

Computational Methods For Functional Motif Identification and Approximate Dimension Reduction in Genomic Data

by

Stoyan Georgiev

Department of Computational Biology and Bioinformatics
Duke University

Date: _____

Approved:

Uwe Ohler, Co-Supervisor

Sayan Mukherjee, Co-Supervisor

Elizabeth Hauser

Philip Benfey

Alexander Hartemink

Dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Department of Computational Biology and Bioinformatics
in the Graduate School of
Duke University

2011

ABSTRACT

Computational Methods For Functional Motif Identification and Approximate Dimension Reduction in Genomic Data

by

Stoyan Georgiev

Department of Computational Biology and Bioinformatics
Duke University

Date: _____

Approved:

Uwe Ohler, Co-Supervisor

Sayan Mukherjee, Co-Supervisor

Elizabeth Hauser

Philip Benfey

Alexander Hartemink

An abstract of a dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Department of Computational Biology and Bioinformatics
in the Graduate School of
Duke University

2011

Copyright © 2011 by Stoyan Georgiev
All rights reserved

Abstract

Uncovering the DNA regulatory logic in complex organisms has been one of the important goals of modern biology in the post-genomic era. The sequencing of multiple genomes in combination with the advent of DNA microarrays and, more recently, of massively parallel high-throughput sequencing technologies has made possible the adoption of a global perspective to the inference of the regulatory rules governing the context-specific interpretation of the genetic code. Extracting useful information and managing the complexity resulting from the sheer volume and the high-dimensionality of the data produced by these genomic assays has emerged as a major challenge. We address this challenge in our work by developing computational methods and tools, specifically designed for the study of the gene regulatory processes in this new global genomic context.

First, we focus on the genome-wide discovery of physical interactions between regulatory sequence regions and their cognate proteins at both the DNA and RNA level. We present a motif analysis framework that leverages the genome-wide evidence for sequence-specific interactions between trans-acting factors and their preferred cis-acting regulatory regions. The utility of the proposed framework is demonstrated on DNA and RNA cross-linking high-throughput data.

A second goal of this thesis is the development of scalable approaches to dimension reduction based on spectral decomposition and their application to the study of population structure in massive high-dimensional genetic data sets. We have developed computational tools and have performed theoretical and empirical analyses of their properties with particular emphasis on the analysis of the individual genetic variation measured by Single Nucleotide Polymorphism (SNP) microarrays.

Contents

| | |
|--|------------|
| Abstract | iv |
| List of Figures | ix |
| List of Abbreviations and Acronyms | ix |
| Acknowledgements | xii |
| 1 Introduction | 1 |
| 1.1 Contributions and Goals | 1 |
| 1.2 Thesis Outline | 1 |
| 2 Background | 3 |
| 2.1 Gene Regulation | 3 |
| 2.1.1 General Overview | 3 |
| 2.1.2 Structure of the Genetic Code | 5 |
| 2.1.3 Biogenesis of the Functional Gene Products | 5 |
| 2.1.4 Transcriptional Regulation | 7 |
| 2.1.5 Post-transcriptional Regulation | 13 |
| 2.2 Dimension Reduction and Inference of Population Structure | 16 |
| 2.2.1 Methods for Studying Population Structure | 17 |
| 2.3 Experimental Assays | 18 |
| 2.3.1 Genome-wide Chromatin Immunoprecipitation | 19 |
| 2.3.2 Genome-wide Detection of DNaseI Hypersensitive Sites | 20 |
| 2.3.3 Transcriptome-wide RNA Cross-linking | 21 |
| 2.3.4 Genome-wide SNP Microarrays | 24 |
| 3 cERMIT: An Approach for Motif Discovery Using Direct Binding Evidence | 25 |

| | | |
|----------|---|-----------|
| 3.1 | <i>De-novo</i> Transcription Factor Binding Site Discovery | 25 |
| 3.1.1 | Classic Formulation of the Motif Finding Problem | 25 |
| 3.1.2 | Genome-wide Quantitative Evidence | 26 |
| 3.1.3 | Reformulation of the Motif Finding Problem | 26 |
| 3.2 | The cERMIT Framework | 27 |
| 3.3 | Evidence of Regulation | 28 |
| 3.3.1 | P-values as Evidence of Regulation: Chip-chip data | 29 |
| 3.3.2 | ChIP-seq Reads as Evidence of Regulation | 30 |
| 3.4 | Integration of Evidence: Definition of the Objective Function | 30 |
| 3.4.1 | Random Set Scoring | 31 |
| 3.4.2 | Linear Regression Scoring | 32 |
| 3.5 | Search Strategy | 34 |
| 3.6 | Post-processing | 36 |
| 3.7 | Integrating Evolutionary Conservation | 37 |
| 3.8 | Significance Evaluation of Motif Enrichment | 38 |
| 3.9 | Conclusions | 38 |
| 4 | Applications of cERMIT to Studying the Transcriptional Regulation | 41 |
| 4.1 | Analysis of Transcriptional Regulation Using High-throughput DNA Binding Evidence | 42 |
| 4.1.1 | Yeast ChIP-chip Compendium | 42 |
| 4.1.2 | Human and Mouse ChIP-seq Experiments | 42 |
| 4.2 | Elucidating Regulatory Sequence from ChIP-chip Experiments | 43 |
| 4.2.1 | Performance Aspects | 46 |
| 4.2.2 | Assessment of False Positive and False Negatives | 46 |
| 4.2.3 | Novel Predictions | 47 |

| | | |
|----------|---|------------|
| 4.3 | Identification of Motifs from Deep Sequencing ChIP-seq Experiments | 48 |
| 4.3.1 | Peak Calling and Processing of Fseq Peaks | 50 |
| 4.3.2 | Analysis Results | 52 |
| 4.4 | Conclusions | 54 |
| 5 | Applications of cERMIT to Studying the Post-transcriptional Regulation | 56 |
| 5.1 | Analysis of RBP Regulation Using Transcriptome-wide RNA Cross-linking Data . | 56 |
| 5.2 | PAR-CLIP Datasets | 56 |
| 5.3 | Data Pre-processing | 57 |
| 5.4 | Analysis of RNA Binding Motifs | 59 |
| 5.4.1 | Sequence-specific RNA Binding Proteins | 60 |
| 5.4.2 | Enrichment Analysis of Argonaute-associated MicroRNAs | 61 |
| 5.5 | Conclusions | 63 |
| 6 | Randomized Dimension Reduction and Inference of Population Structure | 64 |
| 6.1 | Introduction | 64 |
| 6.2 | Statistical Methods and Algorithms | 64 |
| 6.2.1 | Notation | 66 |
| 6.2.2 | Principal Component Analysis | 68 |
| 6.2.3 | Dimension Reduction Via Graph Embeddings | 83 |
| 6.2.4 | Supervised Dimension Reduction | 87 |
| 6.3 | Results | 96 |
| 6.3.1 | Simulation | 96 |
| 6.3.2 | SNP Data | 100 |
| 7 | Conclusions | 103 |
| 7.1 | Direct Evidence of Regulation Improves <i>De Novo</i> Motif Discovery Predictions . . | 103 |

| | | |
|----------|---|------------|
| 7.1.1 | Genome-wide Binding Evidence | 103 |
| 7.1.2 | Transcriptome-wide Binding Evidence | 104 |
| 7.1.3 | Future Directions | 105 |
| 7.2 | Extensions to the Framework for Inference of Population Structure | 106 |
| 7.2.1 | Statistical Inference of the Dimensionality of the Population Structure . . . | 106 |
| 7.2.2 | Theoretical Results for Localized Sliced Inverse Regression | 107 |
| A | Website With Software and Data Sets | 108 |
| A.1 | TF Binding Motif Discovery | 108 |
| A.2 | PAR-CLIP Motif Analysis | 108 |
| A.3 | Randomized Eigendecomposition Analysis | 108 |
| | Bibliography | 109 |
| | Biography | 126 |

List of Figures

| | | |
|-----|---|-----|
| 2.1 | DNA double helix (adapted from http://science.plazza.us/dna-structure). | 3 |
| 2.2 | Gene structure (Courtesy: National Human Genome Research Institute). | 5 |
| 2.3 | Structure similarity between DNA and RNA (adapted from www.accessexcellence.org) | 7 |
| 2.4 | Structure of the eukaryotic promoter (adapted from Levine <i>et al.</i> [110]). | 10 |
| 2.5 | MicroRNA biogenesis (adapted from Bushati <i>et al.</i> [22]). | 15 |
| 2.6 | Agreement between genetic map and geography (adapted after Novembre <i>et al.</i> [146]). | 17 |
| 2.7 | A summary of the ChIP-chip procedure (adapted from Buck and Lieb [21]) | 19 |
| 2.8 | A summary of the PAR-CLIP procedure (adapted from Hafner <i>et al.</i> [74]) | 22 |
| 3.1 | cERMIT motif discovery algorithm | 28 |
| 4.1 | Comparison of cERMIT with other motif finders (adapted after Linhart <i>et al.</i> [119]). | 43 |
| 4.2 | Motif discovery pipeline | 49 |
| 4.3 | cERMIT predictions on human ChIP-seq datasets from [7, 91, 161, 191, 204] | 51 |
| 4.4 | cERMIT predictions on mouse ChIP-seq data from [25] | 53 |
| 5.1 | Top enriched microRNAs based on the Argonaute PAR-CLIP data from [74] | 62 |
| 6.1 | Similar matrix dimensions result in increased runtime gains over exact PCA | 97 |
| 6.2 | Different matrix dimensions result in decreased runtime gains over exact PCA | 98 |
| 6.3 | Application of Randomized PCA to a spiked Wishart covariance structure | 99 |
| 6.4 | Estimated axes of variation (PC2, PC3) based on data from Novembre <i>et al.</i> [146]. | 100 |

List of Abbreviations and Acronyms

Abbreviations:

| | |
|------------|---|
| ChIP | Chromatin immunoprecipitation |
| ChIP-chip | ChIP followed by microarray |
| ChIP-seq | ChIP followed by high-throughput sequencing |
| DNaseI-seq | DNaseI digestion followed by high-throughput sequencing |
| DNA | Deoxyribonucleic acid |
| LPP | Locality Preserving Projections |
| LSIR | Localized Sliced Inverse Regression |
| MCMC | Markov chain Monte Carlo |
| mRNA | Messenger RNA |
| miRNA | Micro RNA |
| PCA | Principal Component Analysis |
| PSSM | Position specific scoring matrix |
| RNA | Ribonucleic acid |
| SIR | Sliced Inverse Regression |
| SNP | Single Nucleotide Polymorphism |
| SDR | Sufficient Dimension Reduction |
| SVD | Singular Value Decomposition |
| TF | Transcription factor |
| TFBS | Transcription factor binding site |
| TSS | Transcription start site |

IUPAC codes:

- A Adenine
- C Cytosine
- G Guanine
- T Thymine
- R Purine (A or G)
- Y Pyrimidine (C or T)
- M C or A
- K T or G
- W T or A
- S C or G
- N A or C or G or T

Acknowledgements

I would first like to thank my advisers, Uwe Ohler and Sayan Mukherjee for their guidance and support throughout my doctoral studies. They have had the patience and perseverance to teach me about the field of computational biology and applied statistical modeling, helping me overcome many challenges along the way. Sayan has introduced me to my mentor and collaborator Nick Patterson, which lead to an exciting project that became part of my thesis work and opened up interesting research directions to explore in the future. Special thanks go to Prof. David Dunson for being a wonderful teacher and an incredible inspiration to pursue statistical modeling in my future research in computational biology.

I would also want to thank the members of my thesis committee Prof. Alexander Hartemink, Prof. Philip Benfey, Prof. Alan Gelfand, and Prof. Elizabeth Hauser for their encouragement and helpful advice.

During my time at Duke I have had the fortune to work with a few other exceptional people who have directly or indirectly helped me a lot grow as a person and a researcher. In particular I would like to thank Bala Rajaratnam, Raluca Gordân, David Corcoran, Greg LaMonte, Neel Mukherjee, Xinarui Cheng, Molly Megraw, Jalean Petricka, Miguel Moireno-Risueno, Gunjan Verma, Karthik Jayasurya, Arpan Roy, Souvik Sen, Rossen Dzhagalov, Ivan Dzhagalov however, there are many others who are not listed here.

I would like to thank my family who have always provided the solid foundation for all my academic and personal achievements, especially my wife Cynthia, my brother Grigor, and my father Georgi. They have been an inspiration and an unwavering support for which I am eternally grateful.

Finally, a most special thanks to my mother–Minka, who is the main reason that this work has come into existence. She was with me at the beginning of my adventurous journey into research but unfortunately could not make it to the end. Thanks mom!

