Prospective Identification of Long-range Transcriptional Regulatory Regions via Integrative Genomics

by

Arvind Rao

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Electrical Engineering:Systems and Bioinformatics)
in The University of Michigan
2008

Doctoral Committee:

Professor James Douglas Engel, Co-Chair
Professor Alfred O. Hero, III, Co-Chair
Professor David J. States, Co-Chair
Professor Jeffrey A. Fessler
Associate Professor Matthias Kretzler

ACKNOWLEDGEMENTS

Over the past few years I have been extremely fortunate to have the company and mentorship of the smartest and nicest people I know. My advisors, Prof. Alfred Hero, Prof. David States and Prof. Doug Engel have shepherded me through my graduate school experience with their constant encouragement, enthusiasm, infinite patience and strong team-spirit. I hope that I am able to carry these learnings forward into my life. I am very grateful to Prof. Jeff Fessler and Prof. Matthias Kretzler for agreeing to be part of my committee as well as for their continuous encouragement of my work. Prof. Peter Woolf has been a wonderful friend – I have learnt tremendously from my interactions with him. Prof. Demosthenis Teneketzis, Prof. Gil Omenn and Prof. Sandeep Pradhan have been very generous with their support and advice about the academic worklife. I have now come to realize that genuine commitment to the research problem is the one thing that can lead us comfortably through life in academic research.

I have had the unique fortune of interacting with three different research groups on campus spanning engineering, bioinformatics and biology. I have learnt a lot from my interactions with each of my fellow group members, who have provided me with several fun-filled experiences both within and outside of work. I would like to particularly acknowledge Dr. Kim Lim and Dr. Takashi Moriguchi of the Engel lab, Dr. Tom Blackwell, Dr. Damian Fermin and Dr. Raji Menon of the States lab, and Dr. Mark Kliger from the Hero lab for their help and guidance through the umpteen

situations in graduate school. Also, a special thanks to Becky, Nancy, Lori, Kristen, Denise, Julia and Yuri who helped me through the various administrative processes, every time with a wonderfully cheerful and encouraging disposition.

Over the years, both at UT and UM, I have been blessed to have the company and friendship of some of the nicest people I know, both amongst faculty and fellow graduate students. Fortunately, they are too numerous to mention – but I will single out those friends who have stood by me since when I started here at UM; Dinesh, Ramji, Aditya, Raghu, Mekhala, Aniket, Nitin, Sarad, Manisha, Divya, Swapnaa, Shyam, Vijay, Somesh, Manoj, Manickam, Aarthi, Bala and of course Siva. They have always provided the true balance in my life, pointing out what is really important, and provided encouragement and mirth during the ups and downs of this journey. Mere words of thanks cannot express the gratitude I have for their presence in my life. A special thanks to all my teachers and friends – right from high school, undergraduate studies at RVCE, Bangalore and Prof. A. G. Ramakrishnan of the Indian Institute of Science, Bangalore who taught me the fundamentals of so many interesting areas. At the University of Texas, I was fortunate to have the guidance and mentorship of Prof. Brian Evans, Prof. Inderjit Dhillon, Prof. Robert Heath, Prof. Gustavo deVeciana, Mohan Sridharan and Vishal Monga in my choice of research direction. I am indebted to each and every one of you.

Finally, my family has been solidly behind me during my educational journey. Their encouragement, belief and emotional support have been extremely important to my day-to-day life as a student. I dedicate this thesis to them.

Support from the following agencies is gratefully acknowledged: the National Institutes of Health (NIH) under grant 5R01-GM028896-21 (to Prof. Doug Engel), the Center for Computational Medicine and Biology (CCMB) Pilot Award, 2007 and

the Rackham Predoctoral Fellowship, 2008.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

CHAPTER I

Introduction

1.1 Thesis Overview

With the advent of knowing the complete genome sequence for many organisms, we now know where each protein-encoding gene is in the human genome. However, the proper function of a cell depends on eliciting appropriate gene expression, in the amount, tissue and time of production of the messenger RNA that encodes the proteins transcribed from every gene. 98% of mammalian genomic sequences do not encode proteins. Previously this non-coding DNA was thought to be "junk"; however, many recent examples in the literature have now shown that a significant portion of this non-coding DNA is involved in the precise regulation of gene expression.

This thesis deals with identifying these critical regulatory elements within the functional non-coding DNA by using computational approaches, specifically using modern methods of statistical learning that can incorporate many different types of experiments - from DNA sequence to gene expression chips to protein-protein and protein-DNA interaction data - that are currently being generated in thousands of high throughput experiments.

Eukaryotic gene expression is regulated by the recruitment of TF (transcription

factor) proteins to the proximal promoter, close to a gene's transcriptional start site, as well as to long-range regulatory elements (LRE). LREs can lie several hundred thousands of base-pairs (kilobase-pairs) away from the actual gene. LREs play important roles in the spatial (tissue-specific) and temporal expression of any gene – and analysis of LREs can be informative for the study of key biological processes, like organ development and disease progression. Current methods to prospectively identify LREs are based on the analysis of inter-species conservation (suggesting evolutionary selection) along the genome sequence and then experimentally examining the role of each conserved sequence element (CSE) in vivo. Because of the large number of such conserved non-coding sequence elements, employing such unselective approaches to identify tissue-specific LREs is experimentally laborious, very costly and unscalable to large genomic loci.

The primary goal of this thesis is to generate a small list of high confidence candidate LREs for any given gene, using the large amount of data that has been generated from high throughput experiments - with the purpose of accelerating discovery and validation. Though the genes considered in this work are *Gata2* & *Gata3*, the methods that we develop are general and can be extended to any gene of interest. The *Gata2* & *Gata3* genes are involved in the development of several important organ systems, such as the urogenital system and central nervous system.

Below, I summarize the main contributions of this thesis.

1. Transcriptional regulatory networks: As suggested above, the recruitment of transcription factors to an LRE that will in turn regulate the expression of a target gene (e.g. *Gata3*) implies the presence of an influence between the "effector" TF (a TF that binds to the LRE) and *Gata3*. Biologists are acutely interested in such influence networks, to look for the presence of such TF binding

site sequences in each CSE, thereby increasing the probability of that CSE being a regulatory element. However, the inference of TF effectors has had two main challenges - previous methods have modeled the dependency between TF and their target-gene as a static phenomenon. Cell processes are, on the contrary, strongly dynamic. Additionally, resolving the direction of influence is an equally important component to delineate the true transcriptional effectors.

In this problem, based on data from gene expression chip (microarray) time series for early development, we propose time-varying and context-specific networks as a framework to model stage-specific transient gene influences. A time-varying network as a model for transcriptional influences has not been examined in the literature so far; such a model is biologically relevant as biological network relationships are rarely ubiquitously active (Chapter 2).

Next, to resolve the direction of dependence between the TF effector and the target gene, we derive a metric to find directed dependency based on gene expression. Inspired from information theory, "directed information" is a new solution in such problems. This metric enables one to examine directed relationships between specific network components, thereby improving on the current state of the art. This metric has strong performance characteristics in spite of the intrinsic non-linearity of gene expression, and outperforms several other metrics that make strong assumptions on the nature of expression data. It is a metric that can resolve direction from highly non-linear gene expression data (Chapter 3).

2. Data Fusion: Given the diversity of genomic data sources that can be potentially mined to identify putative TF effectors, we seek to integrate data from

various modalities (sequence, expression, protein-interaction). We have developed a novel framework to statistically co-embed gene relationships from each modality onto a common space. Such a framework makes it possible to study the combined effect of relationships across different data sources, enabling visualization, interpretation and discovery of high probability LREs. No application for co-embedding data, till recently a purely mathematical problem, has been demonstrated for computational biology before (Chapter 5 and 6).

3. Another perspective to understand the nature of regulatory regions is to examine if their genomic sequences have any specific features. These features can be the over-occurrence of certain key patterns, called "sequence motifs". The availability of experimental DNA sequence data that underlies spatio-temporal expression in some genes enables the design of an approach to discover "motifs" that confer such expression. Using a neutral set of sequences (those that do not confer expression in experiment), we developed a learning paradigm to find 6-nucleotide motifs that can discriminate tissue-specific elements from neutral ones. The identified sequence motifs have a fairly high predictive power for potential specificity of tissue expression and are being used to design statistical classifiers that predict the specificity of new sequences. I have developed two new methodologies - one based on random forest classifiers and the other based on adapting the directed information criterion for feature selection, for this problem (Chapter 4, 5). Such a dataset can be further processed to understand any language-level features (grammar of transcription factor binding, spacing etc) using advanced tools such as structured prediction.

4. Graph Mining: To obtain a mechanistic insight into transcriptional regulation,

we explore the structure of the interaction-graph between TFs at the promoter and those at the enhancer. These graphs are derived from protein interaction databases. Our analyses indicate that several graph metrics, such as centrality, density, heterogeneity are indicative of the cohesive strength that exists between the enhancer and promoter, during distal interaction. We observe that, the higher the cohesive strength of interaction between the TFs, the stronger the chance that the candidate element is truly regulatory (Chapter 5).

5. *In-vivo* experiments are necessary to validate computational predictions. To experimentally validate our model and results, we identified possible LREs from candidate genomic sequences using the integrative model developed here. These regions were cloned upstream of the *Gata3* promoter and assayed for their tissue- and temporally-regulated expression in transgenic mice. This has led to the discovery of a new inner-ear (by Dr. Kim Lim) and pyloric (by Dr. T. Moriguchi) enhancers for *Gata3*. Additionally, the behavior of *Gata2* and *Gata3* enhancers with known spatio-temporal gene expression has also been reconciled through the integration of diverse data sources (Chapter 5, 6).

1.2 Long-range Transcriptional Regulation in *Gata2/Gata3*

Transcription is the process of generation of messenger RNA from the DNA template (or gene). Transcription is initiated by the recruitment of RNA Polymerase II, and regulated by several cis- and trans- elements. These cis- elements, also referred to as enhancers/silencers/insulators are DNA sequences where trans-activating factors (transcription factor proteins) are recruited during formation of the transcriptional machinery that is responsible for transcriptional initiation and elongation. It has been observed that the precise spatio-temporal expression of genes is exquis-

itely regulated at these promoter proximal or distal enhancers, some of which can lie hundreds of kilobases from the transcriptional start site. Hence the identification of these enhancers is critical to understanding gene function and role. Since different enhancers are responsible for conferring expression in different tissues, their precise localization and characterization is crucial to the understanding of fundamental biological processes such as disease and development. In this work, we focus on the regulation of the *Gata2* and *Gata3* genes, which have roles in urogenital, cardiac, hematopoietic, and neural development.

The GATA family of transcription factors (GATA-1 through -6) play critical but diverse roles in biological processes that functionally contribute to mammalian development. Research in our laboratory focuses on the characterization of cis and trans elements that specify how the *Gata2* and *Gata3* genes are controlled in individual tissues. Here, we have analyzed the molecular mechanism(s) that regulate the expression of *Gata3* in the urogenital (UG) and sympathetic nervous /sympathoadrenal systems (SAS). The *Gata3* gene is prominently expressed in the UG/SNS and is centrally implicated in the control of noradrenergic differentiation ([15], [19]) and mammalian kidney morphogenesis [198]. However, the mechanism (i.e. cis regulatory elements and trans-acting factors) by which UG/SAS-specific activation of *Gata3* is achieved is not known.

Tissue-specific expression of genes in metazoa is often governed by cis elements that can lie enormous distances from the structural gene that they regulate ([11], [211]), but there is no current algorithm for identifying tissue-specific enhancer elements with high confidence. An approach that has been recently adopted is based on the identification of DNA sequences that are evolutionarily conserved (conserved sequence elements or CSEs) between syntenic regions of multiple genomes [7]. Al-

though this strategy usually yields many candidate elements (in fact, too many), there is no strategy for distinguishing among these for which might be the most likely to control the activity of a gene in a specific tissue. Furthermore, individually assaying each candidate CSE for regulatory activity is not a viable experimental approach (i.e., both too costly and too time consuming). Currently, we are attempting to extend current bioinformatics approaches to this problem by exploring whether or not additional insight can be gleaned from genomic resources (expression profiles, protein-protein interaction data or phylogenetic studies) to refine the position of functionally relevant gene regulatory elements. The hypothesis explored in the thesis is that integration of these under explored genomic features into a sequence conservation construct should reduce the number of candidate regulatory elements to a small, high confidence set. We explore using this bioinformatics strategy for identifying UG/SAS-specific regulatory elements that control both *Gata3* and other kidney/SAS-specific genes.

We recently localized the positions of UG-specific and SA-specific *Gata3* enhancer element(s) on mouse chromosome 2, between positions 9,065,411 and 9,226,186 and positions 9,226,186 and 9,271,464 respectively (Lim, Moriguchi and Rao, unpublished). A central experimental hypothesis to be tested in this work is that UG-specific (or SA-specific) *Gata3* expression is determined by one or more CSEs located within these intervals. Thus our aim has been to localize these UG/SA-specific gene regulatory element(s) to reveal the molecular basis for *Gata3* urogenital/ sympathoadrenal-specific gene regulation, by testing whether CSEs identified using the bioinformatics strategy are functional. The overall enhancer identification strategy has two main components.

1.2.1 Part 1

Does examination of known enhancer elements that confer tissue-specific gene expression reveal useful 'features' in the context of available genomic data? Known enhancer elements associated with genes that have been identified both in our laboratory and others have been examined. Features responsible for conferring tissue specificity in these enhancer elements were identified by correlating sequence information, gene expression and other genomic data available from public repositories and linking them to the expression pattern they direct (some preliminary work is reported in ref. [24] with respect to the previously-identified urogenital enhancers of the *Gata2* and *Gata3* genes). An integrative classifier that can learn from these features and distinguish between regulatory and non-regulatory elements has been designed, and is described in Chapter 5.

1.2.2 Part 2

Which of the candidate conserved sequence elements within two overlapping BACs direct expression of *Gata3* in the UG/SAS? An experimental strategy has been implemented that involves the cloning of candidate CSEs into a vector employing the *Gata3* proximal promoter to direct lacZ expression, generating transgenic mice, and determining whether any individual CSE directs in vivo reporter gene expression in a pattern that mimics that of endogenous *Gata3*. This step provides an experimental validation for the UG/SAS-specific regulatory activity of the high confidence candidate CSEs identified in the bioinformatics research aim,that is the subject of this thesis. This work has been done by Dr. K.C. Lim (for the UGE case) and Dr. T. Moriguchi (for the SA case) in the Engel Laboratory. This part will be described subsequently in other work and we will only address the various computational meth-

ods in this thesis. At this point we have been able to achieve a fair discrimination of enhancers vs. neutral regulatory elements but we face additional challenges in order to predict the exact tissue-specificity of each identified enhancer. In the last chapter, we also present some recent analysis on identifying a T-cell specific *Gata3* regulatory element. The experimental component of this project is being completed by Dr. S. Hosoya-Ohmura and Dr. T. Kuroha of the Engel laboratory.

1.3 Rationale, motivation, preliminary data

The SAS, and the adrenalin/noradrenalin biosynthetic pathway play a central role in the pathogenesis of hypertension in humans as well as in rodent models. The zinc finger transcription factor *Gata3* plays a vital role in regulating two essential genes of the noradrenalin biosynthetic pathway: tyrosine hydroxylase (TH) and dopamine-β -hydroxylase (DBH) ([19], [22] and T. Moriguchi, personal communication). *Gata3* germ line mutant embryos have reduced accumulation of TH and DBH and consequently suffer mid-embryonic demise as a result of noradrenalin insufficiency ([15], [19]). *Gata3* expression persists in the SAS from early embryonic stages in the sympathetic chain and adrenal chromaffin cells, suggesting a life long contribution of *Gata3* to catecholamine biosynthesis and the control of sympathetic tone ([4], [11], [15], [19]). *Gata3* contains two steroid hormone receptor-like zinc fingers that directly bind to DNA [27], and recognize the consensus motif GATA that is highly conserved amongst all six members (GATA-1 through -6) of this multigene family [10] .

The *Gata3* gene is also expressed during nephrogenesis starting from the expression in the Wolffian duct. Later, in the metanephros, it is expressed in the ureteric bud and is involved throughout development to give rise to the entire col-

lecting system of the developing kidney. *Gata3* thus plays an important role in the differentiation program of the kidney [31]. There have also been detailed studies that point out the critical role of *Gata3* on nephrogenesis in a dosage-dependent manner. *Gata3* loss leads to severe abnormalities in the developing kidney and is reminiscent of the kidney deficiency phenotype associated with the HDR syndrome (hypoparathyroidism, deafness, renal dysplasia).

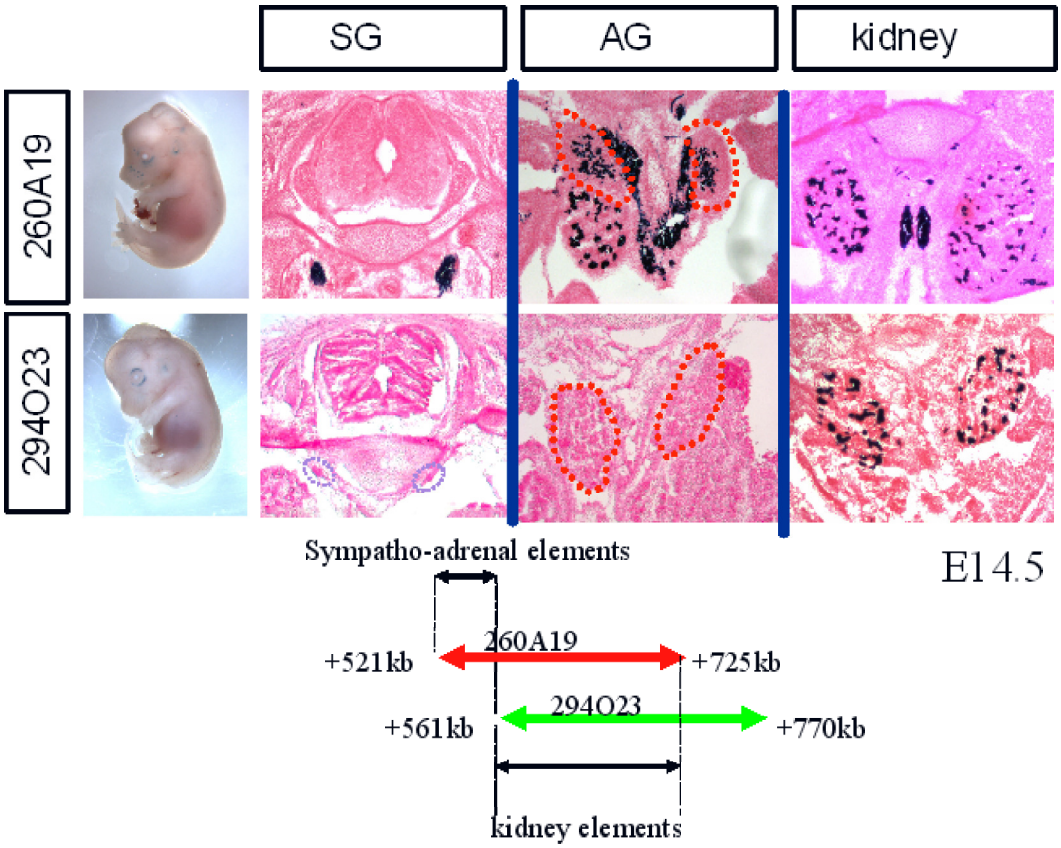


Figure 1.1: Localization of *Gata3* sympathoadrenal and kidney enhancer activity: *Gata3* BAC 260A19 confers sympathetic ganglia (SG), adrenal gland (AG, dotted red outline, arrows) and, at this stage (14.5 dpc), Zukerkandl organ (arrowheads) show beta-galactosidase staining, while overlapping BAC 294O23 does not. However, the region of overlap between these two BACs confers expression in the mesonephros and metanephros in the developing urogenital system, and putatively harbors the urogenital enhancer. (from Dr. T. Moriguchi, Engel laboratory)

Our goal in initiating these studies was to identify the regulatory elements underlying transcriptional regulation of *Gata3* in the UG and SAS. This has been a

daunting experimental task, since our lab showed more than 6 years ago that the UG and SAS-specific element(s) lie more than 400 kbp 5' or 200 kbp 3' to the *Gata3* structural gene on Mm chromosome 2 [11]. The Engel laboratory has recently developed a strategy using bacterial artificial chromosomes (BACs), each encompassing approximately 200 kbp of genomic sequence, to scan the genome in the vicinity of any gene for potential regulatory activity ([12], [204]). We have very recently localized the position of at least one such distal regulatory element on Mm chr2, between genome positions 9226186 and 9271464, which appears to harbor SAS-specific activity (Moriguchi and Rao, unpublished). One BAC, RP23-260A19, is capable of directing lacZ reporter activity in the adrenal gland and sympathetic ganglia, the two main sites of SAS activity, while an overlapping BAC, RP23-294O23, does not (Fig. 1.1), thus localizing *Gata3* SA activity to a region between +521 and +566 kbp 3' to the *Gata3* structural gene. (There are no other SA-specific genes anywhere near this region of the genome). Similarly, the $\sim 150kb$ overlapping region between BACS RP23-260A19 and RP23-294O23 (i.e., chr2: 9065411-9226186), which is +566kbp to +725kbp has UG specific activity (Fig. 1.1).

Without further defining the potential enhancer elements within this 46 kbp (or 150kbp) region, we would need to experimentally test each candidate element. A candidate element is defined as any sequence that is under positive selective pressure (i.e., does not mutate) and thus is evolutionarily conserved. The identification of conserved sequence elements (CSEs) lying in this 46 kbp (or 150kbp) is the first step towards defining an SAS-specific (or UG-specific) enhancer element. Fig. 1.2 (red peaks) defines the CSEs (with more than 70% sequence conservation of length $\geq$ 100 bp) within this 46 kbp interval in the mouse (Mm) relative to Human (Hs), Rat (Rn), and Dog (Cf). None of these are conserved in the pufferfish (Fugu), which

does have a sympathetic nervous system, but not adrenal glands or chromaffin cells, used as a control genome sequence. Fig. 1.4 similarly defines the location of the CSEs which does have UG-specific activity. At least 120 CSEs are found within this 46 kbp ($\geq$ 200 CSEs in the 150kbp) interval. To reduce the number of candidates to a significantly smaller, high confidence subset, other complementary approaches will need to be employed, which can then be finally explored empirically.

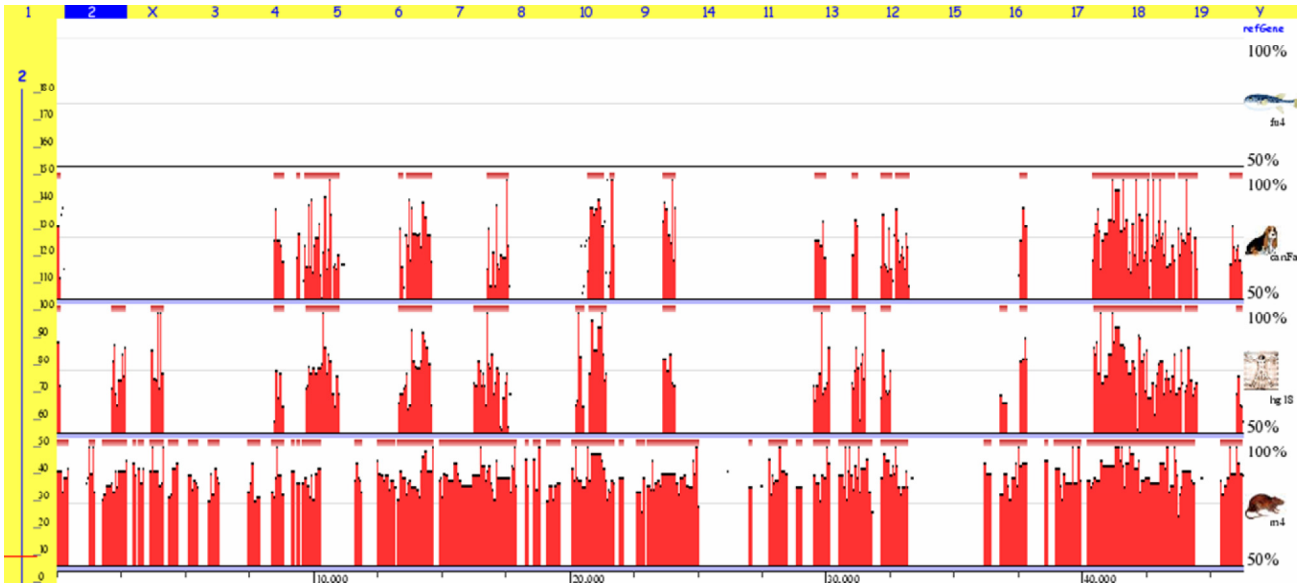


Figure 1.2: Conserved Sequence Elements in a cross-genome comparison between human, mouse, rat, dog and pufferfish (generated using *http://www.ecrbrowser.dcode.org*). Each red region is a candidate enhancer.

A number of computational algorithms to identify enhancers on the basis of conservation studies and transcription factor binding site (TFBS) clustering have been developed and applied, and a subset have been shown to harbor the anticipated *in vivo* regulatory activity [7]. Most methods identify regulatory sequences from interspecies sequence comparisons; such CSEs are empirically defined as regions of high (typically $\geq$85%) DNA sequence identity. The intrinsic assumption is that this conservation reflects conserved regulatory function. Within the regions of strongest conservation, a search for TFBSs from a set of co-regulated genes involved in the bio-

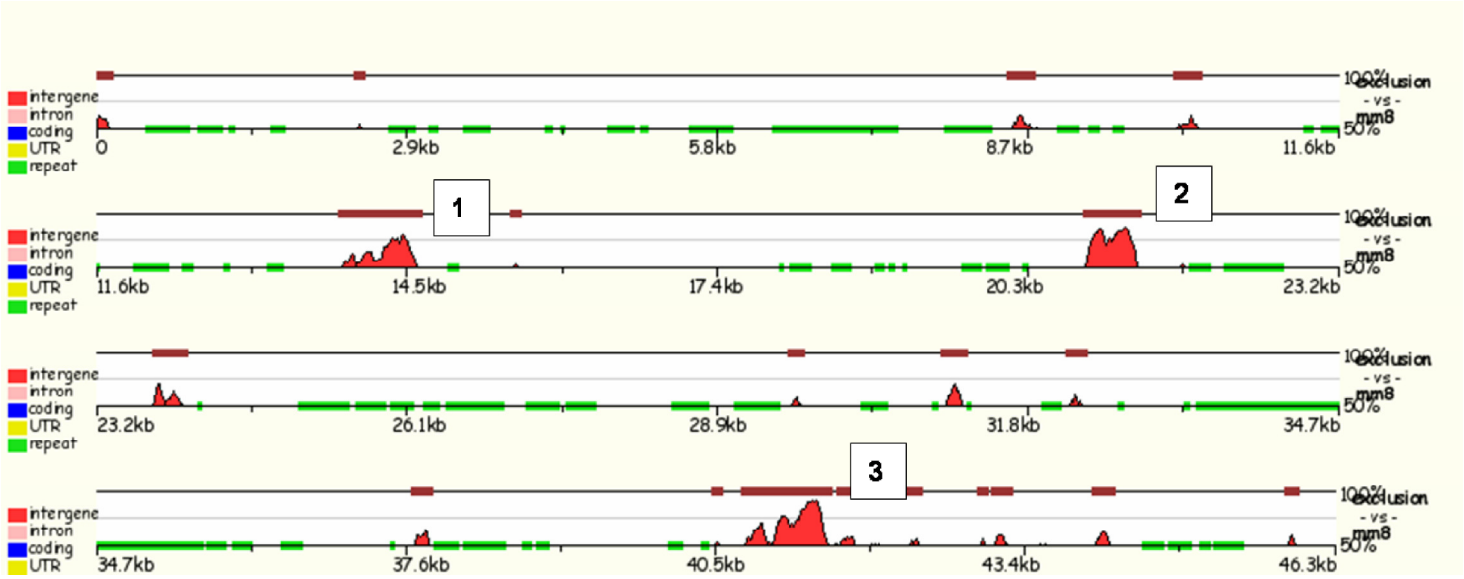


Figure 1.3: Phylogenetic Shadowing of four mammalian genomes to highlight some CSEs under selective pressure(red regions)

logical process of interest usually reveals several candidates. Although simple, these methods have been effective in identifying some mammalian enhancers ([7], [215]). Despite sporadic successes, the efficacy of comparative sequence analyses coupled to binding site cluster detection in regulatory sequence discovery is still unclear, especially in the context of attempting to define organ-specific gene regulation elicited by enhancers lying far from the gene that they control. However, recent results from the ENCODE project have revealed at least one interesting observation – functional regulatory elements are not necessarily highly conserved in primary sequence [188]. Thus, the threshold for conservation might need to be lowered to prospectively identify these potential regulatory elements, leading to an increase in the false positive rate. One way to ensure improved detection is the principled integration of other data sources that could supplement sequence information.

The hypothesis we will explore in this thesis is: given the vast amount of diverse genomic data that is currently available, is it possible to glean additional informa-

tion (features) about potential regulatory activity, to reduce a large set of candidate elements to a smaller, higher confidence set? Publicly available genome data such as microarray expression, protein-protein interaction maps, tissue expression and phylogenetic information could be very useful in identifying features relevant to the identification of such regulatory elements. We present some case studies for understanding the architecture of kidney-specific enhancers of the *Gata2* and *Gata3* genes based on these diverse data sources ([198], [204], [24]). In this work, we explored each of these data repositories (Chapters 2-4) and attempted to integrate them into a single strategy (Chapter 5) to aid in the identification of high confidence candidate regulatory elements. We anticipate that this strategy will lead to the development of an approach that will be applicable to any gene for the purpose of identifying the underlying transcriptional control elements.

1.4 Research Outline

In the first part of this thesis, we will explore the question: does examination of well characterized enhancer elements that control tissue-specific gene regulation reveal any interesting 'features' in the context of available genomic data?

Transcriptional regulation in eukaryotes is a complex process involving the recruitment of multiple transcription factors (TFs) to the basal promoter, as well as to distal regulatory cis elements. These promoter and enhancer elements communicate by looping the DNA between them, over large distances, to drive regulation of the target gene ([2], [1]). These activities are thought to be mediated by protein:protein interactions that take place between the TFs bound to DNA at the basal promoter and at the distal enhancer element to form an active transcriptional complex. Distal elements are evolutionarily selected for function, meaning that they

are conserved among related species [21]. Several studies have reported the presence of other sequence features that contribute to the regulatory role of a conserved sequence element (CSE), ([205], [20]). Individual TFBSs within CSEs are also under evolutionary pressure [21]. In many cases, transcription factor binding to the distal elements is thought to confer tissue-specific expression of a target gene [20].

In this work, I have assessed the utility of the following genomic 'features' in identifying potential regulatory elements: sequence features (phylogenetic shadowing and histone-modification potential), transcription factor features (phylogenetically conserved TFBSs, tissue-specific TFBSs and network linkage) and ontology features. We have examined multiple, publicly available genomic databases and to extract features that contribute to the regulatory elements underlying UG/SAS-specific regulation of *Gata3*. As an example, the steps that we will follow to define a urogenital/ sympathoadrenal enhancer (UGE/SAE) from among the CSEs within the experimentally defined interval of 46 kbp (or 150kbp) (Fig. 1.4 and 1.2) is described below.

1.4.1 Sequence features

The DCODE website (http://www.dcode.org) is used to establish the comparative genomic analyses relevant to this work. We initially aligned the *Gata3* locus (Mm chr2:9774937-9795629, and the SA region 521-566 kbp or 566-725 kbp UG region 3' to the *Gata3* translation initiation site) with several annotated genomes. An appropriate choice of genomes for cross-genome comparison is essential to the identification of potentially functional elements. Since the sympathetic nervous system (SAS) functions similarly in human (Hs), mouse (Mm), rat (Rn) and dog (Cf), we will have used 4 genomes for genome-wide comparisons. For the urogenital function, comparisons up to Chicken (Gg) have been found to be meaningful [30].

- Phylogenetic shadowing [21] , a rebirth of phylogenetic footprinting [6] , identifies highly species-conserved sequences within this 46 kbp (or 150 kbp) interval ($\geq$ 100 bp with $\geq$ 60% sequence identity; a fairly low-stringency criterion that is a useful metric to identify conserved sequences). Fig. 1.3 highlights phylogenetic shadowing over this interval, which identifies (in red) regions conserved across all four genomes (Hs, Mm, Rn and Cf). Along the abscissa, green shading indicates the position of repetitive sequences that lie in the interval. The ordinate axis characterizes the degree of conservation between genomes. Each red region is a potentially important functional regulatory element. The top three regions, in the order of increasing sequence identity, are labeled 1, 2 and 3. While there are several shorter regions which are also of interest, they are not well illustrated in the figure because of the large size of the genomic region that was surveyed.

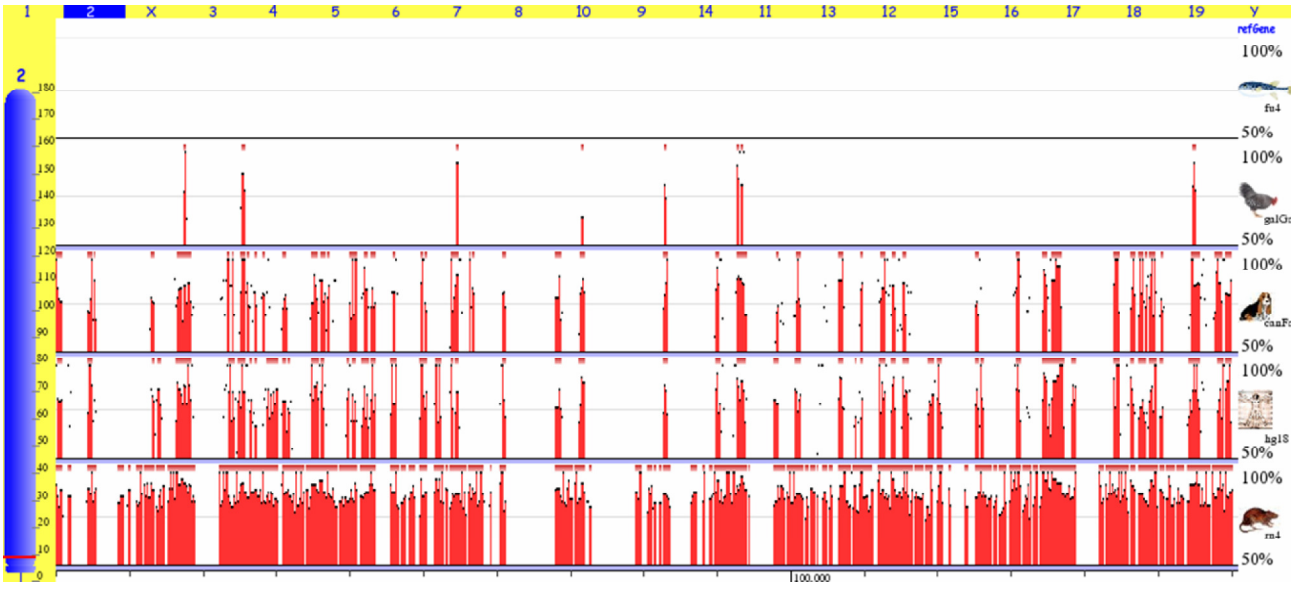


Figure 1.4: Conserved Sequence Elements in a cross-genome comparison between human, mouse, rat, dog, chicken and pufferfish (generated using *http://www.ecrbrowser.dcode.org*). Each red region is a candidate enhancer.

Fig. 1.5 presents the same analysis for the 150 kb UGE region. There are about

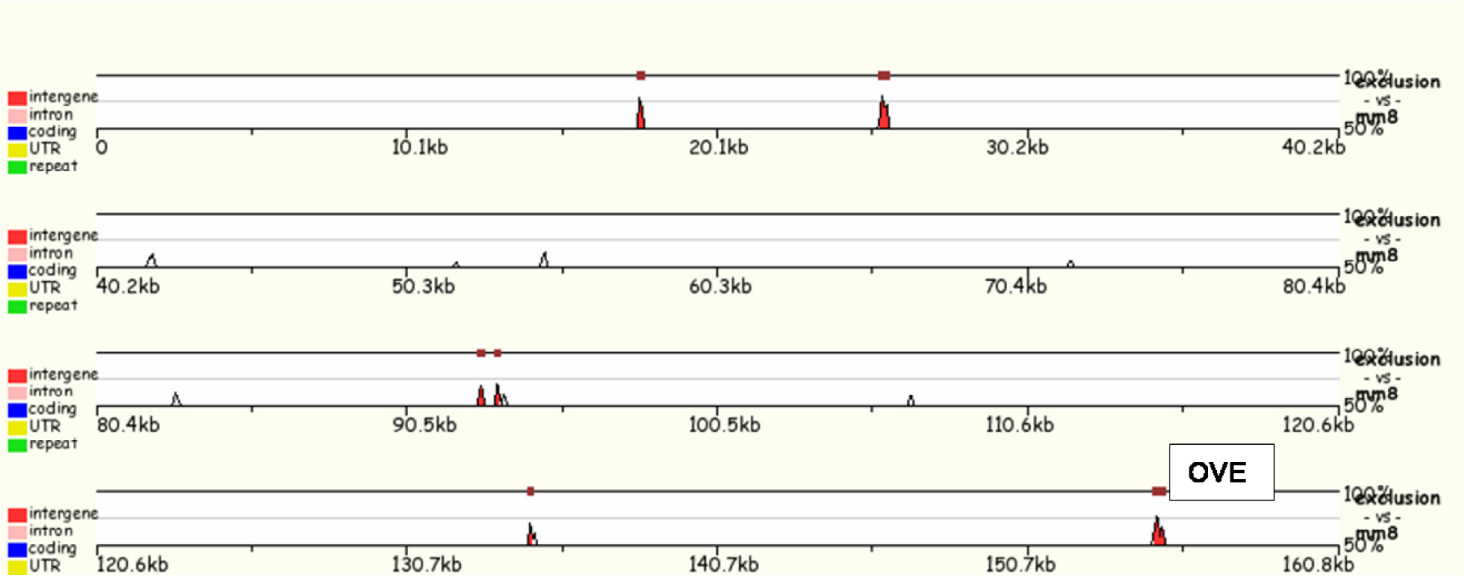


Figure 1.5: Phylogenetic Shadowing of five mammalian genomes to highlight some CSEs under selective pressure(red regions)

8 regions that are conserved between human, opossum, rat, mouse , dog, and chicken and each is a potential UG regulatory region.

- An examination of histone modifications at promoter sequences and some long-range regulatory elements for the ENCODE regions [188] indicate that promoters usually express trimethylated histone H3 lysine 4 residue (i.e. H3K4me3) or H3 acetylation whereas enhancers undergo H3K4 monomethylation. This yields two classes of sequences that have different propensities for histone modification that are amenable to motif discovery approaches outlined below. This is in line with finding a "histone code" related to enhancer characterization. Though this is cell-context dependent, we have observed that the motifs identified from such an approach are fairly discriminatory for nucleosome occupancy and histone modification. We have built a random forest (RF) classifier that yields high classification accuracy and has very good receiver operating characteristic when validated on known urogenital enhancers of *Gata2* and *Gata3* [24]. Such a clas-

sifier is a statistical tool that can discriminate promoter and enhancer sequences by building rules over derived motifs.

Random forest (RF) based histone-modification scores are used to indicate, among a group of CSEs identified by phylogenetic shadowing [24], the probability that a specified sequence is functionally involved in the regulation of a specific target gene. The UCSC genome browser (http://genome.ucsc.edu) is used for this analysis. Among the CSEs identified that might be a sympathoadrenal enhancer (SAE), the RF score of CSE2 (Fig. 1.3) is highest. The implication here is that the higher the RF score, the more likely is the possibility that the sequence is functionally involved in gene regulation. Since this 46 kbp (or 150kbp) interval was shown empirically to harbor SAE (resp. UGE) activity (Fig. 1.1), a high RF score for any CSE in this region recommends its further exploration as a candidate SAE (or UGE). [It should be noted that RF scores were initially established from data describing the enhancers and promoters of a few selected cell lines. It is not established that use of this data set of known regulatory elements leads to unbiased histone-modification scores in other cells, though they seem to have good performance on other promoter and enhancer datasets (ENSEMBL/MGI, Enhancer Browser, and urogenital enhancers of *Gata2*, *Gata3*), [24])].

1.4.2 Transcription factor features

Identification of phylogenetically conserved Transcription Factor Binding Sites (TFBSs): Regulatory elements typically bind clusters of multiple transcription factor binding sites. Databases, such as TRANSFAC and JASPAR are used to analyze sequences for potential TFBS motifs. A transcription factor is a protein which binds

at the promoter (basal transcriptional machinery) or at any of the distal regulatory elements (enhancers or silencers) and is involved in either activating or repressing the expression of transcriptionally-associated genes by binding to sites within candidate CSEs. The identification of phylogenetically conserved TFBS may further refine the array of potential candidate transcription factors required for the localized expression of *Gata3* in the SAS or UG regions ([19], [228]). We note that such methods can discriminate a family of TFs, but not the precise family member that might be involved in mediating transcription in that tissue. To reconcile the behavior of various promoter-specific transcriptional regulatory regions - we decouple the roles of the promoter and the enhancer by examining which set of TFs bind each region and how they interact during formation of the transcriptional machinery.

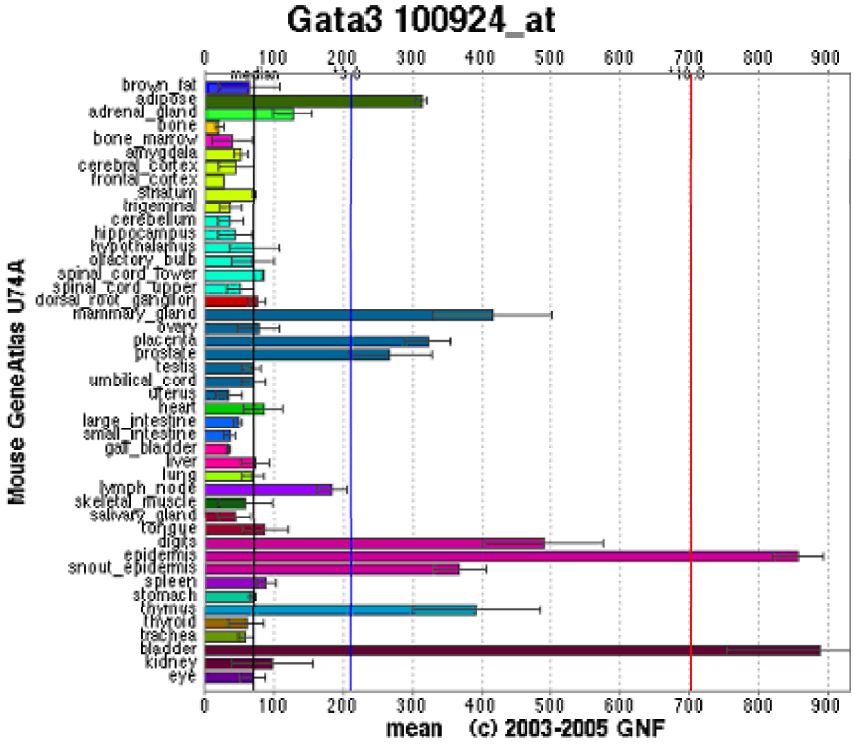


Figure 1.6: Tissue-specific expression of *Gata3* in various tissue types as assayed using the murine U74Av2 microarray chip (from *http://www.symatlas.gnf.org*)

TF binding at the promoter:

1. To identify which set of transcription factor putatively bind at the promoter we use the 'module TF' approach. Module TFs are groups of transcription factors that bind a set of co-expressed genes, given the hypothesis that co-expressed genes are possibly co-regulated. For *Gata3*, we examine the mouse genome informatics (MGI) database to find genes that are co-expressed on day e14.5 in the developing metanephros. We use the TOUCAN tool (http://homes.esat.kuleuven.be/s̃aerts/software/toucan.php/) to find these statistically over-represented TFs. Additionally, we use Gene Ontology information to find TFs with fewer degrees of freedom (i.e. more candidate TFs). These approaches are further clarified in Chapter 6.

2. Another way to find the set of TFs that putatively regulate *Gata3* is to use microarray gene expression data. Using publicly available expression data for the developing kidney ([80], [178], [240]), we adopted several methods from time series analysis (Chapters 2-3) to find transcriptional effectors. Thus, using a combination of TFBS match at the promoter and expression data, we are able to enrich for a list of TFs that bind the *Gata3* proximal promoter.

3. Publicly available gene expression microarray data (GEO:http://www.ncbi.nlm.nih.gov/geo/) for the SAS/UG provides a framework for correlating gene expression profiles and linking them to functional inter-dependence. Genes having high correlation metrics are examined for the presence of 'module regulators' [209] , which are, in essence, combinations of conserved promoter TFBSs. The presence of these 'module-TFs' in the CSEs is likely to identify those CSEs that have a higher statistical significance of being functionally relevant. In the absence of expression data for the devel-

oping SAS (or UG), the search for module-TFs can be accomplished on a set of genes that are co-expressed in the developing SAS/UG at day e14.5 (http://www.informatics.jax.org/). Additionally, we have developed methods for the identification of transcriptional effectors from microarray expression data using network inference techniques ([117], Chapters 2-3). We have explored use of a novel information metric, called directed information, or DTI for inferring gene influence. This metric enables the discovery of putative transcriptional regulatory networks that examine relationships between module TFs and their target genes. As such time series expression data becomes gradually available, our integrative methodology can be modified to incorporate such new data sources.

TF binding at the CSE/enhancer: We have determined the conserved TFBS in each phylogenetically shadowed region. We have examined each of these TFs for their tissue-affiliated expression characteristics by querying the publicly available gene expression atlas (http://www.symatlas.gnf.org) or UNIPROT annotations (http://expasy.org/sprot/). We can then quantitatively determine whether or not the expression of a TF-encoding gene is higher in comparison to its 'basal' (median) level of expression in other tissues. From this database, we can recover a profile for the expression of each transcription factor over all catalogued tissue types. This is illustrated for GATA3 in Figure. 1.6 as an example. As depicted in the figure, murine microarray chipset U74Av2 expression confirms elevated expression of *Gata3* in the adrenal gland and kidney. In a similar manner we can examine the expression profile of every candidate TF that emerges from the analysis. We propose to investigate each CSE that has a statistical over-representation of such tissue-specific TFs as a potential SAE/UGE (some examples of such tissue-restricted TFs are dHAND, PHOX2b, MASH, in the developing SA system, [19]). Additionally, we have adapted

the WebMotifs [29] system to this application, thereby enabling the search for TF family motifs rather than individual members within each candidate CSE. Based on the families identified, we explore the expression of each of the TF family-member genes in the tissues of interest. The enrichment of tissue-specific TFs in any CSE of interest is potential evidence for its regulatory role ([209], [216]). This has been further explained in Chapter 6.

The strategy outlined above requires that a library of tissue-specific TFs be known and publicly available. However, given that the tissue-specificity of many TFs is still under active investigation, as well as the fact that the present version of TRANSFAC is still being annotated, it would be of interest to ask whether there are motifs that are over-represented in adrenal or other SAS-related tissues (similarly for the UGE analysis). The Gene Expression Atlas of the Novartis foundation as well as Mouse Genome Informatics (MGI) databases describe the expression characteristics of an exhaustive compendium of tissue-specific genes. The promoters of these genes can be mined for over-represented sequence motifs. These motif sets are tissue-type dependent. Treating this problem as one that is conceptually similar to searching through a text document by co-clustered interrogation (with random hexanucleotide motifs as words and genetic loci as documents), we continue to explore an approach to the identification of tissue-specific motifs using random forest classifiers. We propose to extend this technique by building sequence models with these motifs; this may allow us to search genome-wide for novel enhancers responsible for directing gene expression in the same tissue represented by the motif set, independent of their relative orientation and spacing.

Our previous analysis on known urogenital enhancers of the *Gata2*, and *Gata3* genes revealed a new feature hitherto unexplored in the context of cis-regulatory el-

ement identification [24]. Examination of the network of protein:protein interactions between module-TFs of a group of co-expressed genes and the phylogenetically conserved TFs at the enhancers reveals a strong "network-linkage" between these TFs (using the MIMI: Michigan Molecular Interactions Database). This is also observed for a study of *Mecp2* enhancers ([215], and unpublished data), and is consistent with the hypothesized long-range interactions between the proximal promoter and enhancer underlying distal gene regulation. On the other hand there is almost no interaction at the TF level between the promoter and a non-enhancer. This "network-linkage" feature is seen to be a potential discriminatory feature to the discovery of possibly functional non-coding elements and may reduce the false positive rate (due to lower conservation thresholds) significantly. Additionally, the examination of gene ontology databases for co-localized transcription factors belonging to the enhancer and promoter is additional evidence for possible linkage. We have examined a structural linkage metric that quantifies the degree of connectedness between the two groups of TFs using the 'Network-Analysis' plugin in Cytoscape (Chapter 5) [173].

