, Raymond Reeves, and David John Tremethick. High-mobility-group protein i can modulate binding of transcription factors to the u5 region of the human immunodeficiency virus type 1 proviral promoter. *Journal of virology*, 74(22):10523–10534, 2000. 3.4

[116] Jochen Hess, Peter Angel, and Marina Schorpp-Kistner. Ap-1 subunits: quarrel and harmony among siblings. *Journal of cell science*, 117(25):5965–5973, 2004. 2.9.1

[117] 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

[118] 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

[119] 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. 3.2.5

[120] 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

[121] 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. 3.2.4

[122] Lu Huang, Guan-lan Xu, Jian-qi Zhang, Ling Tian, Jing-lun Xue, Jin-zhong Chen, and William Jia. Daxx interacts with hiv-1 integrase and inhibits lentiviral gene expression. *Biochemical and biophysical research communications*, 373(2):241–245, 2008. 3.2

[123] Shengping Huang, Jingjing Chen, Huadong Wang, Bing Sun, Hanzhong Wang, Zhiping Zhang, Xianen Zhang, and Ze Chen. Influenza a virus matrix protein 1 interacts with htfiiic102-s, a short isoform of the polypeptide 3 subunit of human general transcription factor iiic. *Archives of virology*, 154(7):1101–1110, 2009. 2.9

[124] Yongsheng Huang, Aimee K Zaas, Arvind Rao, Nicolas Dobigeon, Peter J Woolf, Timothy Veldman, N Christine Øien, Micah T McClain, Jay B Varkey, Bradley Nicholson, et al. Temporal dynamics of host molecular responses differentiate symptomatic and asymp-

tomatic influenza a infection. *PLoS genetics*, 7(8):e1002234, 2011. 2.9

[125] Taisuke Izumi, Katsuhiro Io, Masashi Matsui, Kotaro Shirakawa, Masanobu Shinohara, Yuya Nagai, Masahiro Kawahara, Masayuki Kobayashi, Hiroshi Kondoh, Naoko Misawa, et al. Hiv-1 viral infectivity factor interacts with tp53 to induce g2 cell cycle arrest and positively regulate viral replication. *Proceedings of the National Academy of Sciences*, 107(48):20798–20803, 2010. 3.2

[126] Laurent Jacob and Jean-Philippe Vert. Efficient peptide–mhc-i binding prediction for alleles with few known binders. *Bioinformatics*, 24(3):358–366, 2008. 1.5.1

[127] 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. 1.6.2, 2

[128] Laurence Jeanson, Frédéric Subra, Sabine Vaganay, Martial Hervy, Elizabeth Marangoni, Jean Bourhis, and Jean-François Mouscadet. Effect of ku80 depletion on the preintegrative steps of hiv-1 replication in human cells. *Virology*, 300(1):100–108, 2002. 3.3

[129] Guochun Jiang, Amy Espeseth, Daria J Hazuda, and David M Margolis. c-myc and sp1 contribute to proviral latency by recruiting histone deacetylase 1 to the human immunodeficiency virus type 1 promoter. *Journal of virology*, 81(20):10914–10923, 2007. 3.2

[130] 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

[131] 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

[132] Alexander Karlas, Nikolaus Machuy, Yujin Shin, Klaus-Peter Pleissner, Anita Artarini, Dagmar Heuer, Daniel Becker, Hany Khalil, Lesley A Ogilvie, Simone Hess, et al. Genome-wide rnai screen identifies human host factors crucial for influenza virus replication. *Nature*, 463(7282):818–822, 2010. 2.10

[133] Stefan U Kass, Dmitry Pruss, and Alan P Wolffe. How does dna methylation repress transcription? *Trends in Genetics*, 13(11):444–449, 1997. 4

[134] David R Kelley, Jasper Snoek, and John L Rinn. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome research*, 26(7): 990–999, 2016. 4.1

[135] Dong Soon Kim, Harold R Collard, and Talmadge E King Jr. Classification and natural history of the idiopathic interstitial pneumonias. *Proceedings of the American Thoracic Society*, 3(4):285–292, 2006. 1.3.3

[136] Seyoung Kim and Eric P Xing. Statistical estimation of correlated genome associations to a quantitative trait network. *PLoS genetics*, 5(8):e1000587, 2009. 1.5.1

[137] 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. 1.5.1

[138] Talmadge E King Jr, Williamson Z Bradford, Socorro Castro-Bernardini, Elizabeth A Fagan, Ian Glaspole, Marilyn K Glassberg, Eduard Gorina, Peter M Hopkins, David Kardatzke, Lisa Lancaster, et al. A phase 3 trial of pirfenidone in patients with idiopathic pulmonary fibrosis. *New England Journal of Medicine*, 370(22):2083–2092, 2014. 4.3.1

[139] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1.5.4, 4.3.3

[140] Tomoshige Kino, Alexander Gragerov, Jeffrey B Kopp, Roland H Stauber, George N Pavlakis, and George P Chrousos. The hiv-1 virion-associated protein vpr is a coactivator of the human glucocorticoid receptor. *The Journal of experimental medicine*, 189 (1):51–62, 1999. 3.2

[141] Michael Kleyman, Emre Sefer, Teodora Nicola, Celia Espinoza, Divya Chhabra, James S Hagood, Naftali Kaminski, Namasivayam Ambalavanan, and Ziv Bar-Joseph. Selecting the most appropriate time points to profile in high-throughput studies. *eLife*, 6:e18541, 2017. 3.4