Chapter 1

Introduction

1.1 Contributions and Goals

One of the major goals of this work is the development of tools for the study of gene regulatory processes at the cellular level. Our focus is on leveraging the output from recently introduced high-throughput technologies that allow for genome-wide discovery of physical interactions between regulatory sequence regions and their cognate proteins. A second goal of this thesis is the development of scalable approaches to dimension reduction based on efficient approximate methods for spectral decomposition and their application to the study of population structure in massive high-dimensional genetic data sets.

Both goals deal with one of the major challenges in biological research of the post-genomic era: how to extract useful information from (possibly large number of) high-dimensional genomic measurements from various biological contexts.

1.2 Thesis Outline

The chapters are organized as follows:

- Chapter 2 contains a description of the biological processes in the cell involved in the generation of gene products. I introduce the high-throughput assays used to probe the regulation of these processes that will be of major interest in this work as well as some existing analysis strategies.
- Chapter 3 describes an approach to cis-regulatory pattern discovery based on direct binding evidence for both transcriptional (transcription factors) and post-transcriptional regulators (RNA-binding proteins).

- Chapter 4 describes the application of the proposed motif analysis approach to the study of transcriptional regulation based on genome-wide DNA binding evidence.
- Chapter 5 describes the application of the proposed motif analysis approach to the study of post-transcriptional regulation based on transcriptome-wide cross-linking data, both in the context of sequence-specific binding factors as well as Argonaute-mediated microRNA regulation.
- Chapter 6 describes a framework for approximate dimension reduction using randomized algorithms. A specific application of interest of the proposed methodology is the inference of population structure in genetic data from massive high-dimensional Singular Nucleotide Polymorphism (SNP) data sets.
- Chapter 7 contains a brief summary of the ideas and major contributions of this thesis and a discussion of possible extensions.

Chapter 2

Background

2.1 Gene Regulation

In this section I provide an introduction to the biology of the genetic code contained in the cells of all living organisms with a specific emphasis on the regulatory processes involved in the generation of functional gene products. These processes together modulate the cellular response to environmental stimuli, developmental cues, and disease and characterize the dynamic nature of the cell. I conclude the section with a description of some experimental technologies that can be used to address questions related to the regulation of genes adopting a global quantitative perspective.

2.1.1 General Overview

Cells are the building blocks of life in an organism. Some organisms are unicellular, e.g. bacteria, while other organisms, such as humans, are multicellular. Humans have about 100 trillion cells. Each cell contains a copy of the entire set of hereditary instructions for development, functioning, and passing life on to the next generation for the whole organism in its *genome*.

In this thesis, I focus on eukaryotic organisms which contain a membrane-delineated compartment, the *nucleus*, that houses the cell's genome, separating it from the remaining cellular compo-

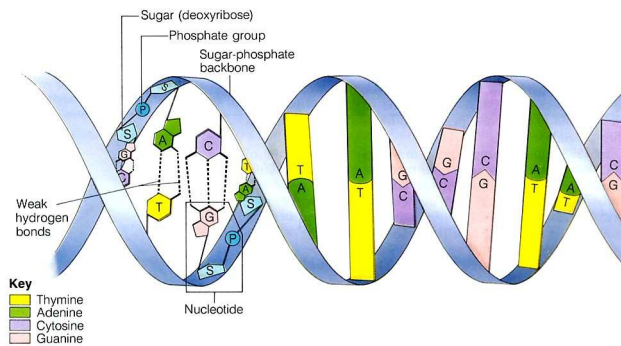


Figure 2.1: DNA double helix (adapted from <http://science.plazza.us/dna-structure>).

nents contained in the intra-cellular space—the *cytoplasm*. The genome is partitioned and packaged into a set of molecules called *chromosomes* which help a cell to keep the large amount of genetic information organized, and compact. Each chromosome is composed of information-carrying double-stranded Deoxyribonucleic acid (DNA) wrapped around structural units called histones. As the organism grows and develops, cells divide and the full set of chromosomes is duplicated in the process of DNA replication, providing each cell its own complement of the hereditary material.

DNA consists of two long polymer strands of repeating units called *nucleotides* held together in a double helix via hydrogen bonds formed by complementary interactions (base-pairing). Each nucleotide has three parts: a sugar molecule, a phosphate molecule, and a nitrogenous base. The nitrogenous base carries the genetic information and comes in four varieties: adenine (A), thymine (T), guanine (G), and cytosine (C), the letters of the genetic alphabet. Each type of base on one strand interacts with just one type of base on the other strand with A bonding only to T, and C bonding only to G. Figure 2.1 depicts the molecular structure of DNA.

Humans have 23 pairs of chromosomes, with 22 *autosomes* and two sex chromosomes. In the case of sexual reproduction, each parent contributes one chromosome to its child, resulting in half of the genetic material from the mother and the other half from the father. Hence, each position on the human DNA has two variants of the complementary base-pairs called *alleles*. In the majority of cases, the combination of variants at each position are the same across individuals. The DNA locations when this is not the case are called Single Nucleotide Polymorphisms (SNPs). In humans there are few such locations (about 3.0×10^7 bp [170]) relative to the total size of the known genomic sequence (2.85×10^9 bp [30]). Hence, these can be enumerated and used to characterize an individual's heterogeneity, with wide-ranging applications in evolutionary history and disease studies. Genomic assays that provide measurements of the genetic heterogeneity are described in more detail in Section 2.3.4.

DNA is segmented into logical regions annotated according to their biological function. The regions that result in functional products are called genes, and can be classified in two broad categories: (protein) *coding* genes and *non-coding* genes.

2.1.2 Structure of the Genetic Code

The structure of a protein-coding gene is illustrated in Figure 2.2(a). The start sequence of the gene is called the 5' untranslated region (5UTR) and terminal sequence is called the 3' untranslated region (3UTR). Both these regions encode regulatory signals responsible for the *post-transcriptional* regulation of the encoded message. The internal part of a gene is composed of an alternating series of two distinct sequence types: *exons* and *introns*, which have specific sequence boundaries and characteristic sequence composition. Depending on the biological context, different combinations of exons are spliced together to form the mature message (*splice isoform*). All introns are excised out and serve no further purpose in the formation of the functional gene product—the *protein*.

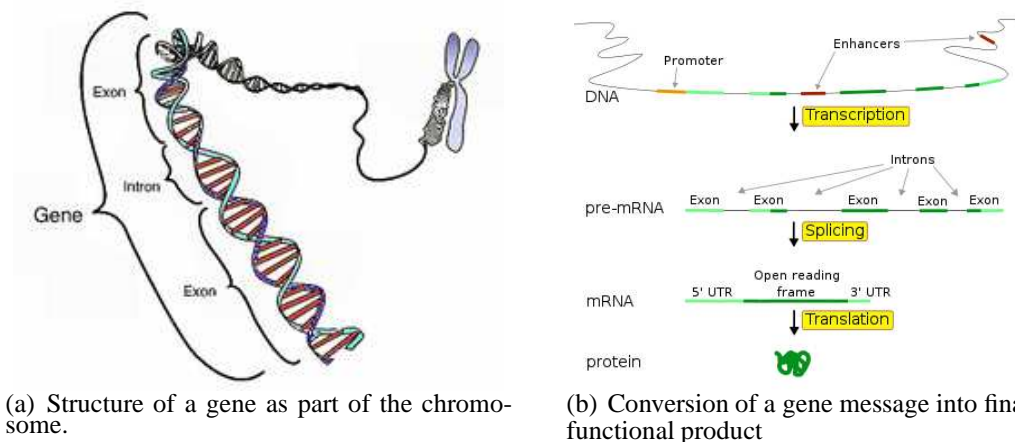


Figure 2.2: Gene structure (Courtesy: National Human Genome Research Institute).

2.1.3 Biogenesis of the Functional Gene Products

In eukaryotes, transcription occurs in the nucleus, where DNA resides. Although a functional gene product may be an RNA (see Section 2.1.5) or a *protein*, in this thesis I focus on studying the mechanisms regulating protein coding genes and hence I consider proteins as the functional unit of interest. After transcription, the DNA message is transferred, by means of a carrier molecule called *messenger RNA* (mRNA), to the cytoplasm where the translation machinery is located. The final step in the generation of a functional gene product is the translation of the mRNA into a protein.

Figure 2.2(b) represents a schematic of this process, which is referred to as *gene expression* and can be summarized by the following steps:

1. *Transcription.* Transcriptional regulatory signals result in the initiation of the copying (transcription) of the DNA sequence coding for the gene into an intermediary single-stranded molecule—pre-mRNA, which is another type of nucleic acid that is very similar to DNA (see Figure 2.3).
2. *Processing:* Inside the nucleus a carefully regulated subset of the the pre-mRNA, the *exons*, is *spliced* together, skipping the intermediate sequence, the *introns*, to produce the mature RNA message. In some cases only a subset of exons is selected to be a part of the final mRNA. This is a carefully controlled process that provides a mechanism for producing different variants (*isoforms*) of the same protein as a function of the specific biological contexts.
3. *Export.* the mRNA is exported from the nucleus into the cytoplasm, where it is subject to further regulatory control that determines if it is degraded, stored in some cellular compartment for later use, or converted into protein.
4. *Translation.* the encoded message in the mRNA is converted into a chain of amino-acids which are then folded into a protein product.

Each stage of the biogenesis of the gene products described above contains multiple checkpoints which are controlled by various regulatory mechanisms to ensure the correct interpretation of the genetic code. Some steps in protein biogenesis occur in parallel and hence many of the regulatory processes are tightly coupled, yet there are two major stages of regulation which are delineated by the transcription of the gene into its mRNA message: *transcriptional* regulation and *post-transcriptional* regulation. In this work I focus on developing computational methodology and tools for the analysis of the regulatory logic at both stages of regulation.

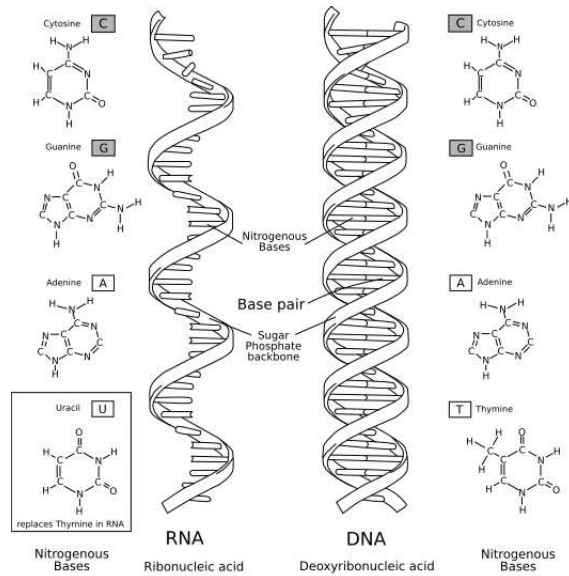


Figure 2.3: Structure similarity between DNA and RNA (adapted from www.accessexcellence.org)

The RNA Message

Even though RNA is very similar to DNA, there are significant differences that have functional consequences. Unlike DNA, RNA is *single stranded* and contains *Uracil* (U) instead of thiamine (T). Another important distinction between RNA and DNA is that they differ in their type of *sugar connector* between the phosphate backbone (see Figure 2.3), which allows RNA to form complex three-dimensional structures and perform a variety of regulatory functions in the cell related to the interpretation of the genetic program embedded in DNA. Finally, mRNA messages are converted into functional products—proteins, serving as necessary intermediaries, while DNA cannot directly serve this purpose.

2.1.4 Transcriptional Regulation

An important checkpoint in the regulation of gene expression is at the stage of transcription initiation, which includes the accessibility of regulatory DNA regions as well as the recruitment of DNA binding factors.

The control of these processes is mediated by a class of proteins called Transcription Factors

(TFs) that directly interact with factor-specific regulatory elements in DNA (a.k.a. *cis-regulatory elements*). These cis-regulatory elements tend to be short (about 6–15 bps in eukaryotes) and often highly degenerate, which makes it difficult to distinguish them from the surrounding sequence [148, 188, 23]. In order for TFs to be able to distinguish the regulatory binding sites from the vast majority of other DNA sequence, the binding locations need to *structurally accessible* and able to form a duplex with the active site of the corresponding TF that has energetically favorable electrostatic properties. Hence, the functional properties of a location within DNA are determined by both *static* and *dynamic* control mechanisms. Static mechanisms correspond to the sequence content which should match the binding preferences of a regulatory TF. Important dynamic modulators of transcription initiation are the specific concentrations of the regulatory TF proteins present in the nucleus as well as the DNA structural state that can either facilitate or impede general binding by regulatory factors.