For each of the features that will be derived using the approach described above, a 'training set' of known enhancers (UG2, UG4, UGE, etc.) were analyzed from multiple organisms. To apply this integrative methodology I have designed a classifier capable of discriminating between regulatory and non-regulatory elements based on each of the various features listed above. The non-regulatory elements, also available from laboratory experimental data [204], are a set of conserved elements that do not have expression in transgenic mice. Each training data object is a known non-coding element that is described by a set of features as described previously (sequence conservation, RF motif scores, network linkage etc.). Based on these features, we have trained a set of classifiers and validate the learned classifier on the test data

set [199]. The test data set comprises of a group of known enhancers that are not present in the training set. Each learned classifier will seek to discriminate between regulatory and non-regulatory DNA based on that feature. Finally the results of the individual classifiers are combined to obtain a "combined belief", across multiple modalities, that indicates if the input sequence is potentially regulatory or not [24].

This study has used some novel methods for cis-regulatory element identification based on histone modification motifs, and network linkage features along with other known strategies and is the first instance of heterogeneous data integration from experimental and computational measures, to understand the architecture of transcriptional regulatory elements. Needless to say, the various aspects of this integrative approach are fairly general and can be easily extended to any gene of interest. This integrative methodology would be an important component of the classifier design step outlined below.

The overall goal of this work is to develop the methodology to create a ranked list of candidate CSEs based on multiple genomic features, without the introduction of any experimental component. Each of these 'feature sources' provides a clue for refining a set of CSEs based on biological relevance to a higher confidence set, which can then be tested in vitro or in vivo. Our laboratory employed a transgenic reporter assay to test the functional relevance of high confidence candidate CSEs, described in greater detail below.

We have used the above methods to reconcile the behavior of known urogenital enhancers of the *Gata2* (Chapter 5) gene [24]. In Chapter 5, we have shown that understanding TF binding at the promoter and enhancer as well as their interactions is extremely meaningful as part of a prospective strategy to localizing such distal regulatory regions.

The last chapter (Chapter 6) describes the discovery of the OVE in the 150kb kidney region and the PE in the 45 kb SA region. We note that these discoveries were made without access to H3K4, network linkage data and hence the predictive capability of these methods is expected to boost the performance of the overall classifier. We have also described several additions to the said methods in Chapter 6, to improve on our model. A frank assessment of our method is that we can distinguish potential enhancers from neutral elements but we are still some way away from picking enhancers that simultaneously account for spatio-temporally specific and promoter-dependent expression.

1.5 Experimental Validation

The experimental strategy is designed to test the regulatory role of the candidate UG/SAS enhancers identified in our research aim. The question we address is: which CSE(s) from the chr2:9229957-9275238 (mm8) interval contains the SAS enhancer? This is tested by generating transgenic mice bearing individual high confidence CSEs, and determining whether these sequences can direct reporter gene expression in a pattern that mimics the expression of *Gata3* in sympathetic ganglia and adrenal chromaffin cells.

Each of the CSEs that is identified as one with a high probability of being the sympathoadrenal enhancer (SAE) will be cloned into a vector that has been modified to include the minimal *Gata3* promoter directing reporter (such as lacZ) expression. The candidate sequence is PCR amplified from the RP23-260A19 BAC, using a suitable choice of restriction sites at the primer ends to ensure compatibility for cloning into the plasmid. Primers are designed using the Primer3 tool available from http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi, and its loca-

tion will be confirmed using the in silico PCR tool at http://genome.cse.ucsc.edu/cgi-bin/hgPcr/. The ligated construct is transformed into competent cells, and transformants will be screened for the presence of the appropriate fragment by restriction enzyme digestion. Purified plasmid DNA will be diluted in and microinjected into early mouse embryos at concentrations of 0.5 to 2 ng/μl using standard techniques. Microinjected embryos are transferred into female pseudo-pregnant mice. Embryos will be isolated at 12.5 gestational days and stained with X-Gal (5-bromo-4-chloro-3-indolyl-β -D-galactopyranoside).

Individual candidate SAEs may confer only a subset of SAS expression (e.g. a single CSE may direct only sympathetic ganglion expression, but not expression in adrenal chromaffin cells). Since the 46 kbp interval (Fig. 1.1) contains both SG and adrenal expression, those elements could be separate, and thus we will continue to assay other, lower probability CSEs for (e.g.) chromaffin cell activity if multiple SAEs are implicated. A similar analysis may be applied to identification of the UGE in chr2: chr2:9069182-9229960 (mm8). Depending on the context, researchers in the Engel laboratory are using validation in cell lines or in transgenic mice as a way to assay for functional role of the regulatory element.

1.6 Contributions of this Thesis:

Most of the results in Chapter 2-6 are for the kidney expression of *Gata2* and *Gata3*, since a lot more literature and data is available for kidney-expression and nephrogenesis studies. Based on this, the chapters are organized as follows:

- Chapter 2: Mixture of Gaussian (MOG) clustering of gene expression data and network identification for the inference of transcriptional effectors.
- Chapter 3: Directed information (DTI)-based TF network identification from

microarray expression data.

- Chapter 4: Generalizing on the use of mutual information for feature selection, we have used DTI for sequence motif discovery to discriminate kidney-specific promoters from housekeeping promoters.

- Chapter 5: We use random forests (RF) classifiers for H3K4 and kidney-specific sequence motif discovery, and combined this with an integrative strategy to reconcile behavior of some known *Gata2* and *Gata3* enhancers, using interaction graphs of promoter-enhancer crosstalk.

- Chapter 6: This is a synopsis of ideas that that have been tried and are possibly useful for augmenting our current *"enhancer prediction"* model. As an example, we have presented a generalization to searches for TF families rather than individual members, as well as module TF discovery after accounting for biological process similarity among genes, thereby improving on the exploratory capability of the approach.

CHAPTER II

Network Inference Using State Space Models

Most current methods for gene regulatory network identification lead to the inference of steady state networks, i.e., networks prevalent over all time, a hypothesis which has been challenged. There has been a need to infer and represent networks in a dynamic (i.e., time-varying) fashion, in order to account for different cellular states affecting the interactions amongst genes. In this work, we present an approach, *regime-SSM*, to understand gene regulatory networks within such a dynamic setting. The approach uses a clustering method based on these underlying dynamics, followed by system identification using a state space model for each learnt cluster, to infer a network adjacency matrix. We finally indicate our results on a mouse embryonic kidney expression dataset as well as a T-cell activation based expression dataset and demonstrate conformity with reported experimental evidence.

2.1 Introduction

Most methods of graph inference work very well on stationary time series data, in that the generating structure for the time series does not exhibit switching. In ([32],[59]), some useful methods to learn network topologies using linear *state space models (SSM)*, from T-cell gene expression data have been presented. However it is known that regulatory pathways do not persist over all time. An important recent

finding in which the above is seen to be true is following examination of regulatory networks during the yeast cell cycle [36], wherein topologies change depending on underlying (endogeneous or exogeneous) cell conditions. This brings out the need to identify the variation of the 'hidden states' that regulate gene network topologies and to incorporate them into a network inference framework [37]. This hidden state at time t (denoted by x_t) might be related to the level of some key metabolite(s) governing the activity (g_t) of the gene(s). These present a notion of condition specificity which influence the dynamics of various genes that are active during that regime (condition). From time series microarray data, we aim to partition each gene's expression profile into such regimes of expression, during which the underlying dynamics of the gene's controlling state (x_t) can be assumed to be stationary. In [53], the powerful notion of context sensitive boolean networks for gene relationships has been presented. However, at least for short time series data, such a boolean characterization of gene state requires a one bit quantization of the continuous state, which is difficult without expert biological knowledge of the activation threshold and knowledge of the precise evolution of gene expression. Here, we work with gene profiles as continuous variables conditioned on the regime of expression. Each regime is related to the state of a state-space model that is estimated from the data.

Our method (*regime-SSM*) examines three components: To find the switch in gene dynamics, we use a Change point detection (CPD) approach using Singular Spectrum Analysis (SSA). Following the hypothesis that the mechanism causing the genes to switch at the same time came from a common underlying input ([36], [47]) we group genes having similar change points. This clustering borrows from a Mixture of Gaussian (MoG) model [33]. The inference of the network adjacency matrix follows from a state space representation of expression dynamics among these

co-clustered genes [32],[59]). Finally, we present analyses on the publicly available embryonic kidney gene expression dataset [35] and the T-cell activation dataset [32], using a combination of the above developed methods and validate our findings with previously published literature as well as experimental data.

For the embryonic kidney dataset, the biological problem motivating our network inference approach is one of identifying gene interactions during mammalian nephrogenesis (kidney formation). Nephrogenesis, like several other developmental processes, involves the precise temporal interaction of several growth factors, differentiation signals and transcription factors for the generation and maturation of progenitor cells. One such key set of transcription factors is the GATA family, comprising six members, all containing the (–GATA–) binding domain. Among these, *Gata2* and *Gata3* have been shown to play a functional role ([35], [39]) in nephric development between days 10-12 after fertilization. From a set of differentially expressed genes pertinent to this time window (identified from microarray data), our goal is to prospectively discover regulatory interactions between them and the *Gata2/3* genes. These interactions can then be further resolved into transcriptional, or signaling interactions on the basis of additional biological information.

In the T-cell activation dataset, the question is if events downstream of T-cell activation can be partitioned into early and late response behaviors and if so, which genes are active in a particular phase. Finally, can a network level influence be inferred among the genes of each phase and do they correlate with known data? We note here that we are not looking for the behavior of any particular gene, but only interested in genes from each phase.

As will be shown in this chapter, *regime-SSM* generates biologically relevant hypotheses regarding time varying gene interactions during nephric development and

T-cell activation. Several interesting transcripts are seen to be involved in the process and the influence network hereby generated resolves cyclic dependencies.

The main assumption for the formulation of a linear state space model to examine the possibility of gene-gene interactions is that gene expression is a function of the underlying cell state and the expression of other genes at the previous time step. If longer range dependencies are to be considered, the complexity of the model would increase. Another criticism of the model might be that non-linear interactions cannot be adequately modeled by such a framework. However, around the equilibrium point (steady state), we can recover a locally linearized version of this non-linear behavior.

2.2 SSA and Change Point Detection

First we introduce some notation. Consider N gene expression profiles, $g^{(1)}, g^{(2)}, \ldots, g^{(N)} \in \mathbb{R}_{+}^{\mathbb{T}}$; T being the length of each gene's temporal expression profile (as obtained from microarray expression). The j^{th} time instant of gene i's expression profile will be denoted by $g_j^{(i)}$.

State space partitioning is done using *Singular Spectrum analysis* [34] (SSA). SSA identifies "structural change points in time" series data using a sequential procedure [40]. We will briefly review this method.

Consider the 'windowed'(width N_W) time series data given by $\{g_1^{(i)}, g_2^{(i)}, \ldots, g_{N_W}^{(i)}\}$, with $M(M \leq \frac{N_W}{2})$ as some integer valued lag parameter, and a replication parameter $K = N_W - M + 1$. The SSA procedure in CPD involves the following:

- Construction of an l dimensional subspace: Here, a 'trajectory matrix' for the time series, over the interval $[n+1, n+T]$ is constructed:

$$\mathbf{G}_B^{i,(n)} = \begin{pmatrix} g_{n+1}^{(i)} & g_{n+2}^{(i)} & g_{n+3}^{(i)} & \cdots & g_{n+K}^{(i)} \\ g_{n+2}^{(i)} & g_{n+3}^{(i)} & g_{n+4}^{(i)} & \cdots & g_{n+K+1}^{(i)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ g_{n+M}^{(i)} & g_{n+M+1}^{(i)} & g_{n+M+2}^{(i)} & \cdots & g_{n+N_W}^{(i)} \end{pmatrix} \tag{2.1}$$

where $K = N_W - M + 1$. The columns of the matrix $\mathbf{G}_B^{i,(n)}$ are the vectors $G_j^{i,(n)} = (g_{n+j}^{(i)}, \ldots, g_{n+j+M-1}^{(i)})^T$, with $j = 1, \ldots, K$.

- Singular Vector Decomposition of the lag covariance matrix $\mathbf{R}^{i,n} = \mathbf{G}_B^{i,(n)}(\mathbf{G}_B^{i,(n)})^T$ yields a collection of singular vectors - a grouping of l of these Singular vectors, corresponding to the l highest eigenvalues - denoted by $I = \{1, \ldots, l\}$; establishes a subspace $\mathscr{L}_{n,I}$ of $\mathbb{R}^M$.

- Construction of the **test matrix**: Using $\mathbf{G}_{test}^{i,(n)}$ defined by

$$\mathbf{G}_{test}^{i,(n)} = \begin{pmatrix} g_{n+p+1}^{(i)} & g_{n+p+2}^{(i)} & \cdots & g_{n+q}^{(i)} \\ g_{n+p+2}^{(i)} & g_{n+p+3}^{(i)} & \cdots & g_{n+q+1}^{(i)} \\ \vdots & \vdots & \ddots & \vdots \\ g_{n+p+M}^{(i)} & g_{n+p+M+1}^{(i)} & \cdots & g_{n+q+M-1}^{(i)} \end{pmatrix}$$

  Here, we use the length (p) and location (q) of test sample. We choose $p \geq K$, with $K = N_W - M + 1$. Also $q > p$, here we take $q = p + 1$. From this construction, the matrix columns are the vectors $G_j^{i,(n)}, j = p+1, \ldots, q$. The matrix has dimension $M \times Q, \quad Q = (q - p) = 1$.

- Computation of the detection statistic:

  The detection statistics used in the CPD are:

- The normed Euclidean distance between the column span of the *test matrix* i.e. $G_j^{i,(n)}$ and the l-dimensional subspace $\mathscr{L}_{n,I}$ of $\mathbb{R}^M$. This is denoted by $\mathscr{D}_{n,I,p,q}$.
- The normalized sum of squares of distances, denoted by $S_n = \frac{\mathscr{D}_{n,I,p,q}}{MQ\mu_{n,I}}$, with $\mu_{n,I} = \mathscr{D}_{m,I,0,K}$, where m is the largest value of $m \leq n$ so that the hypothesis of no change is accepted.
- A cumulative sum (CUSUM) type statistic $W_1 = S_1$, $W_{n+1} = max\{(W_n + S_{n+1} - S_n - \frac{1}{3MQ}), 0\}, n \geq 1$.

The CPD procedure declares a structural change in the time series dynamics if for some time instant n, we observe $W_n > h$ with the threshold: $h = (2t_\alpha/(MQ))\sqrt{\frac{1}{3}q(3MQ - Q^2 + 1)}$, t_α being the $(1-\alpha)$ quantile of the standard normal distribution.

- Choice of algorithm parameters:
  - Window Width (N_W): Here, we choose $N_W \simeq T/5$, T being the length of the original time series, because for this choice the algorithm provides a reliable method of extracting most structural changes. Choosing a much smaller N_W might lead to some outliers being classified as potential change points, but in our set-up this is preferred in contrast to losing genuine structural changes based on choosing larger N_W.
  - Choice of lag M: In most cases, we choose $M = \frac{N_W}{2}$.

2.3 Mixture of Gaussians (MoG) Clustering

Having found change points (and thus, regimes) from the gene trajectories of the differentially expressed genes, our goal is to now group (cluster) genes with similar

temporal profiles within each regime. In this section, we derive the parameter update equations for a *Mixture of Gaussian* clustering paradigm. As will be seen later, the Gaussian assumptions on the gene expression permit the use of co-clustered genes for the SSM based network parameter estimation.

We now consider the group of gene expression profiles $\mathscr{G} = \{\boldsymbol{g^{(1)}}, \boldsymbol{g^{(2)}}, \ldots, \boldsymbol{g^{(n)}}\}$, all of which share a common change point (time of switch) - c_1. Consider gene profile i, $\boldsymbol{g^{(i)}} = [g_1^{(i)}, g_2^{(i)}, \ldots, g_{T_{c_1}}^{(i)}]^T$, a T_{c_1}-dimensional random vector which follows a k-component finite mixture distribution described by:

$$p(\mathbf{g}|\boldsymbol{\theta}) = \sum_{m=1}^{k} \alpha_m p(\mathbf{g}|\boldsymbol{\phi_m}) \tag{2.2}$$

where $\alpha_1, \ldots, \alpha_k$ are the mixing probabilities, each ϕ_m is the set of parameters defining the m^{th} component, and $\boldsymbol{\theta} \equiv \{\phi_1, \ldots, \phi_k, \alpha_1, \ldots, \alpha_k\}$ is the set of complete parameters needed to specify the mixture. We have,

$$\alpha_m \geq 0, m = 1, \ldots, k, \quad and \quad \sum_{m=1}^{k} \alpha_m = 1 \tag{2.3}$$

For a set of n independently and identically distributed samples,

$$\mathscr{G} = \{\boldsymbol{g^{(1)}}, \boldsymbol{g^{(2)}}, \ldots, \boldsymbol{g^{(n)}}\}, \tag{2.4}$$

the log-likelihood of a k-component mixture is given by:

$$\log p(\mathscr{G}|\boldsymbol{\theta}) = \log \prod_{i=1}^{n} p(\boldsymbol{g^{(i)}}|\boldsymbol{\theta}) \tag{2.5}$$

$$= \sum_{i=1}^{n} log \sum_{m=1}^{k} \alpha_m p(\boldsymbol{g^{(i)}}|\phi_m) \tag{2.6}$$

- Treat the labels, $\mathscr{Z} = \{\boldsymbol{z^{(1)}}, \ldots, \boldsymbol{z^{(n)}}\}$, associated with the n samples - as missing data. Each label is a binary vector $\boldsymbol{z^{(i)}} = [z_1^{(i)}, \ldots, z_k^{(i)}]$, where $z_m^{(i)} = 1$

and $z_p^{(i)} = 0$, for $p \neq m$ indicates that sample $\boldsymbol{g^{(i)}}$ was produced by the m^{th} component.

In this setting, the **Expectation Maximization** algorithm can be used to derive the cluster parameter ($\boldsymbol{\theta}$) update equations.

In the *E step* of the *EM algorithm*, the function $Q(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}(t)) \equiv E[\log p(\mathscr{G}, \mathscr{Z}|\boldsymbol{\theta})|\mathscr{G}, \hat{\boldsymbol{\theta}}(t)]$,is computed. This yields,

$$w_m^{(i)} \equiv E[z_m^{(i)}|\mathscr{G}, \hat{\theta}_t] = \frac{\hat{\alpha}_m(t) p(\boldsymbol{g^{(i)}}|\hat{\boldsymbol{\theta}}_{\boldsymbol{m}}(t))}{\sum_{j=1}^{k} \hat{\alpha}_j(t) p(\boldsymbol{g^{(i)}}|\hat{\boldsymbol{\theta}}_{\boldsymbol{j}}(t))} \tag{2.7}$$

where $w_m^{(i)}$ is the posterior probability of the event $z_m^{(i)} = 1$, on observing $g_m^{(i)}$.

The estimate of the number of components (k) is chosen using a Minimum Message Length (MML) criterion [33]. The MML criterion borrows from algorithmic information theory and serves to select models of lowest complexity to explain the data. As can be seen below, this complexity has two components - the first encodes the observed data as a function of the model and the second encodes the model itself. Hence, the MML criterion in our setup becomes, :

$$\hat{k}_{MML} = \text{argmin}_k\{-\log p(\mathscr{G}|\hat{\boldsymbol{\theta}}(k)) + \frac{k(N_p + 1)}{2} \log n\}, \tag{2.8}$$

N_p is number of parameters per component in the k component mixture, given the number of clusters $k_{min} \leq k \leq k_{max}$.

In the *M step*: For $m = 0, 1, \ldots, k$, $\hat{\theta}_m(t+1) = \arg\max_{\phi_m} Q(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}(t))$, for m : $\hat{\alpha}_m(t+1) > 0$, the elements $\hat{\phi}$'s of the parameter vector estimate $\hat{\boldsymbol{\theta}}$ are typically not closed form and depend on the specific parametrization of the densities in in the mixture, i.e. $p(\boldsymbol{g^{(i)}}|\boldsymbol{\phi_m})$. If $p(\mathbf{g}^{(i)}|\phi_m)$ belongs to the Gaussian density $\mathcal{N}(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$ class, we have, $\phi = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and EM updates yield [2],

$$\hat{\alpha}_m(t+1) = \frac{\sum_{i=1}^{n} w_m^{(i)}}{n}, \tag{2.9}$$

$$\boldsymbol{\mu}_m(t+1) = \frac{\sum_{i=1}^{n} w_m^{(i)} \mathbf{g}^{(i)}}{\sum_{i=1}^{n} w_m^{(i)}}, \tag{2.10}$$

$$\boldsymbol{\Sigma}_m(t+1) = \frac{\sum_{i=1}^{n} w_m^{(i)} (\mathbf{g}^{(i)} - \mu_m(t+1))(\mathbf{g}^{(i)} - \mu_m(t+1))^T}{\sum_{i=1}^{n} w_m^{(i)}} \tag{2.11}$$

The equations 2.3.9-11 are the parameter update equations for each of the $m = 1, \ldots k$ cluster components.

For the kidney expression data, since we are interested in the role of *Gata2* and *Gata3* during early kidney development, we consider all the genes which have similar change points as the *Gata2* and *Gata3* genes respectively. We perform a MoG clustering within such genes and look at those co-clustered with *Gata2* or *Gata3*. Co-clustering within a regime potentially suggests that the governing dynamics are the same, even to the extent of co-regulation. Just because a gene is co-clustered with *Gata2* in one regime, it does not mean that it will co-cluster in a different regime. This approach suggests a way to localize regimes of correlation instead of the traditional global correlation measure that can mask transient and condition-specific dynamics. For this gene expression data, the MML penalized criterion indicates that an adequate number of clusters to describe this data is two (k=2). In Tables. 2.1 and 2.2, we indicate some of the genes with similar co-expression dynamics as *Gata2/Gata3* and a cluster assignment of such genes. We observe that this clustering corresponds to the first phase of embryonic development (day 10-12 dpc), the phase

where *Gata2* and *Gata3* are perhaps most relevant to kidney development ([41], [45],[46], [50]).

In Table. 2.1, the entries in each column of a row (gene) indicate the change points (as found by the SSA-CPD procedure) in the time series of the interpolated gene expression profile. Our simulation studies with the T-cell data indicate that the SSM and CoD performance is not much worse with the interpolated data compared to the original time series (Table. 2.7). Because of the present choice of parameters N_W, we might detect some false positive change points, but this is preferable to the loss of genuine change points. An examination of the change points of the various genes in Table. 2.1 indicates three regimes - between points approximately 1-5, 5-11 and 12-20. The missing entries mean that there was no change point identified for a certain regime and are thus treated as such. Since our focus is early *Gata3* behavior, we are interested in time points 1-12, and hence examine the evolution of network-level interactions over the first two regimes for the genes co-clustered in these regimes.

To clarify the validity of the presented approach, we present a similar analysis on another data set – the T-cell expression data presented in [32]. This data looks at the expression of various genes after T-cell activation using stimulation with phorbolester PMA and ionomycin [72]. This data has the profiles of about 58 genes over 10 time points with 44 replicate measurements for each time point. Since here we have no specific gene in mind (unlike earlier where we were particularly interested in *Gata3* behavior), the change point procedure (CPD) yields two distinct regimes – one from time point 1 to 4 and the other from time point 5 to 10. Following the MoG clustering procedure yields the optimal number of clusters to be 1 (from MML) in each regime. We therefore call these two clusters 'early response' and 'late response' genes and

Gene Symbol	Change point I	Change point II	Change point III
Bmp7	6	10	12
Rara	5	11	16
Pax2	6	12	15
Gata3	5	9	12
Gata2			18
Gdf11		10	20
Npnt		12	16
Cd44	5	11	15
Pgf	5	11	
Pbx1	5	12	20
Ret		10	

Table 2.1: Change Point Analysis of some key genes, prior to clustering (annotations in Table. 2.8). The numbers indicate the time points at which regime changes occur for each gene.

Genes with the same dynamics as *Gata3*	Genes with the same dynamics as *Gata2*
Bmp7	Lamc2
Nrtn	Cldn3
Pax2	Ros1
Ros1	Ptprd
Pbx1	Npnt
Rara	Cdh16
Gdf11	Cldn4

Table 2.2: Some of the genes co-clustered with *Gata2* and *Gata3* after MoG Clustering (annotations in Table. 2.8)

then proceed to learn a network relationship amongst them, within each cluster. The CPD and cluster information for the early and late response is summarized in Table. 2.1.

2.4 State Space Model

For a given regime, we treat gene expression as an observation related to an underlying hidden cell state (x_t), which is assumed to govern regime-specific gene expression dynamics for that biological process, globally within the cell. Suppose there are N genes whose expression is related to a single process. The i^{th} gene's expression vector is denoted as $g_t^{(i)}, t = 1, \ldots T$, where T is the number of time points for which the data is available. The *State Space Model (SSM)* is used to model the gene expression ($g_t^{(i)}, i = 1, 2, ..N$ and $t = 1, 2, ..T$) as a function of this

Genes related to early response (time points: 1-4)	Genes related to late response (time points: 5-10)
CD69	CCNA2
Mcp1	CDC2
Mcl1	EGR1
EGR1	IL2r gamma
JunD	IL6
CKR1	

Table 2.3: Some of the genes related to early and late response in T-cell activation (annotations in Table. 2.9)

underlying cell state (x_t) as well as some external inputs. A notion of influence among genes can be integrated into this model by considering the SSM inputs to be the gene expression values at the previous time step. The state and observation equations of the *State Space Model* [48] are:

- State equation:

$$\mathbf{x_{t+1}} = A\mathbf{x_t} + B\mathbf{g_t} + \mathbf{e_{s,t}}; \quad \mathbf{e_{s,t}} \sim \mathcal{N}(0, Q); \quad i = 1, \ldots, N; \quad t = 1, \ldots T \tag{2.12}$$

- Observation equation:

$$\mathbf{g_t} = C\mathbf{x_t} + D\mathbf{g_{t-1}} + \mathbf{e_{o,t}}; \quad \mathbf{e_{o,t}} \sim \mathcal{N}(0, R) \tag{2.13}$$

With $\mathbf{x_t} = [x_t^{(1)}, x_t^{(2)}, \ldots, x_t^{(K)}]^T$ and $\mathbf{g_t} = [g_t^{(1)}, g_t^{(2)}, \ldots, g_t^{(N)}]^T$. A likelihood method [32] is used to estimate the state dimension K . The noise vectors $\mathbf{e_{s,t}}$ and $\mathbf{e_{o,t}}$ are Gaussian distributed with means 0 and covariance matrices Q and R respectively.

From the state and observation equations (2.12) and (2.13), we notice that the matrix-valued parameter $D = [D_{i,j}]_{i=1,\ldots N}^{j=1,\ldots N}$ quantifies the influence among genes i and j from one time instant to the next, within a specific regime. To infer a biological

network using D, we use bootstrapping to estimate the distribution of the strength of association estimates amongst genes and infer network linkage for those associations that are observed to be significant.

Within this proposed framework, we segment the overall gene expression time trajectories into smaller, approximately stationary, gene expression regimes. We note that the MoG clustering framework is a non-linear one in that the regime-specific state space is partitioned into clusters. These cluster assignments of correlated gene expression vectors can change with regime, allowing us to capture the sets of genes that interact under changing cell condition.

2.5 System Identification

We consider the case where we have $R_g = B \times P$ realizations of expression data for each gene available. Arguably, mRNA level is a measure of gene expression, $B(= 2)$ denotes the number of biological replicates and $P(\ = 16$ perfect match probes) denotes the number of probes per gene transcript. Each of these R_g realizations are T time points long and are obtained from Affymetrix U74Av2 murine microarray raw .CEL files. In the section below, we derive the update equations for maximum likelihood estimates of the parameters A, B, C, D, Q and R (in equations (2.12) and (2.13)) using an EM algorithm, based on ([43], [48]). The assumptions underlying this model are outlined in Table. 2.4. A sequence of T output vectors $(\mathbf{g_1}, \mathbf{g_2}, \ldots, \mathbf{g_T})$ is denoted by $\{\mathbf{g}\}$, and a subsequence $\{\mathbf{g_{t_0}}, \mathbf{g_{t_0+1}}, \ldots, \mathbf{g_{t_1}}\}$ by $\{\mathbf{g}\}_{t_0}^{t_1}$. We treat the $(\mathbf{x_t}, \mathbf{g_t})$ vector as the complete data and find the log-likelihood $\log P(\{\mathbf{x}\}, \{\mathbf{g}\})$ under the above assumptions. The complete E and M steps involved in the parameter update steps are outlined in Tables. 2.5 and 2.6.

Symbol	Interpretation	Expression
T	Number of Time points	
R_g	Number of Replicates	
$P(\mathbf{g_t}\|\mathbf{x_t})$	$\equiv$	$\prod_{t=2}^{T}\{e^{-\frac{1}{2}[\mathbf{g_t}-C\mathbf{x_t}-D\mathbf{g_{t-1}}]'R^{-1}[\mathbf{g_t}-C\mathbf{x_t}-D\mathbf{g_{t-1}}]}\}\cdot(2\pi)^{-p/2}\det(R)^{-1/2}$
$P(\mathbf{x_t}\|\mathbf{x_{t-1}})$		$\prod_{t=2}^{T}\{e^{-\frac{1}{2}[\mathbf{x_t}-A\mathbf{x_{t-1}}-B\mathbf{g_{t-1}}]'Q^{-1}[\mathbf{x_t}-A\mathbf{x_{t-1}}-B\mathbf{g_{t-1}}]}\}\cdot(2\pi)^{-k/2}\det(Q)^{-1/2}$
$P(\mathbf{x_1})$	Initial state density assumption	$e^{-\frac{1}{2}[\mathbf{x_1}-\pi_\mathbf{1}]'V_1[\mathbf{x_1}-\pi_\mathbf{1}]}\cdot(2\pi)^{-k/2}\det(V_1)^{-1/2}$
$P(\{\mathbf{x}\},\{\mathbf{g}\})$	Markov property	$\prod_{i=1}^{R_g}(P(\mathbf{x_1}^{(i)})\prod_{t=2}^{T}P(\mathbf{x_t}^{(i)}\|\mathbf{x_{t-1}}^{(i)},\mathbf{g_{t-1}}^{(i)}) \cdot\prod_{t=1}^{T}P(\mathbf{g_t}^{(i)}\|\mathbf{x_t}^{(i)},\mathbf{g_{t-1}}^{(i)}))$
$\log P(\{\mathbf{x}\},\{\mathbf{g}\})$	joint log	$-\sum_{i=1}^{R_g}\{\sum_{t=2}^{T}(\frac{1}{2}[\mathbf{g_t}^{(i)}-C\mathbf{x_t}^{(i)}-D\mathbf{g_{t-1}}^{(i)}]'R^{-1}[\mathbf{g_t}^{(i)}-C\mathbf{x_t}^{(i)}-D\mathbf{g_{t-1}}^{(i)}])$
	probability	$-\quad(\frac{T}{2})\log(\det(R))-\quad\sum_{t=1}^{T}(\frac{1}{2}[\mathbf{x_t}^{(i)}-A\mathbf{x_{t-1}}^{(i)}-B\mathbf{g_{t-1}}^{(i)}]'Q^{-1}$ $\cdot[\mathbf{x_t}^{(i)}-A\mathbf{x_{t-1}}^{(i)}-B\mathbf{g_{t-1}}^{(i)}])$ $-\frac{T-1}{2}\log(\det(Q))-\frac{1}{2}[\mathbf{x_1}-\pi_\mathbf{1}]V_1^{-1}[\mathbf{x_1}-\pi_\mathbf{1}]-\frac{1}{2}\log(\det(V_1))$ - $\frac{T(p+k)}{2}\log(2\pi)\}$

Table 2.4: Assumptions and Log-likelihood calculations in the State Space Model.The ($\equiv$) symbol indicates a definition.

2.6 Bootstrapped Confidence Intervals

As suggested above, the entries of the D matrix indicate the strength of influence among the genes, from one time step to the next (within each regime). We use Bootstrapping to find confidence intervals for each entry in the D matrix and if it is significant, we assign a positive or negative direction (+1 or -1) to this influence.

The bootstrapping procedure [44] is adapted to our situation as follows:

- Suppose there are R regimes in the data with change points $(c_1, c_2, \ldots, c_R)$ identified from SSA. For the r^{th} regime, generate B independent bootstrap samples of size N(the original number of genes under consideration), - $(\mathbf{Y}_1^*, \mathbf{Y}_2^*, \ldots, \mathbf{Y}_B^*)$ from original data, by random resampling from $\boldsymbol{g^{(i)}} = [g_{c_r}^{(i)}, \ldots, g_{c_{r+1}}^{(i)}]^T$.

Matrix Symbol	Interpretation	Expression
M Step :		
${\pi_1}^{new}$	Initial State Mean	$\hat{\mathbf{x_1}}$
V_1^{new}	Initial State Covariance	$P_1 - \hat{\mathbf{x_1}}\hat{\mathbf{x_1}}' + \frac{1}{R_g}\sum_{i=1}^{R_g}[\hat{\mathbf{x}_1'}^{(i)} - \overline{\hat{\mathbf{x_1}}}][\hat{\mathbf{x}_1'}^{(i)} - \overline{\hat{\mathbf{x_1}}}]'$
C^{new}	Output Matrix	$(\sum_{i=1}^{R_g}\sum_{t=1}^{T}\mathbf{g_t}^{(i)}\hat{\mathbf{x_t}}' - D\sum_{i=1}^{R_g}\sum_{t=1}^{T}\hat{\mathbf{x_t}}^{(i)}{\mathbf{g_{t-1}}}'^{(i)}) \quad \cdot(\sum_{i=1}^{R_g}\sum_{t=1}^{T}P_t^{(i)})^{-1}$
R^{new}	Output Noise Covariance	$\frac{1}{R_g \times T}[\sum_{i=1}^{R_g}\sum_{t=1}^{T}(\mathbf{g_t}^{(i)}{\mathbf{g_t}}'^{(i)}) - \quad C^{new}(\hat{\mathbf{x_t}}^{(i)}{\mathbf{g_t}}'^{(i)}) - D^{new}\mathbf{g_{t-1}}^{(i)}{\mathbf{g_t}}'^{(i)}]$
A^{new}	State Dynamics Matrix	$\sum_{i=1}^{R_g}\sum_{t=2}^{T}[P_{t,t-1}^{(i)} - B\hat{\mathbf{x_t}}^{(i)}{\mathbf{g_{t-1}}}'^{(i)}] \quad \cdot(\sum_{i=1}^{R_g}\sum_{t=2}^{T}P_{t-1}^{(i)})^{-1}$
D^{new}	Input to Observation	$\sum_{i=1}^{R_g}\sum_{t=1}^{T}[\mathbf{g_t}^{(i)}{\mathbf{g_{t-1}}}'^{(i)} - \mathbf{g_t}^{(i)}{\hat{\mathbf{x}_t}}'^{(i)}(\sum_{i=1}^{R_g}\sum_{t=1}^{T}P_t^{(i)})^{-1} \quad \hat{\mathbf{x}_t}^{(i)}{\mathbf{g_{t-1}}}'^{(i)})]\cdot$ $[\sum_{i=1}^{R_g}\sum_{t=1}^{T}(\mathbf{g_{t-1}}^{(i)}{\mathbf{g_{t-1}}}'^{(i)} - \mathbf{g_{t-1}}^{(i)}{\hat{\mathbf{x}_t}}'^{(i)} \cdot \quad (\sum_{i=1}^{R_g}\sum_{t=1}^{T}P_t^{(i)})^{-1}\hat{\mathbf{x}_t}^{(i)}{\mathbf{g_{t-1}}}'^{(i)})]^{-1}$
B^{new}	Input to State Matrix	$\sum_{i=1}^{R_g}\sum_{t=2}^{T}[P_{t,t-1}^{(i)}(\sum_{i=1}^{R_g}\sum_{t=2}^{T}P_t^{(i)})^{-1}\hat{\mathbf{x}_t}^{(i)}{\mathbf{g_{t-1}}}'^{(i)} - \quad \hat{\mathbf{x}_t}^{(i)}{\mathbf{g_{t-1}}}'^{(i)}]\cdot$ $[\sum_{i=1}^{R_g}\sum_{t=2}^{T}\mathbf{g_{t-1}}^{(i)}{\mathbf{x_t}}'^{(i)}(\sum_{i=1}^{R_g}\sum_{t=2}^{T}P_t^{(i)})^{-1} \quad \cdot\mathbf{x_t}^{(i)}{\mathbf{g_{t-1}}}'^{(i)} - \mathbf{g_{t-1}}{\mathbf{g_{t-1}}}'^{(i)}]^{-1}$
Q^{new}	State Noise Covariance	$\frac{1}{R_g \times (T-1)}(\sum_{i=1}^{R_g}\sum_{t=2}^{T}P_t^{(i)} - A^{new}\sum_{i=1}^{R_g}\sum_{t=2}^{T}P_{t-1,t}^{(i)} - B\sum_{i=1}^{R_g}\sum_{t=2}^{T}\mathbf{g_{t-1}}^{(i)}{\hat{\mathbf{x}_t}}'^{(i)})$

Table 2.5: M-step of the EM algorithm for State Space parameter estimation. The ($\equiv$) symbol indicates a definition.

- Using the EM algorithm for parameter estimation, estimate the value of D (the influence parameter). Denote the estimate of D for the i^{th} bootstrap sample by D_i^*.

- Compute the sample mean and sample variance of the estimates of D over all

E Step :		
Forward :		
$\mathbf{x_1}^0$	$\equiv$	$\pi_\mathbf{1}$
V_1^0	$\equiv$	V_1
$\mathbf{x_t}^{t-1}$	update	$A\mathbf{x_{t-1}}^{t-1} + B\mathbf{g_{t-1}}$
V_t^{t-1}	update	$AV_{t-1}^{t-1}A' + Q$
K_t	update	$V_t^{t-1}C'(CV_t^{t-1}C' + R)^{-1}$
$\mathbf{x_t}^t$	update	$\mathbf{x_t}^{t-1} + K_t(\mathbf{g_t} - C\mathbf{x_t}^{t-1} - D\mathbf{g_{t-1}})$
V_t^t	update	$V_t^{t-1} - K_tCV_t^{t-1}$
Backward :		
$V_{T,T-1}^T$	*Initialization*	$(I - K_TC)AV_{T-1}^{T-1}$
$\hat{\mathbf{x_t}}$	$\equiv$	$\mathbf{x_t}^\tau$
P_t	$\equiv$	$V_t^T + \mathbf{x_t}^T\mathbf{x_t}^{T'}$
J_{t-1}	update	$V_t^{t-1}A'(V_t^{t-1})^{-1}$
$\mathbf{x_{t-1}}^T$	update	$\mathbf{x_{t-1}}^{t-1} + J_{t-1}(\mathbf{x_1}^T - A\mathbf{x_{t-1}}^{t-1} - Bg_{t-2})$
V_t^T	update	$V_{t-1}^{t-1} + J_{t-1}(V_t^T - V_t^{t-1})J'_{t-1}$
$P_{t,t-1}$	$\equiv$	$V_{t,t-1}^T + \mathbf{x_t}^T\mathbf{x_{t-1}}^{T'}$
$V_{t-1,t-2}^T$	update	$V_{t-1}^{t-1}J'_{t-2} + J_{t-1}(V_{t,t-1}^T - AV_{t-1}^{t-1})J'_{t-2}$

Table 2.6: E-step of the EM algorithm for State Space parameter estimation.

the B bootstrap samples. That is:

$$Mean = \overline{D}^* = \frac{1}{B}\sum_{i=1}^{B}(D_i^*), \tag{2.14}$$

$$Variance = \frac{1}{B-1}\sum_{i=1}^{B}(D_i^* - \overline{D}^*)^2 \tag{2.15}$$

- Using the above obtained sample mean and variance, estimate confidence intervals for the elements of D. If D lies in this *bootstrapped* confidence interval we infer a potential influence and if not, we discard it. Note that even though we write D, we carry out this hypothesis test for each $D_{i,j}, i = 1, \ldots n; j = 1, \ldots, n;$ for each of the n genes under consideration in every regime.

2.7 Summary of Algorithm

Within each regime identified by CPD, we model gene expression as Gaussian distributed vectors. We cluster the genes using a *Mixture of Gaussians (MoG)* clustering

algorithm [33] to identify sets of genes which have similar 'dynamics of expression' - in that they are correlated within that regime. We then proceed to estimate the dynamic system parameters (matrices A, B, C, D, Q and R) for the State Space Model (SSM) underlying each of the clusters. We note two important ideas:

- That we might obtain different cluster assignments for the genes depending on the regime and
- That since all these genes (across clusters within a regime) are still related to the same biological process, the hidden state $\mathbf{x_t}$ is shared among these clusters.

Therefore, we estimate the SSM parameters in an alternating manner by updating the estimates from cluster to cluster while still retaining the form of the state vector $\mathbf{x_t}$. The estimation is done using an *Expectation - Maximization* type algorithm. The number of components during regime-specific clustering are estimated using a *Minimum Message Length* criterion. Typically, $O(N)$ iterations suffice to infer the mixture model in each regime with N genes under consideration. Thus, our proposed approach is as follows:

- Identify the N key genes based on required phenotypical characteristic using fold change studies. Preprocess the gene expression profiles by standardization and cubic spline interpolation.
- Segment each gene's expression profile into a sequence of state dependent trajectories (regime change points), from underlying dynamics, using SSA.
- For each regime (as identified in step 2):

  Cluster genes using a MoG model so that genes with correlated expression trajectories cluster together. Estimate the SSM parameters ([43], [48]) for

each cluster (from IV.1 and IV.2 for estimation of the mean and covariance matrices of the state vector), within that regime. The input to observation matrix (D) is indicative of the topology of the network in that regime.

- Examination of the network matrices D, (by bootstrapping to find thresholds on strength of influence estimates) across all regimes to build the time varying network.

The discussion of the network inference procedure would be incomplete in the absence of any other algorithms for comparison. For this purpose we implement the CoD (Coefficient of Determination) based approach ([62], [63] along with the models proposed in [32] (SSM) and [67] (GGM). The CoD method allows us to determine the association between two genes within a regime via a R^2 goodness of fit statistic. The methods of ([32], [67]) are implemented on the time series data (with regard to underlying regime). Such a study would be useful to determine the relative merits of each approach. We believe that no one procedure can work for every application and the choice of an appropriate procedure would be governed by the biological question under investigation. Each of these methods use some underlying assumptions and if these are consistent with the question that we ask, then that method has great utility. These individual results, their evaluation and their comparison is summarized in the Results section.

2.8 Results

2.8.1 Application to the GATA Pathway

To illustrate our approach (*regime-SSM*) we consider the embryonic kidney gene expression dataset [35] and study the set of genes known to have a possible role in early nephric development. An interruption of any gene in this signaling cascade

potentially leads to early embryonic lethality or abnormal organ development. An influence network among these genes would reveal which genes (and their products) become important at a certain phase of nephric development. The choice of the $N(= 47)$ genes is done using FDR fold change studies [56] between Ureteric Bud and Metanephric Mesenchyme tissue types, since this spatial tissue expression is of relevance during early embryonic development. The dataset is obtained by daily sampling of the mRNA expression ranging from 11.5-16.5 days post coitus (dpc). Detailed studies of the phenotypes characterizing each of these days is available from the Mouse Genome Informatics Database at http://www.informatics.jax.org/. We follow [57] and use interpolated expression data pre-processing for cluster analysis. We resample this interpolated profile to obtain twenty points per gene expression profile. Two key aspects were confirmed after interpolation [57], [73] - (1), that there were no negative expression values introduced and (2), that the differences in fold change were not smoothed out.

Initial experimental studies have suggested that the 10.5-12.5 dpc are relatively more important in determination of the course of metanephric development. We chose to explore which genes (out of the 47 considered) might be relevant in this specific time window. The SSA-CPD procedure identified several genes which exhibit similar dynamics (have approximately same change points, for any given regime) in the early phase and distinctly different dynamics in later phases (Table. 2.1).

Our approach to influence determination using the state space model yields upto three distinct regimes of expression over all the 47 genes identified from fold change studies between Bud and Mesenchyme [56]. MoG clustering followed by state space modeling yield three regime topologies of which we are interested in the early regime (day 10.5-12.5). This influence topology is shown in Fig. 2.1.

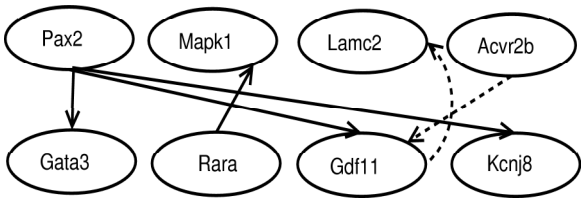


Figure 2.1: Network topology over regimes (solid lines represent the first regime, and the dotted lines indicate the second regime).

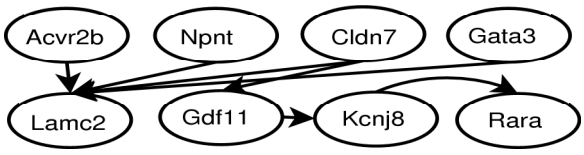


Figure 2.2: Steady state network inferred over all time, using [32].

We compare our obtained network (using *regime-SSM*) with one obtained using the approach outlined in [32], shown in Fig.2.2. We note that the network presented in Fig. 2.2 extends over all time i.e. days 10.5-16.5 for which basal influences are represented but transient and condition specific influences may be missed. Some of these transient influences are recaptured in our method (Fig. 2.1) and are in conformity (lower false positives in network connectivity) with pathway entries in *Entrez Gene* [50] as well as in recent reviews on kidney expression [35], [41] (also, Table. 2.8). For example, the *Mapk1-Rara* [71] or the *Pax2-Gdf11* [69] interactions are completely missed in Fig. 2.2 – this is seen to be the case since these interactions only occur during the 10.5-12.5 dpc regime. We also see that the *Acvr2b-Lamc2* [70] interaction is observed in the steady state but not in the first regime. This interaction becomes active in the second regime (first via the *Acvr2b-Gdf11* and then the *Gdf11-Lamc2*), indicating that it might not have particular relevance in the day 10.5-12.5 dpc stage. Several of these predicted interactions need to be experimentally characterized in the laboratory. It is especially interesting to see the *Rara* gene in this network, because it is known that *Gata3* [64], [65] has tissue-specific expression

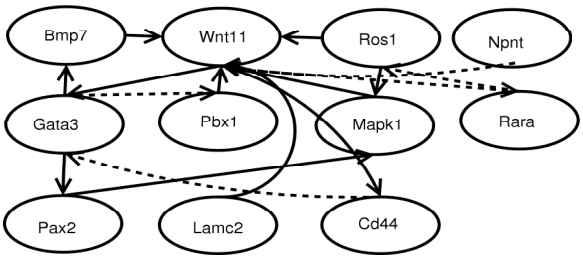


Figure 2.3: Steady state network inferred using CoD (solid lines represent the first regime, and the dotted lines indicate the second regime).

in some cells of the developing eye. Also *Gdf11* exhibits growth factor activity and is extremely important during organ formation.

In Fig. 2.3, we give the results of the CoD approach of network inference. Here the *Gata3-Pax2* interaction seems reversed and counterintuitive. As can be seen, some of the interactions (e.g: *Pax2-Gata3*) can be seen here (via other nodes:*Mapk1-Wnt11*), but there is a need to resolve cycles (*Ros1–Wnt11-Mapk1*) and feedback/feedforward loops (*Bmp7-Gata3-Wnt11*). Both of these topologies can convey potentially useful information about nephric development. Thus a potentially useful way to combine these two methods is to 'seed' the network using CoD and then try to resolve cycles using *regime-SSM*.

2.8.2 T-cell Activation

The *regime-SSM* network is shown in Fig. 2.4. The corresponding network learnt in each regime using CoD is also shown (Fig. 2.5). The study of this network using GGM (for the whole time series data) is already available in [67]. Though there are several interactions of interest discovered in both the SSM and CoD procedures, we point out a few of interest. It is already known that synergistic interactions between IL-6 and IL-1 are involved in T-cell activation [61]. IL-2 receptor transcription is affected by EGR1 [66]. The topology of these two networks (CoD and SSM) indicates some matches and are worth pursuing for experimental investigation. However, as al-

Method (T-cell data)	Edges inferred	f_{new}	f_{lost}
SSM on original data	14		
SSM on undersampled data		3	3
SSM on interpolated data		4	2
CoD on original data	12		
CoD on undersampled data		3	2
CoD on interpolated data		4	2

Table 2.7: Results of Network inference on original, subsampled and interpolated data

ready alluded to above, we have to find a way to resolve cycles from the CoD network [68]. Several of these match the interactions reported in ([32], [67]). However, the additional information that we can glean is that some of the key interactions occur during 'early response' to stimulation and some occur subsequently (Interleukin-6 mediated T-cell activation) in the 'late phase'.

An examination of the Gene Ontology (GO) terms represented in each cluster as well as the functional annotations in *Entrez Gene* shows concordance with literature findings (Table. 2.8). Because this dataset has been the subject of several interesting investigations, it would be ideal to ask other questions related to network inference procedures, for the purpose of comparison. One of the primary questions we seek to answer is: What is the performance of the network inference procedure if a subsampled trajectory is used instead?

In Table. 2.7, the performance of the CoD and SSM algorithms are summarized. Using the T-cell (10 points, 44 replicates) data, we infer a network using the SSM procedure. With the identified edges as the gold standard for comparison, we now use SSM network inference on a undersampled version of this time series (5 points, 44 replicates) and check for any new edges (f_{new}) or deletion of edges (f_{lost}). Ideally, we would want both these numbers to be zero. f_{new} is the fraction of new edges added to the original set and f_{lost}, is no of edges lost from the original data network over both regimes. Further, we now interpolate this undersampled data to 10 points

and carry out network inference. This is done for each of the identified regimes. The same is done for the CoD method. We note that this is not a comparison between SSM and CoD (both work with very different assumptions), but of the effect of undersampling the data and subsequently interpolating this undersampled data to the original data length (via resampling). The above Table .2.7 suggests that as expected, there is degradation in performance (SSM/CoD) in the absence of all the available information. However, it is preferred to infer some false positives rather than lose true positive edges. This also indicates that interpolated data does not do worse than the undersampled data in terms of true positives (f_{lost}).