[142] Mladen Kolar, Le Song, Amr Ahmed, and Eric P Xing. Estimating time-varying networks. *The Annals of Applied Statistics*, pages 94–123, 2010. 1.6.2

[143] Renate König, Yingyao Zhou, Daniel Elleder, Tracy L Diamond, Ghislain Bonamy, Jeffrey T Irelan, Chih-yuan Chiang, Buu P Tu, Paul D De Jesus, Caroline E Lilley, et al. Global analysis of host-pathogen interactions that regulate early-stage hiv-1 replication. *Cell*, 135(1):49–60, 2008. 3.3

[144] Renate König, Silke Stertz, Yingyao Zhou, Atsushi Inoue, H-Heinrich Hoffmann, Suchita Bhattacharyya, Judith G Alamares, Donna M Tscherne, Mila B Ortigoza, Yuhong Liang, et al. Human host factors required for influenza virus replication. *Nature*, 463(7282): 813–817, 2009. 2.10

[145] Tony Kouzarides. Chromatin modifications and their function. *Cell*, 128(4):693–705, 2007. 1.4.3

[146] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 1.5.4

[147] Meghana Kshirsagar, Jaime Carbonell, and Judith Klein-Seetharaman. Multitask learning for host–pathogen protein interactions. *Bioinformatics*, 29(13):i217–i226, 2013. 1.5.1

[148] 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

[149] Ashok Kumar, Sunil K Manna, Subhash Dhawan, and Bharat B Aggarwal. Hiv-tat protein activates c-jun n-terminal kinase and activator protein-1. *The Journal of Immunology*, 161 (2):776–781, 1998. 3.3

[150] Mondira Kundu, Alagarsamy Srinivasan, Roger J Pomerantz, and Kamel Khalili. Evidence that a cell cycle regulator, e2f1, down-regulates transcriptional activity of the human immunodeficiency virus type 1 promoter. *Journal of virology*, 69(11):6940–6946,

1995. 3.2

[151] Andrew T. Kwon, David J. Arenillas, Rebecca W. Hunt, and Wyeth W. Wasserman. oPOSSUM-3: Advanced Analysis of Regulatory Motif Over-Representation Across Genes or ChIP-Seq Datasets. *G3: Genes—Genomes—Genetics*, 2(9):987–1002, September 2012. ISSN 2160-1836. doi: 10.1534/g3.112.003202. URL `http://dx.doi.org/10.1534/g3.112.003202`. 2.10.1

[152] 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. 3.2.5

[153] Alex Lan, Ilan Y Smoly, Guy Rapaport, Susan Lindquist, Ernest Fraenkel, and Esti Yeger-Lotem. Responsenet: revealing signaling and regulatory networks linking genetic and transcriptomic screening data. *Nucleic acids research*, 39(suppl 2):W424–W429, 2011. 1.6.1

[154] Peter Langfelder and Steve Horvath. Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):559, 2008. 1.6.3

[155] Alan Lau, Karra M Swinbank, Parvin S Ahmed, Debra L Taylor, Stephen P Jackson, Graeme CM Smith, and Mark J O'Connor. Suppression of hiv-1 infection by a small molecule inhibitor of the atm kinase. *Nature cell biology*, 7(5):493–500, 2005. 3.3

[156] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553): 436–444, 2015. 4.2.2

[157] Jan K Lenstra. *Local search in combinatorial optimization*. Princeton University Press, 1997. 1.5.2

[158] Chengjun Li, Armand Bankhead, Amie J Eisfeld, Yasuko Hatta, Sophia Jeng, Jean H Chang, Lauri D Aicher, Sean Proll, Amy L Ellis, G Lynn Law, et al. Host regulatory network response to infection with highly pathogenic h5n1 avian influenza virus. *Journal of virology*, 85(21):10955–10967, 2011. 2.9

[159] 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

[160] KS Li, Y Guan, J Wang, GJD Smith, et al. Genesis of a highly pathogenic and potentially pandemic h5n1 influenza virus in eastern asia. *Nature*, 430(6996):209, 2004. 1.3.1

[161] Ying Li, Silvia Innocentin, David R Withers, Natalie A Roberts, Alec R Gallagher, Elena F Grigorieva, Christoph Wilhelm, and Marc Veldhoen. Exogenous stimuli maintain intraepithelial lymphocytes via aryl hydrocarbon receptor activation. *Cell*, 147(3):629–640, 2011. 2.10.1

[162] Erez Lieberman-Aiden, Nynke L Van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science*, 326(5950):289–293, 2009. 1.4.4

[163] Thomas Linnemann, Yong-Hui Zheng, Robert Mandic, and B Matija Peterlin. Interaction

between nef and phosphatidylinositol-3-kinase leads to activation of p21-activated kinase and increased production of hiv. *Virology*, 294(2):246–255, 2002. 3.2

[164] Ryan Lister, Mattia Pelizzola, Robert H Dowen, R David Hawkins, Gary Hon, Julian Tonti-Filippini, Joseph R Nery, Leonard Lee, Zhen Ye, Que-Minh Ngo, et al. Human dna methylomes at base resolution show widespread epigenomic differences. *nature*, 462 (7271):315, 2009. 1.4.3

[165] Di Liu, XiaoLing Liu, JingHua Yan, Wen-Jun Liu, and George Fu Gao. Interspecies transmission and host restriction of avian h5n1 influenza virus. *Science in China Series C: Life Sciences*, 52(5):428–438, 2009. 2.9