Constructing models of TF binding sites (*motifs*) and the identification of their functional occurrences within a pre-specified subset of regulatory regions in the genome is termed *motif discovery* and will be of major interest in the current work. In particular, I focus on the *genome-wide* discovery of such functional DNA elements, leveraging the recently introduced technologies for measuring the *in vivo* (inside the living cells) general accessibility of DNA as well as the degree of occupancy by a specific TF of interest (see Section 2.3.2 and Section 2.3.1). Knowledge of the full complement of cis-regulatory elements bound in different biological contexts will help improve our understanding of context-specific transcriptional regulation and is an important input for the reconstruction of the gene regulatory networks in the cell.

About 5-7% of an eukaryotic genome encodes for TFs [202], and they can be divided in two major functional sub-categories based on their effect on expression of their target genes: *activators* (binding results in increase of the target's expression) and *repressors* (binding results in decrease of the target's expression).

Next I describe in detail the different types of regulatory DNA regions encoding the transcription initiation regulatory logic for individual genes. There are four major types of such regions: *pro-*

moters (contains DNA signals necessary for the initiation of transcription), *enhancers* (responsible for context-specific up-regulation of gene expression), *silencers* (responsible for context-specific down-regulation of gene expression), and *insulators* (limits scope of regulatory interactions between genes).

Promoters

Promoter regions correspond to the stretches of DNA a few hundred base pairs immediately upstream of the gene Transcription Start Site (TSS) and contain signals responsible for the recruitment of the *general transcription machinery* common to all genes. A promoter is further subdivided into a *core* promoter region, immediately adjacent to the TSS, and a *proximal* promoter region, >50bp upstream of TSS.

A gene can have more than one promoter, resulting in gene copies of different length [133]. Promoters tend to have some characteristic features in common that allow for experimental and computational techniques to be used for their detection in the genomic DNA [148, 189, 100]. Recently, it has become clear that promoters can be divided into two classes: those that have a single TSS and those that have a broad or dispersed range of TSS over a 50-100bp region [156, 87]. Next-generation sequencing technologies in combination with novel experimental protocols have allowed for comprehensive mapping at an improved resolution of the different promoter types [141]. Most genes in higher eukaryotes seem to be under the transcriptional control of dispersed promoters, but further studies of these two classes of promoters is necessary to elucidate their distinct functional characteristics.

The core promoter regions are enriched in a number of cis-regulatory elements that aid in the anchoring and assembly of the transcription initiation machinery—the *pre-initiation complex* (PIC) e.g. TATA box, Initiator element (INR), TFIIB recognition element (BRE), Downstream Promoter Element (DPE) etc. [181, 182]. Some of these are remarkably well-conserved across different eukaryotic species, yet there is no single binding site that is universally present in all promoter regions. Even the ubiquitously distributed TATA box, involved in the recruitment of the TATA-

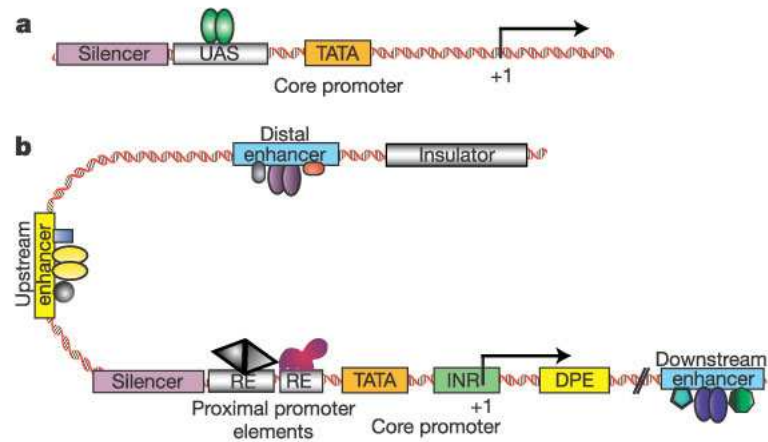


Figure 2.4: Structure of the eukaryotic promoter (adapted from Levine *et al.* [110]).

binding protein (TBP) a central component of the transcription initiation process, is present in at most 30% of eukaryotic promoters [38]. Rather, there seem to functionally equivalent combinations of factors that are required for the initiation of transcription to proceed.

Enhancers

In eukaryotes, many promoters by themselves are insufficient to drive the *in vivo* production of biologically relevant levels of gene transcripts and they need additional assistance from activation signals that are encoded in the *enhancer* regions. Enhancers are found mostly inside intergenic sequence, but also within genes sometimes even as far as 10,000bp (in *Drosophila*) or 100,000bp (in human and mouse) away from the boundaries of the gene they regulate [110, 29, 104]. Still they are able to confer the specificity to the gene transcription initiation process required for the normal functioning of the cell. A schematic representation of a gene with its associated regulatory sequence is illustrated in Figure 2.4 for both the case of unicellular eukaryotes (a) and more complex, higher eukaryotes (b). The exact mechanisms governing the functioning of enhancers is still under active investigation, but it is well known that there is a strong combinatorial aspect to the encoded regulatory logic, as the activation process typically requires the binding of several TFs to their cognate cis-regulatory motifs. Many enhancers have been found within non-coding sequences

that are highly conserved through evolution (e.g. fish to mouse) [144] and regulate gene expression in highly temporal or tissue-specific manner [178].

Reliably identifying functional enhancers in higher eukaryotes has been a major challenge due to the large amount of candidate intergenic sequence (in human over 98.5% of the total size of the DNA). Recent progress has been made in that direction thanks to the advent of high-throughput sequencing technologies which have enabled the profiling of the whole human genome for regions bound by proteins characteristic of most enhancers [193] and also regions enriched in general features typical of enhancers, like open chromatin [19].

Silencers

Silencer regions serve a role that is complementary to the enhancer regions, as they are involved in the repression of gene expression. Two distinct classes of silencers exist: short, position-independent motifs that via their bound repressor TFs actively obstruct the assembly of the Transcription pre-initiation complex and are typically found upstream of TSS. The second category of silencer sequence motifs are position-dependent that passively prevent the binding of activator TFs to their cis-regulatory elements and can be found both upstream and downstream of the TSS [147].

Insulators

Both silencers and enhancers can act on multiple genes, but sometimes their regulatory effects need to be localized. The insulator elements are cis-regulatory sequences that can block such interactions. Two distinct classes of insulator elements have been discovered: *enhancer-blocking* insulators that interfere with enhancer-promoter interaction if located in between the two and *barrier* insulators that interfere with DNA accessibility of promoter and enhancer regions, demarcating the boundaries between active and inactive DNA regions [60]. Enhancer blocking insulators I first described by [86] who studied the *Drosophila* insulator element *gypsy*. In higher eukaryotes a large number of the currently known enhancer-blocking insulator elements contain the cis-regulatory motif for the CTCF protein [12] (for further discussion of CTCF see Section 4.3).

Evolutionary Conservation and Function

The genetic code of living organisms has evolved over time allowing for changes to occur in different locations within DNA. Identification of binding motifs that match the preferences of TFs and are in fact recognized and bound *in vivo* can be aided by the observation that not all DNA has evolved at the same rate and parts of the genome containing functional elements tend to be constrained to change less than other regions [142]. Hence, evolutionarily related species contain quite similar regulatory code which can be used to identify functional regions within DNA. *Phylogenetic footprinting* [14, 15, 109, 139] is one such approach specifically designed to aid in the identification of functional occurrences of cis-regulatory motifs.

Applying phylogenetic footprinting on a single gene level was first introduced by Tagle and colleagues in 1988, who applied the technique to find evolutionarily conserved cis-regulatory elements responsible for embryonic ϵ and γ globulin gene expression in primates [186]. More recently, a large number of computational approaches have been proposed that incorporate evolutionary information as part of functional motif discovery [15, 109, 177, 97]. Given a set of putatively co-regulated genes within an organism of interest those approaches use regulatory regions from orthologous genes (inherited from a common ancestor, but evolved separately within each species) in related species to look for motifs that are overrepresented and highly conserved.

Phylogenetic footprinting does not provide a universal approach to motif discovery as it has some inherent limitations: not all functional cis-regulatory elements fall within strongly conserved DNA regions and even when they do, the aligning of the sequences from related species to produce the evolutionary “footprints” that contain functional motifs could be a very challenging task. Nevertheless, “ensemble” approaches that carefully incorporate evolutionary conservation information as a part of the motif discovery, have tended to produce fewer false positive predictions than comparable approaches that do not use conservation (sometimes at the cost of fewer true positives).

2.1.5 Post-transcriptional Regulation

As it is evident from Figure 2.2(b) the protein production is a multi-step process that can be regulated not only at the *transcription initiation* step but also *post-transcriptionally*, especially when fine-tuning of the level of gene transcripts is to be achieved. RNA binding proteins (RBPs), sometimes with the help of non-coding RNAs (microRNAs) are important mediators of the post-transcriptional control of mRNA messages. Next I describe in more detail the nature and function of these important regulators.

RNA Binding Proteins

RNA binding proteins (RBPs) play important roles in the life cycle of the protein-coding transcript, from its transcription based on a DNA template until its decay by RNases [130]. All steps of RNA processing and function including splicing, nuclear export, localization, stability, and small RNA-mediated regulation are controlled by different RNA-binding proteins (RBPs) and ribonucleoproteins (RNPs) (For a review, see [94]). The identification of which RBPs or RNPs interact with which transcripts, how they interact, and where the interaction occurs, has been the focus of many studies [183, 18].

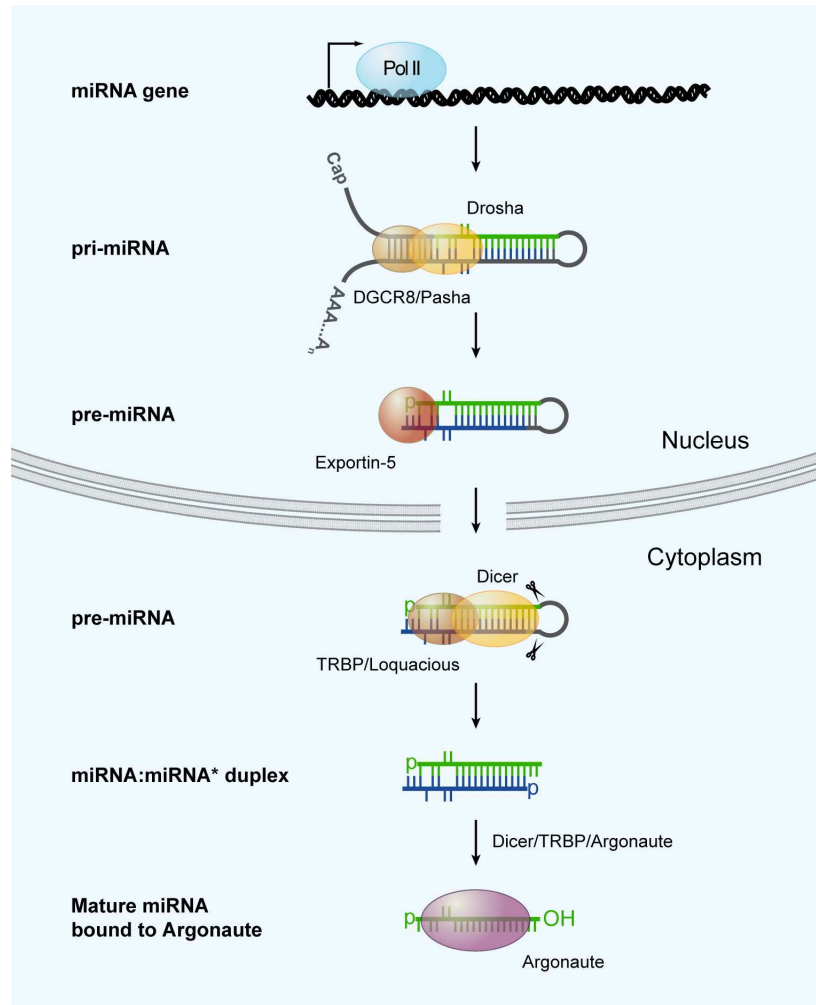
Architecture of RBPs RBPs are defined through their ability to interact with mRNAs via RNA-binding domains. Similarly to TFs interacting with the regulatory regions of gene in order to control transcription initiation, RBPs interact with specific RNA sequence patterns to effect post-transcriptional control. Currently, there are over 1 100 human genes in the Pfam database that are annotated to be RBPs [52], yet there are only about 40 different types of RNA-binding domains [122]. This observation suggests that the combinatorial complexity of different domain arrangement is key to providing the functional diversity necessary for the regulation of the various steps of RNA metabolism and function. The next section focuses on an important class of RBP-mediated regulation—with the participation of small regulatory RNAs.