We make three observations regarding this method of network inference,

- It is not necessary for the target gene (*Gata2/Gata3*) to be present as part of the inferred network. We can obtain insight into the mechanisms underlying transcription in each regime even if some of the genes with *similar co-expression dynamics* as the target gene(s) are present in the inferred network.

- Probe level observations from a small number of biological replicates, seem to be very informative for network inference. This is because the LDS parameter estimation algorithm uses these *multiple* expression realizations to iteratively estimate the state mean, covariance and other parameters, notably D [48]. Hence inspite of few time points, we can use multiple measurements (biological, technical and probe-level replicates) for reliable network inference. This follows similar observations in [60], that probe-level replicates are very useful for understanding inter-gene relationships.

- Following [57], it would seem that several network hypotheses can individually explain the time evolution behavior captured by the expression data. The

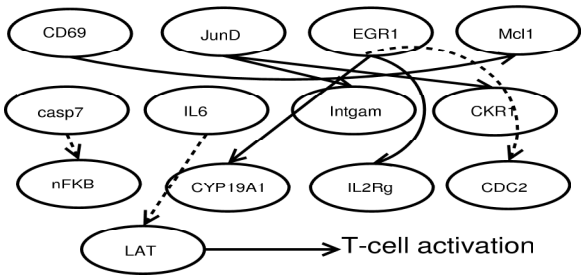


Figure 2.4: Steady state network inferred using SSM (solid lines represent the first regime, and the dotted lines indicate the second regime).

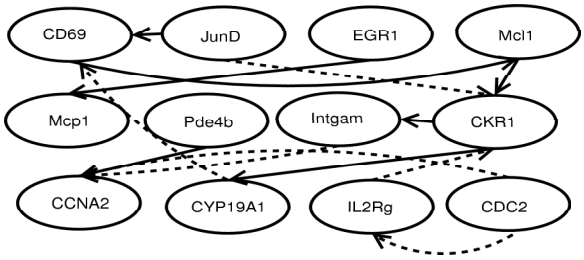


Figure 2.5: Steady state network inferred using CoD (solid lines represent the first regime, and the dotted lines indicate the second regime).

LDS parameter estimation procedure seeks to find a maximum-likelihood (ML) estimate of the system parameters A, B, C and D and then finally uses bootstrapping to only infer high confidence interactions. This ML estimation of the parameters uses an EM algorithm with multiple starts to avoid initialization-related issues [48], and thus finds the 'most consistent' hypothesis which would explain the evolution of expression data. It is this network hypothesis that we investigate. Since this network already contains our gene of interest *Gata3*, we can proceed to verify these interactions from literature and experimentally.

2.9 Discussion

One of the primary motivations for computational inference of state specific gene influence networks is to understand transcriptional regulatory mechanisms [58]. The networks inferred via this approach are fairly general and thus, there is a need to 'decompose' these networks into transcriptional, signal transduction or metabolic

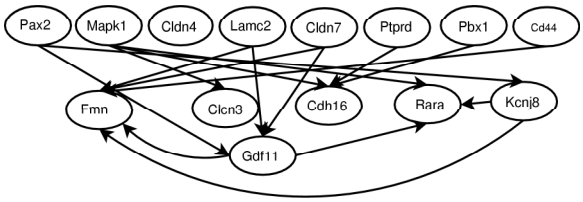


Figure 2.6: Steady state network inferred using GGMs.

Gene Symbol	Gene Name	Possible Role in Nephrogenesis (Function)
Bmp7	Bone morphogenetic protein	Cell signaling
Rara	Retinoic Acid Receptor	Retinoic acid pathway, related to eye phenotype
Gata2	GATA binding protein 2	Hematopoiesis, Urogenital development
Gata3	GATA binding protein 3	Hematopoiesis, Urogenital development
Pax2	Paired Homeobox-2	Direct target of Gata2
Lamc2	Laminin	Cell adhesion molecule
Npnt	Nephronectin	Cell adhesion molecule
Ros1	Ros1 proto-oncogene	Signaling epithelial differentiation
Ptprd	protein tyrosine phosphatase	Cell adhesion
Ret-Gdnf	Ret proto-oncogene, Glial neutrophic factor	Metanephros development
Gdf11	Growth development factor	Cell-cell signaling and adhesion
Mapk1	Mitogen activated protein kinase 1	Role in growth factor activity, cell adhesion
Kcnj8	potassium inwardly-rectifying channel, subfamily J, member 8	Potassium ion transport
Acvr2b	Activin receptor IIB	Transforming growth factor beta receptor activity

Table 2.8: Functional annotations (*Entrez Gene*) of some of the genes co-clustered with *Gata2* and *Gata3*

using a combination of biological knowledge and chemical kinetics. Depending on the insights expected, the tools for dissection of these predicted influences might vary.

For comparison, we additionally investigated a *graphical gaussian model (GGM)* approach as suggested in [38] using partial correlation as a metric to quantify influence (Fig. 2.6). This method works for short time series data but we could not find a way to incorporate previous expression values as inputs to the evolution of state or individual observations – something we could explicitly do in the state-space approach. However, we are now in the process of examining the networks inferred by the GGM approach over the regimes that we have identified from SSA. Again, we observe that the network connections reflect a steady state behavior and that tran-

Gene Symbol	Gene Name	Possible Role in T-cell activation (Function)
CD69	CD69 antigen	early T-cell activation antigen
Mcl1	Myeloid cell leukemia sequence 1 (BCL2-related)	mediates cell proliferation and survival
IL6	Interleukin 6	Accessory factor signal
LAT	Linker for activation of T cells	membrane adapter protein involved in T-cell activation
EGR1	Early Growth Response gene 1	activates nFKB signaling
CDC2	Cell Division Control protein 2	Involved in cell-cycle control
Casp7	Caspase 7	Involved in apoptosis
JunD	Jun D proto-oncogene	regulatory role of in T lymphocyte proliferation and Th cell differentiation
CKR1	Chemokine Receptor 1	negative regulator of the antiviral CD8+ T cell response
CYP19A1	Cytochrome P450, member 19	cell proliferation
Intgam	Integrin alpha M	mediates phagocytosis induced apoptosis
nFKB	nFKB protein	Signaling transduction activity
IL2Rg	Interleukin 2 receptor gamma	signaling activity
Pde4b	Phosphodiesterase 4B, cAMP-specific	mediator of cellular response to extracellular signal
Mcp1	Monocyte Chemotactic protein 1	Cytokine gene involved in immunoregulation
CCNA2	Cyclin A2	Involved in cell-cycle control

Table 2.9: Functional annotations of some of the co-clustered genes (early and late response) following T-cell activation

Method	direction	regime -specific	resolve cycles	higher lags(>1)	non-linear/ locally linear
CoD [31,32]	Y	Y	N	N	Y
GGM [7]	Y	N	N	N	Y
SSM [1]	Y	N	Y	Y	Y
regime-SSM	Y	Y	Y	Y	Y

Table 2.10: Comparison of various network inference methods (Y-Yes, N- No)

sient (state specific) changes in influence are not fully revealed. The same is observed in the case of the T-cell data, from the results reported in [67]. A comparison of all the presented methods, along with *regime-SSM* has been presented in Table. 2.10. The comparisons are based on if these frameworks permit the inference of directional influences, regime specificity, resolution of cycles, and modeling of higher lags.

2.10 Conclusions

In this work, we have developed an approach (*regime-SSM*) to infer the time varying nature of gene influence network topologies, using gene expression data.

The proposed approach integrates change point detection to delineate phases of gene co-expression, MoG clustering implying possible co-regulation and network inference amongst the regime-specific co-clustered genes using a state space framework. We can thus incorporate condition specificity of gene expression dynamics for understanding gene influences. Comparison of the proposed approach with other current procedures like GGM or CoD reveals some strengths and would very well complement existing approaches (Table. 2.10). We believe that this approach, in conjunction with sequence and transcription factor binding information can give very valuable clues to understanding the mechanisms of transcriptional regulation in higher eukaryotes.

CHAPTER III

Network Inference Using Directed Information

3.1 Introduction

Computational methods for inferring dependencies between genes ([106], [111], [127]) using probabilistic techniques have been used for quite some time now. However the biological significance of these recovered networks has been a topic of debate, apart from the fact that such approaches mostly yield networks of significant influences as 'observed/inferred' from the underlying structure of data. Alternatively, other biological data (such as sequence information) might suggest the examination of the probabilistic dependence of one gene on another gene through the transcription factor (TF) encoded by the first gene. What if we were interested in the transcriptional influences on a certain gene 'A' but our prospective network inference technique was unable to recover them? We propose a technique with an eye on two of these challenges: biological significance and influence determination between *'any'* two variables of interest. Such an approach is increasingly necessary to integrate and understand multiple sources of data (sequence, expression etc.).

The method that we propose builds on an information theoretic criterion referred to as the directed information (DTI). The DTI is a variant of mutual information (MI) that attempts to capture the direction of information flow. It is widely used

in the analysis of communication systems with feedback or feedforward ([218], [219], [125]) as well as in economic time series analysis ([194], [125]). The DTI ([218], [117]) can be interpreted as a directed version of mutual information, a criterion used quite frequently in other related work [106]. As we demonstrate, the DTI gives a sense of directional association for the principled discovery of biological influence networks.

The contributions of this work are as follows. Firstly, we present a short theoretical treatment of DTI and an approach to the supervised and unsupervised discovery of influence networks, using microarray expression data. Secondly, we examine two scenarios - the inference of large scale gene influence networks (in mammalian nephrogenesis and T-cell development) as well as potential effector genes for *Gata3* transcriptional regulation in distinct biological contexts. We find that this method outperforms other methods in several aspects and leads to the formulation of biologically relevant hypotheses that might aid subsequent experimental investigation. Finally, we conclude with the application of DTI to two important questions in bioinformatics: TF module discovery and higher-order network inference. TF module discovery is the identification of common regulatory modules (groups of TFs) whose binding sites co-occur on the promoters of co-expressed genes. Higher-order network inference, in this work, examines the resolution of three-way interactions rather than only pairwise relationships among genes [110].

3.2 Organization

This chapter is organized as follows: In section 3.3, the working definition of transcriptional gene networks is given. Based on this definition, four main research problems are posed, pertaining to: *supervised* and *unsupervised* network inference, TF module-gene interactions, and inference of higher order influence networks. Di-

rected information (DTI) is proposed as part of a general framework to answer these questions (section: 5.10.1) and a methodology for determination of influence and its significance is examined (sections: Appendix and 5.10.1). The chapter concludes with results applicable to each of the questions posed above (section: 3.8), using a combination of synthetic and real biological data.

3.3 Gene Networks

Transcription is the process of generation of messenger RNA (mRNA) from the DNA template representing the gene. It is the intermediate step before the generation of functional protein from messenger RNA. During gene expression (Fig. 4.1), transcription factor proteins are recruited at the proximal promoter of the gene as well as at distal sequence elements (enhancers/silencers) which can lie several hundreds of kilobases from the gene's transcriptional start site [204]. Since transcription factors are also proteins (or their activated forms) which are in turn encoded for by other genes, the notion of an influence between a transcription factor gene and the target gene can be considered.

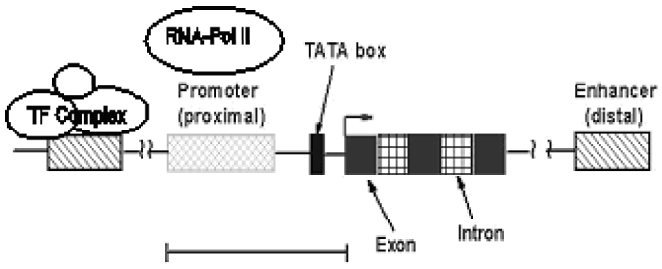


Figure 3.1: Schematic of Transcriptional Regulation. Sequence motifs at the promoter and the distal regulatory elements together confer specificity of gene expression via TF binding.

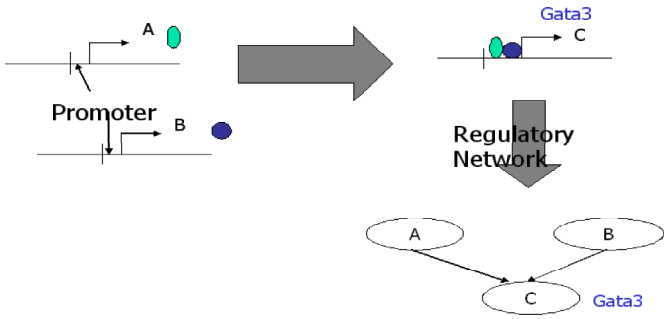


Figure 3.2: A transcriptional regulatory network with genes A and B effect C. An example of C that we study here is the *Gata3* gene.

In Fig. 3.2, a characterization of transcriptional regulatory networks, as relevant to this work, is given. As the name suggests, gene A is connected by a link to gene C if a product of gene A, say protein A, is involved in the transcriptional regulation of gene C. This might mean that protein A is involved in the formation of the complex that binds at the basal transcriptional machinery of gene C to drive gene C regulation.

The components of the transcription factor (TF) complex recruited at the gene promoter are the products of several genes. Therefore, the incorrect inference of a transcriptional regulatory network can lead to false hypotheses about the actual set of genes affecting a target gene. Since biologists are increasingly relying on computational tools to guide experiment design, a principled approach to biologically relevant network inference can lead to significant savings in time and resources in downstream experimental design. In this chapter we try to combine some of the other available biological data (phylogenetic conservation of binding sites across genomes and expression data) to build network topologies with a lower false positive rate of linkage. Some previous work in this regard has been reported in ([216], [209]).

3.4 Problem Setup

In this work, we also study the mechanism of gene regulation, with the *Gata3* gene as an example. This gene has important roles in several processes in mammalian development ([100], [204]), like in the developing urogenital system (nephrogenesis), central nervous system, and T-cell development. To find which TFs regulate the tissue-specific transcription of *Gata3* (either at the promoter or long-range regulatory elements), a commonly followed approach ([209], [216]) is to look for phylogenetically conserved transcription factor binding sites (TFBS). The hypothesis underlying this strategy is that the interspecies-conservation of a TFBS suggests a possibly functional binding of the TF at the motif (from evolutionary pressure for function). With a view to understanding gene regulatory mechanisms, this work primarily addresses the following questions:

- Biologists are also interested in the network of relationships among genes expressed under a certain set of conditions, which uses several network inference procedures, such as Bayesian networks [127], mutual information [106] etc. However, there has been lack of a common framework to do both supervised *and* unsupervised *directed* network inference within these settings to detect directed non-linear gene-gene interactions. We present directed information as a potential solution in both these scenarios. Supervised network inference pertains to finding the strengths of directed relationships between two specific genes. Unsupervised network inference deals with finding the most probable network structure to explain the observed data (like in Bayesian structure learning using expression data). This is addressed in sections 3.8.2 and 3.8.3.

- Which transcription factors are potentially active at the target gene's promoter

during its tissue-specific regulation ? - this question is primarily answered by examining the phylogenetically conserved TFBS at the promoter and asking if microarray expression data suggests the presence of an influence between the TF encoding gene and the target gene (i.e. *Gata3*). This approach thus integrates sequence and expression information (section: 3.8.4).

- Which transcription factors underlie the tissue-specific expression of a group of co-expressed/co-regulated genes (eg: *Gata3* and others)? - one common approach is to search the proximal promoters of all such tissue specific genes, and look for 'modules' of TFs that control tissue-specific expression ([209], [216]). For the *Gata3* example, we ask if there are any TFs underlying ureteric bud (UB) specific expression for *Gata3*, during nephrogenesis. For this purpose, we find modules from co-expressed gene promoters and use microarray expression to point out possible effectors of target gene expression (section: 3.8.5).

- Gene interactions during processes such as development and disease progression are rarely pairwise, and occur in cliques such as pathways. Additionally, cross-talk between components of different pathways is essential in the progression of such dynamic processes. Towards this end, the inference of higher order interactions (more than only two-way gene relationships) is seen to be a useful approach [110]. Using DTI, it would be interesting to find directed interactions between differentially expressed genes of the developing kidney to determine pathway cross-talk (section: 3.8.6).

3.4.1 Phylogenetic Conservation of Transcription Factor Binding Sites (TFBS)

As mentioned above, the mechanism of regulation of a target gene is via the binding site of the corresponding transcription factor (TF). It is believed that several

TF binding motifs might have appeared over the evolutionary time period due to insertions, mutations, deletions etc. in vertebrate genomes. However, if we are interested in the regulation of a process which is known to be similar between several organisms (say Human, Chimp, Mouse, Rat and Chicken), then we can look for the conservation of functional binding sites over all these genomes. This helps us isolate the putatively functional binding sites, as opposed to those which might have randomly arisen. This however, does not suggest that those other TF binding sites have no functional role. If we are interested in the mechanism of regulation of the *Gata3* gene (which is known to be implicated in mammalian nephrogenesis), we examine its promoter region for phylogenetically conserved TFBS (Fig. 3.3). Such information can be obtained from most genome browsers [112]. We see that even for a fairly short stretch of sequence (1 kilobase) upstream of the gene, there are several conserved sequence elements which are potential TFBS (light grey regions in Fig. 3.3).

In this figure, we present the alignment of the mouse *Gata3* promoter region with its human and rat counterparts. The height of each of the dark gray regions indicates the extent of conservation between these species. Furthermore, it indicates that several transcription factors bind at these conserved regions. To test their functional role in-vivo or in-vitro, it is necessary to select only a subset of these TFs, because of the great reliance on resources and effort. Hence the genes coding for these conserved TFs are the ones that we examine for possible influence determination via expression-based influence metrics. If we are able to infer an influence between the TF-coding gene and the target gene at which its TF binds, then this reduces the number of candidates to be tested. To examine *Gata3*'s role in kidney development, we use microarray expression data from a public repository of kidney microarray

data (*http://genet.chmcc.org/*, *http://spring.imb.uq.edu.au/* and *http://kidney.scgap.org/index.html*). Each of these sources contain expression data profiling kidney development from about day 10.5 dpc to the neonate stage. Some of these studies also examine expression in the developing ureteric bud (UB), metanephric mesenchyme (MM) apart from the whole kidney.

Our approach thus integrates several aspects:

- Using phylogenetic information to infer which TF binding sites upstream of a target gene may be functional.
- Identifying if any of the TF genes influence a target gene by coding for a transcription factor that binds at the site discovered from conservation studies. This directed influence is captured using an influence metric (like directed information) in conjunction with expression data ([178], [122])and explained in Section: 3.5.

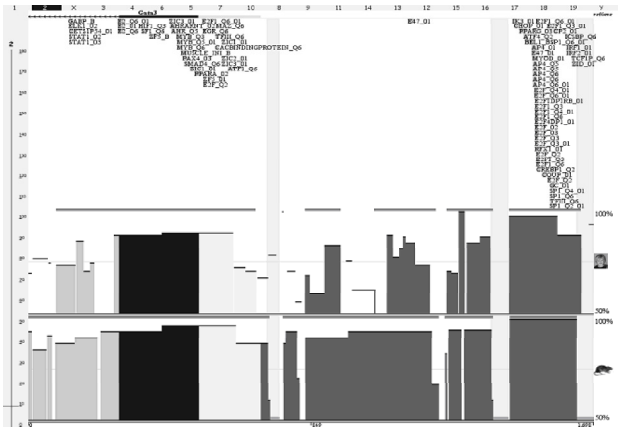


Figure 3.3: TFBS conservation between Human, Mouse and Rat, upstream (x-axis) of *Gata3*, from *http://www.ecrbrowser.dcode.org/*.

3.5 DTI Formulation

As alluded to above, there is a need for a viable influence metric that can find relationships between the TF "effector" gene (identified from phylogenetic conservation) and the target gene (like *Gata3*). Several such metrics have been proposed, notably, correlation, coefficient of determination (CoD), mutual information etc. To alleviate the challenge of detecting non-linear gene interactions, an information theoretic measure like mutual information has been used to infer the conditional dependence among genes by exploring the structure of the joint distribution of the gene expression profiles [106]. However, the absence of a directed dependence metric has hindered the utilization of the full potential of information theory. In this work, we examine the applicability of one such metric - the directed information criterion (DTI), for the inference of non-linear, directed gene influences.

The DTI - which is a measure of the directed dependence between two N-length random processes $X \equiv X^N$ and $Y \equiv Y^N$ is given by [219]:

$$I(X^N \to Y^N) = \sum_{n=1}^{N} I(X^n; Y_n | Y^{n-1}) \tag{1}$$

Here, Y^n denotes $(Y_1, Y_2, .., Y_n)$, i.e., a segment of the realization of a random process Y and $I(X^N; Y^N)$ is the Shannon mutual information [181].

An interpretation of the above formulation for DTI is in order. To infer the notion of influence between two time series (mRNA expression data) we find the mutual information between the entire evolution of gene X (up to the current instant n) and the current instant of Y (Y_n), given the evolution of gene Y up to the previous instant $n-1$ (i.e. Y^{n-1}). This is done for every instant, $n \in (1, 2, \ldots, N)$, in the N - length expression time series.

As already known, $I(X^N; Y^N) = H(X^N) - H(X^N|Y^N)$, with $H(X^N)$ and $H(X^N|Y^N)$

being the entropy of X^N and the conditional entropy of X^N given Y^N, respectively. Using this definition of mutual information, the DTI can be expressed in terms of individual and joint entropies of X^N and Y^N. The task of N-dimensional entropy estimation is an important one and due to computational complexity and moderate sample size, histogram estimation of multivariate density is unviable. However, several methods exist for consistent entropy estimation of multivariate small sample data ([189], [221], [225], [243]). In the context of microarray expression data, wherein probe-level and technical/biological replicates are available, we use the method of [189] for entropy estimation.

From (1), we have,

$$I(X^N \to Y^N) = \sum_{n=1}^{N} [H(X^n|Y^{n-1}) - H(X^n|Y^n)]$$

$$= \sum_{n=1}^{N} \{[H(X^n, Y^{n-1}) - H(Y^{n-1})] - [H(X^n, Y^n) - H(Y^n)]\} \quad (2)$$

- To evaluate the DTI expression in $Eqn.2$, we need to estimate the entropy terms $H(X^n, Y^{n-1})$, $H(Y^{n-1})$, $H(X^n, Y^n)$ and $H(Y^n)$. This involves the estimation of marginal and joint entropies of n random variables, each of which are R dimensional, R being the number of replicates (probe-level, biological and technical).

- Though some approaches need the estimation of probability density of the R-dimensional multivariate data (X^n) prior to entropy estimation, one way to circumvent this is to the use the method proposed in [189]. This approach uses a Voronoi tessellation of the R-dimensional space to build nearly uniform partitions (of equal mass) of the density. The set of Voronoi regions $(V^1, V^2, \ldots, V^n)$ for each of the n points in R-dimensional space is formed by associating with each point X_k, a set of points V^k that are closer to X_k than any other point

X_l, where the subscripts k and l pertain to the k^{th} and l^{th} time instants of gene expression.

- Thus, the entropy estimator is expressed as : $\hat{H}(X^n) = \frac{1}{n}\sum_{i=1}^{n} \log(nA(V^i))$, where $A(V^i)$ is the R-dimensional volume of Voronoi region V^i. $A(V^i)$ is computed as the area of the polygon formed by the vertices of the convex hull of the Voronoi region V^i. This estimate has low variance and is asymptotically efficient [190].

To obtain the DTI between any two genes of interest (X and Y) with N-length expression profiles X^N and Y^N respectively, we plug in the entropy estimates computed above into the above expression (2).

From the definition of DTI, we know that $0 \leq I(X_i^N \to Y^N) \leq I(X_i^N; Y^N) < \infty$. For easy comparison with other metrics, we use a normalized DTI metric (see Appendix) given by $\rho_{DTI} = \sqrt{1 - e^{-2I(X^N \to Y^N)}} = \sqrt{1 - e^{-2\sum_{i=1}^{N} I(X^i; Y_i | Y^{i-1})}}$. This maps the large range of DTI, ($[0, \infty]$) to lie in $[0, 1]$. Another point of consideration is to estimate the significance of the 'true' DTI value compared to a null distribution on the DTI value (i.e. what is the chance of finding the DTI value by chance from the series X and Y). This is done using empirical p-value estimation after bootstrap resampling (Sec: 5.10.1). A threshold p-value of 0.05 is used to estimate the significance of the true DTI value in conjunction with the the density of a random data permutation, as outlined below.

3.6 Significance Estimation of DTI

We now outline a procedure to estimate the empirical p-value to ascertain the significance of the normalized directed information $\hat{I}(X^N \to Y^N)$ between any two N-length time series $X \equiv X^N = (X_1, X_2, \ldots, X_N)$, and $Y \equiv Y^N = (Y_1, Y_2, \ldots, Y_N)$.

In our case, the detection statistic is $\Theta = \hat{I}(X^N \to Y^N)$ and the chosen acceptable p-value is α.

The overall bootstrap based test procedure is ([186],[229],[75]):

- Repeat the following procedure $B(= 1000)$ times (with index $b = 1, \ldots, B$):
  - Generate resampled (with replacement, or reordering) versions of the times series X^N, Y^N, denoted by X_b^N, Y_b^N respectively.
  - Compute the statistic $\theta^b = \hat{I}(X_b^N \to Y_b^N)$.
- Construct an empirical CDF (cumulative distribution function) from these bootstrapped sample statistics, as $F_\Theta(\theta) = P(\Theta \leq \theta) = \frac{1}{B}\sum_{b=1}^{B} I_{x \geq 0}(x = \theta - \theta^b)$, where I is an indicator random variable on its argument x.
- Compute the true detection statistic (on the original time series) $\theta_0 = \hat{I}(X^N \to Y^N)$ and its corresponding p-value ($p_0 = 1 - F_\Theta(\theta_0)$) under the empirical null distribution $F_\Theta(\theta)$.
- If $F_\Theta(\theta_0) \geq (1 - \alpha)$, then we have that the true DTI value is significant at level α, leading to rejection of null-hypothesis (no directional association).

3.7 Summary of Algorithm

We now present two versions of the DTI algorithm, one which involves an inference of general influence network between all genes of interest (*unsupervised-DTI*) and another, a focused search for effector genes which influence one particular gene of interest (*supervised-DTI*).

Our proposed approach using (*supervised-DTI*) for determining the effectors for gene B is as follows:

- Identify the G genes $(A_1, A_2, \ldots, A_G)$, based on required phenotypical characteristic using fold change studies. Preprocess the gene expression profiles by normalization and cubic spline interpolation. Assuming that there are N points for each gene, entropy estimation is used to compute the terms in the DTI expression (Eqn. 2).

- For each pair of genes A_i and B among these G genes :

  {

  - Look for a phylogenetically conserved binding site of TF encoded by gene A_i in the upstream region of gene B.
  - Find $DTI(A_i, B) = I(A_i^N \rightarrow B^N)$, and the normalized DTI from A_i to B, $\rho_{DTI}(A_i, B) = \sqrt{1 - e^{-2I(A_i^N \rightarrow B^N)}}$.
  - Bootstrap resampling over the data points of A_i and B yields a null distribution for $DTI(A_i, B)$. If the true $DTI(A_i, B)$ is greater than the 95% upper limit of the confidence interval (CI) from this null histogram, infer a potential influence from A_i to B.
  - The value of the normalized DTI from A_i to B gives the putative strength of interaction/influence.
  - Every gene A_i which is potentially influencing B is an 'effector'. This search is done for each gene A_i among these G genes $((A_1, A_2, \ldots, A_G))$.

  }

Note: As can be seen, phylogenetic information is inherently built into the influence network inference step above. We note that, in *supervised-DTI*, the choice of potential effectors for a target gene is based on only those TFs that have a binding site at the

target gene's promoter. In this sense, *supervised-DTI* aims to reduce the overall search space based on biological prior knowledge.

For *unsupervised DTI*, we adapt the above approach for every pair of genes (A, B) in the list, noting that $DTI(A, B) \neq DTI(B, A)$. In this case we are not looking at any interaction in particular, but are interested in the entire influence network that can be potentially inferred from the given time series expression data. The network adjacency matrix has entries depending on the direction of influence and is related to the strength of influence as well as control of false discovery rate (FDR). The Benjamini-Hochberg procedure [132] is used to screen each of the $M(= G(G-1))$ hypotheses (both directions) during network discovery amongst G genes.

Briefly, the FDR procedure controls the expected proportion of false positives among the total number of rejections rather than just the chance of false positives [120]. It tolerates more false positives, and allows fewer false negatives.

- The p-values of the various edges $(1, 2, \ldots, M)$ are ranked from lowest to highest, all satisfying the original significance cut-off of $p = 0.05$. The ranked p-values are designated as $p_{(1)}, p_{(2)}, \ldots, p_{(M)}$.
- For $j = 1, 2, \ldots, M$, the null hypothesis (no edge) H_j is rejected at level α if $p_{(j)} \leq \frac{j}{M}\alpha$.
- All the edges with p-value $\leq p_{(j)}$ are retained in the final network.

In Table. 3.7, we compare the various contemporary methods of directed network inference. Recent literature has introduced several interesting approaches such as graphical gaussian models (GGMs), coefficient of determination (CoD), state space models (SSMs) for directed network inference. This comparison is based primarily on expectations from such inference procedures - that we would like any such

metric/procedure to:

- Resolve cycles in recovered interactions.
- Be capable of resolving directional and potentially non-linear interactions. This is because interactions amongst genes involve non-linear kinetics.
- Be a non-parametric procedure to avoid distributional assumptions (noise etc).
- Be capable of recovering interactions that a biologist might be interested in. Rather than use a method that discovers interactions underlying the data purely, the biologist should be able to use prior knowledge (from literature perhaps). For example, a biologist can examine the strength and significance of a known interaction and use this as a basis for finding other such interactions.

From the above comparisons, we see that DTI is one metric which can recover interactions under all these considerations.

Table 3.1: Comparison of various network inference methods.

Method	Resolve Cycles	Non -linear framework	Search for interaction	Non -parametric framework
SSM ([116], [77])	Y	Y	N	Y
CoD ([94])	N	N	Y	N
GGM ([111])	N	Y	N	N
DTI ([117])	Y	Y	Y	Y

3.8 Results

In this section, we give some scenarios where DTI can complement existing bioinformatics strategies to answer several questions pertaining to transcriptional regula-

tory mechanisms. We address four different questions.

- To infer gene influence networks between genes that have a role in early kidney development and T-cell activation, we use *unsupervised DTI* with relevant microarray expression data, noting that these influence networks are not necessarily transcriptional regulatory networks.

- To find transcription factors that might be involved in the regulation of a target gene (like *Gata3*) at the promoter, a common approach is to first look for phylogenetically conserved TFBS sequences across related species. These species are selected based on whether the particular biological process is conserved in them. To add additional credence to the role of these conserved TFBSes, microarray expression can be integrated via *supervised DTI* to check for evidence of an influence between the TF encoding gene and the target gene.

- Thirdly, we examine the promoters of several genes that have a documented role in ureteric bud development. The idea is to look for common transcription factor modules that govern the combined co-expression and co-regulation of these genes [216]. Again, expression data and *supervised DTI* can be used to check for influences between the module components and the target gene(s).

- Finally, the problem of inferring higher-order dependencies between various genes using a combination of mutual and directed information is presented in the context of differentially expressed UB-specific genes of the developing kidney.

Before proceeding, we examine the performance of this approach on synthetic data.

3.8.1 Synthetic Network

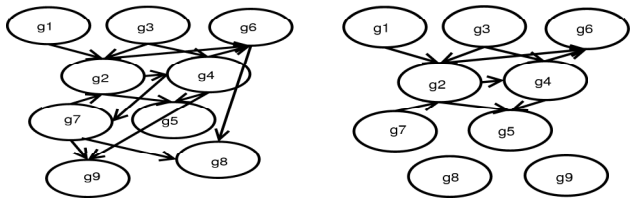
A synthetic network is constructed in the following fashion: We assume that there are three genes g_1, g_3 and g_7 (modeled as uniform random variables) which drive the remaining genes of a nine gene network. The evolution equations are as below:

$$
\begin{aligned}
g_{2,t} &= \frac{1}{2}g_{1,t-1} + \frac{1}{3}g_{3,t-2} + g_{7,t-1} + \epsilon_t; \\
g_{4,t} &= g_{2,t-1}^2 + g_{3,t-1}^{1/2} + \epsilon_t; \\
g_{5,t} &= g_{2,t-2} + g_{4,t-1} + \epsilon_t; \\
g_{6,t} &= g_{4,t-1} + g_{2,t-2}^{1/2} + \epsilon_t; \\
g_{7,t} &= \frac{1}{2}g_{4,t-1}^{1/3} + \epsilon_t; \\
g_{8,t} &= \frac{1}{2}g_{6,t-1}^{1/3} + \frac{1}{3}g_{7,t-1}^{1/2} + \epsilon_t; \\
g_{9,t} &= \frac{2}{3}g_{4,t-1}^{2/3} + \frac{1}{4}g_{7,t-2}^{1/2} + \epsilon_t;
\end{aligned}
$$

ϵ_t is the noise term associated with stochastic gene expression and is modeled as a gaussian random variable $\mathcal{N}(0, \sigma^2)$, $\sigma = 0.01$.

For the purpose of comparison, we study the performance of the Coefficient of Determination (CoD) approach for directed influence network determination. The CoD allows the determination of association between two genes via a R^2 goodness of fit statistic. The methods of ([94], [103]) are implemented on the time series data. Such a study would be useful to determine the relative merits of each approach. We believe that no one procedure can work for every application and the choice of an appropriate method would be governed by the biological question under investigation. Each of these methods use some underlying assumptions and if these are consistent

with the question that we ask, then that method has utility.



(With DTI) **(with CoD)**

Figure 3.4: The synthetic network as recovered by (a) DTI and (b) CoD.

As can be seen (Fig. 3.4), though CoD can detect linear lag influences, the strongly non-linear ones are missed. DTI detects these influences and exactly reproduces the synthetic network. Given the non-linear nature of transcriptional kinetics, this is essential for reliable network inference. DTI is also able to resolve loops and cycles ($g_3, [g_2, g_4], g_5$ and g_2, g_4, g_7, g_2). Based on these observations, we examine the networks inferred using DTI in both the supervised and unsupervised settings.

3.8.2 Directed Network Inference: *Gata3* regulation in early kidney development

Biologists have an interest in influence networks that might be active during organ development. Advances in laser capture microdissection coupled with those in microarray methodology have enabled the investigation of temporal profiles of genes putatively involved in these embryonic processes. Forty seven genes are expressed differentially between the ureteric bud and metanephric mesenchyme [122] and putatively involved in bud branching during kidney development. The expression data [178] temporally profiles kidney development from day 10.5 dpc to the neonate stage. The influence network amongst these genes is shown below (Fig. 3.5). Several of the presented interactions are biologically validated and there is an interest to confirm

the novel ones pointed out in the network. The annotations of some of these genes are given below (Table. 3.2).

Some of the interactions that have been experimentally validated include the *Rara-Mapk1* [76], *Pax2-Gata3* [92] and *Agtr-Pax2* [128] interactions. We note that this result clarifies the application of DTI for network inference in an unsupervised manner - i.e. discovering interactions revealed by data rather than examining the strengths of interactions known a priori. Such a scenario will be explored later (Sec: 3.8.4). We note that though several interaction networks are recovered, we only show the largest network including *Gata3*, because this is the gene of interest in this study.

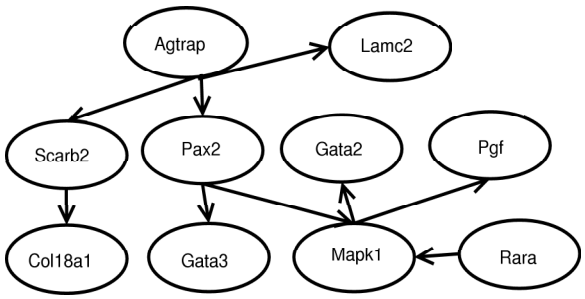


Figure 3.5: Overall Influence network using DTI during early kidney development.

3.8.3 Directed Network Inference: T-cell Activation

To clarify the validity of the presented approach, we present a similar analysis on another data set - the T-cell expression data [116], in Fig. 3.6. This data represents the expression of various genes after T-cell activation using stimulation with phorbolester PMA and ionomycin. The dataset contains the profiles of about 58 genes over 10 time points with 44 replicate measurements for each time point.

Several of these interactions are confirmed in earlier studies ([116], [90], [129], [118]) and again point to the strength of DTI in recovering known interactions. The annotation of some of these genes are given in Table. 3.3. We note that the network of Fig. 3.6 shows the largest influence network (containing *Gata3*) that can be

recovered. *Gata3* is involved in T-cell development as well as kidney development and hence it is interesting to see networks relevant to each context in Figs. 3.5 and 3.6. Also, these 58 genes relevant to T-cell activation are very different from those for kidney development, with fairly low overlap. For example this list does not include *Pax2* (which is relevant in the kidney development data).

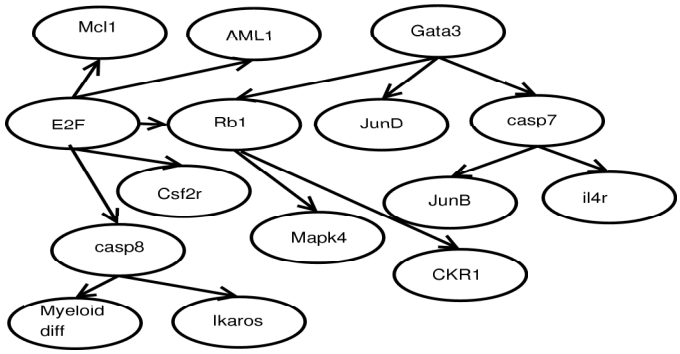


Figure 3.6: DTI based T-cell network.

3.8.4 Phylogenetic conservation of TFBS effectors

A common approach to the determination of "functional" transcription factor binding sites in genomic regions is to look for motifs in conserved regions across various species. Here we focused on the interspecies conservation of TFBS (Fig. 3.3) in the *Gata3* promoter to determine which of them might be related to transcriptional regulation of *Gata3*. Such a conservation across multiple-species suggests selective evolutionary pressure on the region with a potential relevance for function.

As can be seen in Fig. 3.3, we examine the *Gata3* gene promoter and find at least forty different transcription factors that could putatively bind at the promoter as part of the transcriptional complex. Some of these TFs, however, belong to the same family.

Using *supervised DTI*, we examined the strength of influence from each of the TF-encoding genes (A_i) to *Gata3*, based on expression level ([178], *http://spring.imb.uq.edu.au/*).

These "strength of influence" DTI values are first checked for significance at a p-value of 0.05 and then ranked from highest to lowest (noting that the objective is to maximize $I(A_i \rightarrow \textit{Gata3})$).

Based on this ranking, we indicate some of the TFs that have highest influence on *Gata3* expression (Fig. 3.7). Obviously, this information is far from complete, because of examination only at the mRNA level for both effectors as well as *Gata3*.

Table. 3.4 shows the embryonic kidney-specific expression of the TFs from Fig. 3.7. This is an independent annotation obtained from UNIPROT (*http://expasy.org/sprot/*). To understand the notion of kidney-specific regulation of *Gata3* expression by various transcription factors, we have integrated three different criteria. We expect that the TFs regulating expression would have an influence on *Gata3* expression, be expressed in the kidney and have a conserved binding site at the *Gata3* promoter. This is clarified in part by Fig. 3.7 and Table. 3.4. As an example, we see that the TFs *Pax2, PPAR, SP1* have high influence via DTI and are expressed in embryonic kidney (Table. 3.4), apart from having conserved TFBS. This lends good computational evidence for the role of these TFs in *Gata3* regulation, and presents a reasonable hypothesis worthy of experimental validation.

Additionally, we examined the influence for another two TFs - *STE12* and *HP1*, both of which have a high co-expression correlation with *Gata3* as well as conserved TFBS in the promoter region. The DTI criterion gave us no evidence of influence between these two TFs and *Gata3's* activity. This information coupled with the present evidence concerning the non-kidney specificity of *STE12* and *HP1*, presents an argument for the non-involvement of these TFs in kidney specific regulation of *Gata3*. Thus, the DTI criterion can be used to guide more focused experiments to identify the true transcriptional effectors underlying *Gata3* expression.

Table 3.2: Functional annotations (*Entrez Gene*) of some of the genes co-expressed with *Gata2* and *Gata3* during nephrogenesis.

Gene Symbol	Gene Name	Possible Role in Nephrogenesis (Function)
Rara	Retinoic Acid Receptor	crucial in early kidney development
Gata2	*GATA* binding protein 2	several aspects of urogenital development
Gata3	*GATA* binding protein 3	several aspects of urogenital development
Pax2	Paired Homeobox-2	conversion of MM precursor cells to tubular epithelium
Lamc2	Laminin	Cell adhesion molecule
Pgf	Placental Growth Factor	Arteriogenesis, Growth factor activity during development
Col18a1	collagen, type $XVIII$, alpha 1	extracellular matrix structural constituent, cell adhesion
Agtrap	Angiotensin II receptor-associated protein	Ureteric bud cell branching

Table 3.3: Functional annotations of some of the genes following T-cell activation.

Gene Symbol	Gene Name	Possible Role in T-cell activation (Function)
Casp7	Caspase 7	Involved in apoptosis
JunD	Jun D proto-oncogene	regulatory role of in T lymphocyte proliferation and Th cell differentiation
CKR1	Chemokine Receptor 1	negative regulator of the antiviral CD8+ T cell response
Il4r	Interleukin 4 receptor	inhibits *IL4*-mediated cell proliferation
Mapk4	Mitogen activated kinase 4	Signal transduction
AML1	acute myeloid leukemia 1; aml1 oncogene	CD4 silencing during T-cell differentiation
Rb1	Retinoblastoma 1	Cell cycle control

This application shows the utility of DTI to specifically explore the expression-level influence of a TF of interest to any target gene. This result coupled with the unsupervised network inference methods in kidney and T-cell data, establish the DTI-based methodology as a common framework for both types of analysis.

3.8.5 Module TFs in co-regulated genes

We examine another interesting scenario for the principled application of the DTI criterion. The co-expression of genes in a biological context is a complex phenomenon involving the combinatorial regulation of such genes by several transcription factors (TFs). Such co-expression occurs during processes like development and disease progression. This is also observed in co-clustered genes from the output of hierarchical clustering algorithms (signatures). The underlying hypothesis is that co-clustered/co-expressed genes might be under the control of some common TFs

(modules) that underlie the co-ordinated expression of all these implicated genes.

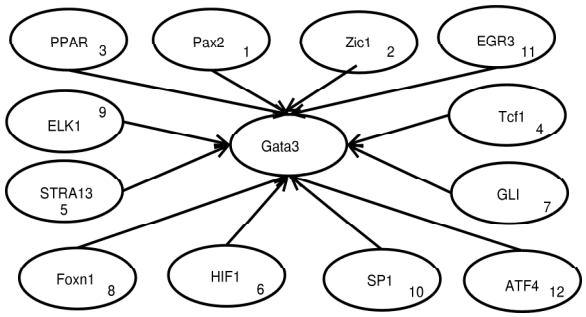


Figure 3.7: Putative upstream TFs using DTI for the *Gata3* gene. The numbers in each TF oval represent the DTI rank of the respective TF.

Several tools (Genomatix [83], CREME [114], Toucan [74]) allow the inference of such transcription factor modules from sets of genes. However, the next logical question is if any of the TFs comprising the module indeed have an expression-level influence on these target gene(s). Supervised DTI can be used in this context to rank the most likely "effector TFs" for each gene in the gene-set.

Table 3.4: Functional annotations of some of the transcription factor genes putatively influencing *Gata3* regulation in kidney.

Gene Symbol	Description	Expressed in Kidney
PPAR	peroxisome proliferator-activated receptor	Y
Pax2	Paired Homeobox-2	Y
HIF1	Hypoxia-inducible factor 1	Y
SP1	SP1 transcription factor	Y
GLI	GLI-Kruppel family member	Y
EGR3	early growth response 3	Y

To illustrate this application, genes that have expression in the developing Ureteric Bud (UB) in the kidney are examined. The inductive signals between the ureteric

bud and metanephric mesenchyme causes the differentiation of fetal kidney stem cells into nephrons, the basic unit of function of the kidney. An examination of these UB-specific genes (obtained from the Mouse Genome Informatics repository at: *http://www.informatics.jax.org/*), ([123], [122]) reveals some modules. The UB-specific genes as well as the modules are listed in Tables. 3.5 and 3.6 respectively.

Briefly, the modules are obtained as follows: the various UB-specific gene sequences are mined for their proximal promoter (from $\sim$ 2000bp upstream to 200bp downstream from the transcription start site). The various promoters are then aligned and a search for significantly over-represented TFs is done using the position weight matrices derived from the TRANSFAC/JASPAR database (MotifScanner). From this set of TFs, modules of TFs (with potentially overlapping sites) are obtained (ModuleSearcher). The TOUCAN 3.0.2 tool [74] allows for the entire sequence of steps from sequence extraction to module searches. The list of all TFs in the various modules identified are listed in Table. 3.6.

Table 3.5: Genes expressed in the developing ureteric bud (day e10.5-11.0), as reported in Mouse Genome Informatics database.

Ensembl Gene ID	Gene Name
ENSMUSG00000015619	*Gata3*
ENSMUSG00000032796	*Lama1*
ENSMUSG00000015647	*Lama5*
ENSMUSG00000026478	*Lamc1*
ENSMUSG00000018698	*Lhx1*
ENSMUSG00000008999	*Bmp7*
ENSMUSG00000023906	*Cldn6*
ENSMUSG00000059040	*Eno1*
ENSMUSG00000004231	*Pax2*
ENSMUSG00000030110	*Ret*
ENSMUSG00000022144	*Gdnf*
ENSMUSG00000031681	*Smad1*
ENSMUSG00000024563	*Smad2*
ENSMUSG00000074227	*Spint2*
ENSMUSG00000015957	*Wnt11*
ENSMUSG00000039481	*Nrtn*
ENSMUSG00000063358	*Mapk1*
ENSMUSG00000063065	*Mapk3*

Table 3.6: Annotation of the module TFs from UB-specific genes.

TFs in module	Annotation	Kidney-specificity (Y/N) (GNF/literature)
SP1	trans-acting TF 1	Y
LMO2	LIM domain only 2	N
OCT1	POU domain, class 2, TF 1	Y
ZIC1	zinc finger protein of the cerebellum 1	N
MZF1	myeloid zinc finger 1	Y
AP2	TF AP-2	Y
AP4	TF AP-4	Y
YY1	YY1 transcription factor	Y
TAL1	T-cell acute lymphocytic leukemia 1	Y (cell line)
PAX2	paired box gene 2	Y
HNF4	Hepatocyte Nuclear Factor 4	Y

The list of module TFs is obtained by combining expression annotations (from MGI) and sequence analysis. For the purpose of integrating heterogeneous data and to reduce the number of candidate TFs that are putatively involved in regulating UB-specific genes, we can use DTI to find influences between the TF-genes and the UB-specific genes using expression data. As an example, one of the TFs in the module list is *Pax2* and has an important role in UB differentiation [92]. Another gene expressed in the developing UB is *Gata3*. We now examine if the DTI, $I(Pax2 \rightarrow Gata3)$ is significant and ranks high in the list. This is highlighted in Fig. 3.8.

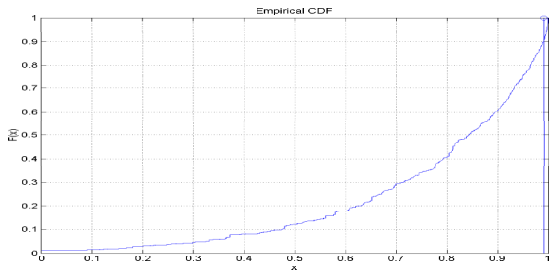


Figure 3.8: Cumulative Distribution Function for bootstrapped $I(Pax2 \rightarrow Gata3)$. The true value of $I(Pax2 \rightarrow Gata3) = 0.9911$.

For the *Pax2-Gata3* interaction, we show the cumulative distribution function of the bootstrapped detection statistic (Fig. 3.8) as well as the position of the true DTI estimate in relation to the overall histogram. With the obtained density estimate of the *Pax2-Gata3* interaction, shown in Fig. 3.8, we can find significance values of the true DTI estimate in relation to the bootstrapped null distribution.

An experimental validation of this is presented in ([184], [92]). Thus, we can look at each module member for possible role in *Gata3* regulation. As can be seen, this approach integrates sequence information, phylogeny, and expression to look for upstream effectors for genes of interest (those that share some pattern of co-expression/co-regulation).