[166] Gang Liu, Arnaud Friggeri, Yanping Yang, Jadranka Milosevic, Qiang Ding, Victor J Thannickal, Naftali Kaminski, and Edward Abraham. mir-21 mediates fibrogenic activation of pulmonary fibroblasts and lung fibrosis. *Journal of Experimental Medicine*, 207 (8):1589–1597, 2010. 4.3.3

[167] Liang Liu, Guangxu Jin, and Xiaobo Zhou. Modeling the relationship of epigenetic modifications to transcription factor binding. *Nucleic acids research*, 43(8):3873–3885, 2015. 4.1

[168] Tak W. Mak and Mary E. Saunders. *The Immune Response: Basic and Clinical Principles*, volume 1. 2006. 2.9.1, 2.9.1, 3.2.4

[169] Giuseppe Mameli, Satish L Deshmane, Mohammad Ghafouri, Jianqi Cui, Kenneth Simbiri, Kamel Khalili, Ruma Mukerjee, Antonina Dolei, Shohreh Amini, and Bassel E Sawaya. C/ebpβ regulates human immunodeficiency virus 1 gene expression through its association with cdk9. *Journal of general virology*, 88(2):631–640, 2007. 3.2

[170] Lara Manganaro, Marina Lusic, Maria Ines Gutierrez, Anna Cereseto, Giannino Del Sal, and Mauro Giacca. Concerted action of cellular jnk and pin1 restricts hiv-1 genome integration to activated cd4+ t lymphocytes. *Nature medicine*, 16(3):329–333, 2010. 3.3

[171] Daniel Marbach, James C Costello, Robert Küffner, Nicole M Vega, Robert J Prill, Diogo M Camacho, Kyle R Allison, Manolis Kellis, James J Collins, Gustavo Stolovitzky, et al. Wisdom of crowds for robust gene network inference. *Nature methods*, 9(8):796–804, 2012. 1.6.2

[172] Joëlle Marchal-Sommé, Yurdagul Uzunhan, Sylvain Marchand-Adam, Dominique Valeyre, Vassili Soumelis, Bruno Crestani, and Paul Soler. Cutting edge: nonproliferating mature immune cells form a novel type of organized lymphoid structure in idiopathic pulmonary fibrosis. *The Journal of Immunology*, 176(10):5735–5739, 2006. 4.3.2

[173] Adam A Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo D Favera, and Andrea Califano. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics*, 7(Suppl 1):S7, 2006. 1.6.2, 1.6.3

[174] David M Margolis, Mohan Somasundaran, and Michael R Green. Human transcription factor yy1 represses human immunodeficiency virus type 1 transcription and virion production. *Journal of virology*, 68(2):905–910, 1994. 3.2

[175] Cyrus Martin and Yi Zhang. The diverse functions of histone lysine methylation. *Nature reviews. Molecular cell biology*, 6(11):838, 2005. 1.4.3

[176] 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

[177] John E McDonough, Ren Yuan, Masaru Suzuki, Nazgol Seyednejad, W Mark Elliott, Pablo G Sanchez, Alexander C Wright, Warren B Gefter, Leslie Litzky, Harvey O Coxson, et al. Small-airway obstruction and emphysema in chronic obstructive pulmonary disease. *New England Journal of Medicine*, 365(17):1567–1575, 2011. 4.3.1

[178] Yulia A Medvedeva, Abdullah M Khamis, Ivan V Kulakovskiy, Wail Ba-Alawi, Md Shariful I Bhuyan, Hideya Kawaji, Timo Lassmann, Matthias Harbers, Alistair RR Forrest, and Vladimir B Bajic. Effects of cytosine methylation on transcription factor binding sites. *BMC genomics*, 15(1):119, 2014. 4.3.3

[179] Alexander Meissner, Andreas Gnirke, George W Bell, Bernard Ramsahoye, Eric S Lander, and Rudolf Jaenisch. Reduced representation bisulfite sequencing for comparative high-resolution dna methylation analysis. *Nucleic acids research*, 33(18):5868–5877, 2005. 1.4.3

[180] Masaya Miyazaki, Hiroshi Nishihara, Hideki Hasegawa, Masato Tashiro, Lei Wang, Taichi Kimura, Mishie Tanino, Masumi Tsuda, and Shinya Tanaka. Ns1-binding protein abrogates the elevation of cell viability by the influenza a virus ns1 protein in association with crkl. *Biochemical and biophysical research communications*, 441(4):953–957, 2013. 2.9.1

[181] 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, 3.2.1, 3.4

[182] Patrick S Moore and Yuan Chang. Why do viruses cause cancer? highlights of the first century of human tumour virology. *Nat Rev Cancer*, 10(12):878–89, 2010. ISSN 1474-1768. URL `http://www.biomedsearch.com/nih/Why-do-viruses-cause-cancer/21102637.html`. 2.10.1

[183] Masako Moriuchi, Hiroyuki Moriuchi, David M Margolis, and Anthony S Fauci. Usf/c-myc enhances, while yin-yang 1 suppresses, the promoter activity of cxcr4, a coreceptor for hiv-1 entry. *The Journal of Immunology*, 162(10):5986–5992, 1999. 3.4