MicroRNA Regulation

MicroRNAs (miRNAs) control gene expression post-transcriptionally by regulating the mRNA *stability* or *translation* in the cytoplasm. They are approximately 21nt long RNA regulators of gene expression which, by pairing to the 3'UTR of mRNAs of protein-coding genes, direct their repression. The discovery of miRNAs [107, 197, 158] has revealed a new dimension in our understanding of post-transcriptional regulation of eukaryotic gene expression in the cell. More than a thousand new miRNAs have been discovered in plants, animals and viruses [103]. In mammals, miRNAs are predicted to control the activity of more than 60% of all protein-coding genes [57] and have been shown to participate in the regulation of a wide variety of cellular process and disease [9].

miRNA Biogenesis Next I describe a summary of the known mechanism of miRNAs biogenesis illustrated in Figure 2.5 (For a review, see [22]).

1. Similar to protein-coding genes, miRNA are transcribed from a DNA template to generate a stem structure called the *primary miRNA* (pri-miRNA) that can vary in size from hundreds to thousands of bp.
2. The pri-miRNA is processed within the nucleus to form a ~70nt double-stranded hairpin precursor called *pre-miRNA*.
3. The pre-miRNA is exported outside of the nucleus with the help of transporter proteins that recognize the characteristic 2nt 3' overhang present at the end of the hairpin.
4. Inside the cytoplasm the pre-miRNA is cleaved to produce a 21nt miRNA:miRNA duplex.
5. Once incorporated into the RNA Induced Silencing Complex (RISC), with the help of its active catalytic unit—the Argonaute protein complex—the miRNA guides the complex to its targets by complementary base-pair interaction. In case of perfect complementarity (common in plants) the target is cleaved, while partial complementarity (common in animals) results in reduction of gene expression by means of translation inhibition or transcript destabilization.



 Bushati N, Cohen SM. 2007. *Annu. Rev. Cell Dev. Biol.* 23:175–205

Figure 2.5: MicroRNA biogenesis (adapted from Bushati *et al.* [22]).

miRNA Function Early studies in the worm *Caenorhabditis elegans* [172] and more recent studies in mammalian cells [127, 145, 152] have provided strong evidence that miRNAs repress protein synthesis at the level of translation initiation as they were associated with mRNA that are being actively translated. Another mechanism for inhibition of the production of protein synthesis was demonstrated to be taking place at the mRNA level by degradation of the target mRNA prior to or in cooperation with translational inhibition [64, 199, 50]. In support of this observation, upon inhibition of the miRNA pathway by depletion of RISC complex members [64, 157, 171] or upon deletion of specific miRNA members [5], the levels of predicted and validated miRNA targets increase. Conversely, overexpression of specific miRNAs results in lower abundance of the transcripts containing binding sites for those miRNAs. Even though, recent high-throughput studies combining proteomic and transcriptomic data suggest that the predominant reason for reduced protein product due to miRNA regulation is due to destabilization of target mRNAs rather than translational repression [73], the mechanistic details of the functioning of the miRNA is still an active area of research. For a recent review on miRNA regulatory function and target prediction see [9].

2.2 Dimension Reduction and Inference of Population Structure

Studying the genetic information encoded into DNA, as discussed so far, provides a microscopic view of the functioning of a living organism, at the cellular level and can be used to gain insight into the cellular processes and their regulation. The focus in such analyses is typically on the interaction among genes, from the viewpoint of their regulatory connections in the gene regulatory network.

An alternative view of the genome is to understand variation in the genome across a population of individuals, in this case the individual is the unit of interest. In Figure 2.6 Principal Components Analysis (PCA [149]) was used to produce a two-dimensional visual summary of the observed genetic variation on 1,387 European individuals. The resemblance between this genetic summary and the geographic map of Europe is clear. Major subdivisions into closely clustered subpopulations

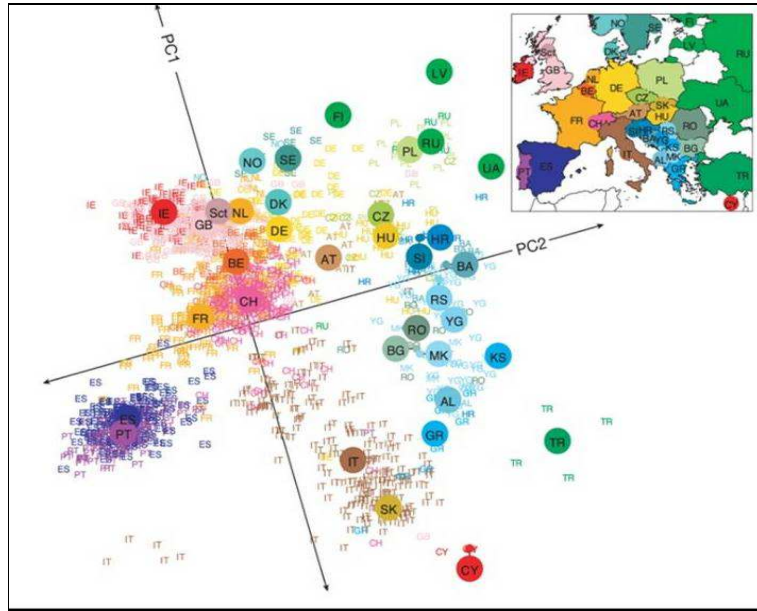


Figure 2.6: Agreement between genetic map and geography (adapted after Novembre *et al.* [146]).

from similar geographic areas are distinguishable. Even within some countries individuals are well differentiated along the principal component (PC) axes. This is an illustration that DNA sequence variation typically reflects the *subpopulation structure* in a sample of individuals which can be formally defined as the difference in allele frequency distributions between the subgroups forming the sample. Population structure has been a classic subject of study for geneticist for decades [129, 114]. It has had multiple applications to the study the demographic histories of populations of individuals [165] and is recognized as a confounder to the inference of disease associated alleles in case-control studies [56].

2.2.1 Methods for Studying Population Structure

There has been two dominant approaches to the analysis of population structure in genetic data. *Structure* [154] is a Bayesian model-based clustering method which assigns individuals to discrete clusters corresponding to subpopulations based on a multinomial probabilistic model for the allele frequencies at each locus of variation. This approach has become very popular and has been successfully applied in many genetic studies [165, 56, 180]. With the increased availability of SNP

microarrays, that probe the genetic variants of each individual sample at hundreds of thousands of bi-allelic loci, it has been problematic to practically apply Structure to such high-dimensional genotype data due to the inherent computational limitations of the sampling procedure underlying the Bayesian inference.

A more recent approach, based on Principal Component Analysis (PCA), *Eigenstrat*, proposed in [149], has successfully addressed this issue. In addition to a computationally fast procedure, the authors in [149] have proposed a principled statistical inference method to test for the presence of population structure, based on recent results from random matrix theory applied to the eigenvalues of the sample covariance matrix. Both theoretical and empirical arguments in [149] strongly suggest that the top few principal components (PCs) tend to capture distinct axes of variation that separate different subpopulations.

The computation time of Eigenstrat is $O(m^2n)$, where m is the number of individuals and n is the number of SNPs. Hence, the inference procedure scales favorably with increasing number of SNPs, which is a clear computational advantage over Structure, yet it scales *quadratically* with the number of individuals. This poses a computational challenge when analyzing *large* high-dimensional genetic data. In this work I address this challenge by proposing to adapt an approximate PCA procedure based on a *randomized* dimension reduction algorithm. The runtime of the proposed approach is $O(kmn)$, where k is on the order of the dimension of the population structure, which can typically be assumed to be small—not more than 10-20. Details regarding the proposed approach to approximate inference of the axes of variation in large high-dimensional data as well as extensions to other dimension reduction approaches based on eigendecomposition: Sliced Inverse Regression (SIR), Localized Sliced Inverse Regression (LSIR), Locality Preserving Projections (LPP), is the main topic of Chapter 6.

2.3 Experimental Assays

The advent of microarrays and the next-generation sequencing technology has opened up opportunities to start learning about the gene regulatory processes and the individual's genetic variation on

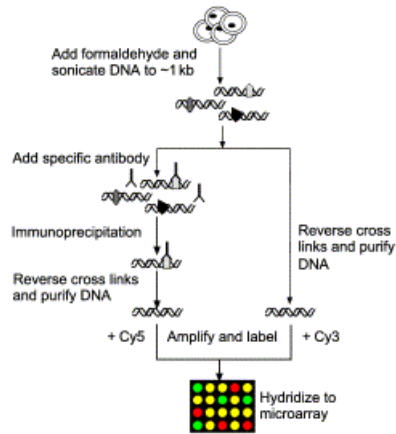


Figure 2.7: A summary of the ChIP-chip procedure (adapted from Buck and Lieb [21])

a global scale. In this section I describe four such assays.

2.3.1 Genome-wide Chromatin Immunoprecipitation

Chromatin immunoprecipitation (ChIP) is a method used to determine the location of DNA binding sites on the genome for a particular protein of interest [16, 164] within a pre-specified stretch of genomic DNA (using on PCR primers). This is an *in vivo* procedure that provides a way to assay the DNA-protein interactions in the cell nucleus.

ChIP-chip

More recently, a key additional step was added to the experimental protocol: hybridization against a microarray, which allows for the characterization of the binding interactions at the genome-wide level [21]. The ChIP-chip technology has been successfully used to characterize the binding preferences for a number of DNA binding proteins [159, 90, 108, 77].

Figure 2.7 illustrates the steps in a typical ChIP-chip experiment. The steps proceed as follows:

1. Cross-link the protein of interest to DNA *in-vivo*, using formaldehyde fixation.
2. Shear DNA, using sonication or micrococcal digestion, into fragments of size approximately 1kb.

3. Enrich in DNA fragments associated with the protein of interest using a protein-specific antibody.
4. Reverse the formaldehyde cross-linking. Purify, amplify and label DNA with a fluorescent dye (e.g. Cy5).
5. Purify, amplify and label *genomic* DNA using a different color fluorescent dye (e.g. Cy3).
6. Hybridize the two differentially tagged probes (e.g. Cy3 and Cy5) against a microarray containing all putatively bound sequence regions (typically genome-wide set) [21].

The output from the ChIP-chip genomic assay is a set of short sequence regions (approx 1kb) with assigned fluorescence intensities for the bound probe and the genomic control. These scores are typically reduced to a single binding score per region, which could be a (normalized) log-ratio of dye intensities or a p-value based on a null statistical model of binding intensities [108, 77].

ChIP-seq

Instead of hybridization to a microarray the Chromatin immunoprecipitation (ChIP) step can be followed by high-throughput sequencing instead, which was recently proposed as an alternative approach to genome-wide successfully used to characterize the binding preferences and locations of multiple DNA binding proteins [7, 91, 161, 191, 204].

2.3.2 Genome-wide Detection of DNaseI Hypersensitive Sites

In the nucleus, the vast majority of genomic DNA is wrapped around structural proteins called histones to form DNA-histone octamer complexes called nucleosomes. The nucleosomes are regularly spaced and serve the role of packaging DNA into more compact form. The degree of packaging has a strong impact on the regulation of gene transcription [51]. The DNaseI enzyme can be used to study regions of DNA where the nucleosomes have been displaced (such as for the activation of promoters) which allow for easier cleavage [128, 96]. Such locations are called DNase I hypersensitive (DH) sites. It has been shown that DH sites strongly correlate with the location of

active regulatory regions within the genome, including promoters, enhancers, silencers, insulators [71, 51]. Traditionally individual DNase I HS sites have been identified using Southern blot assays [198], which is a time-consuming process difficult to apply on a large scale. A recent application of the DHS protocol in combination with high-throughput sequencing (DNaseI-seq) has allowed [19] to create the first comprehensive genome-wide open chromatin map. A brief summary of the DNase-seq protocol contains the follows steps [39, 19]:

1. Start with chromatin that is digested with a small amount of DNase I that preferentially cuts at a DNaseI HS site.
2. Attach a biotinylated linker to the DNase I-digested ends.
3. Use the biotinylated linker to extract short adjacent DNA fragments.
4. Sequence the extracted DNA fragments using next-generation sequencing platform.

2.3.3 Transcriptome-wide RNA Cross-linking

Recent advancements in high-throughput genomic technologies have resulted in profiles of transcriptome-wide RNA-protein interactions *in vivo*. Two of the most established methods for the investigation of these interactions are RIP-Chip [95] and CLIP [190] or a combination of both, known as RIP-seq [174, 205].