Extending this further, the strength and significance of the DTI can be found between every pair of TF and UB-specific gene of Tables. 3.5 and 3.6. This can be visualized as a 'bipartite graph' of TF-gene interactions, shown in Fig. 3.9. The graph summarizes the degree of interactions between the various transcription factors in the modules and the co-expressed genes, and is the overall integration of annotation, sequence and expression data. Additionally, the embryonic kidney specificity of the various module TFs is listed, based on literature and tissue-specificity annotation (*http://symatlas.gnf.org/SymAtlas/*). It is to be noted that some transcription

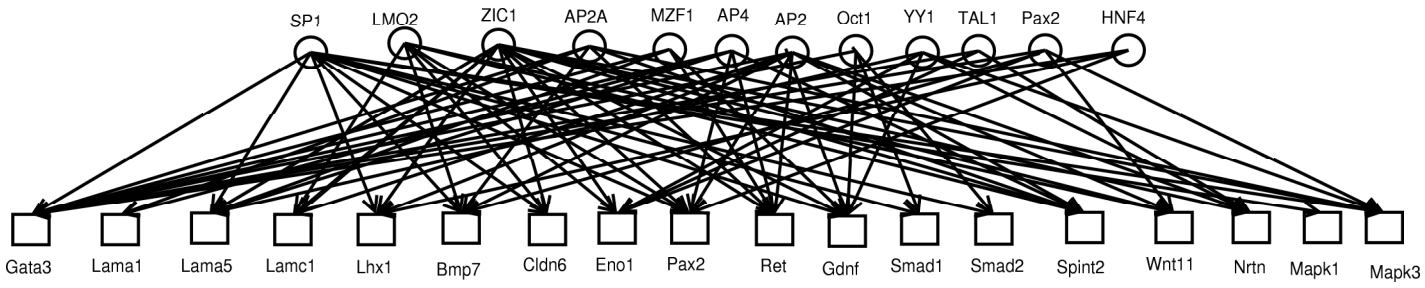


Figure 3.9: A bipartite graph between the group of module TFs and genes co-expressed in the developing ureteric bud (MGI:e10.5-11.0).

factors such as *SP1* have ubiquitous expression across most tissues ([84], [119]), and are not as informative as kidney-specific ones like *Pax2* [92] or *HNF4a* [124].

3.8.6 Higher-order MI and DTI

The final part of this work highlights that directed information (DTI) and mutual information (MI) can together aid in the discovery of higher order interactions amongst genes. Higher order MI ([221], [110]) has been used successfully for the discovery of interactions among triples of genes. Following work done in [121], we use the 'triplet information' given by

$$\begin{aligned} I_3(x_i;x_j;x_k) &= \sum_i H(x_i) - \sum_{i<j} H(x_i,x_j) + H(x_i,x_j,x_k) \\ &= I(x_1;x_2;x_3) - \sum_{i<j} I(x_i;x_j) \\ &= [I(x_1;x_3) + I(x_2;x_3)] - I(\{x_1,x_2\};x_3) \end{aligned}$$

From the above definition, it is clear that the use of triplet information helps resolve the pairwise-joint dependencies between x_i, x_j and x_k versus the synergistic dependence of any variable on the 'combination' of the other two variables. A positive value of $I_3(x_i;x_j;x_k)$ indicates pairwise-dependence and hence DTI can be used to infer directional association between x_i, x_j and x_k. A negative value indicates synergy

and needs to be resolved further.

For the network shown in Fig. 3.5, we aim to recover any synergistic interactions of various genes using higher-order entropy methods, that are potentially missed due to consideration of only pairwise interactions.

For the synergy framework presented above, we seek to determine the direction of association of $\{x_i, x_j\}$ and x_k, for all genes i, j, k. For this purpose, $I(\{x_i, x_j\} \rightarrow x_k)$ is determined, using methods presented earlier. Depending on the relative magnitude of $I(\{x_i, x_j\} \rightarrow x_k)$ and $I(x_k \rightarrow \{x_i, x_j\})$, the direction of association can be determined.

We now consider the set of genes common to those profiled in the microarray expression ([80], [178], [122]) study as well as the annotated genes from MGI. For these 12 genes (*Bmp7, Cldn7, Gata3, Gdnf, Lamc2, Mapk1, Mapk3, Nrtn, Pax2, Ret, Spint1, Wnt11*), we study the dependencies discovered using 'triplet information'. Also, for the purposes of this work, we only present those dependencies wherein the triplet information is negative indicating possible synergistic interactions. These interactions are indicated below (Table. 3.7).

Several of the pathways, such as the *Gdnf-Ret*, *Wnt*, and *Mapk* are implicated in ureteric bud differentiation ([105], [86]). However, most studies have focussed on interaction within a pathway and not so much on cross-talk between various pathways. Organ development is a complex phenomenon and needs several reciprocal interactions to control the growth of various cell populations. It is interesting to see several known cross-interactions picked up using higher-order information, based on expression data alone (Table. 3.7). Given that co-operation/synergies between various pathways is essential in most other biological processes, we believe that using a combination of higher-order MI and DTI would aid in the experimental resolution

of such interactions.

Table 3.7: Some triplet interactions (discovered using DTI) that have putative biological role. Biological validation from literature is given in parentheses.

			UB-Specificity & Citation (*http://symatlas.gnf.org/SymAtlas/*)
Gdnf	*Ret*	*Gata3*	Y [92]
Ret	*Bmp7*	*Gata3*	Y [86]
Pax2	*Gata3*	*Ret*	Y [82]
Ret	*Wnt11*	*Gdnf*	Y [105]
Pax2	*Wnt11*	*Gata3*	Y [92]
Pax2	*Ret*	*Gdnf*	Y ([82],[79])

Conclusions

In this work, we have presented the notion of directed information (DTI) as a reliable criterion for the inference of influence in gene networks. After motivating the utility of DTI in discovering directed non-linear interactions, we present two variants of DTI that can be used depending on context. One version, *unsupervised-DTI*, like traditional network inference, enables the discovery of influences (regulatory or non-regulatory) among any given set of genes. The other version (*supervised-DTI*) aids the modeling of the strength of influence between two specific genes of interest - questions arising during transcriptional influence. It is interesting that DTI enables the use of a common framework for both these purposes as well as is general enough to accommodate arbitrary lag, non-linearity, and resolution of cycles, loops and direction.

We see that the above presented combination of supervised and unsupervised variants enable their applicability to several important problems in bioinformatics

(upstream TF discovery, module-gene interactions, and higher-order influence determination), some of which are presented in the results section. The network inference approach can also allow incorporation of additional biophysical knowledge - both pertaining to physical mechanisms as well as protein interactions that exist during transcription. We point out that given the diverse nature of biological data of varying throughput, one has to adopt an approach to integrate such data to make biologically relevant findings and hence the DTI metric fits very naturally into such an integrative framework.

Acknowledgements

The authors gratefully acknowledge the support of the NIH under award 5R01-GM028896-21 (J.D.E). We would like to thank Prof. Sandeep Pradhan and Mr. Ramji Venkataramanan for useful discussions on Directed Information. We are very grateful to Prof. Erik Learned-Miller for sharing his code for higher-order entropy estimation, and Prof. Bruce Aronow for kidney expression data.

APPENDIX: A NORMALIZED DTI MEASURE

In this section, an expression for a 'normalized DTI coefficient' is derived. This is useful for a meaningful comparison across different criteria during network inference. The purpose of this section is to establish some connections between quantities like MI, DTI, and correlation. In this section, we use X, Y, Z for X^N, Y^N and Z^N interchangeably, i.e $X \equiv X^N$, $Y \equiv Y^N$, and $Z \equiv Z^N$.

By the definition of DTI, we can see that $0 \leq I(X^N \to Y^N) \leq I(X^N; Y^N) < \infty$. The normalized measure ρ_{DTI} should be able to map this large range ($[0, \infty]$) to $[0, 1]$. We recall that the multivariate canonical correlation is given by [[93]]: $\rho_{X^N;Y^N} = \Sigma_{X^N}^{-1/2} \Sigma_{X^N Y^N} \Sigma_{Y^N}^{-1/2}$ and this is normalized having eigenvalues between 0

and 1. We also recall that, under a Gaussian distribution on X^N and Y^N, the joint entropy $H(X^N;Y^N) = -\frac{1}{2}\ln(2\pi e)^{2N}|\Sigma_{X^NY^N}|$, where $|A|$ is the determinant of matrix A, $\Sigma_{X^NY^N}$ denotes the covariance matrix, computed as $\Sigma_{X^NY^N} = \frac{1}{R-1}X^TY$, indicating that there are R replicates of the X, Y time series, each of length N.

Thus, for $I(X^N;Y^N) = H(X^N) + H(Y^N) - H(X^N,Y^N)$, the expression for mutual information, under jointly Gaussian assumptions on X^N and Y^N, becomes, $I(X;Y) = -\frac{1}{2}\ln(\frac{|\Sigma_{X^NY^N}|^2}{|\Sigma_{X^N}|.|\Sigma_{Y^N}|}) = -\frac{1}{2}\ln(1-\rho^2_{X^N;Y^N})$. Hence, a straightforward transformation is normalized MI, $\rho_{MI} = \sqrt{1-e^{-2I(X^N;Y^N)}} = \sqrt{1-e^{-2\sum_{i=1}^{N} I(X^N;Y_i|Y^{i-1})}}$. A connection with [97], can thus be immediately seen.

With this, ρ_{MI} is normalized between $[0,1]$ and gives a better absolute definition of dependency that does not depend on the unnormalized MI. We will use this definition of normalized information coefficients in the present set of simulation studies.

For constructing a normalized version of the DTI, we can extend this approach, from [[194]]. Consider three random vectors $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$, each of which are identically distributed as $\mathcal{N}(\mu_X, \Sigma_{XX})$, $\mathcal{N}(\mu_Y, \Sigma_{YY})$, and $\mathcal{N}(\mu_Z, \Sigma_{ZZ})$ respectively. We also have,

$$(\mathbf{X},\mathbf{Y},\mathbf{Z}) \sim \mathcal{N}\left[\begin{pmatrix}\mu_X\\ \mu_Y\\ \mu_Z\end{pmatrix}, \begin{pmatrix}\Sigma_{XX} & \Sigma_{XY} & \Sigma_{XZ}\\ \Sigma_{YX} & \Sigma_{YY} & \Sigma_{YZ}\\ \Sigma_{ZX} & \Sigma_{ZY} & \Sigma_{ZZ}\end{pmatrix}\right]$$

Their partial correlation $\delta_{YX|Z}$ is then given by, $\delta_{YX|Z} = \sqrt{\frac{a_2^2}{a_1 a_3}}$ with, $a_1 = \Sigma_{YY} - \Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZY}$, $a_2 = \Sigma_{YX} - \Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX}$, $a_3 = \Sigma_{XX} - \Sigma_{XZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX}$.

Recalling results from conditional Gaussian distributions, these can be denoted by: $a_1 = \Sigma_{Y|Z}, a_2 = \Sigma_{XY|Z}$ and $a_3 = \Sigma_{X|Z}$. Thus, $\delta_{YX|Z} = \Sigma_{Y|Z}^{-1/2}\Sigma_{XY|Z}\Sigma_{X|Z}^{-1/2}$. Extending the above result from the mutual information to the directed information

case, we have, $\rho_{DTI} = \sqrt{1 - e^{-2\sum_{i=1}^{N} I(X^i;Y_i|Y^{i-1})}}$.

We recall the primary difference between MI and DTI, (note the superscript on X):

MI: $I(X^N;Y^N) = \sum_{i=1}^{N} I(X^N;Y_i|Y^{i-1})$.

DTI: $I(X^N \to Y^N) = \sum_{i=1}^{N} I(X^i;Y_i|Y^{i-1})$.

Having found the normalized DTI, we ask if the obtained DTI estimate is significant with respect to a 'null DTI distribution' obtained by random chance. This is addressed in Section 5.10.1.

We note that though the normality assumption was used to show the connection between information and correlation, this distributional assumption is not used anywhere in the original DTI metric formulation and computation during its application to network inference.

CHAPTER IV

Finding Motifs underlying Tissue - Specific Expression

4.1 Introduction

Understanding the mechanisms underlying regulation of tissue-specific gene expression remains a challenging question. While all mature cells in the body have a complete copy of the human genome, each cell type only expresses those genes it needs to carry out its assigned task. This includes genes required for basic cellular maintenance (often called "housekeeping genes") and those genes whose function is specific to the particular tissue type that the cell belongs to. Gene expression by way of transcription is the process of generation of messenger RNA (mRNA) from the DNA template representing the gene. It is the intermediate step before the generation of functional protein from messenger RNA. During gene expression (Fig. 4.1), transcription factor (TF) proteins are recruited at the proximal promoter of the gene as well as at sequence elements (enhancers/silencers) which can lie several hundreds of kilobases from the gene's transcriptional start site (TSS). The basal transcriptional machinery at the promoter coupled with the transcription factor complexes at these distal, long-range regulatory elements (LREs) are collectively involved in directing tissue-specific expression of genes.

One of the current challenges in the post-genomic era is the principled discov-

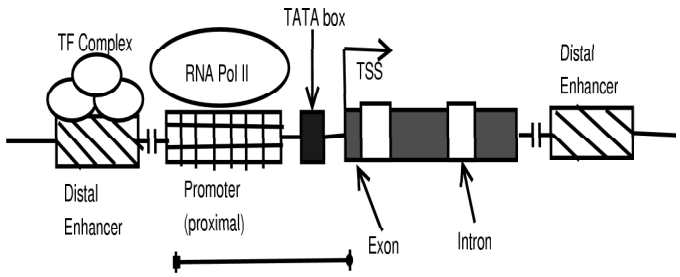


Figure 4.1: Schematic of Transcriptional Regulation. Sequence motifs at the promoter and the distal regulatory elements together confer specificity of gene expression via TF binding.

ery of such LREs genome-wide. Recently, there has been a community-wide effort (*http://www.genome.gov/ENCODE/*) to find all regulatory elements in 1% of the human genome. The examination of the discovered elements would reveal characteristics typical of most enhancers which would aid their principled discovery and examination on a genome-wide scale. Some characteristics of experimentally identified distal regulatory elements ([216],[209]) are:

- Non-coding elements: Distal regulatory elements are non-coding and can either be intronic or intergenic regions on the genome. Hence previous models for gene finding [177] are not directly applicable. With over 98% of the annotated genome being non-coding, the precise localization of regulatory elements that underlie tissue-specific gene expression is a challenging problem.

- Distance/orientation independent: an enhancer can act from variable genomic distances (hundreds of kilobases) to regulate gene expression in conjunction with the proximal promoter, possibly via a looping mechanism [163]. These enhancers can lie upstream or downstream of the actual gene along the genomic locus.

- Promoter dependent: Since the action at a distance of these elements involves the recruitment of TFs that direct tissue-specific gene expression, the promoter

that they interact with is critical.

Although there are instances where a gene harbors tissue-specific activity at the promoter itself, the role of long range elements (LREs) remains of interest, e.g: for a detailed understanding of their regulatory role in gene expression during biological processes like organ development and disease progression [206]. We seek to develop computational strategies to find novel LREs genome-wide that govern tissue specific expression for any gene of interest. A common approach for their discovery is the use of motif-based sequence signatures. Any sequence element can then be scanned for such a signature and its tissue-specificity can be ascertained [159].

Thus, our primary question in this regard is: is there a discriminating sequence property of LRE elements that determine tissue-specific gene expression - more particularly, are there any sequence motifs in known regulatory elements that can aid discovery of new elements [205]. To answer this, we examine known tissue-specific regulatory elements (promoters and enhancers) for motifs that discriminate them from a background set of neutral elements (such as housekeeping gene promoters). For this study, the datasets are derived from the following sources:

- Promoters of tissue-specific genes: Before the widespread discovery of long-range regulatory elements (LREs), it was hypothesized that promoters governed gene expression alone. There is substantial evidence for the binding of tissue-specific transcription factors at the promoters of expressed genes. This suggests that, in spite of newer information implicating the role of LREs, promoters also have interesting motifs that govern tissue-specific expression.

  Another practical reason for the examination of promoters is that their locations (and genomic sequences) are more clearly delineated on genome databases (like

UCSC or Ensembl). Sufficient data (*http://symatlas.gnf.org/*) on the expression of genes is also publicly available for analysis.

Sequence motif discovery is set up as a feature extraction problem from these tissue-specific promoter sequences. Subsequently, a support vector machine (SVM) classifier is used to classify new promoters into specific and non-specific categories based on the identified sequence features (motifs). Using the SVM classifier algorithm, 90% of tissue-specific genes are correctly classified based upon their upstream promoter region sequences alone.

- Known long range regulatory elements (LRE) motifs: To analyze the motifs in LRE elements, we examine the results of the above approach on the Enhancer Browser dataset (*http://enhancer.lbl.gov/*) which has results of expression of ultraconserved genomic elements in transgenic mice [227]. An examination of these ultraconserved enhancers is useful for the extraction of discriminatory motifs to distinguish the regulatory elements from the non-regulatory (neutral) ones. Here the results indicate that up to 95% of the sequences can be correctly classified using these identified motifs.

We note that some of the identified motifs might not be transcription factor binding motifs, and would need to be functionally characterized. This is an advantage of our method - instead of constraining ourselves to the degeneracy present in TF databases (like TRANSFAC/JASPAR), we look for all sequences of a fixed length.

4.2 Contributions

The use of microarray gene expression data ([143],[234]) suggests an approach to assign genes into tissue-specific and non-specific categories using an entropy criterion. Variation in expression and its divergence from ubiquitous expression (uniform

distribution across all tissue types) is used to make this assignment. Based on such assignment, several features like CpG island density, frequency of transcription factor motif occurrence, can be examined to potentially discriminate these two groups. Other work has explored the existence of key motifs (transcription factor binding sites) in the promoters of tissue-specific genes ([170],[172]). Based on the successes reported in these methods, it is expected that a principled examination and characterization of every sequence motif identified to be discriminatory might lead to improved insight into the biology of gene regulation. For example, such a strategy might lead to the discovery of newer TFBS motifs, as well as those underlying epigenetic phenomena.

For the purpose of identifying discriminative motifs from the training data (tissue-specific promoters or LREs), our approach is as follows:

- *Variable selection*: Firstly, sequence motifs that discriminate between tissue-specific and non-specific elements are discovered. In machine learning, this is a feature selection problem with features being the counts of sequence motifs in the training sequences. Without loss of generality, six-nucleotide motifs (hexamers) are used as motif features. This is based on the observation that most transcription factor binding motifs have a 5-6 nucleotide core sequence with degeneracy at the ends of the motif. A similar setup has been introduced in ([180], [201],[241]). The motif search space is, therefore a $4^6 = 4096$ dimensional one. The presented approach, however, does not depend on motif length and can be scaled according to biological knowledge.

  For variable (motif) selection, a novel feature selection approach (based on an information theoretic quantity called **directed information -** DI) is proposed. The improved performance of this criterion over using mutual information for

motif selection is also demonstrated.

- *Classifier design*: After discovering discriminating motifs using the above DI step, a SVM classifier that separates the samples between the two classes (specific and non-specific) from this motif space, is constructed.

Apart from this novel feature selection approach, several questions pertaining to bioinformatics methodology can be potentially answered using this framework. Some of these are:

- Are there common motifs underlying tissue-specific expression that are identified from tissue-specific promoters and enhancers?. In this chapter, an examination of motifs (from promoters and enhancers) corresponding to brain-specific expression is done to address this question.

- Do these motifs correspond to known motifs (transcription factor binding sites)?. We show that several motifs are indeed consensus sites for transcription factor binding, although their real role can only be identified through experimental evidence.

- Is it possible to relate the motif information from the sequence and expression perspectives to understand regulatory mechanisms?, This question is addressed in Section 4.11.C.

- How useful are these motifs in predicting new tissue-specific regulatory elements?. This is explained further in the results of SVM classification.

This work differs from that in ([180], [201]), in several aspects. We present the DI based feature selection procedure as part of an overall unified framework to answer several questions in bioinformatics, not limited to finding discriminating motifs between two classes of sequences. Particularly, one of the advantages is the ability to

examine any particular motif as a potential discriminator between two classes. Also, this work accounts for the notion of tissue-specificity of promoters/enhancers (in line with more recent work in [210],[227],[143], [234],[204]). Also, this framework enables the principled integration of various data sources to address the above questions. These are clarified in the Results (Section: 4.11).

4.3 Rationale

The main approaches to finding common motifs driving tissue-specific gene regulation are summarized in ([209], [216]). The most common approach is to look for TFBS motifs that are statistically over-represented in the promoters of the co-expressed genes based on a background (binomial or Poisson) distribution of motif occurrence genomewide.

In this work, the problem of motif discovery is set up as follows. Using two annotated groups of genes, tissue-specific (*'ts'*) and non-tissue specific (*'nts'*), hexamer motifs that best discriminate these two classes are found. The goal would be to make this set of motifs as small as possible - i.e. to achieve maximal class partitioning with the smallest feature subset.

Several metrics have been proposed to find features with maximal class label association. From information theory, mutual information is a popular choice [160]. This is a symmetric association metric and does not resolve the direction of dependency (i.e., if features depend on the class label or vice versa). It is important to find features that induce the class label. Feature selection from data implies selection (control) of a feature subset that maximally captures the underlying character (class label) of the data. There is no control over the label (a purely observational characterization).

With this motivation, a new metric for discriminative hexamer subset selection, termed "directed information" (DI), is proposed. Based on the selected features, a classifier is used to classify sequences to tissue-specific or non-tissue-specific categories. The performance of this DI based feature selection metric is subsequently evaluated in the context of the SVM classifier.

4.4 Overall Methodology

The overall schematic of the proposed procedure is outlined below (Fig. 4.2).

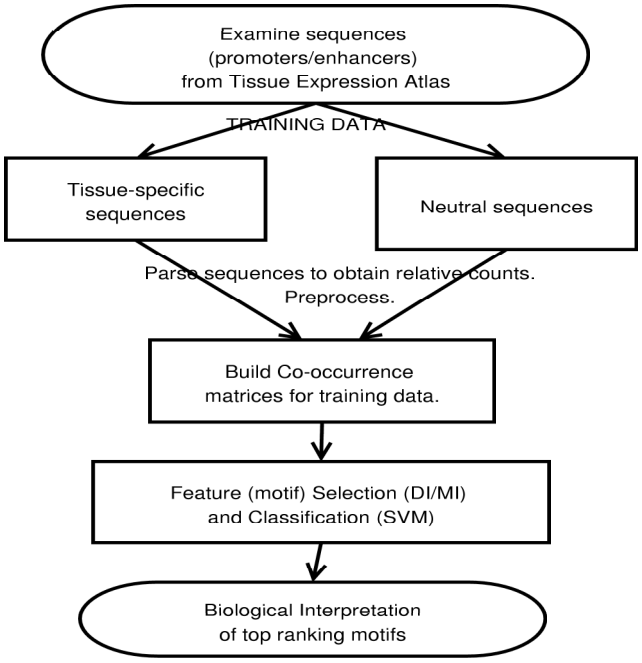


Figure 4.2: An overview of the proposed approach. Each of the steps are outlined in the following sections.

Below we present our approach to find promoter-specific or enhancer-specific motifs.

4.5 Motif Acquisition

4.5.1 Promoter motifs:

Microarray Analysis

Raw microarray data is available from the Novartis Foundation (GNF) [*http://symatlas.gnf.org/*]. Data is normalized using RMA from the Bioconductor packages for R [*cran.r-project.org/*]. Following normalization, replicate samples are averaged together. Only 25 tissue types are used in our analysis including: Adrenal Gland, Amygdala, Brain, Caudate Nucleus, Cerebellum, Corpus Callosum, Cortex, Dorsal Root Ganglion, Heart, HUVEC, Kidney, Liver, Lung, Pancreas, Pituitary, Placenta, Salivary, Spinal Cord, Spleen, Testis, Thalamus, Thymus, Thyroid, Trachea, and Uterus.

In this context, the notion of tissue-specificity of a gene needs clarification. Suppose there are N genes, $g_1, g_2, \ldots, g_N$ and T tissue types (in GNF: $T = 25$), we construct a $N \times T$ tissue specificity matrix : $M = [0]_{N \times T}$. For each gene $g_i, 1 \leq i \leq N$, let $g_{i,[0.5T]} = \text{median}(g_{i,k}), \forall k \in 1, 2, \ldots, T$; $g_{i,k}$ being the expression level of gene $'i'$ in tissue $'k'$. Define, each entry $M_{i,k}$ as,

$$M_{i,k} = \begin{cases} 1 & \text{if } g_{i,k} \geq 2g_{i,[0.5T]}; \\ 0 & \text{otherwise.} \end{cases}$$

Now consider the N dimensional vector $m_i = \sum_{k=1}^{T} M_{i,k}, 1 \leq i \leq N$ i.e. summing all the columns of each row. The inter-quartile range of $'m'$ can be used for *'ts'*/*'nts'* assignment. Gene indices $'i'$ that are in quartile 1 (=3), are labeled as *'ts'*, and those in quartile 4 (= 22), are labeled as *'nts'*.

With this approach, a total of 1924 probes representing 1817 genes were classified

as tissue-specific, while 2006 probes representing 2273 genes were classified as non tissue-specific. In this work, genes which are either heart-specific or brain-specific are considered. From the tissue-specific genes obtained from the above approach, 45 brain-specific gene promoters and 118 heart-specific gene promoters are obtained. As mentioned in Section II, one of the objectives is to find motifs that are responsible for brain/heart specific expression and also correlate them with binding profiles of known transcription factor binding motifs.

Sequence Analysis

Genes (*'ts'* or *'nts'*) associated with candidate probes are identified using the Ensembl Ensmart [*http://www.ensembl.org/*] tool. For each gene, sequence from 2000bp upstream and 1000bp down-stream upto the start of the first exon relative to their reported TSS is extracted from the Ensembl Genome Database (Release 37). The relative counts of each of the 4^6 hexamers are computed within each gene-promoter sequence of the two categories ('ts' and 'nts') - using the *'seqinr'* library in the R environment. A t-test is performed between the relative counts of each hexamer between the two expression categories (*'ts'* and *'nts'*) and the top 1000 significant hexamers ($\overrightarrow{\mathbf{H}} = H_1, H_2, \ldots, H_{1000}$) is obtained. The relative counts of these hexamers is recomputed for each gene individually. This results in two hexamer-gene co-occurrence matrices, - one for the *'ts'* class (dimension $N_{train,+1} \times 1000$) and the other for the *'nts'* class (dimension $N_{train,-1} \times 1000$). Here $N_{train,+1}$ and $N_{train,-1}$ are the number of positive training and negative training samples, respectively.

The input to the feature selection procedure is a gene promoter - motif frequency table (Table 5.1). The genes relevant to each class are identified from tissue microarray analysis, following steps 1 and 2 above, and the frequency table is built by parsing the gene promoters for the presence of each of the $4^6 = 4096$ possible

hexamers.

Ensembl Gene ID	AAAAAA	AAAAAG	AAAAAT	AAAACA
ENSG00000155366	0	0	1	4
ENSG000001780892	6	5	5	6
ENSG00000189171	1	2	1	0
ENSG00000168664	6	3	8	0
ENSG00000160917	4	1	4	2
ENSG00000163655	2	4	0	1
ENSG000001228844	8	6	10	7
ENSG00000176749	0	0	0	0
ENSG00000006451	5	2	2	1

Table 4.1: The 'motif frequency matrix' for a set of gene-promoters. The first column is their ENSEMBL gene identifiers and the other 4 columns are the motifs. A cell entry denotes the number of times a given motif occurs in the upstream (-2000 to +1000bp from TSS) region of each corresponding gene.

4.5.2 LRE motifs:

To analyze long range elements which confer tissue specific expression, the Mouse Enhancer database (*http://enhancer.lbl.gov/*) is examined. This database has a list of experimentally validated ultraconserved elements which have been tested for tissue specific expression in transgenic mice [227], and can be searched for a list of all elements which have expression in a tissue of interest. In this work, we consider expression in tissues relating to the developing brain. According to the experimental protocol, the various regions are cloned upstream of a heat shock protein promoter (*hsp68-lacz*), thereby not adhering to the idea of promoter specificity in tissue-specific expression. Though this is of concern in that there is loss of some gene-specific information, we work with this data since we are more interested in tissue expression and also due to a paucity of public promoter-dependent enhancer data .

This database also has a collection of ultraconserved elements that do not have any transgenic expression in-vivo (caveat: in the context of the *wrong* promoter). This is used as the neutral/background set of data which corresponds to the *'nts'* (non-tissue specific class) for feature selection and classifier design.

As in the above (promoter) case, these sequences (seventy four enhancers for brain-specific expression) are parsed for the absolute counts of the 4096 hexamers, a co-occurrence matrix ($N_{train,+1} = 74$) is built and then t-test $p-values$ are used to find the top 1000 hexamers ($\overrightarrow{\mathbf{H'}} = H'_1, H'_2, \ldots, H'_{1000}$) that are maximally different between the two classes (brain-specific and brain-non-specific).

The next three sections clarify the preprocessing, feature selection and classifier design steps to mine these co-occurrence matrices for hexamer motifs that are strongly associated with the class label. Though this work is illustrated using two class labels, the approach can be extended in a straightforward way to the multi-class problem.

4.6 Preprocessing

From the above, $N_{train,+1} \times 1000$ and $N_{train,-1} \times 1000$ dimensional co-occurrence matrices are available for the tissue-specific and non-specific data, both for the promoter and enhancer sequences. Before proceeding to the feature (hexamer motif) selection step, the counts of the $M = 1000$ hexamers in each training sample need to be normalized to account for variable sequence lengths. In the co-occurrence matrix, let $gc_{i,k}$ represent the absolute count of the k^{th} hexamer, $k \in 1, 2, \ldots, M$ in the i^{th} gene. Then, for each gene g_i, the quantile labeled matrix has $X_{i,k} = l$ if $gc_{i,[\frac{l-1}{K}M]} \leq gc_{i,k} < gc_{i,[\frac{l}{K}M]}, K = 4$. Matrices of dimension $N_{train,+1} \times 1001$, $N_{train,-1} \times 1001$ for the specific and non-specific training samples are now obtained. Each matrix contains the quantile label assignments for the 1000 hexamers ($X_i, i \in (1, 2, \ldots, 1000)$), as stated above, and the last column has the corresponding class label ($Y = -1/+1$).

4.7 Directed Information and Feature Selection

The primary goal in feature selection is to find the minimal subset of features (from hexamers: $\overrightarrow{\mathbf{H}}/\overrightarrow{\mathbf{H'}}$) that lead to maximal discrimination of the class label ($Y_i \in (-1/+1)$), using each of the $i \in (1, 2, \ldots, (N_{train,+1} + N_{train,-1}))$ genes during training. We are looking for a subset of the variables $(H_{i,1}, \ldots, H_{i,1000})$ which are directionally associated with the class label (Y_i). These hexamers putatively influence/induce the class label (Fig. 4.3). As can be seen from [162], there is considerable interest in discovering such dependencies from expression and sequence data. Following [139], we search for features (in *measurement* space) that induce the class label (in *observation* space).

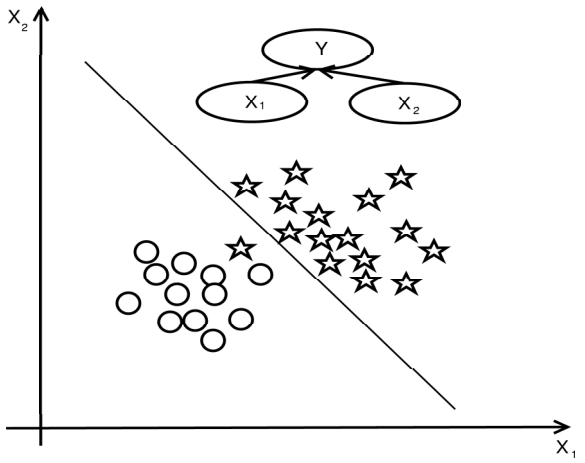


Figure 4.3: Causal Feature discovery for two class discrimination, adapted from [139]. Here the variables X_1 and X_2 discriminate Y, the class label.

One way to interpret the feature selection problem is the following: Nature is trying to communicate a source symbol ($Y \in \{-1/+1\}$), corresponding to the gene class label (*'nts/ts'*), to us. In this setup, an encoder that extracts frequencies of a particular hexamer (H_i) maps the source symbol (Y) to $H_i(Y)$. The decoder outputs the source reconstruction $\hat{Y}$ based on the received codeword $c_i(Y) = H_i(Y)$.

We observe that there are several possible encoding schemes $c_i(Y)$ that the encoder

could potentially use ($i = 1, 2, \ldots, 1000$), each corresponding to feature extraction via a different hexamer H_i. An encoder is the mapping rule $c_i : Y \rightarrow H_i$. The ideal encoding scheme is one which induces the most discriminative partitioning of the code (feature) space, for successful reconstruction of Y by the decoder. The ranking of each encoder's performance over all possible mappings yields the most discriminative mapping. This measure of performance is the amount of information flow from the mapping (hexamer) to the class label. Using mutual information as one such measure indeed identifies the best features [160], but fails to resolve the direction of dependence due to its symmetric nature $I(H_i; Y) = I(Y; H_i)$. The direction of dependence is important since it pinpoints those features that induce the class label (not vice-versa). This is necessary since these class labels are predetermined (given to us by biology) and the only control we have is the feature space onto which we project the data points, for the purpose of classification. This loosely parallels the use the directed edges in bayesian networks for inference of feature-class label associations [139].

Unlike mutual information (MI), directed information (DI) is a metric to quantify the directed flow of information. It was originally introduced in ([218], [219]) to examine the transfer of information from encoder to decoder under feedback/feedforward scenarios and to resolve directivity during bidirectional information transfer. Given its utility in the encoding of sources with memory (correlated sources), this work demonstrates it to be a competitive metric to MI for feature selection in learning problems. DI answers which of the encoding schemes (corresponding to each hexamer H_i) leads to maximal information transfer from the hexamer labels to the class labels (i.e. directed dependency).

The DI is a measure of the directed dependence between two vectors $X_i =$

$[X_{1,i}, X_{2,i}, \ldots X_{n,i}]$ and $Y = [Y_1, Y_2, \ldots, Y_n]$. In this case, $X_{j,i}$ = quantile label for the frequency of hexamer $i \in (1, 2, \ldots, 1000)$ in the j^{th} training sequence. $Y = [Y_1, Y_2, \ldots, Y_n]$ are the corresponding class labels $(-1, +1)$. For a block length N, the DI is given by [219]:

$$I(X_i^N \to Y^N) = \sum_{n=1}^{N} I(X_i^n; Y_n | Y^{n-1}) \tag{1}$$

Using a stationarity assumption over a finite-length memory of the training samples, a correspondence with the setup in ([219], [169]) can be seen. As already known [135], the mutual information $I(X^N; Y^N) = H(X^N) - H(X^N|Y^N)$, where $H(X^N)$ and $H(X^N|Y^N)$ are the Shannon entropy of X^N and the conditional entropy of X^N given Y^N, respectively. With this definition of mutual information, the Directed Information simplifies to,

$$\begin{aligned} I(X^N \to Y^N) &= \sum_{n=1}^{N} [H(X^n|Y^{n-1}) - H(X^n|Y^n)] \\ &= \sum_{n=1}^{N} \{[H(X^n, Y^{n-1}) - H(Y^{n-1})] - [H(X^n, Y^n) - H(Y^n)]\} \end{aligned} \tag{2}$$

Using (2), the Directed information is expressed in terms of individual and joint entropies of X^n and Y^n. This expression implies the need for higher-order entropy estimation from a moderate sample size. A Voronoi tessellation [189] based adaptive partitioning of the observation space can handle $N = 5/6$ without much complexity.

The relationship between MI and DI is given by [219],

DI: $I(X^N \to Y^N) = \sum_{i=1}^{N} I(X^i; Y_i | Y^{i-1})$.

MI: $I(X^N; Y^N) = \sum_{i=1}^{N} I(X^N; Y_i | Y^{i-1})$

$= I(X^N \to Y^N) + I(0Y^{N-1} \to X^N)$.

To clarify, $I(X^N \to Y^N)$ is the directed information from X to Y, whereas

$I(0Y^{N-1} \rightarrow X^N)$ is the directed information from a (one-sample) delayed version of Y^N to X^N. From [169], it is clear that DI resolves the direction of information transfer (feedback or feedforward). If there is no feedback/feedforward, $I(X^N \rightarrow Y^N) = I(X^N; Y^N)$.

From the above chain-rule formulations for DI and MI, it is clear that the expression for DI is permutation-variant (i.e., the value of the DI is different for a different ordering of random variables). Thus, we instead find the $I_p(X^N \rightarrow Y^N)$, a DI measure for a particular ordering of the N random variables (r.vs). The DI value for our purpose, $I(X^N \rightarrow Y^N)$ is an average over all possible sample permutations given by, $I(X^N \rightarrow Y^N) = \frac{1}{N!} \sum_{p=1}^{N!} I_p(X^N \rightarrow Y^N)$. For MI, however, $I_p(X^N; Y^N) = I(X^N; Y^N)$ because, MI is permutation-invariant (i.e., independent of r.v ordering). As can be readily observed, this problem is combinatorially complex, and hence, a monte-carlo sampling strategy (1000 trials) is used for computing $I(X^N \rightarrow Y^N)$. This is because we find that about 1000 trials yields a DI confidence interval (CI) that is only 20% more than the corresponding CI obtained from $10,000$ trials of the data, a far more exhaustive number.

To select features, we maximize $I(X^N \rightarrow Y^N)$ over the possible pairs $(\overrightarrow{\mathbf{X}}, Y)$. This feature selection problem for the i^{th} training instance reduces to identifying which hexamer ($k \in (1, 2, \ldots, 4096)$) has the highest $I(X_k \rightarrow Y)$.

The higher dimensional entropy can be estimated using order statistics of the observed samples [189] by iterative partitioning of the observation space until nearly uniform partitions are obtained. This method lends itself to a partitioning scheme that can be used for entropy estimation even for a moderate number of samples in the observation space of the underlying probability distribution. Several such algorithms for adaptive density estimation have been proposed ([243],[221],[225]) and can find

potential application in this procedure. In this methodology, a Voronoi tessellation approach for entropy estimation because of the higher performance guarantees as well as the relative ease of implementation of such a procedure.

The above method is used to estimate the true DI between a given hexamer and the class label for the entire training set. Feature selection comprises of finding all those hexamers (X_i) for which $I(X_i^N \to Y^N)$ is the highest. From the definition of DI, we know that $0 \leq I(X_i^N \to Y^N) \leq I(X_i^N; Y^N) < \infty$. To make a meaningful comparison of the strengths of association between different hexamers and the class label, we use a normalized score to rank the DI values. This normalized measure ρ_{DI} should be able to map this large range ($[0, \infty]$) to $[0, 1]$. Following [203], an expression for the normalized DI is given by:
$\rho_{DI} = \sqrt{1 - e^{-2I(X^N \to Y^N)}} = \sqrt{1 - e^{-2\sum_{i=1}^{N} I(X^i; Y_i | Y^{i-1})}}$.

Another point of consideration is to estimate the significance of the DI value compared to a null distribution on the DI value (i.e. what is the chance of finding the DI value by chance from the N-length series X_i and Y). This is done using confidence intervals after permutation testing (Sec: $VIII$).

4.8 Bootstrapped Confidence Intervals

In the absence of knowledge of the true distribution of the DI estimate, an approximate confidence interval for the DI estimate ($\hat{I}(X^N \to Y^N)$), is found using bootstrapping [186]. Density estimation is based on kernel smoothing over the bootstrapped samples [229].

The kernel density estimate for the bootstrapped DI (with $n = 1000$ samples), $Z \triangleq \hat{I}_B(X^N \to Y^N)$ becomes,
$\hat{f}_h(Z) = \frac{1}{nh} \sum_{i=1}^{n} \frac{3}{4}[1 - (\frac{z_i - z}{h})^2] I(\left|\frac{z_i - z}{h}\right| \leq 1)$ with $h \approx 0.267\hat{\sigma}_z$ and $n = 1000$.

$\hat{I}_B(X^N \to Y^N)$ is obtained by finding the DI for each random permutation of the X, Y series, and performing this permutation B times. As is the clear from the above expression, the Epanechnikov kernel is used for density estimation from the bootstrapped samples. The choice of the kernel is based on its excellent characteristics - a compact region of support, the lowest AMISE (asymptotic mean squared error) and favorable bias-variance tradeoff [229].

We denote the cumulative distribution function (over the bootstrap samples) of $\hat{I}(X^N \to Y^N)$ by $F_{\hat{I_B}(X^N \to Y^N)}(\hat{I_B}(X^N \to Y^N))$. Let the mean of the bootstrapped null distribution be $I^*_B(X^N \to Y^N)$. We denote by $t_{1-\alpha}$, the $(1-\alpha)^{th}$ quantile of this distribution i.e. $\{t_{1-\alpha} : P([\frac{\hat{I_B}(X^N \to Y^N) - I^*_B(X^N \to Y^N)}{\hat{\sigma}}] \leq t_{1-\alpha}) = 1-\alpha\}$. Since we need the true $\hat{I}(X^N \to Y^N)$ to be significant and close to 1, we need $\hat{I}(X^N \to Y^N) \geq [I^*_B(X^N \to Y^N) + t_{1-\alpha} \times \hat{\sigma}]$, with $\hat{\sigma}$ being the standard error of the bootstrapped distribution,
$\hat{\sigma} = \sqrt{\frac{[\Sigma_{b=1}^{B} \hat{I}_b(X^N \to Y^N) - I^*_B(X^N \to Y^N)]^2}{B-1}}$; B is the number of bootstrap samples.

This hypothesis test is done for each of the 1000 motifs, in order to select the top $'d'$ motifs based on DI value, which is then used for classifier training subsequently. This leads to a need for multiple-testing correction. Because the Bonferroni correction is extremely stringent in such settings, the Benjamini-Hochberg procedure [132], which has a higher false positive rate, but a lower false negative rate is used in this work.

4.9 Support Vector Machines

From the top $'d'$ features identified from the ranked list of features having high DI with the class label, a support vector machine classifier in these $'d'$ dimensions is designed. A SVM is a hyperplane classifier which operates by finding a maximum margin linear hyperplane to separate two different classes of data in high

dimensional ($D > d$) space. The training data has $N(= N_{train,+1} + N_{train,-1})$ pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$, with $x_i \in \mathscr{R}^d$ and $y_i \in \{-1, +1\}$.

An SVM is a maximum margin hyperplane classifier in a non-linearly extended high dimensional space. For extending the dimensions from d to $D > d$, a radial basis kernel is used.

The objective is to minimize $||\beta||$ in the hyperplane $\{x : f(x) = x^T\beta + \beta_0\}$, subject to

$y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i \forall i$, $\xi_i \geq 0, \sum \xi_i \leq$ constant [199].

4.10 Summary of Overall Approach

Our proposed approach is as follows. Here, the term 'sequence' can pertain to either tissue-specific promoters or LRE sequences, obtained from the GNF SymAtlas and Ensembl databases or the Enhancer Browser.

1. The sequence is parsed to obtain the relative counts/frequencies of occurrence of the hexamer in that sequence and to build the hexamer-sequence frequency matrix. The *'seqinr'* package in R is used for this purpose. This is done for all the sequences in the specific (class "$+1$") and non-specific (class "-1") categories. The matrix thus has $N = N_{train,+1} + N_{train,-1}$ rows and $4^6 = 4096$ columns.

2. The obtained hexamer-sequence frequency matrix is preprocessed by assigning quantile labels for each hexamer within the i^{th} sequence. A hexamer-sequence matrix is thus obtained where the $(i, j)^{th}$ entry has the quantile label of the j^{th} hexamer in the i^{th} sequence. This is done for all the N training sequences consisting of examples from the -1 and $+1$ class labels.

3. Thus, two submatrices corresponding to the two class labels are built. One

matrix contains the hexamer-sequence quantile labels for the positive training examples and the other matrix is for the negative training examples.

4. To select hexamers that are most different between the positive and negative training examples, a t-test is performed for each hexamer, between the *'ts'* and *'nts'* groups. Ranking the corresponding t-test p-values yields those hexamers that are most different distributionally between the positive and negative training samples. The top 1000 of these hexamers are chosen for further analysis. This step is only necessary to reduce the computational complexity of the overall procedure - computing the DI between each of the 4096 hexamers and the class label is relatively expensive.

5. For the top $K = 1000$ hexamers which are most significantly different between the positive and negative training examples, $I(X_k^N \to Y^N)$ and $I(X_k^N; Y^N)$ reveals the degree of association for each of the $k \in (1, 2, \ldots, K)$ hexamers. The entropy terms in the directed information and mutual information expressions are found using a higher-order entropy estimator. Using the procedure of Section: 4.7, the raw DI values are converted into their normalized versions. Since the goal is to maximize $I(X_k \to Y)$, we can rank the DI values in descending order.

6. The significance of the DI estimate is obtained based on the bootstrapping methodology. For every hexamer, a $p = 0.05$ significance with respect to its bootstrapped null distribution yields potentially discriminative hexamers between the two classes. The Benjamini-Hochberg procedure is used for multiple-testing correction. Ranking the significant hexamers by decreasing DI value yields features that can be used for classifier (SVM) training.

7. Train the Support Vector Machine classifier (SVM) on the top $'d'$ features from the ranked DI list(s). For comparison with the MI based technique, we use the hexamers which have the top $'d'$ (normalized) MI values. The accuracy of the trained classifier is plotted as a function of the number of features (d), after ten-fold cross-validation. As we gradually consider higher $'d'$, we move down the ranked list. In the plots below, the misclassification fraction is reported instead. A fraction of 0.1 corresponds to 10% misclassification.

Note: An important point concerns the training of the SVM classifier with the top $'d'$ features selected using DI or MI (step 7 above). Since the feature selection step is decoupled from the classification step, it is preferred that the top $'d'$ motifs are consistently ranked high among multiple draws of the data, so as to warrant their inclusion in the classifier. However, this does not yield expected results on this data set. Briefly, a Kendall rank correlation coefficient [144] was computed between the rankings of the motifs between multiple data draws (by sampling a subset of the entire dataset), for both MI and DI based feature-selection. It is observed that this coefficient is very low in both MI and DI, indicating a highly variable ranking. This is likely due to the high variability in data distribution across these multiple draws (due to limited number of data points), as well as the sensitivity of the data-dependent entropy estimation procedure to the range of the samples in the draw. To circumvent this problem of inconsistency in rank of motifs, a *median* DI/MI value is computed across these various draws and the top $'d'$ features based on the median DI/MI value across these draws are picked for SVM training [139].

4.11 Results

4.11.1 Tissue specific promoters

We use DI to find hexamers that discriminate brain-specific and heart-specific expression from neutral sequences. The negative training sets are sequences that are not brain or heart-specific, respectively. Results using the MI and DI methods are given below (Figs. 4.5 and 4.7). The plots indicate the SVM cross-validated misclassification accuracy (ideally 0) for the data as the number of features using the metric (DI or MI) is gradually increased. We can see that for any given classification accuracy, the number of features using DI is less than the corresponding number of features using MI. This translates into a lower misclassification rate for DI-based feature selection. We also observe that as the number of features $'d'$ is increased the performance of MI is the same as DI. This is expected since, as we gather more features using MI or DI, the differences in MI vs. DI ranking are compensated.

An important point needs to be clarified here. There is a possibility of sequence composition bias in the tissue-specific and neutral sequences used during training. This has been reported in recent work [241]. To avoid detecting GC rich sequences as hexamer features, it is necessary to confirm that there is no significant GC-composition bias between the specific and neutral sets in each of the case studies. This is demonstrated in Figs. 4.4, 4.6 and 4.8. In each case, it is observed that the mean GC-composition is almost same for the specific vs. neutral set. However, in such studies, it is necessary to select for sequences that do not exhibit such bias. In Figs. 4.6 and 4.8, even the distribution of GC-composition is similar among the samples. For Fig. 4.4, even though the distributions are slightly different, the box plots indicate similarity in mean GC-content.

Next, some of the motifs that discriminate between tissue-specific and non-specific

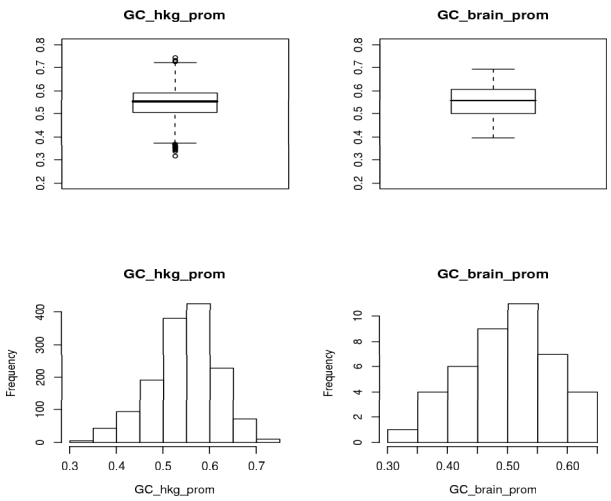


Figure 4.4: GC sequence composition for brain-specific promoters and housekeeping (hkg) promoters.

categories for the brain promoter, heart promoter and brain enhancer cases respectively are listed in Table II. Additionally, if the genes encoding for these TFs are expressed in the corresponding tissue [155], a (*) sign is appended. In some cases, the hexamer motifs match the consensus sequences of known transcription factors (TF). This suggests a potential role for that particular TF in regulating expression of tissue-specific genes. This matching of hexamer motifs with TFBS consensus sites is done using the MAPPER engine (*http://bio.chip.org/mapper/*). A hexamer-TFBS match does not necessarily imply the functional role of the TF in the corresponding tissue (brain or heart). However, such information would be useful to guide focused experiments to confirm their role in-vivo (using techniques such as chromatin immunoprecipitation).