[184] 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

[185] Shiraz Mujtaba, Yan He, Lei Zeng, Amjad Farooq, Justin E Carlson, Melanie Ott, Eric Verdin, and Ming-Ming Zhou. Structural basis of lysine-acetylated hiv-1 tat recognition by pcaf bromodomain. *Molecular cell*, 9(3):575–586, 2002. 3.2

[186] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010. 4.3.3

[187] Xinsheng Nan, Huck-Hui Ng, Colin A Johnson, Carol D Laherty, et al. Transcriptional repression by the methyl-cpg-binding protein mecp2 involves a histone deacetylase complex. *Nature*, 393(6683):386, 1998. 4.3.3

[188] Srinivas D Narasipura, Lisa J Henderson, Sidney W Fu, Liang Chen, Fatah Kashanchi, and Lena Al-Harthi. Role of β-catenin and tcf/lef family members in transcriptional activity of hiv in astrocytes. *Journal of virology*, 86(4):1911–1921, 2012. 3.2

[189] 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, 1.4.6, 2.9, 3.1.2, 3.2.1

[190] David L Nelson, Albert L Lehninger, and Michael M Cox. *Lehninger principles of biochemistry*. Macmillan, 2008. 1.4.3

[191] Sam Nightingale, Alan Winston, Scott Letendre, Benedict D Michael, Justin C McArthur, Saye Khoo, and Tom Solomon. Controversies in hiv-associated neurocognitive disorders. *The Lancet Neurology*, 13(11):1139–1151, 2014. 3.3

[192] Noa Novershtern, Aviv Regev, and Nir Friedman. Physical module networks: an integrative approach for reconstructing transcription regulation. *Bioinformatics*, 27(13):i177–i185, 2011. 1.6.3

[193] Karen E Ocwieja, Troy L Brady, Keshet Ronen, Alyssa Huegel, Shoshannah L Roth, Torsten Schaller, Leo C James, Greg J Towers, John AT Young, Sumit K Chanda, et al. Hiv integration targeting: a pathway involving transportin-3 and the nuclear pore protein ranbp2. *PLoS pathogens*, 7(3):e1001313, 2011. 3.3

[194] Kyoung-Su Oh and Keechul Jung. Gpu implementation of neural networks. *Pattern Recognition*, 37(6):1311–1314, 2004. 1.5.4

[195] Kenzo Ohtsuki, Toshiro Maekawa, Shigeyoshi Harada, Atsushi Karino, Yuko Morikawa, and Masahiko Ito. Biochemical characterization of hiv-1 rev as a potent activator of casein kinase ii in vitro. *FEBS letters*, 428(3):235–240, 1998. 3.2

[196] Melanie Ott, Martina Schnölzer, Jerry Garnica, Wolfgang Fischle, Stephane Emiliani, Hans-Richard Rackwitz, and Eric Verdin. Acetylation of the hiv-1 tat protein by p300 is important for its transcriptional activity. *Current Biology*, 9(24):1489–1493, 1999. 3.2

[197] Sara Pagans, Angelika Pedal, Brian J North, Katrin Kaehlcke, Brett L Marshall, Alexander Dorr, Claudia Hetzer-Egger, Peter Henklein, Roy Frye, Michael W McBurney, et al. Sirt1 regulates hiv transcription via tat deacetylation. *PLoS biology*, 3(2):e41, 2005. 3.3

[198] Peter J Park. Chip–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10(10):669–680, 2009. 1.4.2

[199] Sybil P Parker. *McGraw-Hill Encyclopedia of Science & Technology*. McGraw-Hill Com-

panies, 1997. 1.4.4

[200] 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. 1.6.2, 3, 3.2.4, 3.4

[201] Dana Peer, Aviv Regev, Gal Elidan, and Nir Friedman. Inferring subnetworks from perturbed expression profiles. *Bioinformatics*, 17(suppl_1):S215–S224, 2001. 1.6.1

[202] Eric M Phizicky and Stanley Fields. Protein-protein interactions: methods for detection and analysis. *Microbiological reviews*, 59(1):94–123, 1995. 1.4.5

[203] Ruth Pidsley, Elena Zotenko, Timothy J Peters, Mitchell G Lawrence, Gail P Risbridger, Peter Molloy, Susan Van Djik, Beverly Muhlhausler, Clare Stirzaker, and Susan J Clark. Critical evaluation of the illumina methylationepic beadchip microarray for whole-genome dna methylation profiling. *Genome biology*, 17(1):208, 2016. 4.3.3

[204] Valeria Poli. The role of c/ebp isoforms in the control of inflammatory and native immunity functions. *Journal of Biological Chemistry*, 273(45):29279–29282, 1998. 2.10.1

[205] Anna Portela and Manel Esteller. Epigenetic modifications and human disease. *Nature biotechnology*, 28(10):1057–1068, 2010. 4

[206] 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, 1.4.5, 3.1.2, 3.2.1

[207] Stanley B Prusiner. Prions. *Proceedings of the National Academy of Sciences*, 95(23): 13363–13383, 1998. 1.4.3

[208] 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. 1.5.1

[209] Daniel Quang and Xiaohui Xie. Extreme: an online em algorithm for motif discovery. *Bioinformatics*, 30(12):1667–1673, 2014. 4.2.2, 4.3.3

[210] Daniel Quang and Xiaohui Xie. Danq: a hybrid convolutional and recurrent deep neural network for quantifying the function of dna sequences. *Nucleic acids research*, 44(11): e107–e107, 2016. 4.1, 4.2.2, 4.3.3, 4.3.3, 4.4