RIP-Chip was the first method to use immunoprecipitation to identify RNA targets bound by specific RBPs at genome-wide scale. Associated mRNAs are isolated, and then quantified using mRNA arrays or, more recently, subjected to high-throughput sequencing. This allows for the identification of all transcripts targeted by a particular RBP, but not for direct identification of where, or how many, RNA-protein interactions occur within a transcript. The second method, CLIP, typically uses short wave UV 254 nm crosslinking followed by immunoprecipitation and partial RNase digestion of the bound transcript. Conversion of the residual RNA segments into cDNA libraries and characterization by high-throughput sequencing yields small size windows in which the RNA-protein crosslinking occurred.

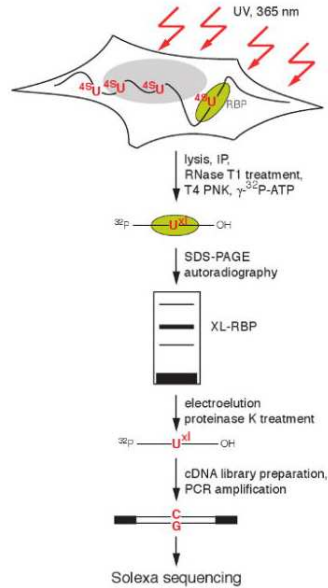


Figure 2.8: A summary of the PAR-CLIP procedure (adapted from Hafner *et al.* [74])

PAR-CLIP (Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation) is a powerful modification of the CLIP technology for the isolation of protein bound RNA segments [74]. Cells are first cultured with a photoreactive ribonucleoside analogue, typically 4-thiouridine (4SU), to boost RNA-protein crosslinking. This is followed by high-throughput sequencing of cDNAs generated from the crosslinked immunopurified RNA fragments. During cDNA generation, preferential base pairing of the 4SU crosslink product to a guanine instead of an adenine results in a thymine (T) to cytosine (C) transition in the PCR-amplified sequence, serving as a diagnostic mutation at the site of contact. The pattern of T-to-C conversions, coupled with read density, can thus provide a strong signal to generate a high-resolution map of confident RNA-protein interaction sites.

Figure 2.8 illustrates the steps in a typical PAR-CLIP experiment. The steps proceed as follows:

1. Culture cells in media containing 4-SU (biochemically modified Uracils)
2. Irradiate cells with UV, 365nm to induce RNA-protein cross-linking
3. Immunoprecipitate and select the gel band of the size corresponding to the protein of interest

4. Convert the RNA from the selected gel band into a cDNA library and amplify using PCR (incorporating U-to-G conversions during the reverse transcription step)
5. Sequence using next-generation sequencing platform [74].

Identifying Bound Regions

Generally speaking, there are three main steps in the analysis of binding data from deep sequencing experiments.

1. *Read alignment.* Sequence reads are aligned against the reference genome
2. *Peak calling.* Genomic regions significantly enriched in aligned reads are identified for further study
3. *Motif analysis.* Peaks are further analyzed to infer the binding affinity and target regions of the trans-acting element under study

The focus of this thesis is on step 3, the Motif analysis of the inferred sequence peaks, taking advantage of the information contained in the quantitative binding evidence provided by the number of reads aligned to each peak.

Current Approaches to Motif Discovery in High-throughput Binding Data

Once sequence peaks have been inferred based on the high-throughput binding data a set of putatively bound sequence regions needs to be defined, which is typically done by selecting an arbitrary subset from the peaks with highest number of aligned ChIP-seq reads. This set of likely co-bound regions is searched for overrepresented motifs that would be good candidates for binding sites for the TF under study. The overrepresentation is often defined with respect to a model for background sequence (e.g. Markov Model of 2nd or 3rd order).

There have been a large number of classic motif discovery tools, originally developed for motif analyses based on a small set of pre-defined co-bound sequence regions, that have been adapted to

the task of motif discovery using CHIP-seq data [6, 166, 150, 44]. More recently, discriminative analyses based on a positive and negative set of sequence regions as well as additional genome-wide information on nucleosome positioning and energetic stability of DNA have been incorporated as part of the motif search, to significantly improve the sensitivity and specificity of the reported results [136, 68].

2.3.4 Genome-wide SNP Microarrays

Single Nucleotide Polymorphism (SNP) arrays enable the study of individual genetic variation and have diverse applications in medical genetics and evolutionary biology. SNP array is a type of DNA microarray which is used to detect genetic variants (*polymorphisms*) within a population of individuals. A SNP in DNA (variation at a single site), is the most frequent type of variation in the genome. For example, there are around 30 million SNPs that have been identified in the human genome [176, 170] based on sequencing studies on multiple human individuals. As SNPs are highly conserved throughout evolution and within a population, they can serve as excellent genetic markers. Detection of population structure based on SNP measurements from a large number of individuals will be the main focus of Section 6.

Chapter 3

cERMIT: An Approach for Motif Discovery Using Direct Binding Evidence

3.1 *De-novo* Transcription Factor Binding Site Discovery

In this Section I describe a computational framework designed for the discovery of the functional binding site as well the occurrences in the genome based on genomic and transcriptomic high-throughput binding data from ChIP-chip, ChIP-seq, or PAR-CLIP experiments.

3.1.1 Classic Formulation of the Motif Finding Problem

The motif finding problem has been traditionally phrased as the following: *given a set of putatively co-regulated genes, find the optimal motif description and the set of occurrence locations in the corresponding regulatory regions*. Many popular approaches are based on iterative updating of a position specific scoring matrix (PSSM) representation of the binding site, which reflects the affinity of the protein to its functional sites. Stochastic searches in the form of Gibbs sampling [105, 166, 120], or EM-based algorithms [6], have been used extensively to address this goal by means of iteratively optimizing a suitable objective function. The use of additional information, such as the sequences from multiple related species, e.g. [177], or priors on the TF binding domain or nucleosome positions [137, 136], has lead to noticeable improvements in the performance of these strategies. As alternative to PSSMs, motifs can be described by a simpler representation of consensus strings over a degenerate alphabet. This representation allows for the exhaustive identification of motifs which are over-represented compared to a genomic background model [192, 82, 179], and frequently use efficient data structures such as suffix arrays to search for overrepresented oligomers [150, 153]. This strategy places the focus directly on optimizing the motif description without having to impose an explicit generative model on the entire DNA sequence, which requires an appropriate model

for background non-motif sequence.

3.1.2 Genome-wide Quantitative Evidence

The detection of functional DNA motifs has been greatly facilitated by the availability of high-throughput functional genomics data that provide direct or indirect evidence for gene regulation. For instance, the genome-wide DNA occupancy by a particular TF can now commonly be measured through in vivo approaches such as chromatin immunoprecipitation microarray experiments (ChIP-chip) [160]. Depending on the setup of the array, these assays map TF binding locations at a resolution on the order of 1kb or better [155]. Recent advances such as genome-wide ChIP-seq (ChIP followed by deep sequencing instead of hybridization to an array) provide a less biased approach to identify candidate regulatory regions, and have been shown to identify hundreds or thousands of enriched regions for individual factors [84]. External sources of information like p-values from ChIP-chip experiments, or the strength of enrichment in ChIP-seq experiments, have strong potential to increase specificity and sensitivity when detecting functional motifs. However, some of the most popular existing approaches scale badly and are computationally infeasible when applied to sets with thousands of candidate regulatory sequences. For instance, the sampling step in PSSM-based approaches is typically performed on the positions of the regulatory sequences, and samples are then used to update the motif model. Due to these limitations, existing approaches have thus often made use of genome-wide quantitative data only to reduce the search space over functional motifs, e.g. by running a standard motif finder on a subset of high-scoring or otherwise pre-filtered regions [121].

3.1.3 Reformulation of the Motif Finding Problem

Instead of modifying traditional approaches, the availability of such data suggests the following re-phrasing of the motif finding problem: *identify enriched sequence motifs, given quantitative experimental evidence for a genome-wide set of regulatory regions*. This formulation allows one to explicitly utilize the total quantitative information from the experiment, rather than to only use

it to define a set of promising target sequences, and then proceed with motif finding as usual. The motif finder REDUCE [24] was an early example of this framework, and applied a linear regression strategy to fit the log expression ratios from microarray experiments to the sum of contributions from a set of putative regulators. This promising approach was later followed up with MatrixReduce [54], which is based on a non-linear statistical mechanics model of TF-DNA interactions fitted to ChIP-chip data. A common feature of approaches in this category is that all the experimental data are used in the model, avoiding the use of an explicit significance threshold. In addition, the utilization of all probes from the high-throughput experiment generally does not require an explicit model for background sequence.

3.2 The cERMIT Framework

To address the above reformulation of the motif finding problem, I propose a new motif discovery approach—(conserved) Evidence-Ranked Motif Identification (cERMIT), which is explicitly designed to be able to analyze current state-of-the-art genomic regulatory datasets such as those from ChIP-chip or ChIP-seq experiments.

In a nutshell, cERMIT takes putative regulatory regions S and scores representing evidence of direct or indirect regulation as input E and searches for an optimal motif of flexible length, represented as a degenerate consensus sequence over the IUPAC alphabet. A post-processing step allows to generate PSSMs from high scoring candidates.

The objective function used to score motif candidates is inspired by [131, 42, 140] and encapsulates the aggregate evidence of regulation for a set of sequence regions. It is used to search for the best partition of S into a *positive* and a *negative* set, where the *positive* set consists of regions that have at least one occurrence of a candidate motif, while the *negative* set contains the remaining sequences regions in S . The search starts from the comprehensive set of all possible non-degenerate 5-mers, each of which defines an initial partition of S . Each of the 5-mers is then “evolved” in a greedy search by varying motif length or degeneracy (see Figure 3.1). cERMIT can take different data as evidence for regulatory interactions, and can optionally utilize orthogonal

Evidence Ranked Motif Identification

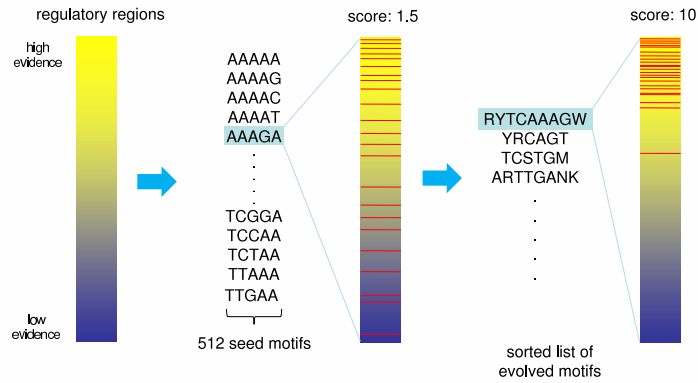


Figure 3.1: cERMIT motif discovery algorithm

sequences from related species to restrict the search to conserved motifs. The following sections outline the process in more detail.

3.3 Evidence of Regulation

The statistical scoring of evidence of regulation in a sequence region will depend on the type of assay used to infer the binding specificities of factors. However, all the statistical scores can be placed in the framework of log-odds ratios

$$e_j = \log \frac{f(s_j | M_1)}{f(s_j | M_0)},$$

where e_j is the evidence for regulatory region s_j and f denotes the likelihood of the experimental evidence for s_j , given that it is in the positive set (M_1) or negative set (M_0).

The types of experimental evidence provided I describe in more detail next are p-values from ChIP-chip experiments, counts of aligned short sequence reads from ChIP-seq experiments, and expression changes from microRNA overexpression assays.

3.3.1 P-values as Evidence of Regulation: Chip-chip data

In the case when p-values are provided as binding evidence for each sequence region in S , the logarithm of the approximate Bayes factors [173] is used as evidence for regulation for each region.

Bayes factors represent the relative evidence between two competing models. In this setting, given a p-value p_j for region s_j the Bayes factor is a ratio of the marginal likelihoods of membership in the positive set (model M_1) vs. the negative set (model M_0)

$$B_{10}^\pi(p_j) = \frac{\int_{\Theta} f(p_j | \theta, M_1) \pi(\theta, M_1) d\theta}{\int_{\Theta} f(p_j | \theta, M_0) \pi(\theta, M_0) d\theta}. \quad (3.1)$$

A direct estimate of the Bayes factor would require a probabilistic model for the p-values of sequences in the positive set (the p-values for the negative set are uniformly distributed) and specification of a prior distribution π . Specifying such a model and integrating out the model parameters θ to arrive at the marginal likelihood estimate could be problematic. Hence, I have decided on a computationally efficient solution, which provides an upper bound to the exact Bayes Factor, and was originally suggested in [173]:

$$B_{10}(p_j) = \begin{cases} \sup_{\pi} B_{10}^\pi(p_j) = -\frac{1}{ep_j \log p_j} & \text{if } p_j \in (0, \frac{1}{e}] \\ 1 & \text{otherwise,} \end{cases}$$

where the supremum is over prior distributions. As observed in [173], this upper bound holds for a broad range of parametric and non-parametric model alternatives. One such possibility is $f(p|\theta_i, M_i) \sim \text{Beta}(\theta_i, 1)$, $\theta_i \in (0, 1)$, $\theta_0 = 1$. Note that M_1 consists of a family of decreasing densities, which includes as a limiting case the uniform distribution of the p-values under M_0 . In the ensuing CHIP-chip analyses $e_j = \log[(B_{10}(p_j))]$ is used as the evidence of regulation for region s_j .