As is clear from the above results, there are several other motifs which are novel or correspond to non-consensus motifs of known transcription factors. Hence, each of the identified hexamers merit experimental investigation. Also, though we identify as many as 200 hexamers in this work (please see Supplementary data), we have

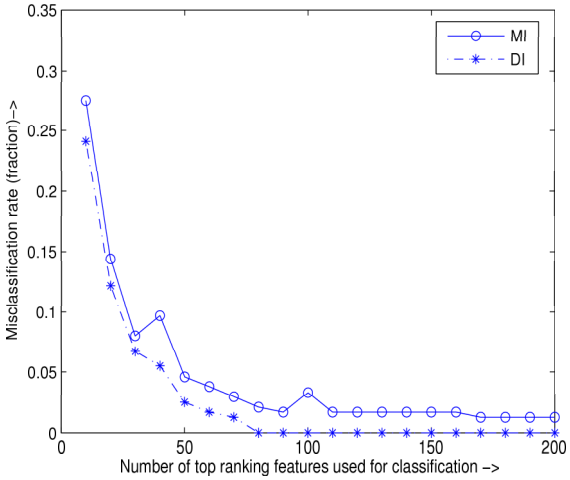


Figure 4.5: Misclassification accuracy for the MI vs. DI case (brain promoter set). Accuracy of classification is ~ 0.9 i.e. 93%.

Brain promoters	Heart promoters	Brain enhancers
Ahr-ARNT (*)	Pax2	HNF-4 (*)
Tcf11-MafG (*)	Tcf11-MafG (*)	Nkx2
c-ETS (*)	XBP1 (*)	AML1
FREAC-4	Sox-17 (*)	c-ETS (*)
T3R-alpha1	FREAC-4	Elk1 (*)
	GATA(*)	

Table 4.2: Comparison of high ranking motifs (by DI) across different data sets. The (*) sign indicates tissue-specific expression of the corresponding TF gene.

reported only a few due to space constraints.

In the context of the heart-specific genes, we consider the cardiac troponin gene (*cTNT*, ENSEMBL:ENSG00000118194), which is present in the heart promoter set. An examination of the high DI motifs for the heart-specific set yields motifs with the GATA consensus site, as well as matches with the MEF2 transcription factor. It has been established earlier that GATA-4, MEF2 are indeed involved in transcriptional activation of this gene [154] and the results have been confirmed by ChIP [131].

4.11.2 Enhancer DB

Additionally, all the brain-specific regulatory elements profiled in the mouse Enhancer Browser database (*http://enhancer.lbl.gov/*), were examined for discriminat-

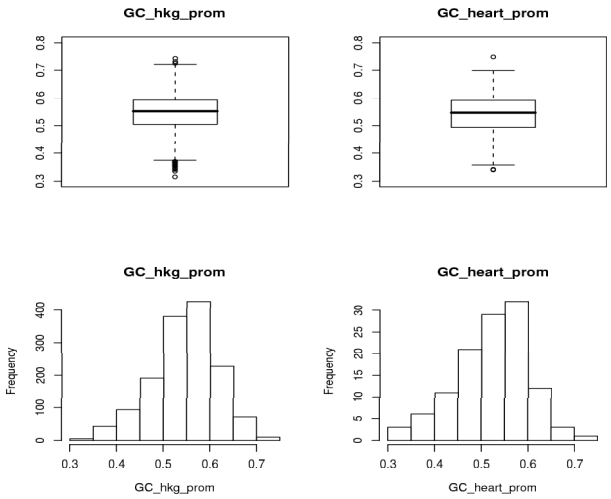


Figure 4.6: GC sequence composition for heart-specific promoters and housekeeping (hkg) promoters.

ing motifs. Fig. 4.8 shows that the two classes have similar GC-composition. Again, the plot of misclassification accuracy vs. number of features in the MI and DI scenarios reveal the superior performance of the DI-based hexamer selection compared to MI (Fig. 4.9).

In this case, the enhancer sequences are ultraconserved, thus obtained after alignment across multiple species. The examination of these sequences identified motifs that are potentially selected for regulatory function across evolutionary distances. Using alignment as a prefiltering strategy helps remove bias conferred by sequence elements that arise via random mutation but might be over-represented. This is permitted in programs like Toucan [172] and rVISTA (*http://rvista.dcode.org/*).

As in the previous case, some of the top ranking motifs from this dataset are also shown in Table *II*. The (*) signed TFs indicate that some of these discovered motifs indeed have documented high expression in the brain. The occurrence of such tissue-specific transcription factor motifs in these regulatory elements gives credence to the discovered motifs. For example, *ELK-1* is involved in neuronal differentiation [168].

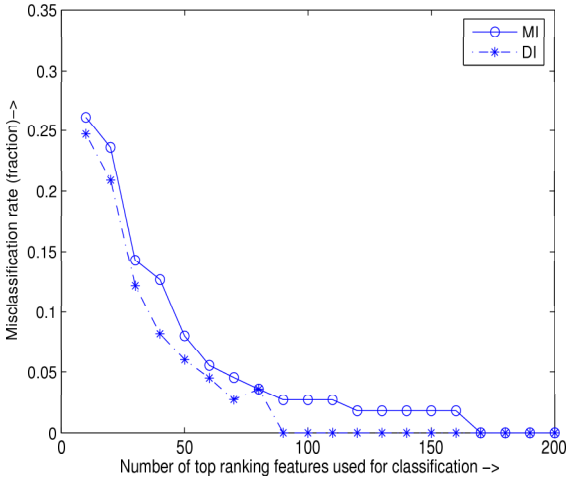


Figure 4.7: Misclassification accuracy for the MI vs. DI case (heart promoter set)

Also, some motifs matching consensus sites of TEF1 and ETS1 are common to the brain-enhancer and brain-promoter set. Though this is interesting, an experiment to confirm the enrichment of such transcription factors in the population of brain-specific regulatory sequences is necessary.

4.11.3 Quantifying *sequence-based* TF influence

A very interesting question emerges from the above presented results. What if one is interested in a motif that is not present in the above ranked hexamer list for a particular tissue-specific set? As an example, consider the case for *MyoD*, a transcription factor which is expressed in muscle and has a putative activity in heart-specific genes [157]. In fact, a variant of its consensus motif - CATTTG is indeed in the top ranking hexamer list. The DI based framework further permits investigation of the directional association of the canonical *MyoD* motif (CACCTG) for the discrimination of heart-specific genes vs. housekeeping genes. This is shown in Fig. 4.10. As is observed, *MyoD* has a significant directional influence on the heart-specific vs. neutral sequence class label. This, in conjunction with the expres-

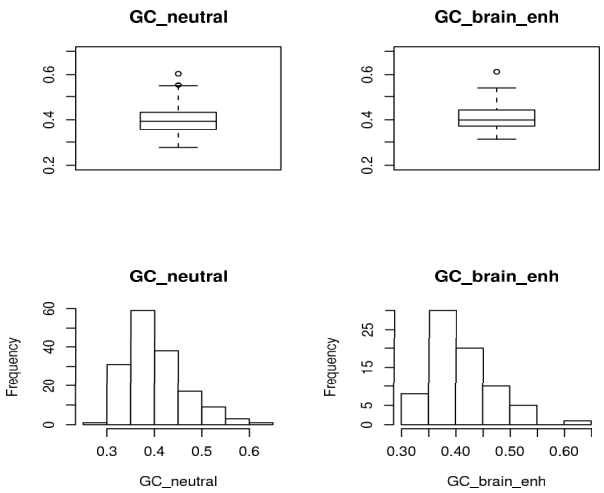


Figure 4.8: GC sequence composition for brain-specific enhancers and neutral non-coding regions.

sion level characteristics of *MyoD*, indicates that the motif CACCTG is potentially relevant to make the distinction between heart-specific and neutral sequences.

Another theme picks up on something quite traditionally done in bioinformatics research - finding key TF regulators underlying tissue-specific expression. Two major questions emerge from this theme.

1. Which putative regulatory TFs underlie the tissue-specific expression of a group of genes?

2. For the TFs found using tools like TOUCAN [172], can we examine the degree of influence that the particular TF motif has in directing tissue-specific expression?

- To address the first question, we examine the TFs revealed by DI/MI motif selection and compare these to the TFs discovered from TOUCAN [172], underlying the expression of genes expressed on day $e14.5$ in the degenerating mesonephros and nephric duct (TS22). This set has about 43 genes (including *Gata2*). These genes are available in the Supplementary data.

  Using TOUCAN, the set of module TFs are combinations of: *E47*, *HNF3B*,

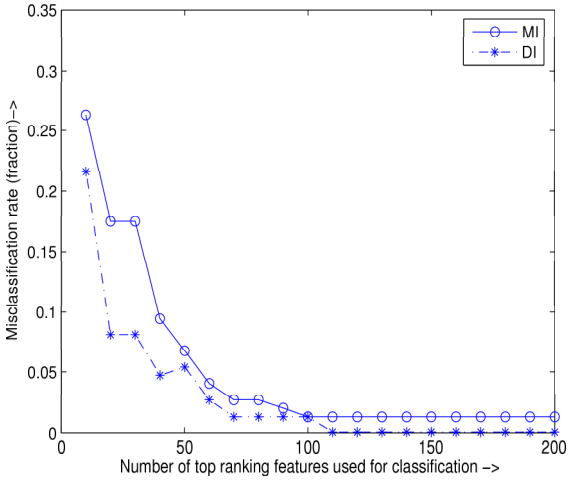


Figure 4.9: Misclassification accuracy for the MI vs. DI case (brain enhancer set).

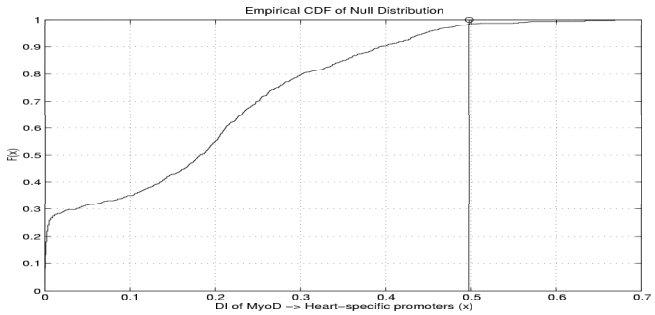


Figure 4.10: Cumulative Distribution Function for bootstrapped $I(MyoD\ motif{:}CACCTG \rightarrow Y)$; Y is the class label (Heart-specific vs. Housekeeping). True $\hat{I}(CACCTG \rightarrow Y) = 0.4977$.

HNF1, *RREB1*, *HFH3*, *CREBP1*, *VMYB*, *GFI1*. These were obtained by aligning the promoters of these 43 genes (−2000bp upstream to +200bp from the TSS), and looking for over-represented TF motifs based on the TRANSFAC/JASPAR databases.

Using the DI based motif selection, a set of 200 hexamers are found that discriminate these 43 gene promoter sequences from the background housekeeping promoter set. They map to the consensus sites of several known TFs, such as (identified from *bio.chip.org/mapper/*) *Nkx*, *Max1*, *c-ETS*, *FREAC4*, *Ahr-ARNT*, *CREBP2*, *E2F*, *HNF3A/B*, *NFATc*, *Pax2*, *LEF1*, *Max1*, *SP1*, *Tef1*, *Tcf11-*

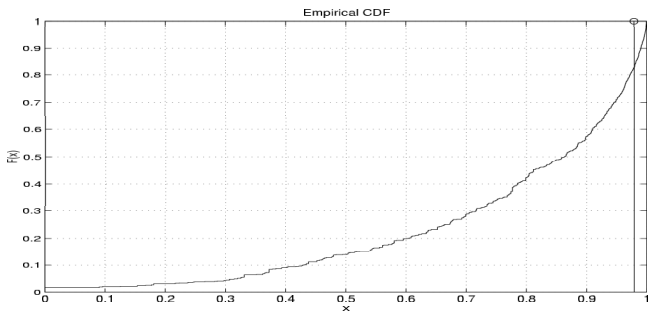


Figure 4.11: Cumulative Distribution Function for bootstrapped $I(Pax2\ motif{:}GTTCC \rightarrow Y)$; Y is the class label (UB/non-UB). True $\hat{I}(GTTCC \rightarrow Y) = 0.9792$.

MafG, many of which are expressed in the developing kidney (*http://www.expasy.org/*). Moreover, we observe that the TFs that are common between the TOUCAN results and the DI based approach: *FREAC4*, *Max1*, *HNF3a/b*, *HNF1*, *SP1*, *CREBP*, *RREB1*, *HFH3* are mostly kidney-specific. Thus, we believe that this observation makes a case for finding all (possibly degenerate) TF motif searches from TRANSFAC, and filtering them based on tissue-specific expression subsequently. Such a strategy yields several more TF candidates for testing and validation of biological function.

- For the second question, we examine the following scenario. The *Gata3* gene is observed to be expressed in the developing ureteric bud (UB) during kidney development. To find UB specific TF regulators, conserved TF modules can be examined in the promoters of UB-specific genes. These experimentally annotated UB-specific genes are obtained from the Mouse Genome Informatics database at *http://www.informatics.jax.org/*. Several programs are used for such analysis, like Genomatix [170] or Toucan [172]. Using Toucan, the promoters of the various UB specific genes are aligned to discover related modules. The top-ranking module in Toucan contains *AHR-ARNT*, *Hox13*, *Pax2*, *Tal1alpha-E47*, *Oct1*. Again, the power of these motifs to discriminate UB-specific and

non-specific genes, based on DI, can be investigated.

For this purpose, we check if the *Pax2* binding motif (GTTCC [184]) indeed induces kidney specific expression by looking for the strength of DI between the GTTCC motif and the class label (+1) indicating UB expression (Fig. 4.11). This once again adds to computational evidence for the true role of *Pax2* in directing ureteric bud specific expression [184]. The main implication here is that, from sequence data, there is strong evidence for the *Pax2* motif being a useful feature for UB-specific genes. This is especially relevant given the documented role of *Pax2* ([138]) directing ureteric-bud expression of the *Gata3* gene, one of the key modulators of kidney morphogenesis. Both the *MyoD* and *Pax2* studies indicate the relevance of principled data integration using expression ([230],[155]) and sequence modalities.

4.11.4 Observations

With regard to the feature selection and classification results, in both studies (enhancers and promoters), we observe that about 100 hexamers are enough to discriminate the tissue-specific from the neutral sequences. Furthermore, some sequence features of these motifs at the promoter/enhancer emerge.

- There is higher sequence variability at the promoter since it has to act in concert with LREs of different tissue types during gene regulation.
- Since the enhancer/LRE acts with the promoter to confer expression in only one tissue type, these sequences are more specific and hence their mining identifies motifs that are probably more indicative of tissue-specific expression.

We however, reiterate that the enhancer dataset that we study uses the *hsp68-lacz* as the promoter driven by the ultraconserved elements. Hence there is no promoter

specificity in this context. Though this is a disadvantage and might not reveal all key motifs, it is the best that can be done in the absence of any other comprehensive repository.

The second aspect of the presented results highlight two important points. Firstly, the identified motifs have a strong predictive value as suggested by the cross-validation results as well as Table II. Moreover, DI provides a principled methodology to investigate any given motif for tissue-specificity as well as for identifying expression-level relationships between the TFs and their target genes, (Section 4.11.3).

4.12 Conclusions

In this work, a framework for the identification of hexamer motifs to discriminate between two kinds of sequences (tissue-specific promoters or regulatory elements vs non-specific elements), is presented. For this feature selection problem, a new metric - the 'directed information' (DI) is proposed. In conjunction with a support vector machine classifier, this method was shown to outperform the state-of-the-art method employing undirected mutual information. We also find that only a subset of the discriminating motifs correlate with known transcription factor motifs and hence the other motifs might be potentially related to non-consensus TF binding or underlying epigenetic phenomena governing tissue-specific gene expression. The superior performance of the directed-information based variable selection suggests its utility to more general learning problems. As per the initial motivation, the discovery of these motifs can aid in the prospective discovery of other tissue-specific regulatory regions.

We have also examined the applicability of DI to prospectively resolve the functional role of any TF motif in a biological process, integrating other sources (litera-

ture, expression data, module searches).

4.13 Future Work

Several opportunities for future work exist within this proposed framework. Multiple sequence alignment of promoter/regulatory sequences across species would be a useful preprocessing step to reduce false detection of discriminatory motifs. The hexamers can also be identified based on other metrics exploiting distributional divergence between the samples of the "$+1$" and "-1" classes. Furthermore, there is a need for consistent high-dimensional entropy estimators within the small sample regime. A very interesting direction of potential interest is the formulation of a stepwise hexamer selection algorithm, using the directed information for maximal relevance selection and mutual information for minimizing between-hexamer redundancy [160].

4.14 Acknowledgements

The authors gratefully acknowledge the support of the NIH under award 5R01-GM028896-21 (J.D.E). We would like to thank Prof. Sandeep Pradhan and Mr. Ramji Venkataramanan for useful discussions on directed information. We are extremely grateful to Prof. Erik Learned-Miller and Dr. Damian Fermin for sharing their code for high-dimensional entropy estimation and ENSEMBL sequence extraction, respectively.

CHAPTER V

Understanding Distal Transcriptional Regulation from Sequence Motif, Network Inference and Interactome Perspectives

5.1 Introduction

Understanding the mechanisms underlying regulation of tissue-specific gene expression remains a challenging question. While all mature cells in the body have a complete copy of the human genome, each cell type only expresses those genes it needs to carry out its assigned task. This includes genes required for basic cellular maintenance (often called "housekeeping genes") and those genes whose function is specific to the particular tissue type that the cell belongs to. Gene expression by way of transcription is the process of generation of messenger RNA (mRNA) from the DNA template representing the gene. It is the intermediate step before the generation of functional protein from messenger RNA. During gene expression, transcription factor (TF) proteins are recruited at the proximal promoter of the gene as well as at sequence elements (enhancers/silencers) which can lie several hundreds of kilobases from the gene's transcriptional start site (Figs. 5.1 and 5.2).

It is hypothesized that the collective set of transcription factors that drive (regulate) expression of a target gene are cell, context and tissue dependent ([227], [242]). Some of these TFs are recruited at proximal regions such as the promoter of the gene,

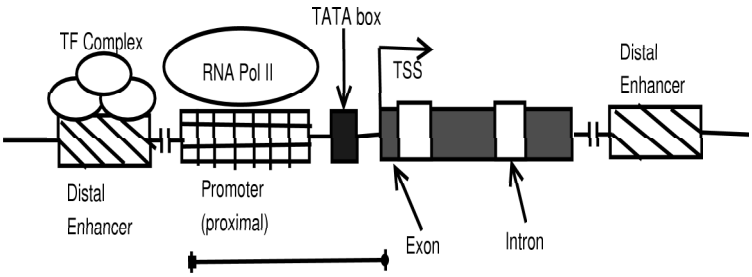


Figure 5.1: Schematic of Transcriptional Regulation. Sequence motifs at the promoter and the distal regulatory elements together confer specificity of gene expression via TF binding.

while others are recruited at more distal regions, such as enhancers. There are several (hypothesized) mechanisms for promoter-enhancer interaction through protein interactions between TFs recruited at these elements during formation of the transcriptional complex [226]. A commonly accepted mechanism of distal interaction, during regulation, is looping ([238], [192], [215]), shown in Fig. 5.2.

To understand the role of various genomic elements in governing gene regulation, functional genomics has played an enabling role in providing heterogeneous data sources and experimental approaches to discern distal interactions during transcription. Each of these experiments have aimed to resolve different aspects (features) of transcriptional regulation focussing on TF binding, promoter modeling and epigenetic preferences for tissue-specific expression in some genomic regulatory elements ([188], [200], [207], [236], [215]). Additionally, some studies have demonstrated that these data sets along with principled statistical metrics can be used to derive such features computationally, with a view to asking questions relevant to the biology of transcriptional regulation ([200], [236], [220], [208]).

There have been several principled yet scattered studies characterizing the role of functional regulatory elements for certain genes (such as *Mecp2*, *Shh*, *Gata2*, *Gata3*) in various organisms ([215], [211], [204], [210], [198], [223]). These reveal an inherent spatio-temporal context of gene expression and regulation. However, there is a need

for a unified set of principles, spanning various genomic attributes, that could account for the behavior of these tissue-specific and gene-specific enhancers. We note that there are promoter-independent enhancers too, and their computational study has been far more principled ([227], [228]); however, their study is outside the scope of this study where we focus on gene-specificity in addition to tissue-specificity.

The results of the ENCODE project (*http://encode.nih.gov/*), ([188], [207]) on 1% of the human genome has established some very interesting results about the nature of transcriptional regulation at the genome scale. Particularly, they report the use of several experimental techniques (Histone ChIP on chip, DNASE1 hypersensitivity assays etc.) analyzing transcribed regions as well as their regulatory regions, genome-wide. A large scale computational effort is developing alongside to "learn" features of such regulatory elements and use of these features for predicting other control elements for genes outside the ENCODE regions, thereby accomplishing a genome-wide annotation. Considering that over 98% of the genome is non-coding, this effort is going to parallel the previous project in gene-annotation at the genome scale in effort and importance. Adding to this complexity is the fact that the same non-coding element can potentially regulate the expression of genes in a spatio-temporal manner, activating different genes at different times in different tissues, and from arbitrarily large distances from the gene. Thus there is a need for the principled "reverse-engineering" of the architectures of these regulatory elements, using features at the sequence, expression and interactome level.

Understanding the mechanism of transcriptional regulation thus entails several aspects:

1. Do regulatory regions like promoters and enhancers have any interesting *sequence properties* depending on their tissue-specificity of gene expression? These

properties can be examined based on their individual sequences or their epigenetic preferences. A common technique of analysis is the identification of tissue-specific motif-signatures ([216], [209]) for such elements.

2. To reduce the vast number of false positives that arise from sequence approaches alone, we appeal to a mechanistic insight from biology. For long-range transcriptional regulation to be enabled, there has to be an enhancer-promoter interaction during formation of the tissue-specific, gene-specific transcriptional machinery. Literature suggests that such interaction is mediated by protein-protein interactions between promoter TFs and enhancer TFs after looping along the chromosomal length ([215], [175], [197], [238]). This insight (Fig. 5.2) poses two further questions:

   - Which TFs bind the promoter and the putative enhancer?
   - Do the resultant interactions between enhancer and promoter TFs have any special characteristic that discriminate functional non-coding regulatory regions from non-functional ones?

As a case study to answer some of these questions, we examine the regulation of *Gata2* regulation in the developing kidney. *Gata2* is a gene belonging to the GATA family of transcription factors (*GATA1-6*), and binds the consensus -WGATAR- motif on DNA [222]. It is located on mouse chromosome 6, and plays an important role in mammalian hematopoiesis, nephrogenesis and CNS development, with important phenotypic consequences. The study of long-range regulatory elements that effect *Gata2* expression has been on for several years now. The most common strategy for identifying possible regulatory elements has hitherto been inter-species conservation studies. Using this approach, all elements flanking the gene that are conserved more

than some threshold are retained for further experimental characterization. The reason underlying this strategy is that truly functional elements are under evolutionary pressure to retain their function across species. Given the technical complexity of associated transgenic experiments, this turns out to be a fairly inefficient strategy, especially since the number of candidate regulatory elements increases as larger genomic regions are examined (to account for distal regulation). Such a scenario prompts the need for an integrative strategy to reduce the number of candidates obtained from a purely conservation-based search strategy using other, complementary genomic modalities.

Recently, our lab reported the characterization of two enhancer elements, conferring urogenital-specific expression of *Gata2*, between 80 and 150kbp away from the gene locus, on chromosome 6 [204]. In this work, we examine if genomic features, other than sequence identity, are predictive of the location of these elements. These feature span sequence, expression and interactome perspectives for such regulatory elements. We will also attempt to motivate the utility of these approaches (metrics and data sources) as well as their biological relevance alongside (how they fit into the biophysics of transcriptional regulation). It must be pointed out that there is scant data available, in that data specific to the developing kidney is hard to come by. Under this constraint, we have made some biologically plausible assumptions so as to obtain maximum information from currently available data sources.

5.2 Rationale and Data Sources:

The overall schematic of distal transcriptional regulation via looping is given in Fig. 5.2. This schematic suggests the decomposition of the regulatory process along three main modalities: sequence, expression and interactome. Our main goal in

this paper is to understand urogenital (kidney) enhancer behavior from these three perspectives. These attributes are discussed below:

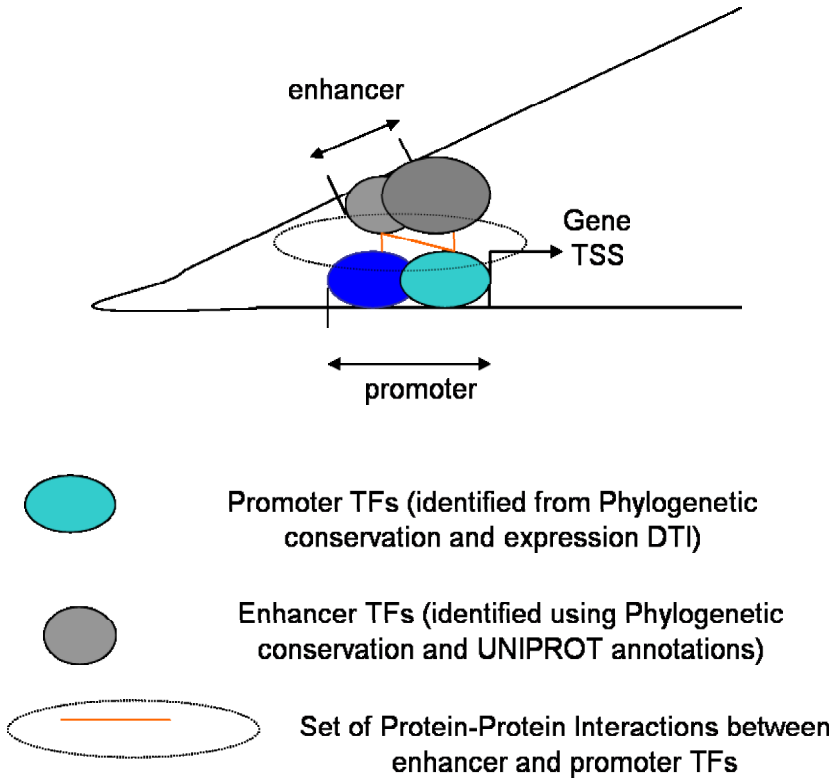


Figure 5.2: Distal enhancer-promoter interaction via looping is mediated via protein interactions during TF complex formation. The set of TFs that are putatively recruited at the proximal promoter and distal enhancer can be found from sequence and expression data [231]. Evidence of protein-interaction between proximal and distal TFs can be found from interaction databases.

1. **Sequence Perspective:** To build motif signatures underlying kidney-specific enhancer activity, it would be best to have a database of previously characterized urogenital (UG) enhancers. However, due to the unavailability of such data, we utilize kidney-specific promoter sequences and histone-modified sequences of enhancers to find motif-signatures of regulatory elements that are potentially UG enhancers.

    - Promoters of kidney-specific genes: A catalog of kidney-specific mouse pro-

moters is available from the GNF Symatlas (*http://symatlas.gnf.org/*). This database contains list of annotated genes and their expression in several tissue types, including the kidney. Since the proximal promoter of such kidney-specific genes harbors the transcriptional machinery for gene regulation, their sequences putatively have motifs that are associated with kidney-specific expression. Additionally, promoters that are spatio-temporally expressed during kidney development are also analyzed (MGI: *http://www.informatics.jax.org*). The GNF dataset profiles mostly adult tissue-types. Since our goal is to study enhancer activity during nephrogenesis, we focus on genes expressed between day $e10$ and $e12$ in the developing kidney - such a list is obtained from the MGI database.

Without loss of generality, we use six-nucleotide motifs (hexamers) as the motifs. This is based on the observation that most transcription factor binding motifs have a $5-6$ nucleotide core sequence with degeneracy at the ends of the motif. A similar strategy was introduced in ([180], [201]). The main difference in our approach from such previous work is that differential hexamer analysis was done for the same class of sequences, and the statistical nature of the "test-set" is, by design, similar to the training set. That is, in [180], differential hexamers are found between known Cis-Regulatory Modules (CRMs) and non-CRMs, and used for the prediction of new CRMs from sequence. On the other hand, [201] deals with finding hexamer features of known promoters and using them to predict new promoters from sequence. In our case, however, we don't have enhancer data (equivalent to CRMs) and we are using promoter-data for the prediction of enhancer (CRM) instead. Thus, the nature of the test sequence is very different. We

demonstrate that our approach is partially useful in the discovery of putative enhancers from sequence.Also,the presented motif-finding approach does not depend on motif length and can be scaled depending on biological knowledge.

We set up the motif discovery as a feature extraction problem from these tissue-specific promoter sequences and then build a random forest (RF) classifier to classify new sequences into specific and non-specific categories based on these identified sequence features (motifs). Based on the motifs derived using a RF classifier algorithm we are able to accurately classify more than 95% (training-error rate) of tissue-specific genes based upon their upstream promoter region sequences alone. Since promoters are non-coding regulatory regions, the derived motifs can be putatively used to find kidney-specificity of other non-coding regions genome-wide (Section: 5.8).

- Chromatin marks in known regulatory elements: Using the recently released ENCODE data, a catalog of sequences that undergo histone modifications such as methylation and acetylation is available for analysis [207]. Reports suggest that mono-methylation of the lysine residue of Histone $H3$ is associated with enhancer activity [200] whereas tri-methylation of $H3K4$ and $H3$ acetylation are associated with promoter activity. Using this set of $H3K4me1$, $H3K4me3$ and $H3ac$ sequences, we aim to find sequence motifs that are indicative of such epigenetic preferences during transcription. Though epigenetic data is available for five different cell lines, we choose the HeLa cell line data because of its widespread use as a model system to understand transcriptional regulation *in-vitro* in the laboratory. Thus, we build a RF classifier to discriminate monomethylated $H3K4$ se-

quences from trimethylated $H3K4$/acetylated $H3$ sequences. We note that this data is *not* kidney-specific, and such data is yet to become available. This yields motifs associated with epigenetic properties of promoters and enhancers, which are potentially predictive of the regulatory potential for novel sequences (section: 5.9).

2. **Expression Perspective:** There is limited expression data for the developing mouse kidney, mainly due to technical reasons concerning small tissue yield at such early time points. For this study, we use microarray expression data from a public repository of kidney microarray data (*http://genet.chmcc.org* [240], *http://spring.imb.uq.edu.au/* [178]). Each of these sources contain expression data profiling kidney development from about day 10.5 dpc to the neonate stage. Such expression data can be mined for potential regulatory influence between upstream TF genes and *Gata2*.

   - *Inference of TF effectors at the promoter region*: The TFs putatively recruited at the proximal promoter are identified using the Directed Information metric, that uses gene-expression level influence in addition to phylogenetic conservation of the corresponding binding site. We have earlier shown that DTI is a good predictor of gene influence and can be used to infer transcriptional regulatory networks [231]. A more detailed explanation is given in sections: 5.10.1 through 5.10.1.

   - *Inference of TF effectors at each non-coding region*:

     At the distal enhancer, it is believed that there is recruitment of tissue-specific transcription factors that co-operate with the basal transcriptional machinery (at the promoter) to direct tissue-specific gene expression ([206],

[216]). Whereas phylogeny and expression-based influence metrics can yield high confidence candidates for promoter TFs, a similar analysis for enhancers is not possible, because of higher order effects ([220], [209]). To this end, the only plausible way to search for enhancer TFs is to combine phylogeny with tissue-specific annotation (from UNIPROT or MGI). Hence, every transcription factor, whose motif is conserved at a non-coding (putative enhancer) region and is tissue-specific in annotation is considered a likely candidate TF at that non-coding region.

3. **Interactome Perspective:** The identification of phylogenetically conserved effector TFs at the promoter (identified via DTI), as also those that are phylogenetically conserved at the putative enhancers, lead to the exploration of protein-interactions between these TFs, during distal enhancer-promoter interaction (Sec:5.10). The STRING database (*http://string.embl.de*) integrates various experimental modalities (genomic context, high-throughput experiments such as co-immunoprecipitation, co-expression and literature) to maintain a list of organism-specific functional protein-association networks that is amenable to such exploration.

In this work, the above questions will be integratively answered for training data as well as in the context of the urogenital enhancers identified in [204]. We aim to show that each of these 'features' have a predictive value for the identification of enhancers and the integration of these heterogeneous data can lead to potential reduction in false positive rate during large-scale enhancer discovery, genome-wide. To date, there has been no comprehensive study for summarizing these various heterogeneous data sources to understand transcriptional regulation.

5.3 Validation/Biological Application

As suggested in Sec: 5.1, we use the recently identified *Gata2* urogenital (UG) enhancers to validate our computational approach. All the data sources (and their analysis) are therefore going to be focused on the kidney.

The experimental characterization of these enhancers was done as follows. Based on BAC transgenic [204] studies, the approximate location of the urogenital enhancer(s) of *Gata2* were localized to a 70 kilobase region on chromosome 6. Using inter-species conservation plots, four elements were selected for transgenic analysis in the mouse. These were designated UG1, 2, 3 and 4. After a lengthy and resource-intensive experimental effort, two out of these four non-coding elements, $UG2$ and $UG4$ were found to be true UG enhancers. Our goal is to find "features" at the sequence, expression and interactome level, that are predictive of this reported behavior of elements $UG1-4$ in the developing kidney.

It is easy to see the utility of such a methodology, since this can be scaled up contextually for other genes of interest. Given the complexity of 1% of the genome, made possible by the ENCODE project, the search for functional elements genome-wide is going to be an important and challenging exercise.

5.4 Organization

With a view to understanding the elements of transcriptional regulation, the first part of this paper (Sections 5.5-5.9) addresses the problem of identifying motif signatures representative of transcriptional control from kidney-specific promoters and epigenetically marked sequences. The second part of this work (Sections 5.10.1 - 5.10.2) integrates phylogeny and expression data to find regulatory TFs at the proximal promoter and enhancer(s) of *Gata2*. Using the notion of TF interactions between

enhancer and promoter, we examine if protein-interaction data (Section: 5.10.3) can offer supporting evidence for the observed *in-vivo* behavior of four putative *Gata2* regulatory elements. Classifiers are designed to discriminate regulatory vs. non-regulatory regions based on each of these multiple modalities. Finally, a probabilistic combination of these classifiers is done to obtain a validation (Section: 5.11) of the *Gata2* UGEs ($UG1-4$). Sections: 5.12 and 5.13 conclude the paper.

5.5 Sequence Data Extraction and Pre-processing

The Novartis foundation tissue-specificity atlas [*http://symatlas.gnf.org/*], has a compendium of genes and their corresponding tissues of expression. Genes have been profiled for expression in about twenty-five tissues, including adrenal gland, brain, dorsal root ganglion, spinal chord, testis, pancreas, liver etc. Considering these diversity of tissue-types, one concern with the interpretation of this data is the variability in expression across tissue-types. To address this concern, we take a fairly stringent approach - if a gene is expressed in less than three tissue types, it is annotated tissue-specific (*'ts'*), and if it is expressed in more than 22 tissue types, it is annotated to be non-specific (*'nts'*). Based on this assignment, we find a list of 86 genes that are tissue-specific as well as have kidney expression (MGI: *http://www.informatics.jax.org/*). For these kidney-specific genes, we extract their promoter sequences from the ENSEMBL database (*http://www.ensembl.org/*), using sequence 2000bp upstream and 1000bp downstream up to the first exon relative to the transcriptional start site reported in ENSEMBL (release 37).

Before proceeding to motif selection, a matrix of motif-promoter correspondences is created. In this matrix, the counts of hexamer (six-nucleotide) motif occurrence in the *'ts'* and *'nts'* promoters is obtained using sequence parsing (*R package: 'se-*

qinr'). The motif length of six is not overly restrictive, since it corresponds to the consensus binding site size of several annotated transcription factor motifs in the TRANSFAC/JASPAR databases. A Welch t-test is then performed between the relative counts of each hexamer in the two expression categories (*'ts'* and *'nts'*) and the top 1000 hexamers with $p-value \leq 10^{-6}$ are selected. This set of discriminating hexamers is designated ($\overrightarrow{\mathbf{H}} = H_1, H_2, \ldots, H_{1000}$). This procedure resulted in two hexamer-gene co-occurrence matrices, - one for the *'ts'* (or +1) class of dimension $N_{train,+1} \times 1000$ and the other for the *'nts'* (or -1) class - dimension $N_{train,-1} \times 1000$. Here $N_{train,+1}$ is the matrix of the 86 kidney-specific genes. $N_{train,-1}$ is the set of *'nts'* that do not have kidney-specific expression.

As an illustration, we show a representative matrix (Table. 5.1).

Ensembl Gene ID	AAAAAA	AAATAG	Class
ENSG00000155366	1	1	+1
ENSG000001780892	4	3	+1
ENSG00000189171	1	2	-1
ENSG00000168664	4	3	-1
ENSG00000160917	2	1	-1
ENSG00000176749	1	1	-1
ENSG00000006451	3	2	+1

Table 5.1: The 'motif count matrix' for a set of gene-promoters. The first column is their ENSEMBL gene identifiers, the next 2 columns are hexamer quantile labels, and the last column is the corresponding gene's class label ($+1/-1$).

All the above steps, from sequence extraction, parsing and quantization to obtain hexamer-promoter counts that are done for the kidney-specific genes can be repeated for the histone-modified sequences. This dataset is obtained from the Sanger ENCODE database (*http://www.sanger.ac.uk/PostGenomics/encode/data-access.shtml*), and contains 298 sequences that undergo modification ($m1/me3/ac$) in histone ChIP assays. 140 of these correspond to $H3K4me1$ (enhancers), and 158 correspond to $H3K4me3/H3ac$ marks (promoters). Here, the 1000 hexamers discriminating $H3K4me1$-sequences (+1 set) and a ($H3K4me3/H3ac$) (-1), are

designated $\overrightarrow{\mathbf{H'}} = H'_1, H'_2, \ldots, H'_{1000}$.

Sequence	AAAATA	AAACTG	Class
chr2:41410492-41411867	2	1	+1
chr6:41654502-41654782	4	2	+1
chr3:41406971-41408059	1	1	-1
chr2:41665970-41667002	2	3	+1
chr4:41476956-41478365	1	2	-1
chr5:41530471-41531046	2	2	-1
chrX:41783327-41784532	1	2	+1

Table 5.2: The 'motif count matrix' for a set of histone-modified sequences. The first column is their genomic locations along the chromosome, the next 2 columns are hexamer quantile labels, and the last column is the corresponding sequence class label $(+1/-1)$.

5.6 Motif-Class Correspondence Matrices

From the above, $N_{train,+1} \times 1000$ and $N_{train,-1} \times 1000$ dimensional count matrices are available both for the kidney-promoter and histone-modified sequences. Before proceeding to the feature (hexamer motif) selection step, the counts of the $M = 1000$ hexamers in each training sample need to be normalized to account for variable sequence lengths. In the co-occurrence matrix, let $gc_{i,k}$ represent the absolute count of the k^{th} hexamer, $k \in 1, 2, \ldots, M$ in the i^{th} gene. Then, for each gene g_i, the quantile labeled matrix has $X_{i,k} = l$ if $gc_{i,[\frac{l-1}{K}M]} \leq gc_{i,k} < gc_{i,[\frac{l}{K}M]}, K = 4$. Matrices of dimension $N_{train,+1} \times 1001$, $N_{train,-1} \times 1001$ for the specific and non-specific training samples are now obtained. Each matrix contains the quantile label assignments for the 1000 hexamers $(X_i, i \in (1, 2, \ldots, 1000))$, as stated above, and the last column would have the corresponding class label $(Y = -1/+1)$. Having constructed two groups of genes for analysis, tissue specific (*'ts'*) and non-tissue specific (*'nts'*) - we seek to find hexamer motifs which are most discriminatory between these two classes. Our goal would be to make this set of motifs as small as possible - i.e. to achieve maximal class partitioning with the smallest feature subset. Towards this goal, we explore the use of random forests (RF) [176] for finding such a discriminative

hexamer subset.

5.7 Random Forest Classifiers

A random forest (RF) is an ensemble of classifiers obtained by aggregating (bagging) several classification trees ([199], [176]). Each data point (represented as an input vector) is classified based on the majority vote gained by that vector across all the trees of the forest. Each tree of the forest is grown in the following way:

- A bootstrapped sample (with replacement) of the training data is used to grow each tree. The sampling for bootstrapped data selection is done individually at each tree of the forest.
- For an M-dimensional input vector, a random subspace of m ($\ll M$)-dimensions is selected, and the best split on this subspace is used to split the node. This is done for all nodes of the tree. Each tree is grown to maximum length, with no pruning.

During the training step, before sampling by replacement, one-third of the cases is kept "out of the training bag". This oob (out-of-bag) data is used to obtain an unbiased estimate of the classification error as trees are added to the forest. It is also used to get estimates of variable importance.

From the above we see that the classifier structure of the random forest is an ensemble of trees. Each tree is trained and built on a different bootstrap sample (split) of the training data. Hence each tree has a different topology. Unlike a tree classifier, therefore, it is not possible to obtain a "consensus topology" of the RF classifier. In the absence of one unifying structure for the purpose of visualization, we can inspect the other outputs like variable importance, confusion matrix, and OOB error rate to ascertain the accuracy and performance of the RF classifier.

The variables selected for optimal partitioning over class labels can be examined from a variable importance plot which indicates which variables are most discriminatory between these two classes ([176], [213]). It is also to be noted that random forests afford the dual advantage of both training and test-set error estimation (through the OOB data) during the overall training procedure. Thus there is no separate procedure for test-set error estimation that needs to be implemented in the case of RFs. Each tree in the ensemble is trained on a $\frac{2}{3}rd - \frac{1}{3}rd$ split of the data. Each tree is grown to get the least oob error before being incorporated into the classifier ensemble.

A confusion matrix is one representative tool to understand the performance of the RF classifier. After the training process, the confusion matrix measures the discordance between true and predicted classes (and can be used for OOB error estimation). Each row represents the instances of the actual class, while each column of the matrix represents the instances in a predicted class. The matrix can then be used for false-positive, false-negative, true-negative and true-positive rate computations.

Several interesting insights into the data are available using random forest analysis. The variable importance plot yields the variables that are most discriminatory for classification under the 'ensemble of trees' classifier. This importance is based on two measures- 'Gini index' and 'decrease in accuracy'. The Gini index is an entropy based criterion which measures the purity of a node in the tree, while the other metric simply looks at the relative contribution of each variable to the accuracy of the classifier. For our studies, we use the 'randomForest' package for R [213]. The classifier performance on the individual data and the related diagnostics are mentioned under each head (Secs: 5.8 and 5.9).

5.8 Random Forests on Kidney-specific promoters

In this section, we aim to find discriminating sequence motifs between a set of kidney-specific promoters and housekeeping promoters with a goal to find sequence motifs underlying kidney-specific regulation. The kidney enriched dataset has 86 genes that are assigned to a tissue specific class and have higher than mean expression in the kidney. For the purpose of training and testing, we consider the set of housekeeping genes identified from the *'nts'* class and reported in literature ([187], [191]). There are almost 1500 genes in the housekeeping gene (*'nts'*) set. Since, this would lead to unbalanced predictions during classifier training, we use a stratified sampling approach [213] to select for a sample size that reduces this effect (the sampling itself is done with a prior on the relative sizes of the two classes). Here, the set of (-1) promoter-sequences are taken to be of the same size as the $(+1)$ class. Using this approach, we obtain a training-error classification accuracy of $> 95\%$ on the kidney enriched tissue-specificity data set.

Before proceeding to motif identification, it is necessary to check for possible sequence bias (GC composition) between the two classes of promoters (kidney-specific vs. housekeeping). Though there are several kinds of sequence bias [232], the composition bias is most closely related to this problem. If there is a significant bias, then the motifs turn out to be just GC rich sequences that are not very biologically informative [241] for regulatory potential. The GC composition of these two classes of sequences is represented in Fig. 5.3. We note that though only a subset of 'nts' gene-promoters were used during the RF analysis, we show the GC-composition for the entire class of 'nts' sequences for completeness. As can be seen, the average GC composition is the same. The ROC space representation and variable importance

plot for the overall classification is indicated below (Fig. 5.11 and Fig. 5.4). The confusion matrices are all explained in the context of the classifier combination in Section:5.11.

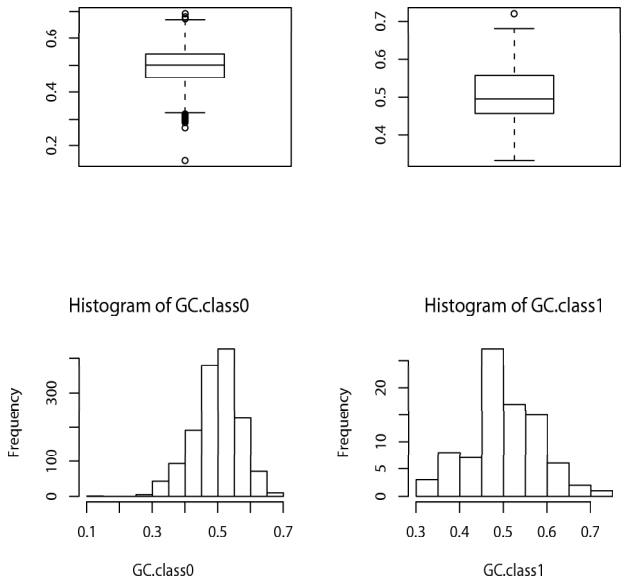


Figure 5.3: GC plots for sequence bias in kidney-specific vs. housekeeping promoters.

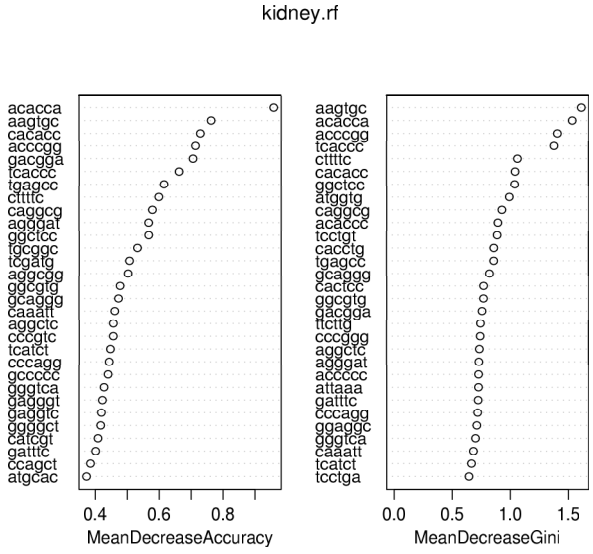


Figure 5.4: Top hexamers which can discriminate between kidney-specific and house-keeping genes.

To address a related question, we examine if the top ranked hexamers in the kidney dataset correspond sequence-wise to known transcription factor binding sites. Using

the publicly available Opossum tool (*http://www.cisreg.ca/cgi-bin/oPOSSUM/opossum/*) or MAPPER (*http://bio.chip.org/mapper*), we found several interesting transcription factors to map to these motifs, such as *Nkx*, *ARNT*, *c-ETS*, *FREAC4*, *NFAT*, *CREBP*, *E2F*, *HNF4A*, *Pax2*, *MSX1*, *SP1* several of which are kidney-specific. Though this is highly consistent with the tissue-specificity of the dataset, the functional relevance of these sites remains to be experimentally validated.

5.9 RFs on chromatin-modified sequences

We train a RF classifier on a set of 298 sequences from chromosome sequence that have varying histone modifications associated with them (namely, $H3K4me1/me3$, and $H3ac$), as mentioned in Section: 5.2. These sequences had a high level of the corresponding histone-modification from ChIP experiments. The other regions that were assayed for but did not have high levels of modification are not considered in this analysis. These are derived from the HeLa cell line and are not necessarily context-specific for kidney development. However, given the widespread use of this cell line for transcriptional studies, we aim to find if the motifs associated with regulatory elements are indeed predictive of enhancer activity.

Here too, we examine the GC-composition bias of these two sequence classes (Fig. 5.5) and confirm that there is no such sequence bias that would skew the discovery and subsequent interpretation of these epigenetic motifs.

The motifs obtained from the random forest analysis indicate the "sequence-preferences" of regulatory elements that are kidney-specific (Fig. 5.4) or nucleosome-free (Fig. 5.6). For the kidney-specific case, the underlying caveat is that co-expression does not imply co-regulation; however we are only using the discovered motifs to understand the "sequence-preferences" of kidney-specific regulatory-regions

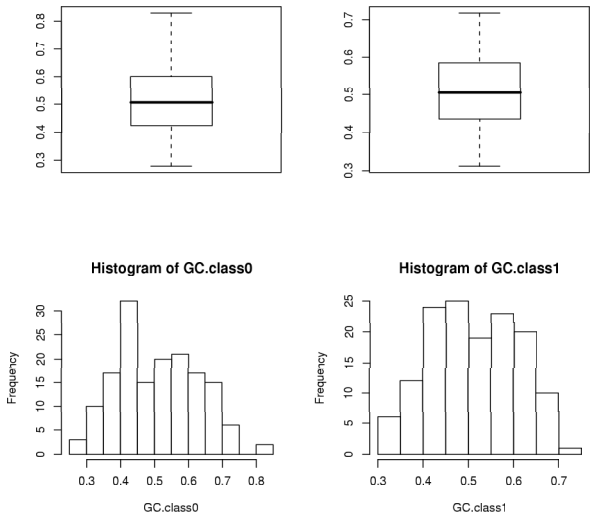


Figure 5.5: GC plots for sequence bias in $H3K4me1$ histone sequences vs. $H3K4me3$ and $H3ac$ sequences.