[211] Daniel Quang and Xiaohui Xie. Factornet: a deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. *bioRxiv*, page 151274, 2017. 4.1

[212] Daniel X Quang, Michael R Erdos, Stephen CJ Parker, and Francis S Collins. Motif signatures in stretch enhancers are enriched for disease-associated genetic variants. *Epigenetics & chromatin*, 8(1):23, 2015. 4.2.2, 4.3.3

[213] Ganesh Raghu, Derek Weycker, John Edelsberg, Williamson Z Bradford, and Gerry Oster. Incidence and prevalence of idiopathic pulmonary fibrosis. *American journal of respiratory and critical care medicine*, 174(7):810–816, 2006. 1.3.3

[214] Ganesh Raghu, Harold R Collard, Jim J Egan, Fernando J Martinez, Juergen Behr, Kevin K Brown, Thomas V Colby, Jean-François Cordier, Kevin R Flaherty, Joseph A Lasky, et al. An official ats/ers/jrs/alat statement: idiopathic pulmonary fibrosis: evidence-based guidelines for diagnosis and management. *American journal of respiratory and critical care medicine*, 183(6):788–824, 2011. 1.3.3

[215] Tamal Raha, SW Grace Cheng, and Michael R Green. Hiv-1 tat stimulates transcription complex assembly through recruitment of tbp in the absence of tafs. *PLoS biology*, 3(2): e44, 2005. 3.2

[216] Christopher V Rao, Denise M Wolf, and Adam P Arkin. Control, exploitation and tolerance of intracellular noise. *Nature*, 420(6912):231, 2002. 1.6.1

[217] 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. 3.2.4

[218] Susanne Rauch, Kati Pulkkinen, Kalle Saksela, and Oliver T Fackler. Human immunodeficiency virus type 1 nef recruits the guanine exchange factor vav1 via an unexpected interface into plasma membrane microdomains for association with p21-activated kinase 2 activity. *Journal of virology*, 82(6):2918–2929, 2008. 3.4

[219] Wolf Reik. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature*, 447(7143):425, 2007. 1.4.3

[220] Luca Richeldi, Roland M du Bois, Ganesh Raghu, Arata Azuma, Kevin K Brown, Ulrich Costabel, Vincent Cottin, Kevin R Flaherty, David M Hansell, Yoshikazu Inoue, et al. Efficacy and safety of nintedanib in idiopathic pulmonary fibrosis. *New England Journal of Medicine*, 370(22):2071–2082, 2014. 4.3.1

[221] William N Rom and Steven B Markowitz. *Environmental and occupational medicine*. Lippincott Williams & Wilkins, 2007. 1.3.2

[222] Sushmita Roy, Jason Ernst, Peter V Kharchenko, Pouya Kheradpour, Nicolas Negre, Matthew L Eaton, Jane M Landolin, Christopher A Bristow, Lijia Ma, Michael F Lin, et al. Identification of functional elements and regulatory circuits by drosophila modencode. *Science*, 330(6012):1787–1797, 2010. 2.1

[223] María Salgado, Pedro López-Romero, Sergio Callejas, Mariola López, Pablo Labarga, Ana Dopazo, Vincent Soriano, and Berta Rodés. Characterization of host genetic expression patterns in hiv-infected individuals with divergent disease progression. *Virology*, 411 (1):103–112, 2011. 1.6.2

[224] Earl T Sawai, Imran H Khan, Phillip M Montbriand, B Matija Peterlin, Cecilia Cheng-Mayer, and Paul A Luciw. Activation of pak by hiv and siv nef: importance for aids in rhesus macaques. *Current Biology*, 6(11):1519–1527, 1996. 3.4

[225] Deanna Saylor, Alex M Dickens, Ned Sacktor, Norman Haughey, Barbara Slusher, Mikhail Pletnikov, Joseph L Mankowski, Amanda Brown, David J Volsky, and Justin C McArthur. Hiv-associated neurocognitive disorderpathogenesis and prospects for treat-

ment. *Nature reviews. Neurology*, 12(4):234, 2016. 3.3

[226] JG Scadding and KFW Hinson. Diffuse fibrosing alveolitis (diffuse interstitial fibrosis of the lungs). *Thorax*, 22(4):291–304, 1967. 4.3.1

[227] Anthony D Schmitt, Ming Hu, Inkyung Jung, Zheng Xu, Yunjiang Qiu, Catherine L Tan, Yun Li, Shin Lin, Yiing Lin, Cathy L Barr, et al. A compendium of chromatin contact maps reveals spatially active regions in the human genome. *Cell reports*, 17(8):2042–2059, 2016. 4.2.2

[228] 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. 3.2.5

[229] Ulrich Schubert, David E Ott, Elena N Chertova, Reinhold Welker, Uwe Tessmer, Michael F Princiotta, Jack R Bennink, Hans-Georg Kräusslich, and Jonathan W Yewdell. Proteasome inhibition interferes with gag polyprotein processing, release, and maturation of hiv-1 and hiv-2. *Proceedings of the National Academy of Sciences*, 97(24):13057–13062, 2000. 3.4

[230] 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. 1.4.2, 1.6.2, 1.6.3, 3.1.2, 3.2.1