3.3.2 ChIP-seq Reads as Evidence of Regulation

The counts of aligned short sequence reads in ChIP-seq experiments can be used to provide evidence of regulation. In [20] a kernel density estimator is used to score binding; this smooths and normalizes the counts of aligned reads. For each sequence region s_j the maximum of the kernel density estimate over all locations in the sequence is considered positive evidence (M_1),

$$f(s_j | M_1) = \max_{t \in s_j} k(t),$$

where t indexes positions in s_j and $k(t)$ is the kernel density estimate at position t . The background binding score b is fixed to be the 75-th percentile of the strictly positive kernel density scores. The evidence of regulation for region s_j is

$$e_j = \log \frac{\max [f(s_j | M_1), b]}{b}$$

3.4 Integration of Evidence: Definition of the Objective Function

I propose two different approaches to integrating the individual putative regulatory region binding evidence into a combined score for a set of co-bound regions. First, I outline an approach based on the assumption of independent contributions of each regulatory regions to the overall set binding score, in which case I can appeal to the Central Limit Theorem and employ a (normalized) average to quantify the cumulative binding evidence. In the second proposed approach I relax the independence assumption, controlling for observable confounders, in a linear regression framework. In this scenario the binding evidence scores for individual genes are independent, conditionally on the values of the covariates for the specific sequence region (e.g. length of region, di-nucleotide frequencies etc.). This provides a more general approach to incorporating additional information into the motif discovery framework, including interaction terms to allow more targeted questions and potentially transitioning to a fully probabilistic solution.

3.4.1 Random Set Scoring

In this section I introduce the sequence region scoring approach that was used in [61] and is best suited to the direct binding evidence provided by chip-chip and chip-seq assays as well as microRNA overexpression.

Given evidence $E = \{e_1, \dots, e_d\}$ for a set of sequence regions S , a motif m_j partitions E into a positive set E^j where the elements of E^j are the evidence for those sequence regions that contain motif m_j , $E^j = \{e_i : m_j \in s_i \text{ for } i = 1, \dots, d\}$, the negative set is the complement.

I assume that there exists a “true” motif m_* that induces a partition of the evidence with a positive set E^* . This partition can be recovered by searching over the discrete space of motifs using an appropriate objective function. This objective function should capture high *aggregate* evidence for regulation in the positive set and low evidence in the negative set. The number of candidate partitions over the set of sequences is very large (at most 2^T) so this objective function must be efficiently computed.

A test statistic introduced in [42, 140] has the above properties and is used as the objective function for cERMIT. Given evidence E^j induced by a motif m_j , define:

$$\begin{aligned}
 J(E^j) &= \frac{\frac{1}{|E^j|}(\sum_{e_i \in E^j} e_i) - \mu}{\sigma_j}, \\
 \mu &= \frac{1}{|E|} \sum_{e \in E} e, \\
 \sigma_j^2 &= \frac{1}{|E^j|} \left(\frac{|E| - |E^j|}{|E| - 1} \right) \left\{ \frac{\sum_{e_i \in E} e_i^2}{|E|} - \left(\frac{\sum_{e_i \in E} e_i}{|E|} \right)^2 \right\},
 \end{aligned} \tag{3.2}$$

where $|E| = d$ and $|E^j|$ are the cardinalities of the total number of regulatory regions and those contained in the positive set, respectively. The resulting optimization problem is

$$\hat{m} = \arg \max_{m_j \in M} J(E^j), \tag{3.3}$$

where \hat{m} is the best guess at the optimal binding motif m_* . To constrain the search, I eliminate implausible candidate k-mers m_i which result in very small or very large target sets. This can be

done by constraining the set size to be in the range $[a, b \times |E|]$, where $a = 20$, $b = .15$ for the CHIP-chip data and $a = 100$, $b = .30$ for the the CHIP-seq data. The values of these parameters may vary depending on the expected number targets for the TF under study. This results in the following objective function

$$J^*(E^j) = \begin{cases} J(E^j), & \text{if } |E^j| \in [a, b \times |E|] \\ -\infty & \text{otherwise.} \end{cases}$$

3.4.2 Linear Regression Scoring

In this section I describe a solution to the problem of evaluating a set of sequence regions, with individually assigned binding evidence, in terms of a summary that reflects their combined binding evidence, assuming that there is some, potentially relevant, confounder information e.g. di-nucleotide frequencies, region length.

Let $s_i, i \in \{1, \dots, n\}$ be a set of sequence regions (e.g. clusters or read-groups as reported in [37]) and y_i be the corresponding binding evidence for each region. A candidate set of putative motifs is defined to be the set $m_j, j \in \{1, \dots, T\}$. Functional motifs are typically of length 6-10 with a limited number of degenerate positions, assuming that the motif has a conserved core of at least 3-5 nt. A match of motif m_j in sequence region s_i is given by the binary indicator variable x_{ij} . Denote the number of motif occurrences in $\{s_i\}_{i=1}^n$ by $n_j = \sum_{i=1}^n x_{ij}$, and let

$$\hat{y} = \frac{1}{n} \sum_{i=1}^n y_i, \quad \hat{\sigma} = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 \quad (3.4)$$

$$y_i^* = y_i - \bar{y}, \quad A_j = \frac{n - n_j}{n - 1} \quad (3.5)$$

$$e_j = \frac{1}{n} \sum_{i:x_{ij}=1} y_i^*, \quad \hat{\sigma}_j^2 = \frac{\hat{\sigma}}{n_j} \quad (3.6)$$

then

$$S_j^{\text{cERMIT}} = A_j \times \frac{e_j}{\hat{\sigma}_j} \quad (3.7)$$

$$m_{\text{cERMIT}}^* = \underset{j \in \{1, \dots, T\}}{\operatorname{argmax}} S_j^{\text{cERMIT}}, \quad (3.8)$$

where m_{cERMIT}^* denotes the top predicted motif using the strategy described in the Section 7.1.1, based on motif enrichment score S_j^{cERMIT} assigned to motif m_j . In the spirit of previously published work [24, 54, 53], I rephrase the above scoring function of a binding motif in a regression model for the sequence region's binding evidence. The baseline version of the model closely resembles the cERMIT score from Section 7.1.1, and provides a framework to easily add additional confounding variables to the primary explanatory variable of motif occurrence. Let the regression coefficient for motif m_j be denoted as β_j , then a simple linear model for the binding evidence is as follows

$$y_i^* = x_{ij}\beta_j + \epsilon_i \quad (3.9)$$

$$\epsilon_i \sim \text{N}(0, \sigma^2). \quad (3.10)$$

Using the classical Ordinary Least Squares (OLS) estimator for the regression coefficient $\beta_j^{\text{OLS}} = \frac{1}{n_j} \sum_{i: x_{ij}=1} y_i^*$, define the motif enrichment score and the resulting top prediction to be

$$S_j^{\text{reg}} = \frac{\beta_j^{\text{OLS}}}{\hat{\sigma}_j} \quad (3.11)$$

$$S_j^{\text{reg}} = \frac{1}{A_j} \times S_j^{\text{cERMIT}} \quad (3.12)$$

$$m_{\text{reg}}^* = \underset{j \in \{1, \dots, T\}}{\operatorname{argmax}} S_j^{\text{reg}}. \quad (3.13)$$

Note that the typical scenario in which the size of the motif target set is small relative to the number of all sequence regions ($n_j < n$) results in $A_j \approx 1$, which implies that $S_j^{\text{reg}} \approx S_j^{\text{cERMIT}}$. Empirical observations based on the re-analyses of ChIP-chip and ChIP-seq datasets from [61] using this slightly modified scoring strategy produces near identical motif predictions, with slightly different enrichment scores.

A natural extension of this framework is to account for features of the sequence regions that expected to be unrelated to motif binding. Two such potential confounders are di-nucleotide counts and region’s sequence length. More generally, let us consider $p-1$ additional confounders to control for when searching for motifs. To generalize the notation, the vector of regression covariates is augmented to a matrix $Z_j = (x_j, c_{1j}, \dots, c_{(p-1)j})$, where x_j denotes the column vector of binary indicators of motif matches and $c_k \in \mathbb{R}^n$, $k \in \{1, \dots, p\}$ denote the additional confounders. With corresponding regression coefficients $\beta_{1j}, \dots, \beta_{pj}$, $\beta_{kj} \in \mathbb{R}$, $k \in \{1, \dots, p\}$ the model becomes (using matrix notation):

$$Y^* = Z_j \beta_j + \epsilon, \quad \epsilon \sim \mathbf{N}(0, \sigma^2 \mathbf{I}_{n \times n}) \quad (3.14)$$

In order to be able to score a large number of motif candidates, it necessary to have a computationally efficient estimator for the regression coefficients, that has good statistical properties. Hence I adopt the simple OLS estimators $\hat{\beta}_j = (Z_j^T Z_j)^{-1} Z_j^T Y^*$, which for small to moderate values of p provide a fast solution. Note there are typically a large number of putative regulatory regions and therefore a large number of sample points to use in the estimation process. The resulting estimator for motif scores is a straightforward generalization of the univariate regression case:

$$S_j^{\text{REG}} = \frac{\hat{\beta}_j^{\text{OLS}}}{\hat{\sigma}_j} \quad (3.15)$$

$$\hat{\sigma}_j = \sqrt{(\hat{\Sigma}_{\hat{\beta}_j})_{11}}, \quad (3.16)$$

where $(\hat{\Sigma}_{\hat{\beta}_j})_{11}$ is the first diagonal element in the covariance matrix for the parameter estimate $\hat{\beta}_j^{\text{OLS}}$. The proposed motif analysis pipeline can be easily adapted to use conservation information to improve target prediction specificity by requiring that matches fall within conserved genomic regions similarly to the strategy described in Section 3.7.

3.5 Search Strategy

An exhaustive search over the space of all potential motifs to optimize the objective function defined in Section 7.1.1 or Section 3.4.2 is not computationally feasible. Instead, I adopt a direct greedy

search strategy that relies on local motif updates to construct candidate motifs.

All possible 5 or 6-mers are used as seed points to start the search (in the case of TFs, pooling reverse complements, hence $T = 512$). Given a motif m , a candidate set of motifs is constructed by locally varying the length and the degeneracy of m . The *extension* move takes a k-mer as input and independently appends or prepends A, G, C, or T generating 8 new $(k + 1)$ -mers. When *reducing* the length the motif is truncated by one letter on either side to produce two new candidate motifs. The *degeneracy* move operates on a single position in the k-mer at a time to produce a new motif candidate. The following update rules are applied to each position j in motif m :

1. $m[j] = A$ then three new k-mers are constructed with $m[j]$ set to M, R, W respectively;
2. if $m[j] = C$ then three k-mers are constructed with $m[j]$ set to M, S, Y respectively;
3. if $m[j] = G$ then three k-mers are constructed with $m[j]$ set to K, R, S respectively;
4. if element $m[j] = T$ then three k-mers are constructed with $m[j]$ set to K, W, Y respectively;
5. if $m[j] = R, Y, S, M, K, W$ then $m[j]$ is set to N ;
6. if $m[j] = N$ then the k-mer is not updated.

A move, reducing the motif degeneracy is also incorporated and it follows the same rules described above but reversed from the more degenerate to the less degenerate symbol e.g. M collapses to A and C and produce two new candidates.

For a k-mer with no degeneracies, these moves will generate $3k$ candidate k-mers. For a k-mer with double degeneracy in all positions, the move will generate k candidate k-mers with the same double degeneracy in all but one position set to N .

For each seed motif m the search algorithm applies the update rules and examines if they result in a higher motif score, in which case the highest scoring candidate is used in the following iteration. This procedure is repeated until the update rules cannot improve the motif score, resulting in a candidate for the best scoring motif evolved from the particular seed.

The result of the search is a set of motifs $\hat{M} = \{\hat{m}_1, \dots, \hat{m}_T\}$ and their corresponding scores $\hat{J} = \{\hat{J}_1, \dots, \hat{J}_T\}$. For each motif a PSSM is constructed based on the empirical counts of occurrences of each of the exact instantiations in the subset of the top 50% of the evidence-ordered list of putative regulatory regions.

3.6 Post-processing

Many of the top scoring motifs will be very similar, varying by a few letters. Hence, to reduce redundancy and improve the interpretability of the reported results, a post-processing step is added: similar motifs are clustered around “cluster centers” defined to be distinct individual k-mers with maximum objective function score J^* .