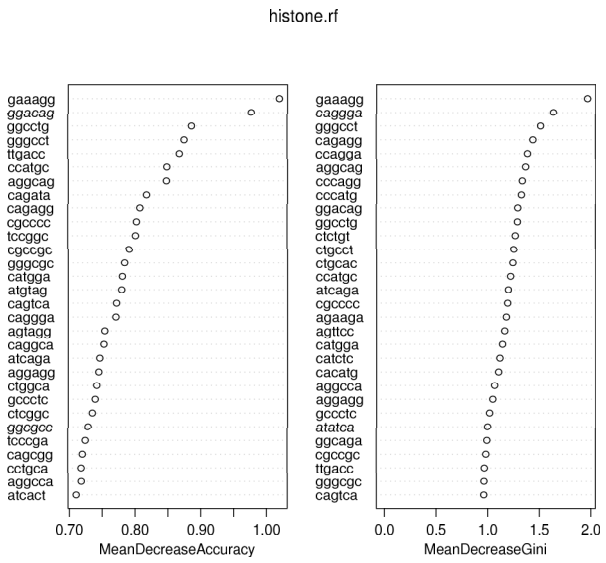


Figure 5.6: Top hexamers which can discriminate between $H3K4me1$ histone sequences vs. $H3K4me3$ and $H3ac$ sequences.

[205] rather than using them for *de-novo* prediction of new genes that are regulated by the same transcriptional machinery. Most of the motifs do not overlap TFBS motifs and might be indicative of more interesting sequence properties. We analyze the performance of these classifiers on the 4 UG enhancers, mentioned previously.

In both cases, $UG2-4$ are classified as kidney-specific enhancers, whereas $UG1$ is correctly classified as not being regulatory. Additionally, a control set of "promoter-independent" enhancers derived from the Mouse Enhancer database [227] was also classified as enhancers based on these chromatin signatures. This high prediction accuracy inspite of non-specificity of cell context (*HeLa* cell line) is very interesting and has potentially high predictive value. This is explored further in Sec: 5.11.

We now proceed to the mechanistic insight (based on TF effector identification and PPI) mentioned in Section. 5.1 to understand the behavior of putative regulatory elements.

5.10 PPI between promoter and enhancer TFs

In order to understand the nature of distal interactions between the enhancer and promoter TFs (Fig. 5.2), we decouple the overall regulation problem into three parts:

1. Identification of putative TF effectors at the promoter (Section: 5.10.1),
2. Identification of enhancer TFs (Section: 5.10.2), and
3. Examination of the interaction-graph formed between enhancer-TFs and promoter TFs (Section: 5.10.3).

5.10.1 TF effector identification at Promoter and Enhancer

Promoter TF identification: TFs that regulate basal transcription at the promoter can be identified from phylogenetic conservation or co-expression studies. In this approach, the promoter sequence (here, the *Gata2* promoter) is aligned across multiple species and the TFBS motifs that are conserved in the multiple alignment are considered to be putative effectors of gene regulation. An additional step involves examining the promoters of all genes that are co-expressed in the same spatio-temporal manner

as the gene of interest (e.g.: *Gata2* in the kidney). Such sequence-based approaches have been examined in literature ([216], [209], [220]).

Since the list of putative TFs (identified above) that potentially bind at the promoter is still large, there have been efforts to incorporate gene-expression data to reduce the set of potential TF effectors. In this scenario, if the gene corresponding to the conserved TF has a high expression-level influence on *Gata2* expression, then that TF has stronger evidence for being a potential regulator ([217], [212]). Recently, we introduced the directed information (DTI) as a metric to infer expression-level influence between any putative transcription factor (TF) gene and a target gene (such as *Gata2*) [231]. We will briefly summarize the utility of DTI for TF effector identification in the following sections (Sec. 5.10.1 and 5.10.1). This seeks to integrate sequence and expression data into the determination of relationships between transcription factors and their target-genes. All additional details (performance on synthetic data, other biological data and comparison with other metrics) are available in [231]. Information-based measures have enabled the investigation of non-linear gene relationships in the presence of measurement noise [217]. An important point to note is that unlike mutual information, the DTI is a *directed* metric that enables the inference of both strength and direction of gene influence.

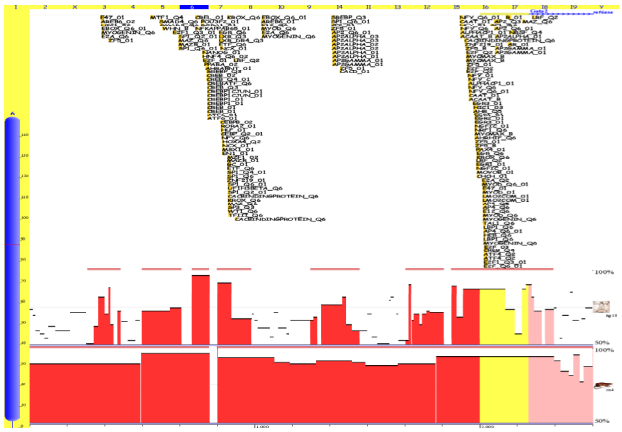


Figure 5.7: TFBS conservation between Human, Mouse and Rat, upstream (x-axis) of *Gata2*, from *http://www.ecrbrowser.dcode.org/*. The mouse sequence is the base sequence and is hence not displayed. The dark and light red regions correspond to potential TF binding regions on DNA.

DTI Formulation

As alluded to above, there is a need for a viable influence metric that can find relationships between the TF "effector" gene (identified from phylogenetic conservation) and the target gene (like *Gata2*). Several such metrics have been proposed, notably correlation, coefficient of determination (CoD), mutual information etc. To alleviate the challenge of detecting non-linear gene interactions, an information theoretic measure like mutual information has been used to infer the conditional dependence among genes by exploring the structure of the joint distribution of the gene expression profiles [217]. However, the absence of a directed dependence metric has hindered the utilization of the full potential of information theory. In this section, we examine the applicability of one such metric - the directed information criterion (DTI), for the inference of non-linear, directed gene influences.

The DTI - which is a measure of the directed dependence between two N-length

random processes $X \equiv X^N$ and $Y \equiv Y^N$ is given by [219]:

$$I(X^N \to Y^N) = \sum_{n=1}^{N} I(X^n; Y_n | Y^{n-1}) \tag{1}$$

Here, Y^n denotes $(Y_1, Y_2, .., Y_n)$, i.e. a segment of the realization of a random process Y and $I(X^N; Y^N)$ is the Shannon mutual information [181].

An interpretation of the above formulation for DTI is in order. To infer the notion of influence between two time series (mRNA expression data) we find the mutual information between the entire evolution of gene X (up to the current instant n) and the current instant of Y (Y_n), given the evolution of gene Y up to the previous instant $n-1$ (i.e. Y^{n-1}). This is done for every instant, $n \in (1, 2, \ldots, N)$, in the N - length expression time series.

As already known, $I(X^N; Y^N) = H(X^N) - H(X^N|Y^N)$, with $H(X^N)$ and $H(X^N|Y^N)$ being the entropy of X^N and the conditional entropy of X^N given Y^N, respectively. Using this definition of mutual information, the DTI can be expressed in terms of individual and joint entropies of X^N and Y^N. The task of N-dimensional entropy estimation is an important one and due to computational complexity and moderate sample size, histogram estimation of this multivariate density is unviable. However, several methods exist for consistent entropy estimation of multivariate small sample data ([189], [221], [225], [243]). In the context of microarray expression data, wherein probe-level and technical/biological replicates are available, we use the method of [189] for entropy estimation.

From (1), we have,

$$I(X^N \rightarrow Y^N) = \sum_{n=1}^{N} [H(X^n|Y^{n-1}) - H(X^n|Y^n)]$$
$$= \sum_{n=1}^{N} \{[H(X^n, Y^{n-1}) - H(Y^{n-1})] -$$
$$[H(X^n, Y^n) - H(Y^n)]\} \quad (2)$$

- To evaluate the DTI expression in (2), we need to estimate the entropy terms $H(X^n, Y^{n-1})$, $H(Y^{n-1})$, $H(X^n, Y^n)$ and $H(Y^n)$. This involves the estimation of marginal and joint entropies of n random variables, each of which are R dimensional, R being the total number of replicates (probe-level, biological and technical).

- Though some approaches need the estimation of probability density of the R-dimensional multivariate data (X^n) prior to entropy estimation, one way to circumvent this is to the use the method proposed in [189]. This approach uses a Voronoi tessellation of the R-dimensional space to build nearly uniform partitions (of equal mass) of the density. The set of Voronoi regions $(V^1, V^2, \ldots, V^n)$ for each of the n points in R-dimensional space is formed by associating with each point X_k, a set of points V^k that are closer to X_k than any other point X_l, where the subscripts k and l pertain to the k^{th} and l^{th} time instants of gene expression.

- Thus, the entropy estimator is expressed as : $\hat{H}(X^n) = \frac{1}{n}\sum_{i=1}^{n} \log(nA(V^i))$, where $A(V^i)$ is the R-dimensional volume of Voronoi region V^i. $A(V^i)$ is computed as the area of the polygon formed by the vertices of the convex hull of the Voronoi region V^i. This estimate has low variance and is asymptotically efficient [190].

To obtain the DTI between any two genes of interest (X and Y) with N-length expression profiles X^N and Y^N respectively, we plug in the entropy estimates computed above into the expression (2).

From the definition of DTI, we know that $0 \leq I(X_i^N \to Y^N) \leq I(X_i^N; Y^N) < \infty$. For easy comparison with other metrics, we use a normalized DTI metric given by $\rho_{DI} = \sqrt{1 - e^{-2I(X^N \to Y^N)}} = \sqrt{1 - e^{-2\sum_{i=1}^{N} I(X^i; Y_i | Y^{i-1})}}$. This maps the large range of DI, ($[0, \infty]$) to lie in $[0, 1]$. Another point of consideration is to estimate the significance of the 'true' DTI value compared to a null distribution on the DTI value (i.e. what is the chance of finding the DTI value by chance from the series X and Y). This is done using empirical p-value estimation after bootstrap resampling (Sec: 5.10.1). A threshold p-value of 0.05 is used to estimate the significance of the true DTI value in conjunction with the density of a random data permutation, as outlined below.

Significance Estimation of DTI

We now outline a procedure to estimate the empirical p-value to ascertain the significance of the normalized directed information $\hat{I}(X^N \to Y^N)$ between any two N-length time series $X \equiv X^N = (X_1, X_2, \ldots, X_N)$, and $Y \equiv Y^N = (Y_1, Y_2, \ldots, Y_N)$. In our case, the detection statistic is $\Theta = \hat{I}(X^N \to Y^N)$ and the chosen acceptable p-value is α.

The overall bootstrap based test procedure is ([186], [229], [196]):

- Repeat the following procedure $B(= 1000)$ times (with index $b = 1, \ldots, B$):
  - Generate resampled (with replacement) versions of the times series X^N, Y^N, denoted by X_b^N, Y_b^N respectively.
  - Compute the statistic $\theta^b = \hat{I}(X_b^N \to Y_b^N)$.

- Construct an empirical CDF (cumulative distribution function) from these bootstrapped sample statistics, as $F_\Theta(\theta) = P(\Theta \leq \theta) = \frac{1}{B}\sum_{b=1}^{B} I_{x\geq 0}(x = \theta - \theta^b)$, where I is an indicator random variable on its argument x.

- Compute the true detection statistic (on the original time series), $\theta_0 = \hat{I}(X^N \rightarrow Y^N)$ and its corresponding p-value ($p_0 = 1 - F_\Theta(\theta_0)$) under the empirical null distribution $F_\Theta(\theta)$.

- If $F_\Theta(\theta_0) \geq (1-\alpha)$, then we have that the true DTI value is significant at level α, leading to rejection of null-hypothesis (no directional association).

Summary of DTI-based TF effector Inference

Our proposed approach using DTI for determining the effectors for gene B (*Gata2* in the enhancer study) is as follows:

- Identify the G genes $(A_1, A_2, \ldots, A_G)$, based on phylogenetic conservation (Fig. 5.7). Preprocess the gene expression profiles by normalization and cubic spline interpolation. Assuming that there are N points for each gene, entropy estimation is used to compute the terms in the DTI expression (Eqn. 2).

- For each pair of genes A_i and B among these G genes :

  {

  - Look for a phylogenetically conserved binding site of TF encoded by gene A_i in the upstream region of gene B.
  - Find $DTI(A_i, B) = I(A_i^N \rightarrow B^N)$, and the normalized DTI from A_i to B, $DTI(A_i, B) = \sqrt{1 - e^{-2I(A_i^N \rightarrow B^N)}}$.
  - Bootstrap resampling over the data points of A_i and B yields a null distribution for $DTI(A_i, B)$. If the true $DTI(A_i, B)$ is significant at level α with

respect to this null histogram, infer a potential influence from A_i to B.

– The value of the normalized DTI from A_i to B gives the putative strength of interaction/influence.

– Every gene A_i which is potentially influencing B is an 'effector'. This search is done for each gene A_i among these G genes $(A_1, A_2, \ldots, A_G)$.

}

Note: As can be seen, phylogenetic information is inherently built into the influence network inference step above. We note that, in this approach, the choice of potential effectors for a target gene is based on only those TFs that have a binding site at the target gene's promoter. This aims to reduce the overall search space based on biological prior knowledge.

As an example, we indicate the significance and strength of the DTI between the *Oct1* TF and *Gata2*. The high strength of influence and its significance coupled with the phylogenetic conservation of the *Oct1* motif indicates expression evidence for the role of *Oct1* in *Gata2* regulation ([178], [185], [224]).

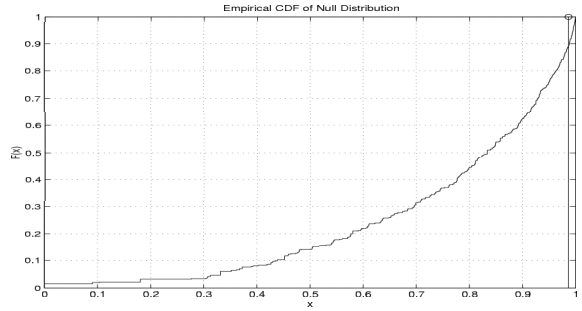


Figure 5.8: Cumulative Distribution Function for bootstrapped $I(\textit{Oct1} \rightarrow \textit{Gata2})$ interaction. True $\hat{I}(\textit{Oct1} \rightarrow \textit{Gata2}) = 0.9866$. Also, $\hat{I}(\textit{Gata2} \rightarrow \textit{Oct1}) = 0.8588$.

Such analysis can be extended to all TFs that are phylogenetically conserved. For *Gata2* regulation in the developing kidney, this set of putative TF effectors (apart from *Oct1*) is shown in Fig. 6.4. However, the functional role of these TFs in

regulating *Gata2* regulation needs to be experimentally confirmed.

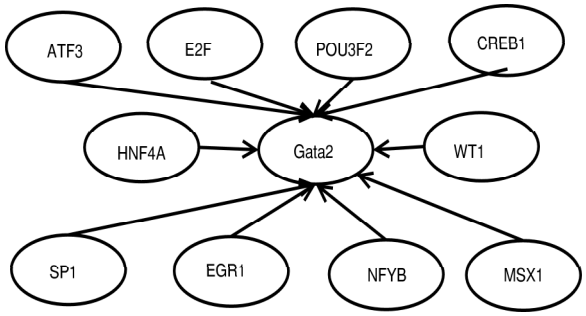


Figure 5.9: Putative upstream TFs using DTI for the *Gata2* gene.

5.10.2 Enhancer TF identification

In the earlier section, we have examined the identification of promoter TFs using phylogenetic sequence conservation of TFBS motifs in conjunction with expression level influence using DTI. The next key step towards determining the nature of promoter-enhancer TF interactions is the identification of enhancer-TFs. As has been alluded to earlier, there is no method to precisely infer which transcription factors bind a certain regulatory element during long-range gene regulation. Thus, we appeal to a traditional approach of finding tissue-specific transcription factors that are phylogenetically conserved at any potential regulatory region ([228], [242]). This is consistent with earlier observations that enhancers recruit tissue-specific transcription factors during the formation of the overall transcriptional machinery during gene expression, whereas promoters recruit components of the basal transcriptional machinery ([206], [216], [209], [242], [238]).

To ascertain the tissue-specificity of each TF that putatively binds a regulatory element (identified via phylogenetic conservation), we examine that TF's annotation in the UNIPROT database. This database is one of the most current sources of TF annotation and has details pertaining to the sequence specificity of the binding

motif, the structure of the TF and its tissue-specificity of expression. For those TFs that do not have a UNIPROT annotation, we look at the tissue-expression of the corresponding gene from the mouse genome informatics (MGI) mRNA annotations. The MGI expression annotations encompass multiple modalities (literature, RNA *in-situ*) to suggest a tissue-restricted or conversely, a ubiquitous expression of the TF gene. Thus, a set of tissue-specific transcription factors that bind any non-coding region of interest (such as an enhancer) can be identified ([223], [242], [228], [209], [216]). For the *Gata2* UGEs, several potential TFs can be found, some of which are highlighted in Fig. 5.10.

5.10.3 Enhancer-Promoter Distal Interaction via Protein-Protein Interactions - A Graph Based Analysis

Using the notion of protein-protein interaction mediating long-distance interactions between promoters and enhancers during looping ([226], [175], [197]), we explore the interactome to look for within-group and between-group interactions in the promoter-TF and the enhancer-TF groups. The resultant interaction-graph can be examined for several "structural" characteristics (like heterogeneity, degree distribution, path length, density, clustering coefficient and connected components) ([173], [183]). The goal is to identify structural features that discriminate true enhancer vs. non-functional element activity based on their interaction-graph.

The interaction-graphs (e.g: Fig. 5.10) are obtained in the following manner:

- One part of the graph (hollow circles) corresponds to the TF effector group at the promoter. These V_p TFs are identified based on phylogenetic conservation and directed information (section: 5.10.1).
- The other part of the graph (filled circles) corresponds to the V_e tissue-specific TFs group at the enhancer, identified based on phylogeny and annotation (sec-

tion: 5.10.2).

- The interaction-graph is defined by the vertices $V = (V_p \cup V_e)$, and the edges $E = e_{i,j}$, $i, j \in (1, 2, \ldots, |V_p \cup V_e|)$. Each bidirectional edge $E = (e_{i,j})$ is derived from an annotated interaction between TFs i and j, based on an interaction database. These edges describe both within-group TF interactions as well as between-group interactions. To obtain the TF interactions, we use protein-interaction information derived from the STRING (*http://string.embl.de/*) and MiMI (*http://mimi.ncibi.org/MiMI/home.jsp*) databases, both of which contain data derived from multiple sources, such as yeast-2-hybrid screens, literature, ChIP etc. Though there is some inherent noise in the accuracy of these high-throughput sources, they permit the use of a confidence threshold to discriminate a potentially true interaction from a spurious one.

Though it would be of great value to use a catalog of gene-specific and tissue-specific regulatory regions (with all possible transcription factors) from which to find such interaction characteristics - such a repository does not yet exist. In this section, we use a few examples (Gata3 OVE, Gata3 KE, Fgf OVE, Mecp2 F21/F6 , Shh FE) of known tissue-specific and gene-specific regulatory elements from literature, as a positive training set. For the negative training set, we consider the set of regions that were reportedly investigated in these transgenic experiments but did not yield gene-specific regulatory activity. Based on which structural metrics are associated with potential regulatory activity for these examples, we will examine if these features are predictive of *Gata2* UGE enhancer behavior, from its interaction-graph.

We have presented a preliminary analysis of enhancer-promoter TF interaction-graphs for some genomic elements with known regulatory or non-regulatory activity ([215], [211], [198], [223]) in Table. 5.3. The table represents the listing of the

structural attributes of these interaction-graphs, following analysis methods from literature ([174], [173], [233]). A brief summary of these attributes are given below. A deeper analysis of other graph topology metrics and their relation to functional enhancer activity is a topic of future interest.

Sequence	Class	Clustering Coefficient	Characteristic path length	Heterogeneity	Centralization	Density
Mecp2 F21[215]	+1	0.208	2.824	0.668	0.184	0.133
Mecp2 F6 [215]	-1	0	1.75	0.342	0.067	0.145
Gata3 OVE [198]	+1	0.036	2.254	0.779	0.359	0.154
Gata3 KE [198]	+1	0.409	2.0	0.813	0.684	0.216
Gata3 NE1 [198]	-1	0.383	2.131	1.139	0.757	0.15
Gata3 NE2 [198]	-1	0.458	2.013	0.872	0.699	0.203
Fgf10 OVE [223]	+1	0.313	2.433	0.72	0.323	0.133
Shh FE [211]	+1	0.394	2.312	0.797	0.49	0.175

Table 5.3: The first column is the various regulatory and non-regulatory elements from literature, the next column corresponds to its class label $(+1/-1)$. The subsequent columns correspond to the attributes of the overall TF-interaction graph (both within-group and between-group interactions).

- Clustering coefficient: In undirected networks, the clustering coefficient C_n of a node n is defined as $C_n = 2e_n/[k_n(k_n-1)]$, where k_n is the number of neighbors of n and e_n is the number of connected pairs between all neighbors of n. Thus C_n, of a node in a graph is the ratio of the number of edges between the neighbors of that node over the total number of edges that could exist among its neighbors. The clustering coefficient of a node is always a number between 0 and 1. The network clustering coefficient is the average of the clustering coefficients for all nodes in the network.

- Characteristic Path length: The length of a path along the graph is the number of hops (or edges) between any two nodes along the graph. Though, there may be multiple paths between two nodes n and m (TFs) along the interaction-graph , the shortest path length $L(n,m) = (L(m,n))$ corresponds to the minimum across these multiple paths. This measure is computed for all pairs of nodes in

the network. The characteristic path length denotes the average shortest-path distance of the graph. This gives the expected distance of any two connected nodes in the graph and is a global indicator of network-connectivity.

- Heterogeneity: Network heterogeneity denotes the coefficient of variation of the degree distribution. A network that is heterogeneous would consist of some nodes that are highly connected (exhibit 'hub' behavior), while the majority of nodes tend to have very few connections. Understanding the heterogeneity of the degree distribution in biological networks is an interesting topic of current research, especially as a way to discover modularity [183].

- Centralization: This refers to the overall connectivity (cohesion) of the graph. It indicates how strongly the graph is organized around its most central point(s). The central point(s) of the graph are the set of nodes which minimize the maximum distance distance from all other nodes in the graph [239]. Networks whose topologies resemble a star/wheel pattern have a centralization close to 1, whereas decentralized networks are characterized by having a centralization close to 0.

- Density: The neighborhood of a given node n is the set of its neighbors. The connectivity of n, denoted by k_n, is the size of its neighborhood. The average number of neighbors indicates the average connectivity of a node in the network. A normalized version of this parameter is the network density $k_n/n(n-1)$. The density is a value between 0 and 1. It is also the average standardized degree. It shows how densely the network is populated with edges (i.e. how "close-knit" an empirical graph is [239], [235]). A network which contains no edges and solely isolated nodes has a density of 0, whereas the density of a clique (completely

connected graph) is 1.

The above mentioned several network properties (as well as clustering coefficients, number of connected components etc.) are examined for the overall interaction-graphs for the reported enhancers from literature [173]. A logistic regression as well as random forest analysis reveals that low values of heterogeneity, characteristic path length and centralization are fairly good predictors of potential enhancer activity. All of these attributes point to the decentralized, homogenous and somewhat tighter connectivity of the interaction-graphs for true enhancers. We note that the OOB error rate of the RF here is about 25%. The quality of this classifier can be expected to improve as we get more data from which to extract features.

We now examine the interaction-graphs for the test set, i.e. the four *Gata2* UGEs. For illustration, we only show the largest connected component of the inter-group edges for each interaction graph (Fig. 5.10).

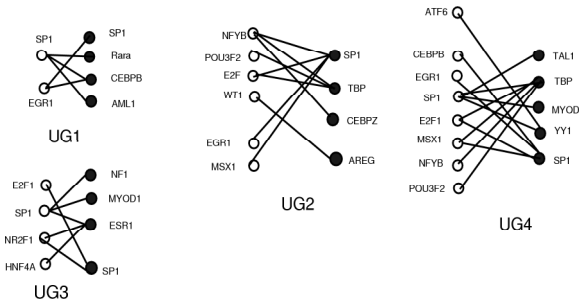


Figure 5.10: Protein-protein interaction between putative *Gata2* TFs (hollow circles) and putative UG element TFs (filled circles). Note: This only shows the connections between two groups for one of the connected components. For our analysis, we consider both *intra-* and *inter-group* connections. From *http://string.embl.de/*

This figure indicates a very interesting property of the real enhancers vis-a-vis the other conserved elements. We see that the TF effectors for *Gata2* such as *SP1*, *POU3F2* (identified in the TF effector network above, Fig. 6.4), are involved in cross-element interactions at the protein level, between the promoter and true enhancer

($UG2/4$). However, the network linkage in the elements that showed no enhancer activity is very sparse suggesting low cross-talk between promoter and enhancer. Also, the TFs at the enhancer nodes (dark circles) have a more uniform degree distribution in the functional elements $UG2/4$ as compared to the non-functional ones. Both these observations suggest lower heterogeneity and centralization of such functional interaction-graphs. Thus, the extent of TF cross-talk is a potential discriminator of possible enhancer function. This shows that superimposing PPI information along with sequence and expression data helps reduce the number of false positives while integrating various aspects of distal regulation.

5.11 Heterogeneous Data Integration and Validation on Gata2 UGEs

As mentioned previously, the primary goal of the various methods developed above is to understand the behavior of known transcriptional elements along different genomic modalities. To validate their predictive potential, we have to demonstrate their application to predicting the behavior of the *Gata2* UGEs (which is our test set). In this section, we present a framework that combines the results of the individual classifiers developed before (kidney-promoter RF, histone RF and interactome-RF) to obtain a integrated prediction. For combining heterogeneous classifiers, we will explore a "probabilistic belief fusion" framework in this paper. Of course, other techniques from literature (like ensemble methods) are also highly amenable for exploration in this context.

The framework involves combining the 'beliefs' of the individual classifiers to obtain a combined belief of prediction. To compute the belief of each classifier we start with examining the confusion matrices for each of the classifiers (kidney-promoter RF, histone-RF and graph-RF), following ([195], [244], [179]). Since each of

the classifiers are random forests, we can obtain their OOB error estimates through these confusion matrices. For the graph-RF, this confusion matrix is as below,

$$\mathbf{CM_{graph-RF}} = \begin{pmatrix} Class & -1 & 1 & class.error \\ -1 & 4 & 1 & 0.20 \\ 1 & 1 & 4 & 0.20 \end{pmatrix},$$

thereby yielding an OOB error estimate of $\sim 20\%$.

Similarly, we have,

$$\mathbf{CM_{promoter-RF}} = \begin{pmatrix} Class & -1 & 1 & class.error \\ -1 & 67 & 19 & 0.22 \\ 1 & 10 & 76 & 0.12 \end{pmatrix},$$

thus yielding an OOB error estimate of $\sim 17\%$.

$$\mathbf{CM_{histone-RF}} = \begin{pmatrix} Class & -1 & 1 & class.error \\ -1 & 134 & 24 & 0.15 \\ 1 & 21 & 119 & 0.15 \end{pmatrix},$$

yielding an OOB error estimate of $\sim 15\%$.

The three random forest classifiers are represented in ROC space (Fig. 5.11). As can be seen, these three classifiers have fairly good performance characteristics. Moreover these are three complementary data sources and can be effectively combined to improve detection reliability. Since they are trained on very different modalities, they can be assumed to be independent.

Each classifier is a function $e_k(x) = j_k$ that maps a data point (x) to the class 'j', with $k = 1, 2, \ldots, K$ and $j_k \in (-1, 1)$. Here, $K = 3$, and $J = 2$ classes.

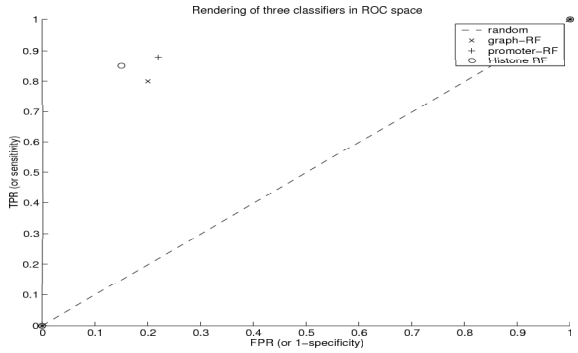


Figure 5.11: Representation of the three RF classifiers in ROC space (RF-promoter in (+), and RF-histone in (o), and graph-RF in ($\times$)). The diagonal line is the classification by random chance.

Thus, the belief of the k^{th} classifier is,

$$bel_k(x \in C_i | e_k(x) = j_k) = P(x \in C_i | e_k(x) = j_k)$$

The overall belief, *bel(i)*, given by,

$$bel(i) = bel(x \in C_i | e_1(x) = j_1, \ldots, e_K(x) = j_K) =$$

$$P(x \in C_i | e_1(x) = j_1, \ldots, e_K(x) = j_K)$$
$$= \frac{P(e_1(x) = j_1, \ldots, e_K(x) = j_K | x \in C_i).P(x \in C_i)}{P(e_1(x) = j_1, \ldots, e_K(x) = j_K)}$$

Further, we have that,

$$\frac{\prod_{k=1}^{K} P(e_k(x) = j_k | x \in C_i)}{\prod_{k=1}^{K} P(e_k(x) = j_k)} = \frac{\prod_{k=1}^{K} P(x \in C_i | e_k(x) = j_k)}{\prod_{k=1}^{K} P(x \in C_i)}$$

Thus,

$$bel(i) = P(x \in C_i).\frac{\prod_{k=1}^{K} P(x \in C_i | e_k(x) = j_k)}{\prod_{k=1}^{K} P(x \in C_i)}$$

(due to independence of the K classifiers,)

In the absence of the posterior probability $P(x \in C_i)$, an approximation is used, leading to [244],

$$bel(C_i) = \frac{\prod_{k=1}^{K} P(x \in C_i | e_k(x) = j_k)}{\sum_{i=1}^{J} \prod_{k=1}^{K} P(x \in C_i | e_k(x) = j_k)}.$$

Note: $J = 2$ and $K = 3$. Depending on the belief value $bel(i)$, the decision rule $(E(x))$ for classifying data point x is,

$$E(x) = j, \text{if } bel(j) = max_i\, bel(i),$$

$$\text{or, } E(x) = j, \text{if } bel(j) = max_i\, bel(i), \text{and, } bel(j) \geq \alpha,$$

where $0 < \alpha \leq 1$, with α being a threshold.

Sequence	True Class	Promoter RF prediction $e_1(x)$	Histone RF prediction $e_2(x)$	Interaction-graph RF prediction $e_3(x)$	P(Class=+1) (Overall Belief)
Gata2 UG1	-1	-1	-1	-1	0.0054
Gata2 UG2	+1	+1	+1	+1	0.9875
Gata2 UG3	-1	+1	+1	-1	0.832
Gata2 UG4	+1	+1	+1	+1	0.9875

Table 5.4: Combined belief generation during heterogeneous classifier integration. The last column represents the combined belief (probability that the UG sequence is an enhancer) as a result of integrating the promoter-RF, histone-RF and graph-RF predictions.

We now show the output classes of each of the 3 classifiers as well as the combined belief on the *Gata2* UGEs in Table. 5.4. More specifically, for the first row in Table. 5.4, the overall belief equation above becomes,

$$bel(ug1 = +1) = \frac{P(ug1 = +1|e_1(x) = -1).P(ug1 = +1|e_2(x) = -1).P(ug1 = +1|e_3(x) = -1)}{P(ug1 = +1|e_1(x) = -1).P(ug1 = +1|e_2(x) = -1).P(ug1 = +1|e_3(x) = -1) + P(ug1 = -1|e_1(x) = -1).P(ug1 = -1|e_2(x) = -1).P(ug1 = -1|e_3(x) = -1}$$

$$\frac{[(1 - prec_{n,1}) \times (1 - prec_{n,2})] \times [(1 - prec_{n,3})}{(1 - prec_{n,1}) \times (1 - prec_{n,2}) \times (1 - prec_{n,3})] + [prec_{n,1} \times prec_{n,2} \times prec_{n,3}]}$$

Here, $prec_{n,k} = \frac{TN_k}{TN_k + FN_k}$. Similarly, $prec_{p,k} = \frac{TP_k}{TP_k + FP_k}$. These are the negative and positive precision values respectively, for the k^{th} classifier. These rates are obtained

from the corresponding confusion matrices shown above. This approach is followed for each of the $UG1-4$ elements.

If we set a threshold of $\alpha = 0.85$ or 0.90, we would get $UG2$ and $UG4$ to be the true enhancers (100% accuracy). However, for a choice of $\alpha = 0.8$, $UG3$ is predicted to be an enhancer in spite of being declared a member of the (-1) class by the graph-RF. This choice of threshold thus determines the performance of the combined classifier.

Under the $\alpha = 0.8$ case, however, the results are not to be interpreted as a 25% error rate since the nature of the test set (*Gata2* UG enhancers) are very different from the training data of each modality (promoters are proximal elements whereas enhancers are distal; histone sequences are for a different cell-context; and interaction-graphs are obtained over different genes). The fact that we are getting such good prediction in spite of the training sets being so different is a strong point in favor of examining and integrating these data sources. The real test-error rates are given by the OOB error estimates of the individual classifiers.

5.12 Summary of Approach

In this work, we have shown that,

- Motif signatures are predictive of regulatory element location. These comprise sequence-motifs derived from tissue-specific gene promoter sequences as well as sequences related to epigenetic preferences during gene regulation.
- Promoter and enhancer TFs that are putatively recruited during gene (*Gata2*) regulation can be identified using a combination of phylogenetic conservation, expression data, and tissue-specificity annotation.
- Effector TFs (via DTI) at the gene proximal promoter have high network linkage

with enhancer TFs in case of functional enhancers. The TF interaction-graphs of truly functional elements are seen to be have a lower centralization, characteristic path length and heterogeneity suggesting higher cross-talk during formation of the transcription factor complex.

These perspectives (based on sequence, expression and interactome data) shed some light on the sequence and mechanistic preferences of true regulatory regions interspersed genome-wide. It is to be noted that this model is data driven and may not directly correspond to the biology of transcription. However, much like markov models for gene sequence annotation, we believe that such data-driven models are useful for model-building during genome-wide study.

5.13 Conclusions

In this work, we have examined the problem of regulatory element identification. Such an effort has implications to understand the genomic basis of key biological processes such as development and disease. Using the biophysics of transcription, this can be modeled as a problem in data integration over various experimental modalities such as sequence, expression, transcription factor binding and interactome-data. Using the case study of enhancers corresponding to the *Gata2* gene, we examine the utility of these heterogeneous data sources for predictive feature selection, using principled methodologies and metrics.

Based on motif signatures, we find that they predict the true enhancers ($UG2$, $UG4$), and the false enhancer $UG1$, but mispredict $UG3$ to be an enhancer. However, a mechanistic insight that analyzes enhancer behavior based on the interactions between distally and proximally recruited transcription factors can greatly improve on prediction accuracy. Additionally, combining heterogeneous classifiers based on

multiple data modalities yields an improved accuracy of prediction.

The novelty of the proposed work spans several areas. Firstly, data sources that are relevant to understand the mechanism of gene regulation (with *Gata2* as an example) have been identified. We have developed methods that reconcile the behavior of known regulatory elements along each of these modalities. The kidney-promoter based classifier aims to discover sequence preferences of kidney-specific regulatory regions. The utilization of histone-modified sequences and their exploration for sequence motifs are indicative of epigenetic-preferences and nucleosome-occupancy patterns. This has not been explored before in the realm of LRE characterization. The use of DTI as a metric to infer putative TF to target-gene influence is a recent one that serves to integrate phylogenetic TFBS conservation along with expression data. Finally, the utilization of graph-based analysis techniques to understand the "structure" of the TF interaction-graph between enhancer and promoter helps us understand true enhancer behavior from a mechanistic viewpoint. The probabilistic combination of multiple classifiers (each deriving from a unique data resource) aims to reconcile the behavior of existing enhancers along multiple modalities. We hope to demonstrate that a principled integration of non-overlapping genomic modalities can be used to interpret the context and specificity of gene regulation.

5.14 Future Work

Some key elements directly emerge for guiding future research. As already alluded to in the motif-signature procedure, specific expression data corresponding to stages and tissues of interest would greatly improve the specificity of regulatory element prediction. Furthermore, as histone modification maps for different cell lines are generated, the false positive rate of prediction would decrease, thereby improving

accuracy. Several other learning paradigms can be introduced into this setting, since we are learning from structured data. Also, methods in joint classifier and feature optimization might likely improve the accuracy of predictions. Additionally, methods that analyse the grammar of these cis-regulatory regions (LREs) and look for motif position, spacing and orientation will be of great utility.

At the expression level, methods for supervised network inference would have a great impact on the discovery of TF effectors. Rapid advances have been made in this area and their relevance to the biological context of the problem has become very principled. At the interactome level, the work presented here can be extended to the investigation of graph-clusters for weighted interaction-graphs. The weighted edges are obtained from the confidence of the individual data sources, as well as the number of species over which that particular edge is conserved ([174], [237]). Such analysis enables the discovery of subgraphs of various degrees of inter-connectedness, thereby discovering functional "graph-motifs".

5.15 Acknowledgements

The authors gratefully acknowledge the support of the NIH under award 5R01-GM028896-21 (J.D.E). We would like to thank Prof. Sandeep Pradhan, Mr. Ramji Venkataramanan and Mr. Dinesh Krithivasan for useful discussions on directed information. Ms. Swapnaa Jayaraman at the University of Michigan is gratefully acknowledged for discussions about network-attributes. We are grateful to Prof. Erik Learned-Miller and Dr. Damian Fermin for sharing their code for high-dimensional entropy estimation and ENSEMBL sequence extraction, respectively.

CHAPTER VI

Some Other Ideas

6.1 Various Ideas

This chapter is meant to be a synopsis of several ideas that were developed in the course of this project. Many of these are potentially useful in that they lead to particularly interesting research directions and improvements to the existing model.

6.2 The story thus far

Based on inter-species conservation, RP scores, TFBS clustering, tissue-specificity of co-clustered TFBSes (i.e., all the knowledge and data prevailing up to that time), we hypothesized the existence of some candidate enhancers in the 150kb UG and 45kb SA regions respectively. One of the candidates in the UG region indeed has *Gata3*-specific regulatory activity, but in the inner ear – and has been named the otic vesicle enhancer (OVE). Similarly, one of the candidates in the SA region has *Gata3* activity in the pylorus – and has been named the pyloric enhancer (PE). The OVE and PE discoveries were made by Dr. Kim Lim and Dr. Takashi Moriguchi respectively, of the Engel laboratory. The positions of these regions along the 150kb UG regions and 45kb SA region is presented in Figs. 6.12 and 6.14 below.

However, since then there have been other data sources pertaining to tissue-specificity of TFs, characterization of TF families, availability of limited histone mod-

ification data (ENCODE), as well as more detailed protein-protein interaction data via STRING (*http://string.embl.de/*) and MiMI (*http://mimi.ncibi.org/MiMI/home.jsp*) databases. We have attempted to reconcile the behavior of existing *Gata2* and *Gata3* enhancers (UG2, UG4, UGE, OVE and PE) as well as some of those that did not work in the context of these new data sources. Some of this has been explained in Chapter 5 – where we have examined the UG2 and UG4 enhancers of the *Gata2* gene [204].

Based on our findings from Chapters 2-5, as well as some modifications explored in this chapter, we are in the process of generating and validating new UG/SA/T-cell-specific candidates regulating *Gata3* expression. This work is currently underway at the Engel laboratory and we are optimistic about the postulation of a suitable model that enables gene-dependent discovery of enhancers that confer stage-specific expression.

Below, we outline some of our recent findings in relation to the various aspects of our "continuously-evolving" model of enhancer behavior.

6.3 TF Modules

One of the earliest strategies to find transcriptional effectors for a gene is to look for the presence of common TF modules in the promoters of co-expressed genes ([216], [209]). For the case of kidney expression of the *Gata2* and *Gata3* genes, we aim to find TF sets that govern the co-ordinated regulation of genes that are spatio-temporally co-expressed with *Gata2* or *Gata3*. To this purpose, we have used Toucan tool, at (*http://homes.esat.kuleuven.be/s̃aerts/software/toucan.php*) for this purpose and examined variants of the module TF approach under various scenarios – this is explained further below.

6.3.1 TOUCAN results on *Gata2* and *Gata3* expression

NCBI DAVID (*http://david.abcc.ncifcrf.gov/home.jsp*) is used to convert these MGI gene names to ENSEMBL identifiers (*http://www.ensembl.org/*). These ENSEMBL identifiers are the inputs to TOUCAN tool, wherein the promoter sequences from the ENSEMBL database are extracted, using sequence 2000bp upstream and 1000bp downstream up to the first exon relative to the transcriptional start site ('First Exon' option) reported in ENSEMBL (release 37). Note: in concordance with newer observations, one can modify the search to look at the -300bp region to be the promoter. Using the MotifScanner tool, we can look for statistically over-represented TFBS motifs from the TRANSFAC/JASPAR databases. After the TF motifs are located, ModuleScanner is used to locate modules of transcription factors (of size 5, with partial overlap). According to the tool, this search enables the de-novo discovery of regulatory modules of TFs that might underlie the co-ordinated regulation of this co-expressed gene set. As is clear from the presented approach, the degrees of freedom within which to do a TF search is biased towards motifs that have to be over-represented in each and every promoter relative to the background set, thereby losing out on motifs that have only subtle over-representation. Also, the biological basis for co-ordinated regulation of co-expressed genes only applies to those genes that have some intrinsic relationship amongst them. Thus the same analysis, presented above, would be far more meaningful for a subset in this co-expressed gene set. To discover such clusters, we use *apriori* knowledge of gene-gene relationships from GO ontology.

6.3.2 Co-embedding gene expression data based on GO ontology (BP)

The main idea here is to find sets of genes that are coupled not only based on co-expression, but also their 'proximity' in ontology space. The rationale behind this approach is that genes that have an apriori knowledge of being "coupled" with *Gata2/Gata3* should be weighted higher when looking for transcriptional machinery that co-ordinately co-regulates their joint expression. This also increases the degrees-of-freedom among the set of possible transcription factors in the regulating module. In an exploratory scenario this is preferable. This approach can be generalized to any combination of weighting matrices (phylogenetic conservation, QTL etc.).

6.3.3 Part I: Building Realism while Clustering

As suggested in the previous sections, it would be useful to have a "space" which respects biological process proximities in addition to expression similarities. This can be enabled by considering a set of annotations that describe the "biological process" (BP) information for each of the genes (in the nephrogenic differentiation program). One set of annotations that is well researched by the bioinformatics community is the Gene Ontology (GO) descriptors (*http://www.geneontology.org/*). This is a controlled hierarchical vocabulary that annotates genes in various organisms by cellular component (CC), molecular function (MF) and biological process (BP), based on literature reports.

The next section examines the generation of a "semantic similarity matrix" between genes based on their GO (BP) descriptors, to quantify the cellular proximity among them. Just like lexical word ontologies for spoken languages (e.g. WordNet at *http://wordnet.princeton.edu/*), this structure imposes a tree structure on the various GO terms, thereby expressing the similarity between any two terms in the

ontology as a function of their parents in the ontology tree.

The next step involves the use of manifold embedding techniques that can integrate such GO similarity along with expression-level similarity to construct an embedding of the genes as points in some space. One such technique is Laplacian Eigenmaps [245], also profiled in Section: 6.3.3 that approximate both these relationships (semantic and expression). This is a generalization of the principal component approach in that the distance measures on such manifolds are not necessarily euclidian.

We remind the reader that the main goal here is to embed genes based on their expression profiles, but additionally weighted based on their BP proximity - this would be more biologically relevant for the discovery of true biological activity. We believe that such an approach is consistent with the rationale of using integrative genomics or principled data integration for stronger hypothesis generation [**?**].

GO Semantic Similarity

To quantify the notion of similarity of terms along an ontology, we appeal to a vast amount of literature that addresses such questions [246]. The semantic similarity of any two GO terms along the ontology hierarchy is based on the number of shared parents and the information content of the individual GO terms (measures: Jiang Conrath, Resnik etc.). Based on the literature, we use the Jiang-Conrath similarity measure, given by,

$W_{i,j} = sim(c_i, c_j) = \frac{1}{jc_{dist}(c_i,c_j)}$, with $jc_{dist}(c_i, c_j) = 2log(p(lso(c_i, c_j))) - [log(p(c_i)) + log(p(c_j))]$

where c_i and c_j are two terms (nodes) along the GO ontology tree ($i, j \in \{1, 2, \ldots, 14\}$).

$lso(c_i, c_j)$ refers to the the information content of the last common parent of these two nodes. The information content is computed based on the probabilities of observing the individual nodes and their last common ancestor in an overall corpus.

For the 14 genes profiled in this study, we use the R package "GOSim" to obtain the semantic similarity matrix (size 14×14) based on GO BP annotation (this information can be found both along mouse and human annotations). This similarity matrix is used to obtain the weight matrix W during the Laplacian Eigenmap embedding procedure [245] below.

LLE (Laplacian Eigenmaps)

- Build the $K \times K$, $(K = 14)$ dimensional weight matrix W from the Gene Ontology ("Biological Process") terms of the genes in the dataset. This distance is the "normalized" semantic similarity alluded to above (section 6.3.3).

- Assign weight $W_{i,j}$, from (1) for each gene pair (i, j), for each of the $\binom{K}{2}$ gene pairs. *Note*: The higher this weight, the closer the genes are.

- Find n nearest neighbors using the euclidian distance in principal component space. The scores of the functional data along the first two principal components can be interpreted as co-ordinates in a euclidian space.

- Form the Graph Laplacian:

$$L_{i,j} = \begin{cases} d_i = \sum_k W_{i,k} & \text{if } i = j; \\ -W_{i,j} & \text{if } i \text{ is connected to } j; \\ 0 & \text{otherwise.} \end{cases}$$

- Solve: $min_y y^T L y = \frac{1}{2} \sum_{i,j} (y_i - y_j)^2 W_{i,j} \quad (2)$,

  subject to:

  - $y^T D y = 1$, and

  - $y^T D \mathbf{1} = 0$,

  where $D_{i,i} = \sum_j W_{j,i}$, a diagonal weight matrix.

- Embed the co-ordinates to a lower dimensional manifold, using the solution (the Laplacian Eigenmap) obtained from the minimization above.

  - The solution to (2) is given by the d generalized eigenvectors associated with the d smallest generalized eigenvalues solving $L\mathbf{y} = \lambda D\mathbf{y}$ (neglecting the zero eigenvalue and its eigenvector).

  - If $\mathbf{y} = [y_1, \ldots, y_d]$ is the collection of these eigenvectors, then the embedding is given by :

    $y_i = (y_{i1}, \ldots, y_{id})^T$, i.e., the d dimensional representation of the i^{th} data

point (gene).

- In our representation, we take dimensionality, $d = 2$ and number of neighbors, $n = 5$. The final embedding of the functional data based on expression and BPmodalities is shown in Fig. 6.2.

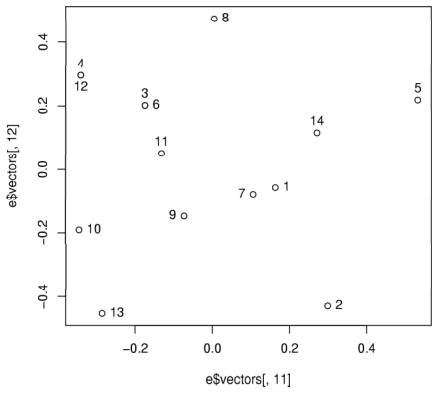


Figure 6.1: Manifold embedding various kidney-specific genes (MGI, e12.5) without GO weighting.

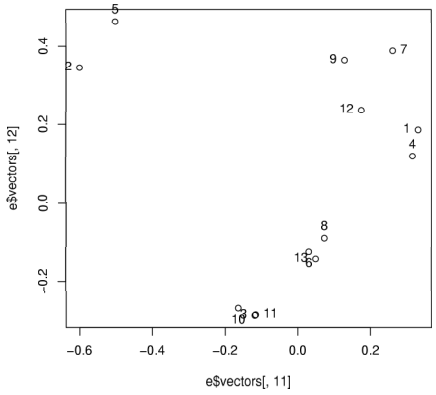


Figure 6.2: Manifold embedding various kidney-specific genes (MGI, e12.5), using GO BP similarity (Mm).

We note that we can also use a literature/prior-knowledge based weighting matrix

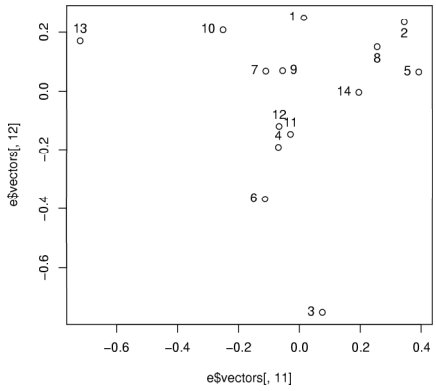


Figure 6.3: Manifold embedding various kidney-specific genes (MGI, e12.5), using GO BP similarity (Hs).

to convey similarities between genes in space. We can also combine several *weight matrices* from various modalities to obtain a combined weight matrix W_{ij} for use during embedding.