[231] Max A Seibold, Russell W Smith, Cydney Urbanek, Steve D Groshong, Gregory P Cosgrove, Kevin K Brown, Marvin I Schwarz, David A Schwartz, and Susan D Reynolds. The idiopathic pulmonary fibrosis honeycomb cyst contains a mucocilary pseudostratified epithelium. *PloS one*, 8(3):e58658, 2013. 4.3.2

[232] Alex K Shalek, Rahul Satija, Xian Adiconis, Rona S Gertner, Jellert T Gaublomme, Raktima Raychowdhury, Schragi Schwartz, Nir Yosef, Christine Malboeuf, Diana Lu, et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*, 498(7453):236, 2013. 5.2.4

[233] Sagi D Shapira, Irit Gat-Viks, Bennett OV Shum, Amelie Dricot, Marciela M de Grace, Liguo Wu, Piyush B Gupta, Tong Hao, Serena J Silver, David E Root, et al. A physical and regulatory map of host-influenza interactions reveals pathways in h1n1 infection. *Cell*, 139(7):1255–1267, 2009. 2.9, 2.10

[234] Ehud Shapiro, Tamir Biezuner, and Sten Linnarsson. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature reviews. Genetics*, 14(9):618, 2013. 5.2.4

[235] Kulbhushan Sharma, Shashank Tripathi, Priya Ranjan, Purnima Kumar, Rebecca Garten, Varough Deyde, Jacqueline M Katz, Nancy J Cox, Renu B Lal, Suryaprakash Sambhara, et al. Influenza a virus nucleoprotein exploits hsp40 to inhibit pkr activation. *PLoS One*, 6(6):e20215, 2011. 2.9

[236] Shikhar Sharma, Theresa K Kelly, and Peter A Jones. Epigenetics in cancer. *Carcinogenesis*, 31(1):27–36, 2010. 4

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

[238] Ilya Shmulevich and Stuart A Kauffman. Activities and sensitivities in boolean network models. *Physical review letters*, 93(4):048701, 2004. 1.6.1

[239] Iart Luca Shytaj and Andrea Savarino. A cure for aids: a matter of timing. *Retrovirology*, 10(1):145, 2013. 1.6.2

[240] Ritambhara Singh, Jack Lanchantin, Arshdeep Sekhon, and Yanjun Qi. Attend and predict: Understanding gene regulation by selective attention on chromatin. *arXiv preprint arXiv:1708.00339*, 2017. 4.1, 5.2.3

[241] 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

[242] Johanna A Smith, Feng-Xiang Wang, Hui Zhang, Kou-Juey Wu, Kevin J Williams, and René Daniel. Evidence that the nijmegen breakage syndrome protein, an early sensor of double-strand dna breaks (dsb), is involved in hiv-1 post-integration repair by recruiting the ataxia telangiectasia-mutated kinase in a process similar to, but distinct from, cellular dsb repair. *Virology journal*, 5(1):11, 2008. 3.1.1, 3.3

[243] Duncan Sproul, Nick Gilbert, and Wendy A Bickmore. The role of chromatin structure in regulating the expression of clustered genes. *Nature reviews. Genetics*, 6(10):775, 2005. 1.4.4

[244] Ralph Stadhouders, Petros Kolovos, Rutger Brouwer, Jessica Zuin, Anita Van Den Heuvel, Christel Kockx, Robert-Jan Palstra, Kerstin S Wendt, Frank Grosveld, Wilfred Van Ijcken, et al. Multiplexed chromosome conformation capture sequencing for rapid genome-scale high-resolution detection of long-range chromatin interactions. *Nature protocols*, 8(3): 509, 2013. 1.4.4

[245] 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. 1.4.5, 3.1.2, 3.2.1

[246] Oliver Stegle, Sarah A Teichmann, and John C Marioni. Computational and analytical challenges in single-cell transcriptomics. *Nature reviews. Genetics*, 16(3):133, 2015. 5.2.4

[247] Amy B Strasner, Malini Natarajan, Tom Doman, Douglas Key, Avery August, and Andrew J Henderson. The src kinase lck facilitates assembly of hiv-1 at the plasma membrane. *The Journal of Immunology*, 181(5):3706–3713, 2008. 3.4

[248] Yi Sun, Yue-Chen Huang, Qin-Zhi Xu, Hui-Ping Wang, Bei Bai, Jian-Li Sui, and Ping-Kun Zhou. Hiv-1 tat depresses dna-pk¡sub¿cs¡/sub¿ expression and dna repair, and sensitizes cells to ionizing radiation. *International Journal of Radiation Oncology* Biology* Physics*, 65(3):842–850, 2006. 3.3

[249] Yu Sun and Edward A Clark. Expression of the c-myc proto-oncogene is essential for hiv-1 infection in activated t cells. *The Journal of experimental medicine*, 189(9):1391–1398,

1999. 3.2

[250] Lajos Szles, Dniel Trcsik, and Lszl Nagy. Ppar in immunity and inflammation: cell types and diseases. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids*, 1771(8):1014 – 1030, 2007. ISSN 1388-1981. doi: http://dx.doi.org/10.1016/j.bbalip.2007.02.005. URL http://www.sciencedirect.com/science/article/pii/S1388198107000364. PPARs. 2.9.1

[251] Lionel Tafforeau, Thibault Chantier, Fabrine Pradezynski, Johann Pellet, Philippe E Mangeot, Pierre-Olivier Vidalain, Patrice Andre, Chantal Rabourdin-Combe, and Vincent Lotteau. Generation and comprehensive analysis of an influenza virus polymerase cellular interaction network. *Journal of virology*, 85(24):13010–13018, 2011. 2.9