In the clustering procedure I use the Harbison metric (at 0.75 cutoff) [78] to compute similarity between two motifs. For motifs a, b of equal length w the distance $D(a, b)$ is

$$D(a, b) = \frac{1}{\sqrt{2}w} \sum_{i=1}^w \sqrt{\sum_{L \in \{A, G, C, T\}} (a_{i,L} - b_{i,L})^2}, \quad (3.17)$$

where $a_{i,L}$ and $b_{i,L}$ are the relative frequencies of base L at position i for the PSSM motif descriptions of a and b , respectively. For motifs of differing lengths define the following metric

$$\text{sim}(a, b) = \max_{a', b'} [1 - D(a', b')],$$

where a', b' are correspond to all possible overlaps of between motifs a, b induced by shifts such that the minimum overlap length is six. This metric is also used in [136].

Two motifs m_1 and m_2 are considered similar if

1. The PSSMs of m_1 and m_2 have Harbison similarity score ≥ 0.75 ;
2. The motifs m_1 and m_2 co-occur in the same sequences significantly more frequently than expected by chance, as measured by the following p-value threshold

$$\text{Hyp}(|S^{\text{co-occur}}|; d, |S^1|, |S^2|) < 10^{-20},$$

where S^1 and S^2 are the positive sets for motifs m_1 and m_2 . The set of co-occurring regions $S^{\text{co-occur}}$ are those regions where the motifs m_1 and m_2 are both present and separated by at most τ nucleotides, where τ is set to be $\tau = \frac{\min(|m_1|, |m_2|)}{2}$.

Then, given the set of redundant output motifs $\hat{M} = \{\hat{m}_1, \dots, \hat{m}_T\}$, the following procedure outputs a set of motif clusters $\{R_i\}$ and smaller indices corresponding to higher motif scores.

1. Initialize the cluster count: $n = 1$;
2. Find the top motif in the set C

$$m^* = \arg \max_{i=1, \dots, k} J^*(m_i);$$

3. Add m^* and all other motifs in C similar to m^* to R_n ;
4. Remove the set R_n from C ;
5. Update cluster count $n = n + 1$;
6. Repeat steps (3) and (4) and (5) until C is empty.

Given a motif cluster R_i a cluster PSSM is computed by averaging the corresponding PSSM columns of each cluster member weighted by their motif score.

3.7 Integrating Evolutionary Conservation

Sequence conservation between related species can be used to help guide the motif search: Defining the positive set of putative regulatory regions based on the motif presence across a set of species can help to increase the signal-to-noise ratio by eliminating false positive matches which occur in individual genomes. I followed the example of previous approaches which utilized pattern co-occurrence without relying on alignments [69, 49]. In cERMIT, conservation information is incorporated by refining the positive set S^j of regulatory regions in which m_j is present. If orthologous regions are available for a given region in one or more of the other species, I remove from S^j those

regions where m_j is not found in all orthologous regions. I.e., rather than restricting the analysis to the subset of genes with clearly defined orthologs, I simply require that patterns must co-occur in available orthologs; in cases where no ortholog is defined, set membership is based on the occurrences in the species with experimental evidence of regulation. This strategy allows for a more complete utilization of the full data set, and I otherwise follow the same motif search procedure, applied to the refined set S^j .

3.8 Significance Evaluation of Motif Enrichment

For the top representative motif predictions $\{m_k\}$ a p-value is provided, using a permutation procedure. For a motif m_ℓ with score J_ℓ the following procedure is used to compute its p-value:

1. Generate $1, \dots, \Pi$ permutations of the evidence E , $\{E_{(\pi)}\}_{\pi=1}^{\Pi}$.
2. For each $\pi = 1, \dots, \Pi$ compute

$$J_\pi = \max_{m_\ell \in M} J(E_{(\pi)}^\ell).$$

3. From $\{J_\pi\}_{\pi=1}^{\Pi}$ fit a Gamma distribution,

$$f(J) = \text{Gamma}(J; \hat{\alpha}, \hat{\beta}).$$

4. The p-value is $f(J(E^\ell))$.

3.9 Conclusions

In the classic motif finding framework the search aims to identify overrepresented short patterns in a pre-defined subset $S' \subset S$ (with S being the genome wide set of regulatory regions), which is assumed to be enriched in functional motif occurrences. In this work I have proposed an effective and highly scalable novel approach to the identification of functional non-coding sequence motifs, which is applicable to the motif finding problem in a rephrased definition, where each regulatory sequence in the whole set S is annotated with quantitative experimental evidence. This

method circumvents the problem of having to define a sequence set enriched in cis-regulatory targets, and makes use of the additional information provided by quantitative evidence from current high-throughput experiments.

Other recent approaches have worked within this rephrased definition; for instance, rank-based algorithms have been described to generate canonical motif descriptions for protein binding arrays [135, 13]. The FIRE algorithm [48] could also be mentioned in this context, as it is based on the idea that the presence of an oligomer in regulatory regions is statistically dependent on a relevant phenotype of interest (e.g. expression level or expression cluster membership). Compared to some of these other rank-order based approaches, it is important to note that cERMIT incorporates the entire genome-wide evidence of regulation into the motif search. This is achieved through a carefully chosen objective function which provides a simple, yet effective quantitative measure for co-regulation of a set of sequences, without the need to define any cutoffs. The inspiration for this overall framework, and the particular function I used [42, 140], draws from gene set enrichment analysis (GSEA) [131, 184], in which the aggregate evidence of a predefined gene set such as a functional pathway is used to increase the power to detect differential gene expression. The proposed approach to motif discovery can be seen as an inverse to GSEA: instead of scoring a pre-defined gene set, cERMIT searches for optimal gene sets defined by shared functional cis-regulatory motifs. The GSEA framework has attracted considerable attention, and other objective functions have been proposed which can be explored as potential alternatives for cERMIT.

Of key computational importance is the fact that the objective function is efficiently computable, which allows cERMIT to determine a putative motif enrichment extremely quickly, making the proposed direct motif search strategy feasible. To score a partition corresponding to a given consensus motif, cERMIT operates on a set of sequences represented in an efficiently searchable suffix array data structure [124, 125]. Hence the overall runtime of the algorithm on a standard single processor workstation is on the order of a minute per typical run for the comprehensive set of upstream sequences for a yeast TF of interest, and 2-5 minutes for the 35,000 regions (1KB) from human ChIP-seq experiments. Instead of directly searching for over-represented short patterns in a pre-

defined set of co-regulated sequences, I update the candidate for optimal partition by modifying the corresponding consensus motif. Thus, I perform a search on the discrete space of IUPAC motifs, which is independent of the number of regulatory regions and scales logarithmically with the total length of the sequences in S .

The adaptation to incorporate additional covariate information provides extra flexibility to the scoring of the putative co-bound sets, retaining the computational efficiency due to the availability of a closed form solution provided by the Ordinary Least Squares.

Chapter 4

Applications of cERMIT to Studying the Transcriptional Regulation

In this section I describe the application of the motif discovery framework introduced in Section 3.1 to studying the processes of transcriptional regulation in the cell. In particular, I will focus on the studying the binding preferences of proteins (Transcription Factor) that recognize specific DNA sequence sites in the genome. Such description would allow for further exploration of the specific genes that are targeted by the protein of interest in the given biological context. The main data sources for the analysis will be high-throughput assays that provide direct evidence for TF binding by means of ChIP followed by hybridization or sequencing.

For a controlled evaluation of a new motif discovery approach, it is desirable to have reliable sets of positive examples for which it is straightforward to compare the success of different strategies. ChIP-chip or ChIP-seq data on factors with known literature binding site consensus sequences provides the most straightforward setting, as it implies direct evidence of binding, presumably mediated by a common sequence motif. To provide a common ground with other recent algorithms, I focus on the genome-wide yeast ChIP-chip data set from the Young lab, which is still the most comprehensive ChIP data set [78], but also demonstrate the application of cERMIT on a compendium of recent mammalian ChIP-seq datasets. Finally, I consider expression and mass spectrometry data collected from microRNA overexpression data sets, to show that the motif finder performs well in cases where the influence of a factor is not determined by a direct binding assay but rather by downstream changes in mRNA or protein expression levels.

cERMIT shows competitive performance on gold-standard high-throughput ChIP-chip datasets. In addition, I present an application on ChIP-seq data sets as final step in an integrated pipeline for motif inference, which includes the alignment of high-throughput sequencing reads [113] and peak calling of enriched locations [19], and additionally integrates genome-wide information on open

chromatin as determined by DNaseI hypersensitive assays.

4.1 Analysis of Transcriptional Regulation Using High-throughput DNA Binding Evidence

This section provides a detailed description of the analysis of ChIP-seq data from two different mammalian organisms (*H. sapiens* and *M. musculus*) as well as a more comprehensive compendium of ChIP-chip data sets assaying the binding preferences of *S. Cerevisiae* transcription factors. In all cases there exists a reasonably good prior knowledge about the target binding site, which makes it possible to implement an unbiased evaluation of the performance of the proposed motif discovery strategy (see Section 3.1) in the different contexts. For the remainder reference to the “cERMIT” approach, correspond to the enrichment scoring strategy that assumes independent contributions by the individual sequence binding evidence as described in section 7.1.1 and originally introduced in [61].

4.1.1 Yeast ChIP-chip Compendium

The 352 ChIP-chip *S. cerevisiae* datasets and the corresponding orthologous probe sequences were extracted as described in [136].

4.1.2 Human and Mouse ChIP-seq Experiments

The six human TFs ChIP-seq datasets were used as provided in the following papers: STAT1 [161], the insulator binding protein CTCF [7], SRF, GABP [191], FoxA1 [204], NRSF [91]. The twelve chip-seq datasets analyzed by cERMIT, cMyc, nMyc, E2f1, CTCF, Esrrb, Klf4, Nanog, Oct4, Sox2, STAT3, Tcfcp2l1, Zfx, were used as provided by [25]. The embryonic stem cell panel additionally included datasets for the factors Suz12 and Smad1 which are not considered for further analysis. The former factor does not interact directly with DNA; the dataset for the latter contained reads of length 36bp instead of the reported 26bp, and successful alignment to the mouse genome was significantly impacted.

| Motif finder | # successes top 1 | Motif finder | # successes top 4 |
|--------------|-------------------|--------------|-------------------|
| AlignACE | 16 | Trawler | 52 |
| MEME | 35 | YMF | 57 |
| MEME-c | 49 | AlignACE | 64 |
| Kellis | 50 | MEME | 76 |
| Converge | 56 | Weeder | 78 |
| PRIORITY-C | 69 | Amadeus | 90 |
| MD-scan | 54 | | |
| PRIORITY-DC | 78 | ERMIT | 92 |
| ERMIT | 77 | cERMIT | 114 |
| cERMIT | 88 | | |

Figure 4.1: Comparison of cERMIT with other motif finders (adapted after Linhart *et al.* [119]).

4.2 Elucidating Regulatory Sequence from ChIP-chip Experiments

The transcription factor data set from [78] consists of genome wide location data for 203 yeast TFs assayed in a total of 352 different experiments. 82 TFs were assayed in more than one condition. The input consists of an upstream sequence for each gene, as well as an associated p-value of binding of a specific TF to each upstream sequence. Previous studies [78, 123] have combined known literature consensus with the results of different motif finders to arrive at a comprehensive list of binding site representations. Knowing the literature consensus provides us with a common basis to compare the performance of motif finders, but different publications use different criteria to define success. In this work I follow the PSSM similarity metric introduced by [78]. For the detailed specification, please, refer to Equation 3.17 in the Methods section.

Following [136], I set a requirement of at least 6 overlapping bases (rather than 7 as initially introduced [78]), as I do consider motifs of more variable length. Varying the similarity threshold cutoff of course influences the absolute number of successful predictions, but for any fixed cutoff, it provides a relatively fair assessment of different algorithms. The similarity cutoff is selected according to previous evaluations of the same data set.

The yeast data set has been used as a starting point for many recent motif finder evaluations,

of which I will use two to assess the proposed motif discovery approach. While yeast is often regarded as “easy” with respect to regulatory sequence analysis, these assessments demonstrated that there was still considerable room for possible improvement. The first evaluation focused on a subset of 156 out of the 352 total experiments for which there was strong evidence of more than 10 bound probes (p-value < 0.001) [136]. This gold standard set for motif finders covers 80 unique TFs for which there is a known literature consensus binding site [123]. With the idea that a ChIP experiment should strongly enrich for sequences sharing the binding site of the TF assayed, a motif was only counted as successfully identified if the top prediction matched the known consensus at a cutoff of 0.75. Applying cERMIT on this data set leads to the results summarized in Figure 4.1, where cERMIT predictions with and without conservation are reported in the context of a comprehensive recent comparison adapted from [69]. The species related to *S. cerevisiae* used here were the remaining four yeast species in the *sensu stricto* clade, the set commonly used in other approaches relying on cross-species conservation.