Based on the above procedure, we make the following observations:

- Neighbors of *Gata3* without GO-embedding, Fig. 6.1 : *Lamc2*, *Ret*, *Cldn7*, *Mapk1*, *Pax2*.
- Neighbors of *Gata3* in GO-embedded space (Mm), Fig. 6.2 : *Lhx1*, *Pax2*, *Ret*, *Wnt11*, *Mapk3*.
- Neighbors of *Gata3* in GO-embedded space (Hs), Fig. 6.3 : *Pax2*, *Ret*, *Lhx1*, *Mapk3*, *Wnt11*.

To find transcription factor effectors that potentially co-regulate co-expressed genes, we only consider those genes that are within a finite neighborhood of *Gata3*. Using TOUCAN tool (*http://homes.esat.kuleuven.be/ saerts/software/toucan.php*), we can look for TFBS modules that are only over-represented in this neighbor-subset. This increases the degrees of freedom for the problem and enables the prospective

nature of this approach to find transcriptional effectors (we need as many candidates that we can find). Below, we indicate the TFs that are discovered under the scenarios with and without embedding. As mentioned in other studies, this is an exploratory exercise and the tissue-specificity and biological plausibility of these TFs as effectors needs to be confirmed.

JASPAR:

1. All genes: GKLF, HFH3, HMGIY, PAX4, RREB1, HFH2, IRF1, COUP.
2. 5 neighbor subset: NRF2, MYF, HEN1, GKLF, AP2gamma, HMGIY, c-FOS, Staf.

TRANSFAC:

1. All genes: SP1, AP2gamma, SP1, ATF, AP2, PAX5, CREB, PAX4, AP2alpha.
2. 5 neighbor subset: HEN1, AP2alpha, AP4, PAX4, AP2, LMO2COM, SP1.

For the SA case: *Gata3* (14462), *Trim37* (68729), *Th* (Mm: 21823), *Tle4* (21888) are the co-expressed genes in day $e14.5$ of the adrenal medulla.

Toucan analysis: We imported sequences based on Entrez ids, used motif scanner from the TRANSFAC/JASPAR databases and 3rd order markov model of mouse proximal promoters as background; after getting TFs, looked for module enrichment using Module Searcher - GA algorithm, 5 elements per module and 10 top scoring modules. Based on this setup, the set of over-represented module TFs are:
HNF1, OCT1, NFY, SP1, NKX22, NKX61, USF, RREB1, CEBP, LHX3,PAX4, AHR-ARNT, HFH3, MRF2, BRN2, E2.
Since there is no public expression data that is available for the developing SA system, and hence we cannot do the embedding type analysis here. However, as such data becomes gradually available the above analysis can be done for this case too.

It is important to realize two aspects:

- The results for the *Gata3* case above convey a need to work with TF families rather than their members since these are more generic and can enable the examination of experimental data to ascertain their tissue-specificity.
- Some of these are tissue-specific whereas some are statistically abundant (like SP1), hence one needs to look for biologically meaningful subsets that increase the degrees of freedom and hence the number of candidate TFs. The GO ontology based embedding approach helps with this need.

6.4 Using Sparsity-penalized Regression for Inferring TF-gene dependencies for *Gata2* and *Gata3*

As seen in the previous chapters, we have tried to characterize the transcription factor effectors underlying *Gata2* and *Gata3* expression in the developing kidney. This approach, using the directed information criterion, serves to integrate sequence information with expression data.

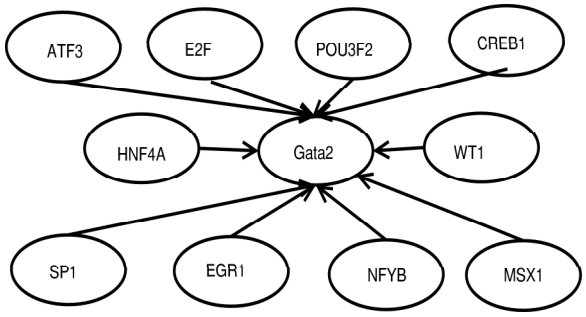


Figure 6.4: Putative upstream TFs using DTI for the *Gata2* gene.

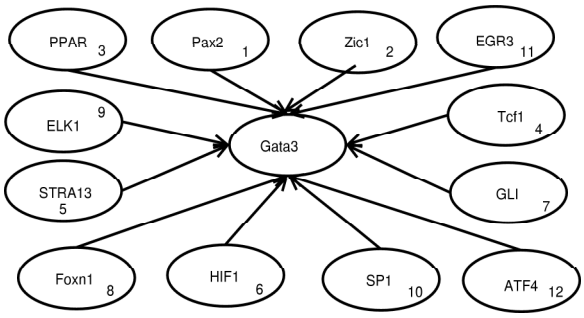


Figure 6.5: Putative upstream TFs using DTI for the *Gata3* gene.

As a means to reduce the number of candidate effectors even further, we explore a linear regression approach among the high DTI TFs while incorporating a sparsity constraint on the total number of possible effectors. Following, ([248], [249]), we explore the LASSO approach for such sparsity penalized regression.

Briefly, LASSO (least angle shrinkage and selection operator) is a method of penalized regression. Like in ordinary least squares (OLS) regression, the objective is to minimize the sum of squared errors, but with a penalty on the number of non-zero regression coefficients,

$$SSE_\lambda \beta = \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p-1} X_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

In our setting, Y_i corresponds to the expression of *Gata3* at the i^{th} time instant of the expression profiling. X_{ij} corresponds to the expression of predictor X_j at the i^{th} time instant. Each of the $(p-1)$ predictors correspond to the TF effectors identified via DTI above. β_0 corresponds to the intercept term of the regression equation. λ is the tuning parameter; $\lambda = 0$ is equivalent to OLS regression, whereas $\lambda = \infty$ implements the sparsity constraint. We once again note that this sparsity-constrained regression formulation serves to understand target gene expression (*Gata3*/*Gata2*) as a *joint* function of the various predictors. Whereas, the DTI based approach

treats each predictor as independent - the LASSO-regression approach can combine the effects of the various predictors for appropriate biological inference.

The approach is detailed in REF and is outlined below. More information is also available at: *http://www-stat.stanford.edu/tibs/lasso/simple.html*. One approach for solution of the lasso regression problem is the forward stepwise regression algorithm:

- Initialize each of the regression coefficients β_j to zero.
- Find the predictor X_j most correlated with the response, Y, and add it into the additive model.
- Find the residuals $r = Y - \hat{Y}$. At each step, add the predictor that is most correlated with the residual, r to the model .
- Iterate until all predictors are in the model.

An extension of the above approach is the least angle regression (LAR) procedure ([248], [249]). Applying this method to the predictors identified in Figs. 6.4 and 6.5, we get the regression models as:

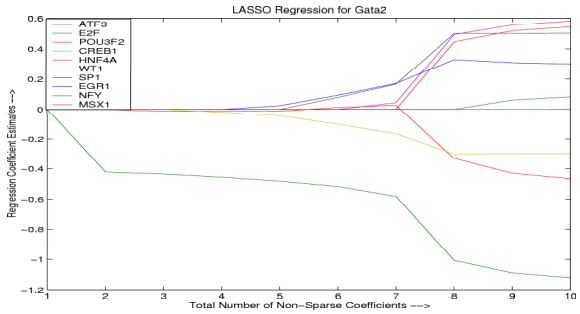


Figure 6.6: Path for the LASSO-regression of *Gata2* along its DTI predictors of Fig.6.4.

Thus, from Fig. 6.6, the primary effectors of *Gata2* are found to be: E2F, POU3F2, HNF4A, WT1, SP1, EGR1, MSX1. The contributions of the various effectors is given in Fig. 6.6.

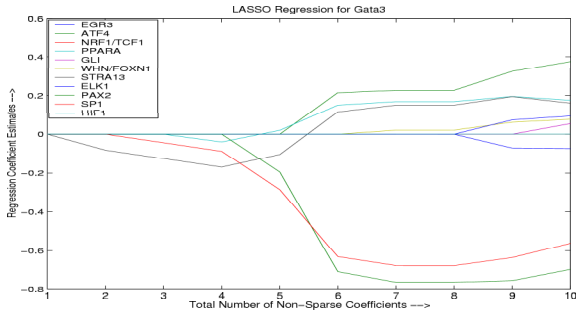


Figure 6.7: Path for the LASSO-regression of *Gata3* along its DTI predictors of Fig. 6.5.

For *Gata3*, the joint effectors are found to be: PPAR, PAX2, ELK1, STRA13 and SP1. Their contributions are given in Fig. 6.7.

6.5 Understanding variation in cis-regulatory regions

6.5.1 SNP TFs in Promoter

Another strategy to enable the discovery of possibly "functional" TFBS in *Gata3* (or any other gene) regulation is to examine the TF sites interrupted by regulatory SNPs associated with diseases in which *Gata3* is implicated (with phenotype in the tissue of interest). For example, *Gata3* haploinsufficiency causes human HDR (hypoparathyroidism, sensorineural deafness, renal anomaly) syndrome. An examination of HDR in the dbSNP database reveals ~ 300 SNPs interspersed on chr:6. In Figs. 6.8- 6.11, we have indicated the set of putative TFs that are likely interrupted by SNPs and are therefore possibly functional.

6.5.2 SNP TFs in Enhancer(s)

The same kind of analysis can be pursued for the various enhancers listed in Chapter 5 – however, we note that this exploratory strategy only serves to increase the confidence in some of the effector TF since they are conserved, have tissue-specific expression in the kidney and are putatively associated with the gene via

disease linkage (eg: *Gata3* and HDR). For some of the enhancers listed in Chapter 5, we examine if there are any rSNPs in these regions that interrupt TF binding sites (Table. 6.1).

Table 6.1: rSNPs in TF families within some enhancers.

Enhancer name	Description (hg18 position)	TF mutation/rSNP (up to Mm)
Gata2 UG1	chr3:129535922-129536271	None
Gata2 UG2	chr3:129518788-129520320	MH1, zf-Dof
Gata2 UG3	chr3:129479811-129484592	PAX, GATA, MH1, ForkHead, HMG
Gata2 UG4	chr3:129579673-129581942	CTF, MH1, GATA
Gata3 UGE	chr10:7983218-7983638	None
Gata3 OVE	chr10:8726703-8727080	HLH, Homeobox, CTF
Gata3 PE	chr10:8675225-8677369	Homeobox

6.6 Looking for TFBS families in CSEs

A predominant concern during the use of databases like TRANSFAC and JASPAR is the high redundancy of the motifs listed therein. A consequence of this is that several motifs from the same family show up as binding on the CSE thereby increasing the spce of transcription factors that needs to be assayed for function (either computationally or experimentally). Towards addressing this problem, JASPAR makes available a list of TFBS family motifs which are useful for an unbiased search across all conserved sequence elements. However, because the ease of use of these matrices is still not favorable, we utilize a set of 36 TFBS family motifs that are made available from the WebMotifs program [29] of the Fraenkel lab at MIT (*http://fraenkel.mit.edu/webmotifs/fbps.html*). The binding of such TF families for

the UG/SA and T-cell regions of *Gata3* expression are shown below Figs. 6.12 - 6.14.

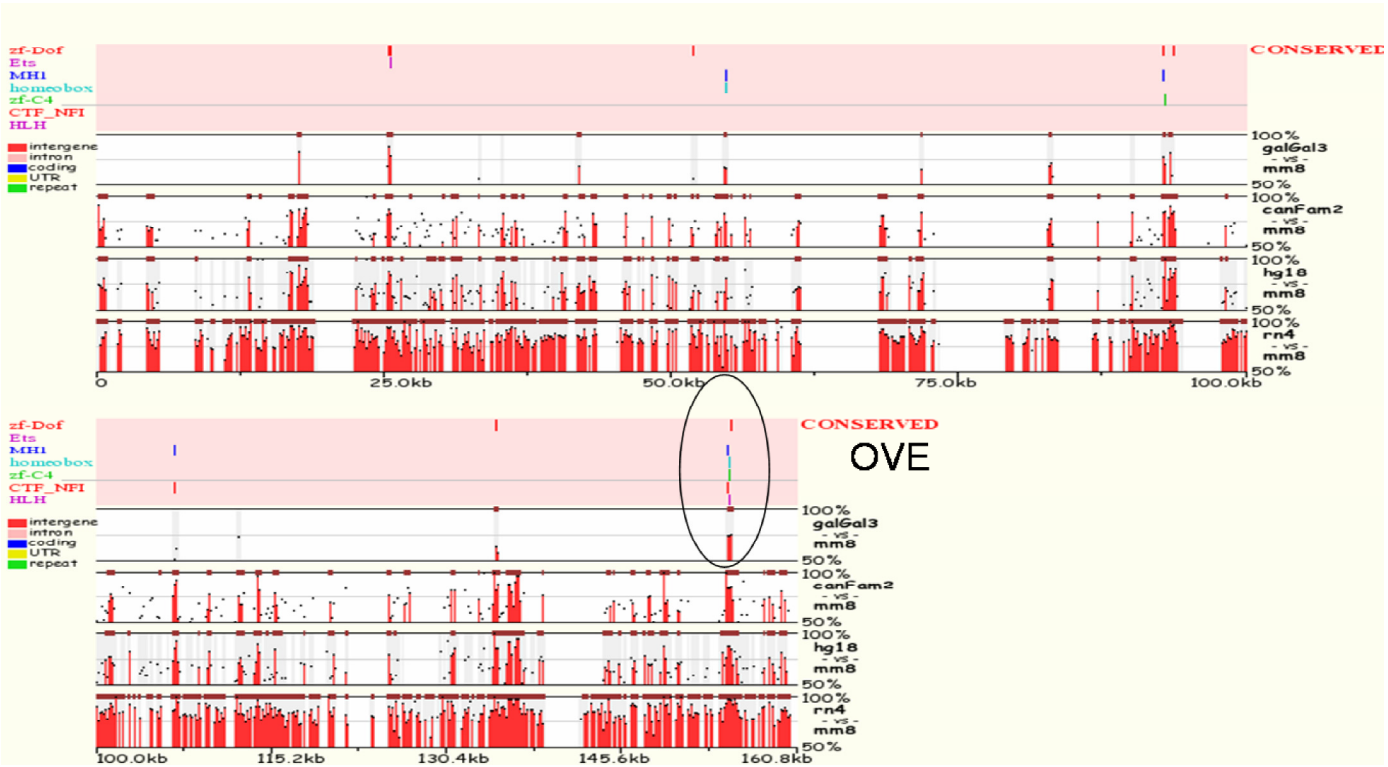


Figure 6.12: Putative TF families in the 160kb UG region.

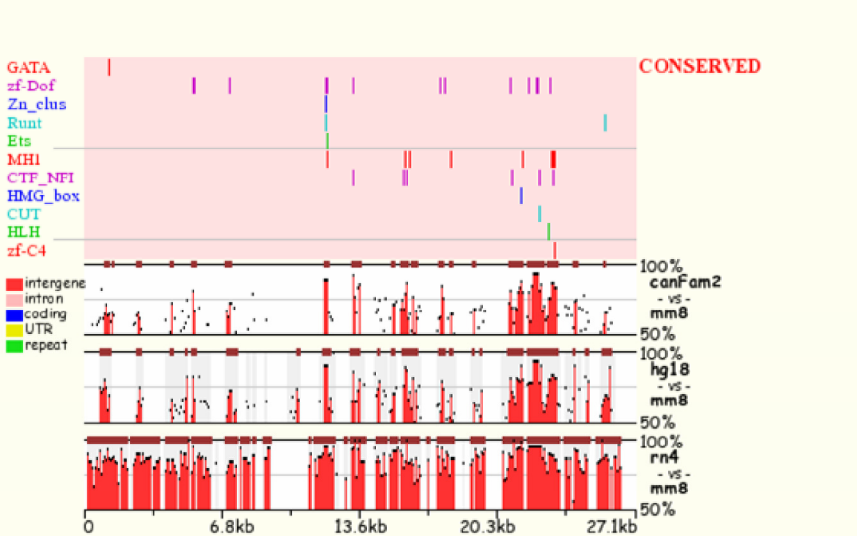


Figure 6.13: Putative TF families in the 27kb T-cell region.

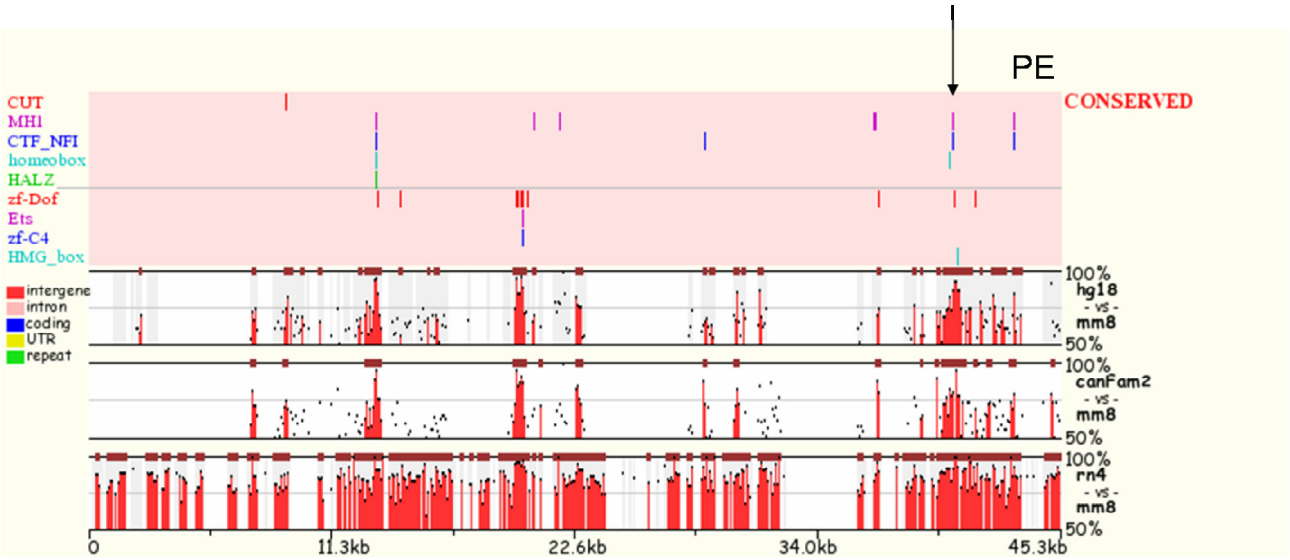


Figure 6.14: Putative TF families in the 45kb SA region.

TF symbol	Motif	Description
AP2	GGgAwNyGTGy	AP2-domain
bZIP	sNTGACGy	bZIP (Basic Leucine Zipper)
bZIP_Maf	GCtgaGTCA	bZIP_MAF (Basic Leucine Zipper/Maf Extended Homology Region transcription factor)
CBFB_NFYA	CCAAysrg	CBFB NFYA (CCAAT-binding transcription factor subunit B)
CTF_NFI	tTGSCNNN	CTF NFI (CCAAT-box-binding factor /Nuclear Factor I)
CUT	NNATyGRT	CUT
E2F_TDP	GCGssAAa	E2F_TDP (E2F/DP family winged-helix DNA-binding domain)
Ets	smGGAagy	Ets (Erythroblast Transformation Specific Domain)
Fork_head	rYAAACAa	Fork_head
GATA	NmGAyArG	GATA (GATA zinc finger)
HALZ	AATNATTG	HALZ (Homeobox associated leucine zipper)
HLH	sNCrsGTG	HLH (Helix-loop-helix DNA-binding domain)
HMG_box	AACAAwRr	HMG_box (High Mobility Group box)
HNF-1_N	gNyRAwNATTAAC	HNF-1 N (Hepatocyte nuclear factor 1, amino terminus)
homeobox	TAAKKrss	Homeobox

continued on next page

continued from previous page

TF symbol	Motif	Description
HSF_DNA-bind	GAANNYTCkmG	HSF DNA-bind (Heat Shock Factor-type DNA-binding)
IRF	gAAANyGAAAs	IRF (Interferon regulatory factor transcription factor)
MH1	tGGCwNNN	MH1 (MAD homology 1 domain)
Myb_DNA-binding	yAACsGNc	Myb DNA-binding (Myb-like DNA-binding domain)
Myc_N_term	CACGTGsNN	Myc N term (Myc amino-terminal region)
PAX	raSCgKGrm	PAX (Paried box domain)
PBX	tGATTGAT	PBX (PBC)
POU	ATGCAAAT	POU
RFX_DNA_binding	GTTGCcrNGNNrm	RFX DNA binding (Regulatory Factor binding to X box DNA binding domain)
RHD	GGrAaNyCCc	RHD (Rel Homology Domain)
Runt	yTGyGGTN	Runt
SRF-TF	CCwwAwaTrG	SRF-TF (Serum Response Factor-type transcription factor)
STAT_bind	TTCyNGGAA	STAT bind (Signal Transducers and Activators of Transcription protein

continued on next page

continued from previous page		
TF symbol	Motif	Description
		DNA-binding domain)
TBP	GNATATAwA	TBP (Transcription factor RFIID/
		TATA-binding protein)
TEA	GGAATGNrr	TEA (Transcriptional Enhancer Activators
		/ATTS domain family)
TF_AP-2	GsSwssgss	TF AP-2 (Transcription Factor Activator Protein 2)
TF_Otx	kgrGaTTAgtg	TF Otx (Otx1 Transcription Factor)
WRKY	cgGtCamcg	WRKY
zf-C4	NNrGGTCA	zf-C4 (Zinc Finger, C4 type/Nuclear Hormone Receptor)
zf-Dof	NNNwAAAGN	zf-Dof (Dof domain, zinc finger)
Zn_clus	CGGNNgNN	Zn Clus (Fungal Zn(2)-Cys(6) binuclear
		cluster domain)

Table 6.2: Functional annotations of some of the transcription factor families from WebMotifs.

6.7 Other directions:

A tissue-specificity index has been formulated using results from the DiRE (distant regulatory elements of co-regulated genes) (*http://dire.dcode.org/*) and the Enhancer Identification (EI) tools [228] from dcode (*http://www.dcode.org/*). However, this treatment needs to be extended to TF families rather than family members.

Based on the observation that tissue-specific TFs are recruited at enhancer regions during spatio-temporal-specific expression, we are formulating a strategy that examines TF specificity in conjunction with their co-ordinated expression and interaction during regulation. This is summarized below,

1. Find the most coherent set of proteins at the promoter that can explain the co-ordinated expression for stage-specific co-expressed genes.

2. Find the most coherent protein set that is tissue-specific at the enhancer.

3. Find the most coherent protein set at the enhancer that is co-enriched with the corresponding module TF set at the promoter – possibly via bipartite graphs across promoter and enhancer.

We call this approach Protein Set Investigation (Ψ) – an extension of Gene Set Enrichment Analysis (GSEA), in conjunction with the interactome level information from MiMI. However, unlike traditional GSEA which only has a validation phase (since gene sets are defined *apriori*), this method has both a discovery and validation component/phase.

The heuristic/schematic idea for Ψ is the following:

- TF family wise search; find top 3 ts members of each family identified in CSE/promoter region; build maximal spanning subgraph across these cliques. Note that these cliques are derived from MiMI and can have several intermediate proteins that are not in the original query set (upto two hops away).

- Can find clique at promoter between module TFs; clique between TFs at enhancer candidates; and the clique from "both" module TFs and CSE TFs to find co-enrichment. The module TFs should be clearly well represented in the highest scoring graph.

- Look for size of span of the resulting graph, the overall tissue-specificity score across all nodes in clique, and interconnectedness of the maximal graph for a way to find the "best" graph.

- All the intermediate nodes identified can be examined for differential expression at mRNA level; pathway enrichment and so on – thereby finding the enrichment of new protein sets for various scenarios (a generalization of GSEA [247], without *apriori* defining gene sets).

- Validate on the set of tissue-specific enhancers (*Gata2*, *Gata3*, *Mecp2*, *Shh*, *GLI3*, *Fgf10*); as well as on the module TF set from biologically relevant sets of co-regulated genes [83].

Based on [228], we obtain tissue-specificity scores for various TFs in the rVista TRANSFAC database (*http://www.dcode.org/EI/*, and *http://dire.dcode.org/*). Each TF (TF_i) is characterized by its coverage C_{TF_i} (i.e. the number of tissue-specific genes in whose promoter TF_i was found) and variable importance, V_{TF_i} (the ability of the TFBS motif to discriminate the set of SymAtlas annotated, tissue-specific genes from a background set of random genes).

For our purpose – we examine each CSE for the tissue- specific enrichment of the various TFBS that putatively bind the sequence. If the CSE binds K TF clusters, then we have the tissue-specificity score of the CSE to be $ts_{CSE} = \frac{\sum_{i=1}^{K} C_{TF_{(i)}} \times V_{TF_{(i)}}}{K}$, where $TF_{(i)}$ is the highest scoring TF (in tissue-specificity) at cluster $i, i = (1, 2, \ldots, K)$. We note that since most sequences have clusters of binding sites, TF_i corresponds to the highest scoring TF among all the members of the cluster. This is mainly due to the fact that at most one TF can bind at any cluster site.

Based on this formulation, we have attempted to reconcile the behavior of some enhancers and neutral regions along some common characteristics: conservation, RP score, TFBS family clustering, tissue-specificity score, enhancer-promoter graph, H3K4 modification.

As can be seen, the tissue-specificity score proposed is insufficient to predict which

CSE is most likely functional (how about combination with RP score and inter-species conservation, put that in table too). Need to combine with interactome analysis, also move on to TFBS family analysis. We note that the set of genes from which these tissue-specificity scores are derived come from SymAtlas, and has gene expression studies mostly for adult tissue. This approach, therefore, needs extension to embryonic stage-specific co-regulated genes (from MGI for example), in which case it becomes another way of doing module TF discovery.

Another approach that has recently been explored for regulatory region discovery is the prediction of nucleosome free regions along DNA using a "nucleosome positioning code" ([252], [251]). However, for efficient translation to the eukaryotic domain these methods will need t be adjusted for cell context. One such recent tool is at: *http://compbio.med.harvard.edu/nuScore/*.

6.8 A 'protocol' for the discovery of putative long-range, promoter-specific regulatory elements (LREs) from sequence

As alluded to in the previous chapters, we extract the following "features" at the promoter and enhancer and then combine these scores across sequence, expression and interactome modalities to obtain a combined prediction for potential enhancer activity.

I. At the promoter,

- Sequence Perspective Find all candidate transcription factors (preferably families) using program of choice: ECR Browser, MatInspector, MAPPER.

- For each potential TFBS family, find putative effector TF using directed information over the kidney-expression data. Augment this set with,

    1. Components of basal transcriptional machinery, such as mediator com-

plexes, histone modification complexes (HDAC/HATs/Swi-Snf), Pol II.

2. SNP TFs (related to possible tissue-specific cis-regulatory variation).
3. Module TFs, obtained by examination of transcription factors that are over-represented in co-expressed genes (either directly from MGI or from the GO similarity based embedding presented above).

note that TF effectors should be on non-overlapping regions of the DNA segment being analyzed.

II. At the putative regulatory element (identified via a reasonably *low* interspecies sequence-identity threshold, $\sim 70\%$),

- Sequence Perspective: For each *candidate* regulatory region within the contig/locus of interest,

  1. Find all candidate transcription factor binding sites (TFBS) using program of choice: ECRBrowser, MatInspector, MAPPER. Alternatively, examine the TF families that bind in the overlapping region (from BAC transgenics). Using PfaM and MGI annotations, examine the tissue-specific expression of each of the family members in the tissue of interest (kidney metanephros, T-cell or adrenal medulla). The CSE with the higher number of non-overlapping tissue-specific TFs ranks higher in the list.
  2. For each potential TFBS, find tissue-specific transcription factors based on annotation (MGI, UNIPROT). Augment this set of TFBS with those that are potentially disrupted by SNPs (*http://cisreg.ca/RAVEN*).
  3. Find the scores for each region based on the kidney-specificity and histone-modification based classifiers.
  4. Find the RP score for each region.

- Combine scores across multiple modalities based on bayesian model averaging or belief combining.
- Examine bipartite graph topology using TF interaction graph between putative effector TFs at promoter and putative TFs at enhancer (Cytoscape). Certain graph characteristics are related to true enhancer activity (as in Chapter 5).
- Additionally, find the interactome level graph that spans the maximal nodes in the sets of non-overlapping TFs at the promoter and enhancer, with highest tissue specificity.

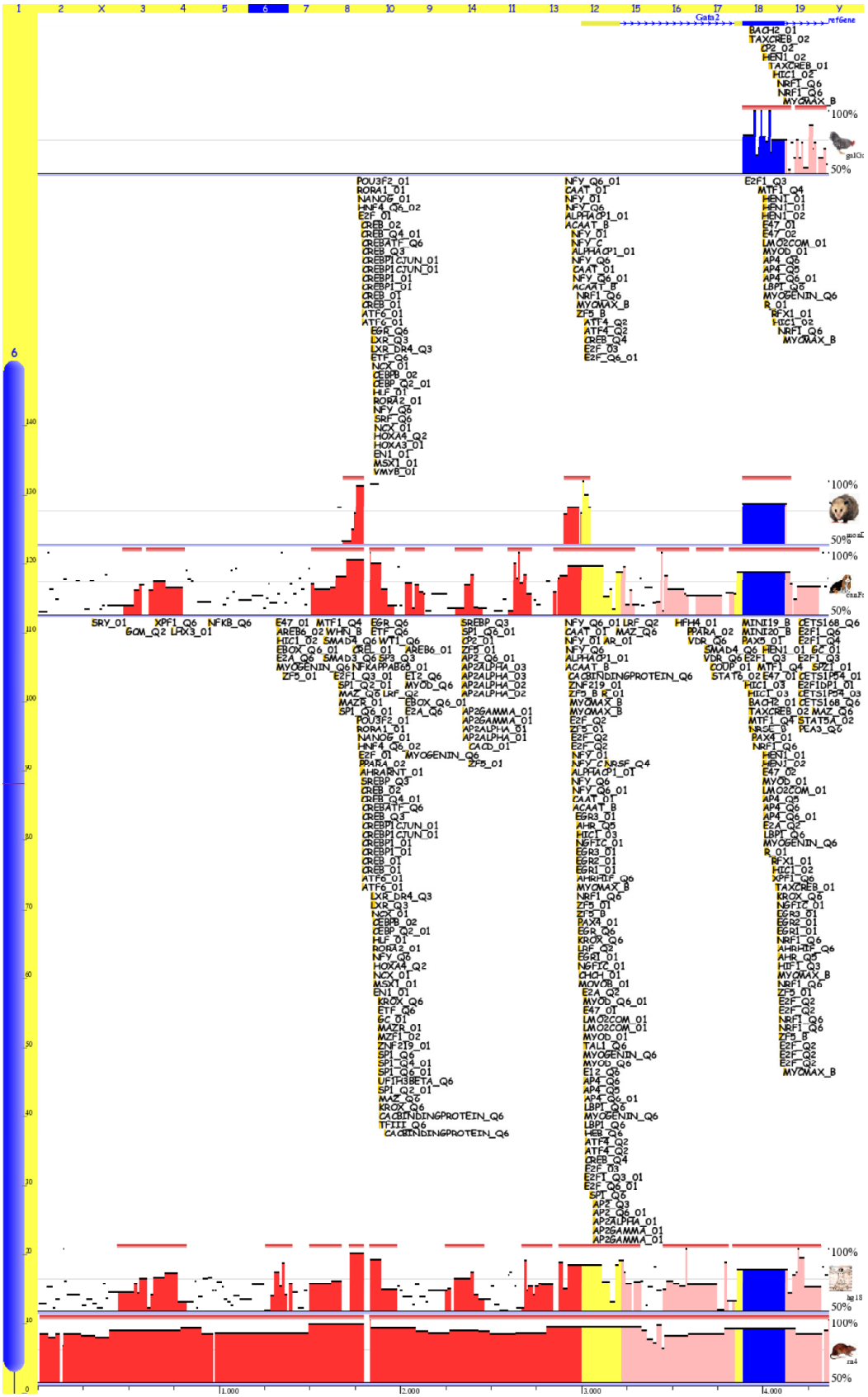


Figure 6.8: Putative upstream from ECRBrowser for the *Gata2* gene.

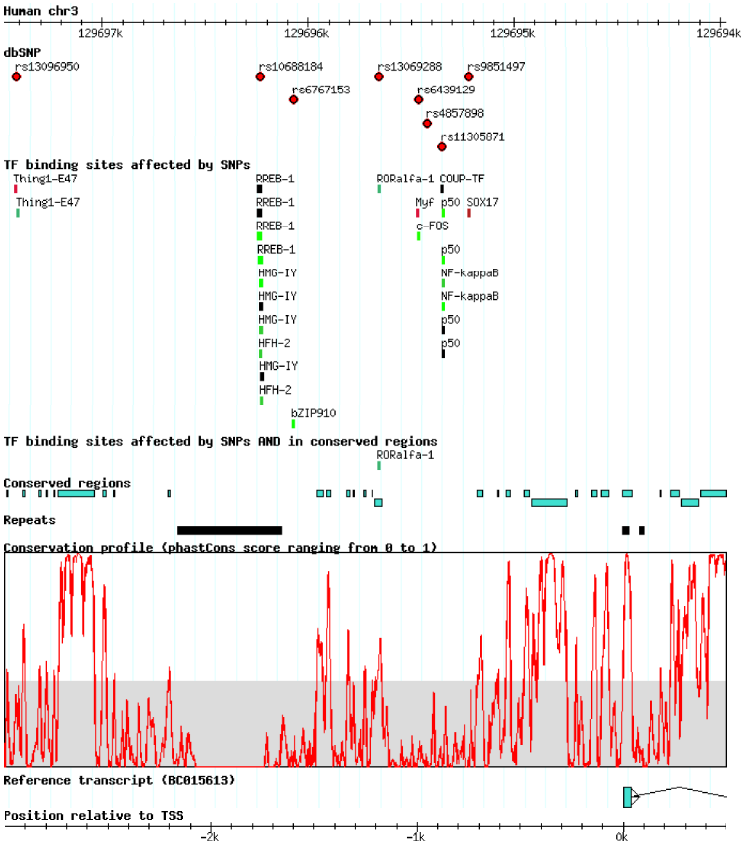


Figure 6.9: Putative upstream SNP TFs in the *Gata2* gene proximal promoter.

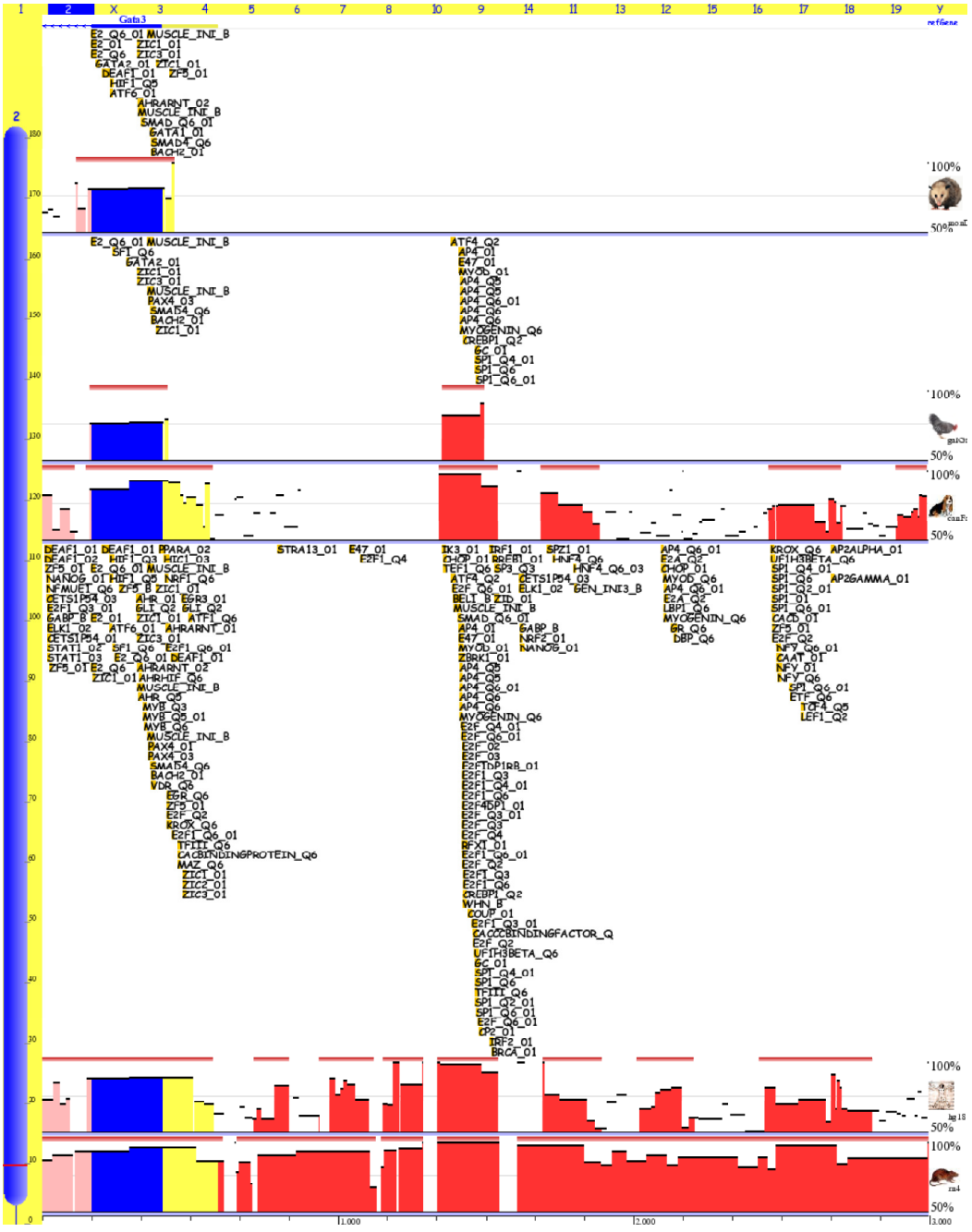


Figure 6.10: Putative upstream TFs from ECRBrowser for the *Gata3* gene.

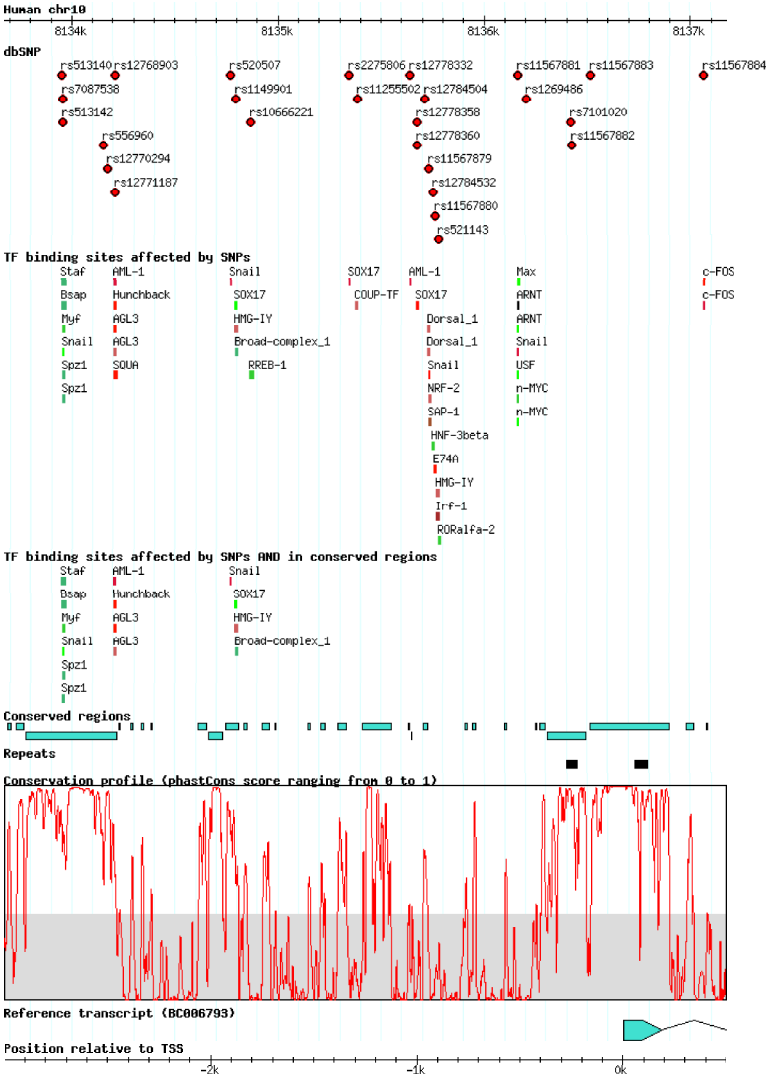


Figure 6.11: Putative upstream SNP TFs in the *Gata3* gene proximal promoter.

CHAPTER VII

Conclusions, Summary and Future Work

7.1 Summary of Previous Work

In this work, we have described the challenge and potential approaches to localizing long-range regulatory elements for the *Gata3* gene, responsible for directing tissue specific expression in the developing urogenital and sympatho-adrenal systems.

Methodologies for the following problems have been discussed:

- Network inference between effectors (upstream TFs) and target genes (*Gata3*), using State Space and Directed Information approaches (Chapters 2 and 3). Some of the networks obtained here (on the kidney expression data) are quite different – this is primarily because SSM approaches are linear-model based and are used for unsupervised network discovery. DTI is a non-linear, information based measure for influence discovery, and has been used in both supervised and unsupervised settings.

- Identification of tissue-specific and histone modification motifs using DTI and Random Forest classifiers (Chapters 4, 5).

- Characterization of enhancer-promoter cross-talk via TF interaction graph analysis (Chapter 5).

To demonstrate the value of these developed methodologies in enhancer prediction, we have analyzed known enhancers of the *Gata2* gene and shown their utility to understand the behavior of why a conserved sequence region might be an enhancer. Furthermore, we have discussed the beginnings of some future ideas in Chapter 6 to point out what else might be useful for the development of a good "enhancer prediction model".

Our efforts in this project have led us to believe that,:

- Enhancer discovery is an art – and will need some detailed knowledge of the spatio-temporal biological process to be incorporated into the model.
- We have found that the traditional ideas of inter-species conservation, RP score and TFBS clustering are insufficient, and lead to several false positives during prospective enhancer discovery. The choice of species (during alignment) is critical, also tissue-specificity scores of TFs for embryonic vs. adult tissues are very different and will need to be suitably imputed.
- DNAse1 hypersensitive site (DHS) region characterization, H3K4 modification maps, protein-protein interaction data for UG-specific or SA-specific contexts have the potential to greatly improve the accuracy of predictions. In the light of unavailability of such data for our particular problem, it might be necessary to generate at least some of this data, depending on feasibility, to improve prediction accuracy.
- It is very important to pay attention to the cell context of the underlying cells/tissue from which experimental data is generated. This particularly matters in the motif discovery and validation steps that is intrinsic to TFBS and generalized motif discovery.

- We have to build an in-house promoter-specific enhancer database to examine all the characteristics based on such heterogeneous data integration – since the mechanisms of promoter-specific and promoter-independent enhancers is quite different ([253], [255], [254]).

- We are optimistic of these new methods for enhancer discovery. However, though we can predict the location of enhancers with some fidelity, predicting the right one responsible for directing precise spatio-temporal specificity is still some distance away.

7.2 Future Work

Some aspects of this work that need to be completed will pertain to the following issues.

- We are in the process of understanding cis-regulatory grammar, i.e., the relevance of the spacing and position of the various motifs that constitute these regulatory regions. To this end we are exploring novel methods in language processing, such as conditional random fields [260], infinite relational models [258], bayesian sets and hierarchical dirichlet processes [259] for such structured prediction scenarios.

- Detailed computational and experimental approach for enhancer discovery for the *Gata3* kidney element (UGE2). This work is jointly with Dr. Kim Lim (Engel laboratory, UM).

- Detailed computational and experimental approach for enhancer discovery for the *Gata3* sympatho-adrenal element (SAE). This work is jointly with Dr. Takashi Moriguchi (formerly of the Engel laboratory, UM).

- Finally, we are very interested to explore a probabilistic kernel setting within which to generalize some of the methods of heterogeneous data integration [256].

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Carter, D., L. Chakalova, C. S. Osborne, Y. F. Dai, and P. Fraser. 2002. Long-range chromatin regulatory interactions in vivo. Nat Genet 32:623-6.

[2] Choi, O. R., and J. D. Engel. 1988. Developmental regulation of beta-globin gene switching. Cell 55:17-26.

[3] Elnitski, L., R. C. Hardison, J. Li, S. Yang, D. Kolbe, P. Eswara, M. J. O'Connor, S. Schwartz, W. Miller, and F. Chiaromonte. 2003. Distinguishing regulatory DNA from neutral sites. Genome Res 13:64-72.

[4] George, K. M., M. W. Leonard, M. E. Roth, K. H. Lieuw, D. Kioussis, F. Grosveld, and J. D. Engel. 1994. Embryonic expression and cloning of the murine GATA-3 gene. Development 120:2673-86.

[5] GuhaThakurta, D., L. Palomar, G. D. Stormo, P. Tedesco, T. E. Johnson, D. W. Walker, G. Lithgow, S. Kim, and C. D. Link. 2002. Identification of a novel cis-regulatory element involved in the heat shock response in Caenorhabditis elegans using microarray gene expression and computational methods. Genome Res 12:701-12.

[6] Gumucio, D. L., H. Heilstedt-Williamson, T. A. Gray, S. A. Tarle, D. A. Shelton, D. A. Tagle, J. L. Slightom, M. Goodman, and F. S. Collins. 1992. Phylogenetic footprinting reveals a nuclear protein which binds to silencer sequences in the human gamma and epsilon globin genes. Mol Cell Biol 12:4919-29.

[7] Hallikas O., K. Palin, N. Sinjushina, R. Rautiainen, J. Partanen, E. Ukkonen, J. Taipale, 2006. Genome-wide Prediction of Mammalian Enhancers Based on Analysis of Transcription-Factor Binding Affinity. Cell, 124,(1): 47-59.

[8] Hasegawa SL, Moriguchi T, Rao A, Kuroha T, Engel JD, Lim KC. 2007. Dosage-dependent rescue of definitive nephrogenesis by a distant Gata3 enhancer. Dev Biol. 301(2):568-77.

[9] Khandekar, M., N. Suzuki, J. Lewton, M. Yamamoto, and J. D. Engel. 2004. Multiple, distant Gata2 enhancers specify temporally and tissue-specific patterning in the developing urogenital system. Mol Cell Biol 24:10263-76.

[10] Ko, L. J., and J. D. Engel. 1993. DNA-binding specificities of the GATA transcription factor family. Mol Cell Biol 13:4011-22.

[11] Lakshmanan, G., K. H. Lieuw, K. C. Lim, Y. Gu, F. Grosveld, J. D. Engel, and A. Karis. 1999. Localization of distant urogenital system-, central nervous system-, and endocardium-specific transcriptional regulatory elements in the GATA-3 locus. Mol Cell Biol 19:1558-68.

[12] Lee, E. C., D. Yu, J. Martinez de Velasco, L. Tessarollo, D. A. Swing, D. L. Court, N. A. Jenkins, and N. G. Copeland. 2001. A highly efficient Escherichia coli-based chromosome engineering system adapted for recombinogenic targeting and subcloning of BAC DNA. Genomics 73:56-65.

[13] Lettice, L. A., S. J. Heaney, L. A. Purdie, L. Li, P. de Beer, B. A. Oostra, D. Goode, G. Elgar, R. E. Hill, and E. de Graaff. 2003. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Hum Mol Genet 12:1725-35.

[14] Lieuw, K. H., G. Li, Y. Zhou, F. Grosveld, and J. D. Engel. 1997. Temporal and spatial control of murine GATA-3 transcription by promoter- proximal regulatory elements. Dev Biol 188:1-16.

[15] Lim, K. C., G. Lakshmanan, S. E. Crawford, Y. Gu, F. Grosveld, and J. D. Engel. 2000. Gata3 loss leads to embryonic lethality due to noradrenaline deficiency of the sympathetic nervous system. Nat Genet 25:209-12.

[16] Liu J, Francke U. 2006. Identification of cis-regulatory elements for MECP2 expression. Hum Mol Genet. 15(11):1769-82.

[17] Mayer, H., M. Bilban, V. Kurtev, F. Gruber, O. Wagner, B. R. Binder, and R. de Martin. 2004. Deciphering regulatory patterns of inflammatory gene expression from interleukin-1-stimulated human endothelial cells. Arterioscler Thromb Vasc Biol 24:1192-8.

[18] Militello, K. T., M. Dodge, L. Bethke, and D. F. Wirth. 2004. Identification of regulatory elements in the Plasmodium falciparum genome. Mol Biochem Parasitol 134:75-88.

[19] 21. Moriguchi, T., N. Takako, M. Hamada, A. Maeda, Y. Fujioka, T. Kuroha, R. Huber, S. Hasegawa, S. Takahashi, K.-C. Lim, and J. Engel. 2006. GATA-3 participates in a complex transcriptional feedback network to regulate sympathoadrenal differentiation. Development 133(19):3871-81.