[252] 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. 3.2.5

[253] Osamu Takeuchi and Shizuo Akira. Innate immunity to virus infection. *Immunological reviews*, 227(1):75–86, 2009. 2.10

[254] 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. 3.2.5

[255] 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

[256] Nevins W Todd, Sergei P Atamas, Irina G Luzina, and Jeffrey R Galvin. Permanent alveolar collapse is the predominant mechanism in idiopathic pulmonary fibrosis. *Expert review of respiratory medicine*, 9(4):411–418, 2015. 4.3.2

[257] Susanne Toepfer, Reinhard Guthke, Dominik Driesch, Dirk Woetzel, and Michael Pfaff. The netgenerator algorithm: reconstruction of gene regulatory networks. In *Knowledge Discovery and Emergent Complexity in Bioinformatics*, pages 119–130. Springer, 2007. 1.6.2

[258] Ioannis P Tomos, Argyrios Tzouvelekis, Vassilis Aidinis, Effrosyni D Manali, Evangelos Bouros, Demosthenes Bouros, and Spyros A Papiris. Extracellular matrix remodeling in idiopathic pulmonary fibrosis. it is the bedthat counts and not the sleepers. *Expert review of respiratory medicine*, 11(4):299–309, 2017. 4.3.2

[259] 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. 2.10.2

[260] 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

[261] William D Travis, Ulrich Costabel, David M Hansell, Talmadge E King Jr, David A Lynch, Andrew G Nicholson, Christopher J Ryerson, Jay H Ryu, Moisés Selman, Athol U Wells, et al. An official american thoracic society/european respiratory society statement: update of the international multidisciplinary classification of the idiopathic interstitial pneumonias. *American journal of respiratory and critical care medicine*, 188(6):733–748, 2013. 1.3.3

[262] Barbara Treutlein, Doug G Brownfield, Angela R Wu, Norma F Neff, Gary L Mantalas, F Hernan Espinoza, Tushar J Desai, Mark A Krasnow, and Stephen R Quake. Reconstructing lineage hierarchies of the distal lung epithelium using single cell rna-seq. *Nature*, 509 (7500):371, 2014. 5.2.4

[263] Terrence M Tumpey, Kristy J Szretter, Neal Van Hoeven, Jacqueline M Katz, Georg Kochs, Otto Haller, Adolfo García-Sastre, and Peter Staeheli. The mx1 gene protects mice against the pandemic 1918 and highly lethal human h5n1 influenza viruses. *Journal of virology*, 81(19):10818–10821, 2007. 2.9.1

[264] Nurcan Tuncbag, Alfredo Braunstein, Andrea Pagnani, Shao-Shan Carol Huang, Jennifer Chayes, Christian Borgs, Riccardo Zecchina, and Ernest Fraenkel. Simultaneous reconstruction of multiple signaling pathways via the prize-collecting steiner forest problem. *Journal of Computational Biology*, 20(2):124–136, 2013. 1.6.1

[265] John J Tyson, Katherine C Chen, and Bela Novak. Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell. *Current opinion in cell biology*, 15(2):221–231, 2003. 1.6.1

[266] WHO UNAIDS. 2007 aids epidemic update. *PDF). December*, 2007. 1.3.2

[267] Pascale C van Weeren, Kim MT de Bruyn, Alida MM de Vries-Smits, Johan Van Lint, M Th Boudewijn, et al. Essential role for protein kinase b (pkb) in insulin-induced glycogen synthase kinase 3 inactivation characterization of dominant-negative mutant of pkb. *Journal of Biological Chemistry*, 273(21):13150–13156, 1998. 5.2.2

[268] Narasimhan J Venkatachari, Jennifer M Zerbato, Siddhartha Jain, Allison E Mancini, Ansuman Chattopadhyay, Nicolas Sluis-Cremer, Ziv Bar-Joseph, and Velpandi Ayyavoo. Temporal transcriptional response to latency reversing agents identifies specific factors regulating hiv-1 viral transcriptional switch. *Retrovirology*, 12(1):85, 2015. 3.3.1

[269] Narasimhan J Venkatachari, Siddhartha Jain, Leah Walker, Shalmali Bivalkar-Mehla, Ansuman Chattopadhyay, Ziv Bar-Joseph, Charles Rinaldo, Ann Ragin, Eric Seaberg, Andrew Levine, et al. Transcriptome analyses identify key cellular factors associated with hiv-1 associated neuropathogenesis in infected men. *AIDS (London, England)*, 31(5):623, 2017. 3.3

[270] 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

[271] Dorothee Viemann, Mirco Schmolke, Aloys Lueken, Yvonne Boergeling, Judith Friesenhagen, Helmut Wittkowski, Stephan Ludwig, and Johannes Roth. H5n1 virus activates signaling pathways in human endothelial cells resulting in a specific imbalanced inflammatory response. *The Journal of Immunology*, 186(1):164–173, 2011. 2.9.1

[272] Uwe Vinkemeier, Ismail Moarefi, James E Darnell, and John Kuriyan. Structure of the amino-terminal protein interaction domain of stat-4. *Science*, 279(5353):1048–1052, 1998. 2.9.1

[273] Donald Voet, Judith G Pratt Voet, W Charlotte, G Voet Judith, and W Pratt Charlotte. *Fundamentals of biochemistry: life at the molecular level*. Number 577.1 VOE. 2013. 5.2.2