[20] Ness, S. A., and J. D. Engel. 1994. Vintage reds and whites: combinatorial transcription factor utilization in hematopoietic differentiation. Current Opin. Genet. Devel. 4:718-724.

[21] Ovcharenko, I., D. Boffelli, and G. G. Loots. 2004. eShadow: a tool for comparing closely related sequences. Genome Res 14:1191-8.

[22] Pandolfi, P. P., M. E. Roth, A. Karis, M. W. Leonard, E. Dzierzak, F. G. Grosveld, J. D. Engel, and M. H. Lindenbaum. 1995. Targeted disruption of the GATA3 gene causes severe abnormalities in the nervous system and in fetal liver haematopoiesis. Nat Genet 11:40-4.

[23] Pennacchio LA, Loots GG, Nobrega MA, Ovcharenko I. 2007. Predicting tissue-specific enhancers in the human genome. Genome Res. 2007 Feb;17(2):201-11. bibitemicassp2006 Rao, A., A. Hero, D. States, and J. Engel. 2006. Biologically Relevant Influence Networks using the Directed Information Criterion, IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP.

[24] Rao, A., A. Hero, D. States, and J. Engel. 2007. Understanding Transcriptional Regulation Using De-novo Sequence Motif Discovery, Network Inference and Interactome Data. IEEE Journal of Selected Topics in Signal Processing, submitted.

[25] Rombauts, S., K. Florquin, M. Lescot, K. Marchal, P. Rouze, and Y. van de Peer. 2003. Computational approaches to identify promoters and cis-regulatory elements in plant genomes. Plant Physiol 132:1162-76.

[26] The ENCODE Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447, 799-816.

[27] Yang, Z., L. Gu, P. H. Romeo, D. Bories, H. Motohashi, M. Yamamoto, and J. D. Engel. 1994. Human GATA-3 trans-activation, DNA-binding, and nuclear localization activities are organized into distinct structural domains. Mol Cell Biol 14:2201-12.

[28] Assenov Y, Ramrez F, Schelhorn SE, Lengauer T, Albrecht M., Computing topological parameters of biological networks., Bioinformatics. 2008 Jan 15;24(2):282-4.

[29] Romer KA, Kayombya GR, Fraenkel E., WebMOTIFS: automated discovery, filtering and scoring of DNA sequence motifs using multiple programs and Bayesian approaches., Nucleic Acids Res. 2007 Jul;35(Web Server issue):W217-20.

[30] James RG, Kamei CN, Wang Q, Jiang R, Schultheiss TM., Odd-skipped related 1 is required for development of the metanephric kidney and regulates formation and differentiation of kidney precursor cells., Development. 2006 Aug;133(15):2995-3004.

[31] Labastie MC, Catala M, Gregoire JM, Peault B., The GATA-3 gene is expressed during human kidney embryogenesis., Kidney Int. 1995 Jun;47(6):1597-603.

[32] Rangel C, Angus J, Ghahramani Z, Lioumi M, Sotheran E, Gaiba A, Wild DL, Falciani F, "Modeling T-cell activation using gene expression profiling and state-space models", *Bioinformatics*, 20(9),1361-72,June 2004.

[33] M. Figueiredo and A.K.Jain, "Unsupervised learning of finite mixture models", *IEEE Transaction on Pattern Analysis and Machine Intelligence - PAMI*, vol. 24(3), 381-396, March 2002.

[34] Nina Golyandina, Vladimir Nekrutkin, Anatoly Zhigljavsky , "Analysis of Time Series Structure - SSA and Related Techniques",*Chapman & Hall/CRC*, 2001.

[35] Stuart RO, Bush KT, Nigam SK, "Changes in gene expression patterns in the ureteric bud and metanephric mesenchyme in models of kidney development", *Kidney International*,64(6),1997-2008,December 2003.

[36] Luscombe NM, Babu MM, Yu H, Snyder M, Teichmann SA, Gerstein M, "Genomic analysis of regulatory network dynamics reveals large topological changes",*Nature*,431(7006),308-312,Sep 2004.

[37] Sontag E, Kiyatkin A, Kholodenko BN, "Inferring dynamic architecture of cellular networks using time series of gene expression, protein and metabolite data", *Bioinformatics*, 20(12),1877-86,Aug 2004.

[38] Schfer, J., and K. Strimmer, "An empirical Bayes approach to inferring large-scale gene association networks",*Bioinformatics*,Oct 2004.

[39] Khandekar M, Suzuki N, Lewton J, Yamamoto M, Engel JD, "Multiple, distant Gata2 enhancers specify temporally and tissue-specific patterning in the developing urogenital system", *Mol Cell Biol*,24(23):10263-76,Dec 2004.

[40] V. Moskvina and A. Zhigljavsky, "Change-point detection algorithm based on the singular-spectrum analysis", *Communications in Statistics*,32(2):319-352, 2003.

[41] Schwab K, Patterson LT, Aronow BJ, Luckas R, Liang HC, Potter SS, "A catalogue of gene expression in the developing kidney",*Kidney Int*,64(5):1588-604,Nov 2003.

[42] Rao A,Hero AO,States DJ,Engel JD, "An Integrated Approach to Understanding Mechanisms of Transcriptional Regulation",*Proc. of the 3rd CSHL meeting on Systems Biology*, 70,March 2005.

[43] Shumway RH, Stoffer DS, "Time Series Analysis and Applications",*Springer Texts in Statistics*, 2000.

[44] Effron B, "An Introduction to the Bootstrap",*Chapman & Hall/CRC*,1993.

[45] Zhou Y, Lim KC, Onodera K, Takahashi S, Ohta J, Minegishi N, Tsai FY, Orkin SH, Yamamoto M, J.D.Engel, "Rescue of the embryonic lethal hematopoietic defect reveals a critical role for GATA-2 in urogenital development,"*EMBO J*,17(22):6689-700,Nov 16 1998.

[46] Challen GA, Martinez G, Davis MJ, Taylor DF, Crowe M, Teasdale RD, Grimmond SM, Little MH, "Identifying the molecular phenotype of renal progenitor cells",*J Am Soc Nephrol*,15(9):2344-57,Sep 2004 .

[47] Li H, Wood CL, Liu Y, Getchell TV, Getchell ML, Stromberg AJ., "Identification of gene expression patterns using planned linear contrasts.",*BMC Bioinformatics.7:245*,2006.

[48] Ghahramani Z, Hinton GE, "Parameter Estimation for Linear Dynamical Systems",*University of Toronto Technical Report*, 1996.

[49] *Proc. of the 3rd CSHL meeting on Identification of Functional Elements in Mammalian Genomes*, Nov 11-13, 2004.

[50] NCBI Pubmed URL: *http://www.ncbi.nlm.nih.gov/entrez/query.fcgi*

[51] Alberts B,Bray D,Lewis J,Raff M,Roberts K, Watson JD, Molecular Biology of the Cell,*Garland Publishing*, 3rd edition (March, 1994).

[52] Loots G and Ovcharenko I , "rVista 2.0: evolutionary analysis of transcription factor binding sites", *Nucleic Acids Research, 32(Web Server Issue)*, W217-W221 (2004).

[53] S. Kim, H. Li, D. Russ, J. Whitmore, Y. Chen, E. R. Dougherty, E. Suh, and M. L. Bittner,"Context-sensitive probabilistic Boolean networks to mimic biological regulation", *Proc. Oncogenomics* 2003.

[54] Scott MP, Matsudaira P, Lodish H, Darnell J, Zipursky L, Kaiser CA, Berk A, Krieger M, "Molecular Cell Biology",*WH Freeman and Company*, 2003.

[55] Casella G. and Berger RL, "Statistical Inference", *Duxbury Press*, 1990.

[56] A. Hero, G. Fleury, A. Mears and A. Swaroop, "Multicriteria Gene Screening for Analysis of Differential Expression with DNA Microarrays", *Special Issue on Genomic Signal Processing, EURASIP Journ. of Applied Signal Processing (EURASIP JASP)* 2004(1), 43-52, 2004.

[57] Bar-Joseph,Z., "Analyzing time series gene expression data", *Bioinformatics*, 20(16),2493-2503, 2004.

[58] Rao,Hero AO,States DJ,Engel JD, "Inference of biologically relevant Gene Influence Networks using the Directed Information Criterion",*Proc. of the IEEE Conference on Acoustics, Speech and Signal Processing*, 2006.

[59] Perrin BE, Ralaivola L, Mazurie A, Bottani S, Mallet J, D'Alche-Buc F., "Gene networks inference using dynamic Bayesian networks",*Bioinformatics. 19 Suppl 2:II138-II148*,2003.

[60] Li C, Wong WH.,"Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection.",*Proc Natl Acad Sci U S A, 98(1):31-6.*. 2001.

[61] Mizutani H, May LT, Sehgal PB, Kupper TS., "Synergistic interactions of IL-1 and IL-6 in T cell activation. Mitogen but not antigen receptor-induced proliferation of a cloned T helper cell line is enhanced by exogenous IL-6.", *J Immunol. 1989 Aug 1;143(3):896-901.*

[62] Dougherty E., Kim S., Chen Y., "Coefficient of determination in nonlinear signal processing",. *Signal Processing, 80:2219-2235*, 2000.

[63] Kim S., Dougherty E. R., Bittner M. L., Chen Y., Sivakumar K., Meltzer P., Trent J. M., "General nonlinear framework for the analysis of gene interaction via multivariate expression arrays".,*J. Biomed. Opt. 5, 411-424.*, 2000.

[64] Mori, M., Ghyselinck, N. B., Chambon, P., and Mark, M., "Systematic immunolocalization of retinoid receptors in developing and adult mouse eyes.", *Invest. Ophthalmol. Vis. 42, 1312-1318*, 2001.

[65] Lim, K-C., Lakshmanan, G., Crawford, S. E., Gu, Y., Grosveld, F., and Engel, J. D. , "Gata3 loss leads to embryonic lethality due to noradrenaline deficiency of the sympathetic nervous system" *Nat. Genet. 25, 209-212*, 2000.

[66] Lin JX, Leonard WJ., "The immediate-early gene product Egr-1 regulates the human interleukin-2 receptor beta-chain promoter through noncanonical Egr and Sp1 binding sites.",*Mol Cell Biol. 17(7):3714-22*,1997.

[67] Opgen-Rhein, R., and Strimmer K., "Using regularized dynamic correlation to infer gene dependency networks from time-series microarray data", *Proc. of Fourth International Workshop on Computational Systems Biology, WCSB*, 2006.

[68] Herrgard MJ, Covert MW, Palsson BO., "Reconciling gene expression data with known genome-scale regulatory network structures".,*Genome Res. 13(11):2423-34*, 2003.

[69] Esquela AF; Lee SJ, "Regulation of metanephric kidney development by growth/differentiation factor 11".,*Dev Biol. 257(2):356-70*, 2003.

[70] Maeshima A, Yamashita S, Maeshima K, Kojima I, Nojima Y., "Activin a produced by ureteric bud is a differentiation factor for metanephric mesenchyme".,*J Am Soc Nephrol. 14(6):1523-34*, 2003.

[71] Balmer JE, Blomhoff R., "Gene expression regulation by retinoic acid",*J Lipid Res., 43(11):1773-808*, 2002.

[72] Zadeh HH, Tanavoli S, Haines DD, Kreutzer DL., "Despite large-scale T cell activation, only a minor subset of T cells responding in vitro to Actinobacillus actinomycetemcomitans differentiate into effector T cells".,*J Periodontal Res. 35(3):127-36*, 2000.

[73] A. Kundaje, O. Antar, T. Jebara and C. Leslie, "Learning Regulatory Networks from Sparsely Sampled Time Series Expression Data"., *Columbia University Technical Report*, 2002.

[74] Aerts S, Thijs G, Coessens B, Staes M, Moreau Y, De Moor B., "Toucan: deciphering the cis-regulatory logic of coregulated genes", *Nucleic Acids Res.* 2003 Mar 15;31(6):1753-64.

[75] Moonen C.T.W. (Ed), Bandettini P.A. (Ed), Functional MRI (Medical Radiology / Diagnostic Imaging), Springer, 2000.

[76] Balmer JE, Blomhoff R., "Gene expression regulation by retinoic acid", *J. Lipid Res.* 2002 Nov;43:11:1773-808.

[77] Beal MJ, Falciani F, Ghahramani Z, Rangel C, Wild DL.,"A Bayesian approach to reconstructing genetic regulatory networks with hidden factors", *Bioinformatics.* 2005 Feb 1;21(3):349-56.

[78] Benjamini, Y. and Hochberg, Y., "Controlling the false discovery rate: A practical and powerful approach to multiple testing", *J. Roy. Statist. Soc. Ser. B.* 1995; 57:289-300.

[79] Brophy PD, Ostrom L, Lang KM, Dressler GR., "Regulation of ureteric bud outgrowth by Pax2-dependent activation of the glial derived neurotrophic factor gene", *Development.* 2001 Dec;128(23):4747-56.

[80] Challen GA, Martinez G, Davis MJ, Taylor DF, Crowe M, Teasdale RD, Grimmond SM, Little MH., "Identifying the molecular phenotype of renal progenitor cells.", *J Am Soc Nephrol.* 2004 Sep;15(9):2344-57.

[81] Challen G, Gardiner B, Caruana G, Kostoulias X, Martinez G, Crowe M, Taylor DF, Bertram J, Little M, Grimmond SM., "Temporal and spatial transcriptional programs in murine kidney development", *Physiol Genomics.* 2005 Oct 17;23(2):159-71.

[82] Clarke JC, Patel SR, Raymond RM, Andrew S, Robinson BG, Dressler GR, Brophy PD., "Regulation of c-Ret in the developing kidney is responsive to Pax2 gene dosage.",*Hum Mol Genet.* 2006 Dec 1;15(23):3420-8.

[83] Cohen CD, Klingenhoff A, Boucherot A, Nitsche A, Henger A, Brunner B, Schmid H, Merkle M, Saleem MA, Koller KP, Werner T, Grone HJ, Nelson PJ, Kretzler M., "Comparative promoter analysis allows de novo identification of specialized cell junction-associated proteins", *Proc Natl Acad Sci U S A.* 2006 Apr 11;103(15):5682-7.

[84] Cohen HT, Bossone SA, Zhu G, McDonald GA, Sukhatme VP.,"Sp1 is a critical regulator of the Wilms' tumor-1 gene", *J Biol Chem.* 1997 Jan 31;272(5):2901-13.

[85] Cover T.M, Thomas J.A, "Elements of Information Theory", Wiley-Interscience, 1991.

[86] Davies J, "Intracellular and extracellular regulation of ureteric bud morphogenesis", *Journal of Anatomy* 198 (3), 257264. (2001)

[87] G. A. Darbellay and I. Vajda, "Estimation of the information by an adaptive partitioning of the observation space," *IEEE Trans. on Information Theory*, vol. 45, pp. 1315–1321, May 1999.

[88] Dressler, G.R. and Douglas, E.C. , "Pax-2 is a DNA-binding protein expressed in embryonic kidney and Wilms tumor", *Proc. Natl. Acad. Sci. USA* 89: 1179-1183, 1992.

[89] Effron B, Tibshirani R.J, An Introduction to the Bootstrap (Monographs on Statistics and Applied Probability), Chapman & Hall/CRC, 1994.

[90] Ezzat S, Mader R, Yu S, Ning T, Poussier P, Asa SL., "Ikaros integrates endocrine and immune system development", *J Clin Invest.* 2005 Apr;115(4):844-8.

[91] Geweke J., "The Measurement of Linear Dependence and Feedback Between Multiple Time Series," *Journal of the American Statistical Association*, 1982, 77, 304-324. (With comments by E. Parzen, D. A. Pierce, W. Wei, and A. Zellner, and rejoinder)

[92] Grote D, Souabni A, Busslinger M, Bouchard M., "Pax 2/8-regulated Gata3 expression is necessary for morphogenesis and guidance of the nephric duct in the developing kidney"., *Development.* 2006 Jan;133(1):53-61.

[93] Gubner J. A., Probability and Random Processes for Electrical and Computer Engineers, Cambridge, 2006.

[94] Hashimoto RF, Kim S, Shmulevich I, Zhang W, Bittner ML, Dougherty ER., "Growing genetic regulatory networks from seed genes",*Bioinformatics.* 2004 May 22;20(8):1241-7.

[95] Hastie T, Tibshirani R, The Elements of Statistical Learning, Springer 2002.

[96] Hudson, J.E., "Signal Processing Using Mutual Information", *Signal Processing Magazine*,23(6):50-54, Nov. 2006.

[97] H. Joe., "Relative entropy measures of multivariate dependence", *J. Am. Statist. Assoc.*, 84:157164, 1989.

[98] Khandekar M, Suzuki N, Lewton J, Yamamoto M, Engel JD., "Multiple, distant Gata2 enhancers specify temporally and tissue-specific patterning in the developing urogenital system", *Mol Cell Biol.* 2004 Dec;24(23):10263-76.

[99] Kreiman G., "Identification of sparsely distributed clusters of cis-regulatory elements in sets of co-expressed genes"., *Nucleic Acids Res.* 2004 May 20;32(9):2889-900.

[100] Lakshmanan G, Lieuw KH, Lim KC, Gu Y, Grosveld F, Engel JD, Karis A., "Localization of distant urogenital system-, central nervous system-, and endocardium-specific transcriptional regulatory elements in the GATA-3 locus", *Mol Cell Biol.* 1999 Feb;19(2):1558-68.

[101] Miller E., "A new class of entropy estimators for multi-dimensional densities", *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2003.

[102] Learned-Miller E., "Hyperspacings and the estimation of information theoretic quantities", UMass Amherst Technical Report 04-104, 2004.

[103] Li H, Sun Y, Zhan M.,"Analysis of Gene Coexpression by B-Spline Based CoD Estimation",*EURASIP J Bioinform Syst Biol.* 2007;:49478.

[104] MacIsaac KD, Fraenkel E., "Practical strategies for discovering regulatory DNA sequence motifs".,*PLoS Comput Biol.* 2006 Apr;2(4):e36.

[105] Majumdar A, Vainio S, Kispert A, McMahon J, McMahon AP.,"Wnt11 and Ret/Gdnf pathways cooperate in regulating ureteric branching during metanephric kidney development", *Development.* 2003 Jul;130(14):3175-85.

[106] Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, Califano A., "ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context", *BMC Bioinformatics.* 2006 Mar 20;7 Suppl 1:S7.

[107] H. Marko, "The Bidirectional Communication Theory - A Generalization of Information Theory", *IEEE Transactions on Communications*, Vol. COM-21, pp. 1345-1351, 1973.

[108] J. Massey, "Causality, feedback and directed information," in *Proc. 1990 Symp. Information Theory and Its Applications (ISITA-90)*, Waikiki, HI, Nov. 1990, pp. 303305.

[109] Nemenman, F Shafee, and W Bialek. "Entropy and inference, revisited.", In TG Dietterich, S Becker, and Z Ghahramani, editors, *Advances in Neural Information Processing Systems* 14, Cambridge, MA, 2002. MIT Press.

[110] Nemenman I,. "Information theory, multivariate dependence, and genetic network inference.", Technical Report NSF-KITP-04-54, KITP, UCSB, 2004.

[111] Opgen-Rhein, R., and Strimmer K., "Using regularized dynamic correlation to infer gene dependency networks from time-series microarray data", *Proc. of Fourth International Workshop on Computational Systems Biology, WCSB*, 2006.

[112] I. Ovcharenko, M.A. Nobrega, G.G. Loots, and L. Stubbs, "ECR Browser: a tool for visualizing and accessing data from comparisons of multiple vertebrate genomes", *Nucleic Acids Research*, 32, W280-W286 (2004).

[113] Paninski, L. , "Estimation of entropy and mutual information", *Neural Computation* 15: 1191-1254, 2003.

[114] Papatsenko D., Levine M., "Computational identification of regulatory DNAs underlying animal development", *Nature Methods* 2, 529 - 534 (2005)

[115] J. Ramsay, B. W. Silverman, Functional Data Analysis (Springer Series in Statistics), Springer 1997.

[116] Rangel C, Angus J, Ghahramani Z, Lioumi M, Sotheran E, Gaiba A, Wild DL, Falciani F, "Modeling T-cell activation using gene expression profiling and state-space models", *Bioinformatics*, 20(9),1361-72, June 2004.

[117] Rao A, Hero AO, States DJ, Engel JD, "Inference of biologically relevant Gene Influence Networks using the Directed Information Criterion", *Proc. of the IEEE Conference on Acoustics, Speech and Signal Processing*, 2006.

[118] Rogoff HA, Pickering MT, Frame FM, Debatis ME, Sanchez Y, Jones S, Kowalik TF., "Apoptosis associated with deregulated E2F activity is dependent on E2F1 and Atm/Nbs1/Chk2", *Mol Cell Biol.* 2004 Apr;24(7):2968-77.

[119] Ryan G, Steele-Perkins V, Morris JF, Rauscher FJ, Dressler GR.,"Repression of Pax-2 by WT1 during normal kidney development", *Development.* 1995 Mar;121(3):867-75.

[120] Schfer, J., and K. Strimmer, "An empirical Bayes approach to inferring large-scale gene association networks", *Bioinformatics*,Oct 2004.

[121] Schneidman E, Still S, Berry MJ II, and Bialek W,"Network information and connected correlations", *Phys. Rev. Lett.*, 91, p. 238701, (2003)

[122] Schwab K, Patterson LT, Aronow BJ, Luckas R, Liang HC, Potter SS., "A catalogue of gene expression in the developing kidney", *Kidney Int.* 2003 Nov; 64(5):1588-604.

[123] Stuart RO, Bush KT, Nigam SK, "Changes in gene expression patterns in the ureteric bud and metanephric mesenchyme in models of kidney development", *Kidney International*,64(6),1997-2008,December 2003.

[124] Taraviras S, Monaghan AP, Schtz G, Kelsey G.,"Characterization of the mouse HNF-4 gene and its expression during mouse embryogenesis", *Mech Dev.* 1994 Nov;48(2):67-79.

[125] Venkataramanan R. , Pradhan S. S., "Source Coding With Feed-Forward: Rate-Distortion Theorems and Error Exponents for a General Source," *IEEE Transactions on Information Theory* , vol.53, no.6, pp.2154-2179, Jun. 2007.

[126] Willett R, Nowak R, "Complexity-Regularized Multiresolution Density Estimation", *Proc. Intl Symp. on Information Theory*, ISIT 2004.

[127] Woolf PJ, Prudhomme W, Daheron L, Daley GQ, Lauffenburger DA., "Bayesian analysis of signaling networks governing embryonic stem cell fate decisions", *Bioinformatics.* 2005 Mar;21(6):741-53.

[128] Zhang SL, Moini B, Ingelfinger JR., "Angiotensin II increases Pax-2 expression in fetal kidney cells via the AT2 receptor", *J Am Soc Nephrol.* 2004 Jun;15(6):1452-65.

[129] Zhang, DH, Yang L, and Ray A., "Differential responsiveness of the IL-5 and IL-4 genes to transcription factor GATA-3", *J Immunol* 161: 3817-3821, 1998.

[130] Aerts S, Van Loo P, Thijs G, Mayer H, de Martin R, Moreau Y, De Moor B., "TOUCAN 2: the all-inclusive open source workbench for regulatory sequence analysis", *Nucleic Acids Res.* 2005 Jul 1;33(Web Server issue):W393-6.

[131] Azakie A, Fineman JR, He Y.,"Myocardial transcription factors are modulated during pathologic cardiac hypertrophy in vivo"., *J Thorac Cardiovasc Surg.* 2006 Dec;132(6):1262-71.

[132] Benjamini, Y. and Hochberg, Y. (1995)., "Controlling the false discovery rate: a practical and powerful approach to multiple testing"., *J. Roy. Statist. Soc. Ser. B* 57 289-300.

[133] Burge C, Karlin S, "Prediction of complete gene structures in human genomic DNA". *J Mol Biol* 1997, 268:78-94.

[134] Chan BY, Kibler D., "Using hexamers to predict cis-regulatory motifs in Drosophila", *BMC Bioinformatics*, 2005 Oct 27;6:262.

[135] Cover TM, Thomas JA, Elements of Information Theory, *Wiley- Interscience*, 1991.

[136] Dressler, G.R. and Douglas, E.C. (1992)., "Pax-2 is a DNA-binding protein expressed in embryonic kidney and Wilms tumor"., *Proc. Natl. Acad. Sci. USA* 89: 1179-1183.

[137] Effron B, Tibshirani R.J, An Introduction to the Bootstrap (Monographs on Statistics and Applied Probability), Chapman & Hall/CRC, 1994.

[138] Grote D, Souabni A, Busslinger M, Bouchard M.,"Pax 2/8-regulated Gata 3 expression is necessary for morphogenesis and guidance of the nephric duct in the developing kidney".,*Development.* 2006 Jan;133(1):53-61.

[139] I. Guyon, A. Elisseeff, "An Introduction to Variable and Feature Selection", *Journal of Machine Learning Research* 3 1157-1182.

[140] Hastie T, Tibshirani R, The Elements of Statistical Learning , Springer 2002.

[141] Hutchinson GB., "The prediction of vertebrate promoter regions using differential hexamer frequency analysis".,*Comput Appl Biosci.* 1996 Oct;12(5):391-8.

[142] H. Joe., "Relative entropy measures of multivariate dependence". *J. Am. Statist. Assoc.*, 84:157164, 1989.

[143] Kadota K, Ye J, Nakai Y, Terada T, Shimizu K., "ROKU: a novel method for identification of tissue-specific genes",.2003. *BMC Bioinformatics.* 2006 Jun 12;7:294.

[144] M. G. Kendall, "A New Measure of Rank Correlation", *Biometrika*, 30 (1/2): 81-93 Jun., 1938.

[145] Khandekar M, Suzuki N, Lewton J, Yamamoto M, Engel JD., "Multiple, distant Gata2 enhancers specify temporally and tissue-specific patterning in the developing urogenital system".,*Mol Cell Biol.* 2004 Dec;24(23):10263-76.

[146] King DC, Taylor J, Elnitski L, Chiaromonte F, Miller W, Hardison RC., "Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences".,*Genome Res.* 2005 Aug;15(8):1051-60.

[147] Kleinjan DA, van Heyningen V., "Long-range control of gene expression: emerging mechanisms and disruption in disease"., *Am J Hum Genet.* 2005 Jan;76(1):8-32.

[148] Kreiman G., "Identification of sparsely distributed clusters of cis-regulatory elements in sets of co-expressed genes".,*Nucleic Acids Res.* 2004 May 20;32(9):2889-900.

[149] Lakshmanan, G., K. H. Lieuw, K. C. Lim, Y. Gu, F. Grosveld, J. D. Engel, and A. Karis. 1999. "Localization of distant urogenital system-, central nervous system-, and endocardium-specific transcriptional regulatory elements in the GATA-3 locus". *Mol. Cell. Biol.* 19:1558-1568.

[150] MacIsaac KD, Fraenkel E., "Practical strategies for discovering regulatory DNA sequence motifs".,*PLoS Comput Biol.* 2006 Apr;2(4):e36.

[151] H. Marko, "The Bidirectional Communication Theory - A Generalization of Information Theory", *IEEE Transactions on Communications*, Vol. COM-21, pp. 1345-1351, 1973.

[152] J. Massey, "Causality, feedback and directed information," *Proc. 1990 Symp. Information Theory and Its Applications (ISITA-90)*, Waikiki, HI, Nov. 1990, pp. 303305.

[153] Miller E, "A new class of entropy estimators for multi-dimensional densities"., *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2003.

[154] Murphy AM, Thompson WR, Peng LF, Jones L.,"Regulation of the rat cardiac troponin I gene by the transcription factor GATA-4"., *Biochem J.* 1997 Mar 1;322 (Pt 2):393-401.

[155] NCBI Pubmed URL: *http://www.ncbi.nlm.nih.gov/entrez/query.fcgi*

[156] Nemenman, F Shafee, and W Bialek. "Entropy and inference, revisited.", In TG Dietterich, S Becker, and Z Ghahramani, editors, *Advances in Neural Information Processing Systems* 14, Cambridge, MA, 2002. MIT Press.

[157] Olson EN., "Regulation of muscle transcription by the MyoD family. The heart of the matter".,*Circ Res.* 1993 Jan;72(1):1-6.

[158] Paninski, L. , "Estimation of entropy and mutual information". *Neural Computation* 15: 1191-1254, 2003.

[159] Pennacchio LA, Loots GG, Nobrega MA, Ovcharenko I., "Predicting tissue-specific enhancers in the human genome"., *Genome Res.* 2007 Jan 8;

[160] Peng H.,Long F.,Ding C.,"Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, No. 8, pp: 1226-1238, August 2005.

[161] Pennacchio, L. A., Ahituv, N., Moses, A., Prabhakar, S., Nobrega, M., Shoukry, M., Minovitsky, A., Dubchak, I., Holt, A., Lewis, K., Plazer-Frick, I., Akiyama, J., DeVal, S., Afzal, V., Black, B., Couronne, O., Eisen, M., Visel, A., and Rubin, E.M. 2006., "In vivo enhancer analysis of human conserved non-coding sequences", *Nature*, 444(7118):499-502.

[162] Proc. NIPS 2006 Workshop on Causality and Feature Selection, available at: *http://research.ihost.com/cws2006/.*

[163] Li Q, Barkess G, Qian H.,"Chromatin looping and the probability of transcription"., *Trends Genet.* 2006 Apr;22(4):197-202.

[164] J. Ramsay, B. W. Silverman, Functional Data Analysis (Springer Series in Statistics), Springer 1997.

[165] Rao A, Hero AO, States DJ, Engel JD, "Inference of biologically relevant Gene Influence Networks using the Directed Information Criterion", *Proc. of the IEEE Conference on Acoustics, Speech and Signal Processing*, 2006.

[166] Schug J., Schuller W-P., Kappen C., Salbaum J.M., Bucan M., Stoeckert C.J. Jr., "Promoter Features Related to Tissue Specificity as Measured by Shannon Entropy"., *Genome Biology* 6(4): R33, March 2005.

[167] Sumazin P, Chen G, Hata N , Smith A D., Zhang T, Zhang M Q., "DWE: Discriminant Word Enumerator", *Bioinformatics*, 21(1):31-38, 2005.

[168] Vanhoutte P, Nissen JL, Brugg B, Gaspera BD, Besson MJ, Hipskind RA, Caboche J., "Opposing roles of Elk-1 and its brain-specific isoform, short Elk-1, in nerve growth factor-induced PC12 differentiation"., *J Biol Chem.* 2001 Feb 16;276(7):5189-96.

[169] Venkataramanan, R.; Pradhan, S. S.,"Source Coding With Feed-Forward: Rate-Distortion Theorems and Error Exponents for a General Source," *IEEE Transactions on Information Theory*, vol.53, no.6, pp.2154-2179, Jun. 2007.

[170] Werner T., "Regulatory networks: Linking microarray data to systems biology".,*Mech Ageing Dev.* 2007 Jan;128(1):168-72.

[171] Willett R, Nowak R, "Complexity-Regularized Multiresolution Density Estimation", *Proc. Intl Symp. on Information Theory*, ISIT 2004.

[172] Aerts S, Van Loo P, Thijs G, Mayer H, de Martin R, Moreau Y, De Moor B., "TOUCAN 2: the all-inclusive open source workbench for regulatory sequence analysis", *Nucleic Acids Res.* 2005 Jul 1;33(Web Server issue):W393-6.

[173] Assenov Y, Ramrez F, Schelhorn SE, Lengauer T, Albrecht M.,'Computing topological parameters of biological networks".*Bioinformatics.* 2007 Nov 15

[174] Bader GD, Hogue CW.,"An automated method for finding molecular complexes in large protein interaction networks".,*BMC Bioinformatics.* 2003 Jan 13;4:2.

[175] E Blackwood and J Kadonaga, Going the distance: a current view of enhancer action. Science 281 (1998), pp. 6063.

[176] L. Breiman., "Random forests"., *Machine Learning*, 45(1): 5.32, 2001.

[177] Burge C, Karlin S, "Prediction of complete gene structures in human genomic DNA". *J Mol Biol* 1997, 268:78-94.

[178] Challen G, Gardiner B, Caruana G, Kostoulias X, Martinez G, Crowe M, Taylor DF, Bertram J, Little M, Grimmond SM.,"Temporal and spatial transcriptional programs in murine kidney development".,*Physiol Genomics.* 2005 Oct 17;23(2):159-71.

[179] Chen, L.; Tang, H.L.,"Improved computation of beliefs based on confusion matrix for combining multiple classifiers", *Electronics Letters* Volume 40, Issue 4, Feb 2004 Page(s):238 - 239.

[180] Chan BY, Kibler D., "Using hexamers to predict cis-regulatory motifs in Drosophila", *BMC Bioinformatics*, 2005 Oct 27;6:262.

[181] Cover TM, Thomas JA, Elements of Information Theory, *Wiley- Interscience*, 1991.

[182] Crawford GE, Holt IE, Whittle J, Webb BD, Tai D, Davis S, Margulies EH, Chen Y, Bernat JA, Ginsburg D, Zhou D, Luo S, Vasicek TJ, Daly MJ, Wolfsberg TG, Collins FS.,"Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS)",.*Genome Res.* 2006 Jan;16(1):123-31.

[183] Dong J, Horvath S (2007) "Understanding Network Concepts in Modules", *BMC Systems Biology* 2007, 1:24

[184] Dressler, G.R. and Douglas, E.C. (1992)., "Pax-2 is a DNA-binding protein expressed in embryonic kidney and Wilms tumor"., *Proc. Natl. Acad. Sci. USA* 89: 1179-1183.

[185] Drummond IA.,"The zebrafish pronephros: a genetic system for studies of kidney development"., *Pediatr Nephrol.* 2000 May;14(5):428-35.

[186] Effron B, Tibshirani R.J, An Introduction to the Bootstrap (Monographs on Statistics and Applied Probability), Chapman & Hall/CRC, 1994.

[187] Eisenberg E, Levanon EY., "Human housekeeping genes are compact".,*Trends Genet.* 2003 Jul;19(7):362-5.

[188] ENCODE Project Consortium, "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project"., *Nature.* 2007 Jun 14;447(7146):799-816.

[189] Erik Miller., "A new class of entropy estimators for multi-dimensional densities",. *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2003.

[190] Erik Learned-Miller.,"Hyperspacings and the estimation of information theoretic quantities".,UMass Amherst Technical Report 04-104, 2004.

[191] Farr D, Bellora N, Mularoni L, Messeguer X, Alb MM., "Housekeeping genes tend to show reduced upstream sequence conservation"., *Genome Biol.* 2007 Jul 13;8(7):R140

[192] Fraser P., "Transcriptional control thrown for a loop"., *Curr Opin Genet Dev.* 2006 Oct;16(5):490-5.

[193] G. A. Darbellay and I. Vajda, "Estimation of the information by an adaptive partitioning of the observation space," *IEEE Trans. on Information Theory*, vol. 45, pp. 1315–1321, May 1999.

[194] Geweke J., "The Measurement of Linear Dependence and Feedback Between Multiple Time Series," *Journal of the American Statistical Association*, 1982, 77, 304-324. (With comments by E. Parzen, D. A. Pierce, W. Wei, and A. Zellner, and rejoinder)

[195] Z. Ghahramani and H-C. Kim., "Bayesian classifier combination"., Gatsby Computational Neuroscience Unit Technical Report, 2003.

[196] P. Golland, F. Liang, S. Mukherjee, D. Panchenko. Permutation Tests for Classification. In Proceedings of COLT: Annual Conference on Learning Theory, LNCS 3559:501-515, 2005.

[197] Gilbert S.F, "Developmental Biology", Sinauer Associates Inc., Publishers Sunderland, Massachusetts, 1997.

[198] Hasegawa SL, Moriguchi T, Rao A, Kuroha T, Engel JD, Lim KC.,"Dosage-dependent rescue of definitive nephrogenesis by a distant Gata3 enhancer".,*Dev Biol.* 2007 Jan 15;301(2):568-77.

[199] Hastie T, Tibshirani R, The Elements of Statistical Learning , Springer 2002.

[200] Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, Wang W, Weng Z, Green RD, Crawford GE, Ren B.,"Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome"., *Nat Genet. 2007* Mar;39(3):311-8.

[201] Hutchinson GB., "The prediction of vertebrate promoter regions using differential hexamer frequency analysis".,*Comput Appl Biosci.* 1996 Oct;12(5):391-8.

[202] Hudson, J.E., "Signal Processing Using Mutual Information", *Signal Processing Magazine*,Volume: 23,no: 6 pp:50-54, Nov. 2006.

[203] H. Joe., "Relative entropy measures of multivariate dependence". *J. Am. Statist. Assoc.*, 84:157164, 1989.

[204] Khandekar M, Suzuki N, Lewton J, Yamamoto M, Engel JD., "Multiple, distant Gata2 enhancers specify temporally and tissue-specific patterning in the developing urogenital system".,*Mol Cell Biol.* 2004 Dec;24(23):10263-76.

[205] King DC, Taylor J, Elnitski L, Chiaromonte F, Miller W, Hardison RC., "Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences".,*Genome Res.* 2005 Aug;15(8):1051-60.

[206] Kleinjan DA, van Heyningen V., "Long-range control of gene expression: emerging mechanisms and disruption in disease"., *Am J Hum Genet.* 2005 Jan;76(1):8-32.

[207] Koch CM, Andrews RM, Flicek P, Dillon SC, Karaz U, Clelland GK, Wilcox S, Beare DM, Fowler JC, Couttet P, James KD, Lefebvre GC, Bruce AW, Dovey OM, Ellis PD, Dhami P, Langford CF, Weng Z, Birney E, Carter NP, Vetrie D, Dunham I.,"The landscape of histone modifications across 1% of the human genome in five human cell lines".,*Genome Res.* 2007 Jun;17(6):691-707.

[208] Komili S, Silver PA., "Coupling and coordination in gene expression processes: a systems biology view"., Nat Rev Genet. 2008 Jan;9(1):38-48.

[209] Kreiman G., "Identification of sparsely distributed clusters of cis-regulatory elements in sets of co-expressed genes".,*Nucleic Acids Res.* 2004 May 20;32(9):2889-900.

[210] Lakshmanan, G., K. H. Lieuw, K. C. Lim, Y. Gu, F. Grosveld, J. D. Engel, and A. Karis. 1999. "Localization of distant urogenital system-, central nervous system-, and endocardium-specific transcriptional regulatory elements in the GATA-3 locus". *Mol. Cell. Biol.* 19:1558-1568.

[211] Lettice LA, Heaney SJ, Purdie LA, Li L, de Beer P, Oostra BA, Goode D, Elgar G, Hill RE, de Graaff E., "A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly".,*Hum Mol Genet.* 2003 Jul 15;12(14):1725-35.

[212] Huai Li, Yu Sun, and Ming Zhan, Analysis of Gene Coexpression by B-Spline Based CoD Estimation, EURASIP Journal on Bioinformatics and Systems Biology, vol. 2007, Article ID 49478, 10 pages, 2007.

[213] A Liaw, M Wiener, "Classification and Regression by randomForest"., *R News.* 2002; 2: 1822

[214] Lieb JD, Beck S, Bulyk ML, Farnham P, Hattori N, Henikoff S, Liu XS, Okumura K, Shiota K, Ushijima T, Greally JM., "Applying whole-genome studies of epigenetic regulation to study human disease".,*Cytogenet Genome Res.* 2006;114(1):1-15.

[215] Liu J, Francke U.,"Identification of cis-regulatory elements for MECP2 expression".,*Hum Mol Genet.* 2006 Jun 1;15(11):1769-82.

[216] MacIsaac KD, Fraenkel E., "Practical strategies for discovering regulatory DNA sequence motifs".,*PLoS Comput Biol.* 2006 Apr;2(4):e36.

[217] Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, Califano A., "ARACNE: an algorithm for the reconstruction of gene regulatory networks in amammalian cellular context"., *BMC Bioinformatics.* 2006 Mar 20;7 Suppl 1:S7.

[218] H. Marko, "The Bidirectional Communication Theory - A Generalization of Information Theory", *IEEE Transactions on Communications*, Vol. COM-21, pp. 1345-1351, 1973.

[219] J. Massey, "Causality, feedback and directed information," *Proc. 1990 Symp. Information Theory and Its Applications (ISITA-90)*, Waikiki, HI, Nov. 1990, pp. 303305.

[220] Mayer H, Bilban M, Kurtev V, Gruber F, Wagner O, Binder BR, de Martin R., "Deciphering regulatory patterns of inflammatory gene expression from interleukin-1-stimulated human endothelial cells"., *Arterioscler Thromb Vasc Biol.* 2004 Jul;24(7):1192-8.

[221] Nemenman, F Shafee, and W Bialek. "Entropy and inference, revisited.", In TG Dietterich, S Becker, and Z Ghahramani, editors, *Advances in Neural Information Processing Systems* 14, Cambridge, MA, 2002. MIT Press.

[222] Minegishi N, Ohta J, Yamagiwa H, Suzuki N, Kawauchi S, Zhou Y, Takahashi S, Hayashi N, Engel JD, Yamamoto M., "The mouse GATA-2 gene is expressed in the para-aortic splanchnopleura and aorta-gonads and mesonephros region"., *Blood.* 1999 Jun 15;93(12):4196-207.

[223] Ohuchi H, Yasue A, Ono K, Sasaoka S, Tomonari S, Takagi A, Itakura M, Moriyama K, Noji S, Nohno T., "Identification of cis-element regulating expression of the mouse Fgf10 gene during inner ear development"., *Dev Dyn.* 2005 May;233(1):177-87.

[224] Oren T, Torregroza I, Evans T.,"An Oct-1 binding site mediates activation of the gata2 promoter by BMP signaling"., *Nucleic Acids Res.* 2005 Aug 1;33(13):4357-67.

[225] Paninski, L. , "Estimation of entropy and mutual information". *Neural Computation* 15: 1191-1254, 2003.

[226] Petrascheck M, Escher D, Mahmoudi T, Verrijzer CP, Schaffner W, Barberis A.,"DNA looping induced by a transcriptional enhancer in vivo"., *Nucleic Acids Res.* 2005 Jul 7;33(12):3743-50.

[227] Pennacchio, L. A., Ahituv, N., Moses, A., Prabhakar, S., Nobrega, M., Shoukry, M., Minovitsky, A., Dubchak, I., Holt, A., Lewis, K., Plazer-Frick, I., Akiyama, J., DeVal, S., Afzal, V., Black, B., Couronne, O., Eisen, M., Visel, A., and Rubin, E.M. 2006., "In vivo enhancer analysis of human conserved non-coding sequences", *Nature*, 444(7118):499-502.

[228] L.A. Pennacchio, G.G. Loots, M.A. Nobrega, and I. Ovcharenko, "Predicting tissue-specific enhancers in the human genome", *Genome Research*, 17(2), 201-11 (2007)

[229] J. Ramsay, B. W. Silverman, Functional Data Analysis (Springer Series in Statistics), Springer 1997.

[230] Rao A, Hero AO, States DJ, Engel JD, "Inference of biologically relevant Gene Influence Networks using the Directed Information Criterion", *Proc. of the IEEE Conference on Acoustics, Speech and Signal Processing*, 2006.

[231] Rao A, Hero AO, States DJ, Engel JD, "Using Directed Information to build Biologically Relevant Influence Networks", *Proc. Computational Systems Bioinformatics (CSB)*, 2007.

[232] Royce TE, Rozowsky JS, Gerstein MB., "Assessing the need for sequence-based normalization in tiling microarray experiments"., *Bioinformatics.* 2007 Apr 15;23(8):988-97.

[233] Said MR, Begley TJ, Oppenheim AV, Lauffenburger DA, Samson LD.,"Global network analysis of phenotypic effects: protein networks and toxicity modulation in Saccharomyces cerevisiae",. *Proc Natl Acad Sci U S A.* 2004 Dec 28;101(52):18006-11.

[234] Schug J., Schuller W-P., Kappen C., Salbaum J.M., Bucan M., Stoeckert C.J. Jr., "Promoter Features Related to Tissue Specificity as Measured by Shannon Entropy"., *Genome Biology* 6(4): R33, March 2005.

[235] "Social Network Analysis: A Handbook" , by John P Scott, Sage Publications Ltd; Second Edition, March 2000.

[236] Segal E, Fondufe-Mittendorf Y, Chen L, Thstrm A, Field Y, Moore IK, Wang JP, Widom J.,"A genomic code for nucleosome positioning".,*Nature.* 2006 Aug 17;442(7104):772-8.

[237] Sharan R, Suthram S, Kelley RM, Kuhn T, McCuine S, Uetz P, Sittler T, Karp RM, Ideker T., "Conserved patterns of protein interaction in multiple species"., *Proc Natl Acad Sci U S A.* 2005 Feb 8;102(6):1974-9.

[238] Simonis M, Kooren J,de Laat W (2007) "An evaluation of 3C-based methods to capture DNA interactions"., *Nature Methods* 4(11): 895.

[239] "Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sciences)", by Stanley Wasserman, Katherine Faust, Cambridge University Press; First Edition, November 25, 1994.

[240] Stuart RO, Bush KT, Nigam SK, "Changes in gene expression patterns in the ureteric bud and metanephric mesenchyme in models of kidney development", *Kidney International*,64(6),1997-2008,December 2003.

[241] Sumazin P, Chen G, Hata N , Smith A D., Zhang T, Zhang M Q., "DWE: Discriminant Word Enumerator", *Bioinformatics*, 21(1):31-38, 2005.

[242] Visel A, Minovitsky S, Dubchak I, Pennacchio LA., "VISTA Enhancer Browser–a database of tissue-specific human enhancers"., *Nucleic Acids Res.* 2007 Jan;35(Database issue):D88-92.

[243] Willett R, Nowak R, "Complexity-Regularized Multiresolution Density Estimation", *Proc. Intl Symp. on Information Theory*, ISIT 2004.

[244] Xu, L.; Krzyzak, A.; Suen, C.Y., Methods of combining multiple classifiers and their applications to handwriting recognition, *Systems, Man and Cybernetics, IEEE Transactions on* Volume 22, Issue 3, May-June 1992 Page(s):418 - 435.

[245] M. Belkin, P. Niyogi, "Laplacian Eigenmaps for Dimensionality Reduction and Data Representation", *Neural Computation*, June 2003; 15 (6):1373-1396.

[246] Budanitsky, A., and G. Hirst, "Semantic Distance in WordNet: An Experimental, Application-oriented Evaluation of Five Measures, Workshop on WordNet and Other Lexical Resources", in the North American Chapter of the Association for Computational Linguistics (NAACL-2001), Pittsburgh, PA, June 2001.

[247] Subramanian, A. and Tamayo, P. Mootha, V. K. and Mukherjee, S. and Ebert, B. L. and Gillette, M. A. and Paulovich, A. and Pomeroy, S. L. and Golub, T. R. and Lander, E. S. and Mesirov, J. P. (2005). "A knowledge-based approach for interpreting genome-wide expression profiles". *Proc. Natl. Acad. Sci. USA* 102, pg 15545-15550.

[248] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.*, Vol. 58, No. 1, pages 267-288).

[249] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004), Least angle regression, *Annals of Statistics* 32(2), 407499. (with discussion).

[250] Andersen MC, Engstrm PG, Lithwick S, Arenillas D, Eriksson P, Lenhard B, Wasserman WW, Odeberg J., "In silico detection of sequence variations modifying transcriptional regulation.", *PLoS Comput Biol.* 2008 Jan;4(1):e5.

[251] Tolstorukov MY, Choudhary V, Olson WK, Zhurkin VB, Park PJ., "nuScore: a web-interface for nucleosome positioning predictions.", *Bioinformatics.* 2008 Apr 29.

[252] Segal E, Fondufe-Mittendorf Y, Chen L, Thstrm A, Field Y, Moore IK, Wang JP, Widom J., "A genomic code for nucleosome positioning.", *Nature.* 2006 Aug 17;442(7104):772-8.

[253] Levine M, Tjian R., "Transcription regulation and animal diversity.", *Nature.* 2003 Jul 10;424(6945):147-51.

[254] Maeda RK, Karch F., "Ensuring enhancer fidelity".,*Nat Genet.* 2003 Aug;34(4):360-1.

[255] Szutorisz H, Dillon N, Tora L., "The role of enhancers as centres for general transcription factor recruitment"., *Trends Biochem Sci.* 2005 Nov;30(11):593-9.

[256] Lanckriet, G.R.G., De Bie, T., Cristianini, N. , Jordan, M.I., Noble, W.S. "A statistical framework for genomic data fusion", *Bioinformatics*, 20, 2626-2635, 2004.

[257] Costa J.A, Hero A.O, "Classification Constrained Dimensionality Reduction", *Proceedings of the IEEE International Conference in Acoustics, Speech and Signal Processing*, ICASSP 2005.

[258] Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T., and Ueda, N., "Learning systems of concepts with an infinite relational model", *Proceedings of the Twenty-First National Conference on Artificial Intelligence* (AAAI-06).

[259] Y. W. Teh, M. I. Jordan, M. J. Beal and D. M. Blei., "Hierarchical Dirichlet processes"., *Journal of the American Statistical Association*, 101, 1566-1581, 2006.

[260] F. Prez-Cruz, Z. Ghahramani and M. Pontil, (2007).,Conditional Graphical Models. In Predicting Structured Data, Edited by G. H. Bakir, T. Hofmann, B. Schlkopf, A. J. Smola, B. Taskar and S. V. N. Vishwanathan, MIT Press, September 2007.