[274] Martin Vogel, Carolynne Schwarze-Zander, Jan-Christian Wasmuth, Ulrich Spengler, Tilman Sauerbruch, and Jürgen Kurt Rockstroh. The treatment of patients with hiv. *Deutsches Ärzteblatt International*, 107(28-29):507, 2010. 1.3.2

[275] Dai Wang, Cynthia de la Fuente, Longwen Deng, Lai Wang, Irene Zilberman, Carolyn Eadie, Marlene Healey, Dana Stein, Thomas Denny, Lawrence E Harrison, et al. Inhibition of human immunodeficiency virus type 1 transcription by chemical cyclin-dependent kinase inhibitors. *Journal of virology*, 75(16):7266–7279, 2001. 3.4

[276] Pui Wang, Wenjun Song, Bobo Wing-Yee Mok, Pengxi Zhao, Kun Qin, Alexander Lai, Gavin JD Smith, Jinxia Zhang, Tianwei Lin, Yi Guan, et al. Nuclear factor 90 negatively regulates influenza virus replication by interacting with viral nucleoprotein. *Journal of virology*, 83(16):7850–7861, 2009. 2.9, 3.2

[277] 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

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

[279] Koichi Watashi, Mohammad Khan, Venkat RK Yedavalli, Man Lung Yeung, Klaus Strebel, and Kuan-Teh Jeang. Human immunodeficiency virus type 1 replication and regulation of apobec3g by peptidyl prolyl isomerase pin1. *Journal of virology*, 82(20): 9928–9936, 2008. 3.3

[280] Robin A Weiss. How does hiv cause aids? *SCIENCE-NEW YORK THEN WASHINGTON-*, 260:1273–1273, 1993. 1.3.2

[281] 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. 1.5.1

[282] Samuel A Williams, Lin-Feng Chen, Hakju Kwon, Carmen M Ruiz-Jarabo, Eric Verdin, and Warner C Greene. Nf-κb p50 promotes hiv latency through hdac recruitment and repression of transcriptional initiation. *The EMBO journal*, 25(1):139–149, 2006. 3.2

[283] 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. 3.2.5

[284] 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. 3.2.5

[285] Laurence A Wolsey and George L Nemhauser. *Integer and combinatorial optimization*. John Wiley & Sons, 2014. 1.5.2

[286] Paul J Wolters, Harold R Collard, and Kirk D Jones. Pathogenesis of idiopathic pulmonary fibrosis. *Annual Review of Pathology: Mechanisms of Disease*, 9:157–179, 2014. 4.3.1

[287] Angela R Wu, Norma F Neff, Tomer Kalisky, Piero Dalerba, Barbara Treutlein, Michael E Rothenberg, Francis M Mburu, Gary L Mantalas, Sopheak Sim, Michael F Clarke, et al. Quantitative assessment of single-cell rna-sequencing methods. *Nature methods*, 11(1): 41–46, 2014. 5.2.4

[288] Ke Xu, Christoph Klenk, Bin Liu, Bjoern Keiner, Jinke Cheng, Bo-Jian Zheng, Li Li, Qinglin Han, Chen Wang, Tianxian Li, et al. Modification of nonstructural protein 1 of influenza a virus by sumo1. *Journal of virology*, 85(2):1086–1098, 2011. 2.9.1

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

[290] Esti Yeger-Lotem, Laura Riva, Linhui Julie Su, Aaron D Gitler, Anil G Cashikar, Oliver D King, Pavan K Auluck, Melissa L Geddie, Julie S Valastyan, David R Karger, et al. Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. *Nature genetics*, 41(3):316–323, 2009. 1.6

[291] Nir Yosef, Alex K Shalek, Jellert T Gaublomme, Hulin Jin, Youjin Lee, Amit Awasthi, Chuan Wu, Katarzyna Karwacz, Sheng Xiao, Marsela Jorgolli, et al. Dynamic regulatory network controlling th17 cell differentiation. *Nature*, 496(7446):461, 2013. 1.6.3

[292] H. P. Young and A. Levenglick. A Consistent Extension of Condorcet's Election Principle. *SIAM Journal on Applied Mathematics*, 35(2), 1978. ISSN 00361399. doi: 10.2307/2100667. URL `http://dx.doi.org/10.2307/2100667`. 2.10.1

[293] Robert M Youngson. *Collins dictionary of human biology*. Collins, 2006. 1.4.3

[294] 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

[295] 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. 3.2.5

[296] Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012. 1.5.4

[297] 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

[298] Antonis S Zervos, Jenő Gyuris, and Roger Brent. Mxi1, a protein that specifically interacts with max to bind myc-max recognition sites. *Cell*, 72(2):223–232, 1993. 3.4

[299] Yijie Zhai, Luis M Franco, Robert L Atmar, John M Quarles, Nancy Arden, Kristine L Bucasas, Janet M Wells, Diane Niño, Xueqing Wang, Gladys E Zapata, et al. Host transcriptional response to influenza and other acute respiratory viral infections–a prospective cohort study. *PLoS pathogens*, 11(6):e1004869, 2015. 1.6.3, 1.8

[300] 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

[301] Shan Zhong. *Computational Study of Transcriptional Regulation-From Sequence To Expression*. PhD thesis, Carnegie Mellon University, 5 2013. 1.4.2, 1.2

[302] 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, 5.2.3