Applying cERMIT results in a dramatic increase in terms of number of recovered motifs as compared to AlignACE and MEME, which make use of only *S. cerevisiae* genomic sequence information and do not exploit quantitative information on binding, or conservation across species. MEME-c, the Kellis approach [98], and Converge [123] are heavily based on conservation information across the four related yeast *sensu stricto* species, yet resulting in a substantially lower number of successfully predicted motifs even when conservation is not used (ERMIT). The proposed motif discovery approach also shown significant improvements over MD-scan which uses the ChIP-chip information. The recently introduced PRIORITY algorithm is a state-of-the-art Gibbs sampling approach which makes use of both conservation (PRIORITY-C) and ChIP data (PRIORITY-DC), by additionally utilizing discriminative counts obtained from bound versus unbound probes [69]. Even PRIORITY-DC produces a smaller number of successful predictions than cERMIT, and overall, the performance improvement compared to other recent approaches is significant.

Another recent assessment of motif finders also included results on this yeast ChIP dataset. The assessment was part of the description of Amadeus [119], a motif finding platform which

introduces multiple strategies for detecting enriched motifs, based on ranking all genes based on evidence of binding. The gold standard defined in this paper was highly similar to the set in Figure 4.1. The intersection set between the two data sets [119, 136] contains 150 experiments (77TFs) out of possible 156. In contrast to the more stringent evaluation in [136], this study defined a success if any of four motifs (the top two predictions obtained by running the motif finder on fixed word lengths of 8 and 10 nucleotides) matched the known consensus. As cERMIT identifies motif of flexible length, I compared the top 4 cERMIT predictions to the results reported in this study in Figure 4.1. I used the results provided on the Amadeus website, which is based on a Harbison similarity threshold of 0.76 almost identical to the one used in Figure 4.1.

As can be seen, results improve consistently for the motif finders also evaluated in Figure 4.1. The superior performance of cERMIT can be explained by two major reasons: cERMIT uses the actual quantitative scores for each sequence as evidence of binding, rather than only the ranks of the genes based on decreasing evidence of binding. This increases cERMIT's power to detect functional binding. Clearly, another advantage of cERMIT is its ability to take into consideration evolutionary information from closely related genomes if available. It is however not a result of the particular similarity cutoff: The authors of Amadeus also report performance results based on a cutoff of 0.82, at which they successfully recovered motifs for 78 conditions covering 53 TFs; cERMIT (100/58) clearly exceeds these numbers.

Finally, I assessed cERMIT in comparison to DRIM [47], a recent motif finder which is likely to be the closest to the proposed motif discovery approach. While DRIM was evaluated on the yeast ChIP-chip data, the authors considered a specific subset, not for all of which a known literature consensus is available. The subset of TFs with known consensus contains 44 conditions out of the set of 156 from Figure 4.1, corresponding to 36 unique TFs. DRIM generally predicts more than one motif, with an average of 2.5 motif predictions per ChIP-chip dataset. For the purpose of a meaningful comparison, I verified whether the cERMIT results from Figure 4.1 which relate to this TFs contained a successful prediction among the top two and three motifs. DRIM successfully predicts motifs for 26 conditions and 19 TFs (a 53% success rate on the level of TFs). cERMIT

identifies the correct motif among the top two predictions in 30 out of the 44 conditions, and 30 of the 36 TFs (83%); for the top three motifs, these numbers increase to 32 conditions/31 TFs.

4.2.1 Performance Aspects

In this section I describe a detailed analysis of the TFBS predictions on the ChIP-chip compendium data set from [78, 123], particularly focusing on the assessment of False Positive and False Negative predictions. This provides additional insight into the inherent difficulty of the problem of identifying de-novo motifs in the in-vivo context of yeast ChIP-chip experiments. Novel predictions are also discussed and could provide valuable source for new hypotheses to be followed with further wetlab experiments.

4.2.2 Assessment of False Positive and False Negatives

In order to obtain significance estimates for the scores of the top cERMIT prediction, a permutation procedure was applied randomly assigning the binding evidence to the putative sequence regions (see Section 3.8). This helps to investigate cases in which the motif search appears to fail, and to pinpoint experiments in which the scores for even the best predicted motifs do not rise above background scores on randomized data. To check the consistency of these estimates, I compared them with results from the motif finder PRIORITY which also included a similar significance analysis [69]. In the following, a stringent p-value cutoff of 10^{-4} was applied to the cERMIT results. Comparing how estimated p-values agree with successful predictions, i.e. the cases in which the top motif corresponded to the known literature consensus, there were 67 True Positives (TPs), 21 False Negatives (FNs), 50 True Negatives (TNs), and 18 False Positives (FPs), which corresponded to a FP rate of 26% and a TP rate of 76%. On the FN side, where cERMIT fails to assign high enough significance to predictions that match the literature, there are 21 cases, for 7 of which PRIORITY also reported a non-significant match. This means that even when the signal from the experiment does not exceed random expectation, motif recovery may still be successful.

There is a significant number of cERMIT prediction in which the literature motif is among the

top three or four reported motifs, but not at the top, and these cases somewhat misleadingly count as False Positives here. I further investigated the top predicted motifs in these cases, where a significant p-value did not match the literature consensus of the factor assayed in the experiment. At least for eight cases (involving the TFs DAL81, INO4, MET32, MSN2, MSN4, and TEC1), there is convincing circumstantial evidence explaining the predictions. These cases are likely due to experimental conditions in which several factors regulate a largely overlapping set of target genes, and this effectively demonstrates cERMIT's ability to predict more than one functional motif. Details are given in the Supplementary Information.

Looking at the overall results from a different angle, there were only 34 conditions in which cERMIT (with or without using evolutionary information from other species) failed to recover the literature consensus motif among the top 3 predictions. When using conservation, 25 of these had comparatively large p-values ($> 10^{-4}$), and may be cases in which the experimental noise may have been too high to successfully recover a functional site, or conditions in which the factor assayed does in fact not directly bind DNA. Furthermore, PRIORITY consistently did not assign a significant p-value and/or predict a matching motif for any of these. In the remaining 9 cases with stringent p-values, 3 concerned the experiments INO4_YPD, TEC1_Alpha, and TEC1_YPD discussed above, which likely corresponded to cases where another protein in a complex is enriched, or in which the reported consensus is similar to the prediction but not called at the predefined threshold. cERMIT failed to report a high ranking matching prediction under any condition for only 6 TFs. Overall, this means that cERMIT predictions are able to explain almost all ChIP-chip experiments. Contrary to the reportedly low success rates of various algorithms on the originally formulated motif finding problem [188], this shows that motif discovery on current genomic datasets has now become a highly successful undertaking.

4.2.3 Novel Predictions

Finally, I ran the cERMIT analysis on the complete set of 352 experiments described by Harbison *et al.*; for 51 of the 196 datasets without known TF consensus, cERMIT predictions had a p-value

less than 10^{-4} . The recent PRIORITY publication [69] reported predictions for a total of 82 out of the 196 experiments. Comparing the cERMIT's novel predictions to significant PRIORITY predictions provides computational support for predicted motifs from two highly different motif finding approaches. 18 out of the 82 PRIORITY predictions met cERMIT's stringent p-value cutoff of 10^{-4} , while cERMIT passed this P value cutoff for 25 motifs out of these 82. Significant predictions overlap on 12 conditions, and the actual predicted top PSSM were similar to each other in 7 out of the 12 cases. This shows a trend for the motif finders to agree on the top motifs if both are supported by stringent P values.

4.3 Identification of Motifs from Deep Sequencing ChIP-seq Experiments

ChIP-chip experiments are in the process of being replaced by ChIP-seq experiments, in which chromatin immunoprecipitation is followed by high-throughput sequencing of the bound DNA fragments. This allows for a cheaper and potentially less biased assay of the whole genome, but like genomic ChIP-chip before it, poses new challenges for motif finding, as the number of bound regions can be in the hundreds or even thousands. Not all motif finders are able to deal with input sets of such a large size efficiently, and some are not applicable at all. cERMIT has been specifically developed to make use of evidence for a genome-wide set of regulatory regions. For the compact yeast genome, ChIP experiments followed the common assumption that binding sites are found in close proximity to genes' transcription start sites. The definition of an appropriate set of *putative* regulatory regions is a more difficult task in multicellular eukaryotes with more complex genomes. For instance, randomly selecting intergenic regions in mammalian genomes will include a large fraction of non-regulatory sequences such as repeats. However, high-throughput sequencing technology has already demonstrated its great promise for the study of gene regulation in such organisms, and currently available experimental measurements can be utilized to extract salient features of gene regulation at a whole-genome scale.

As the main focus of this work is developing global approaches to studying the condition-

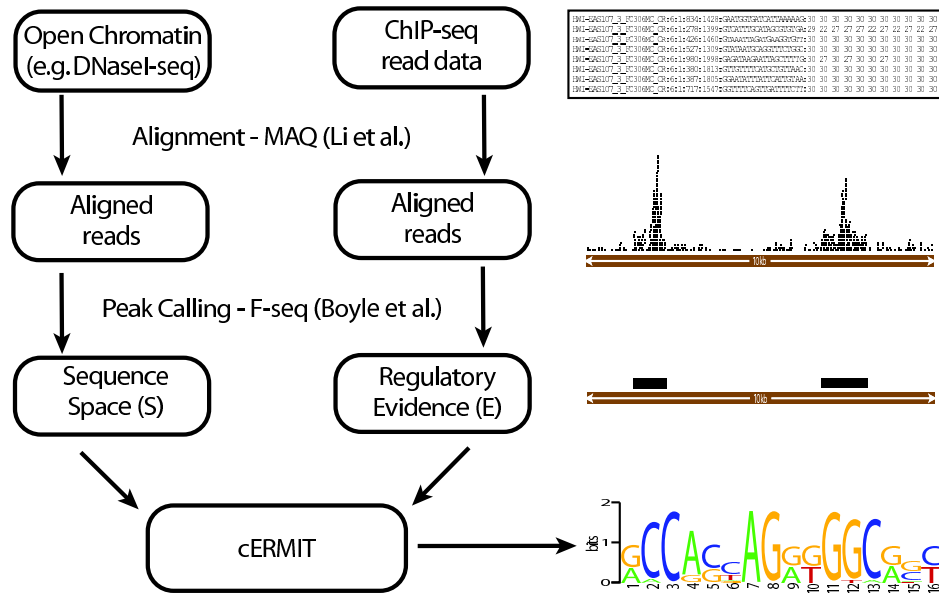


Figure 4.2: Motif discovery pipeline

specific gene regulation, it would be most appropriate to define the search space to be the complete set of enhancer regions in the genome, or at least those active within the specific condition. In a recent paper [193], the authors mapped thousands of *in vivo* target sites of the enhancer-associated protein p300 using ChIP-seq, which provides a large set of enhancer regions conditional on interactions with p300. A perhaps even more comprehensive strategy to defining potential enhancer regions is to use regions known to fall within open chromatin, which tend to be accessible to binding by regulatory factors. This has been assayed e.g. by DNaseI digestion, and DNaseI Hypersensitive Sites (DHS) have been determined by high-throughput sequencing [19].

Starting from such data in its entirety, I can then focus on the more nuanced transcription regulation signals that control condition-specific gene-regulatory programs. Hence, the high-throughput deep sequencing data is utilized in two parallel ways, first from assays defining the space of putative regulatory regions, e.g. as those around DHS peaks, and second from factor-specific binding evidence based on the corresponding ChIP-seq data. A schematic pipeline that intersects different sources of high-throughput regulatory evidence for motif prediction as described above is shown in Figure 4.2. Comprehensive ChIP-seq gold standard data sets like the one in yeast [123] are not yet available, and therefore cERMIT was applied on a number of currently available mammalian

datasets from human and mouse. For all experiments, I started from the deposited raw sequence reads, which were realigned to the genome and then analysed by cERMIT.

Next I outline the peak calling and pre-processing steps implemented prior to the motif analysis.

4.3.1 Peak Calling and Processing of Fseq Peaks

The following two steps are implemented to produce a set of peaks to be used (after some further processing) as input to the motif analysis.

1. Identify discrete ChIP peaks using the kernel density estimation (KDE) procedure implemented in [20].
2. Assign binding score = maximum KDE value across all locations within the peak.
 - (a) Discard regions with binding score scores more than 10, as those are most likely to be pile-ups within repeat regions.
 - (b) Extend/Trim peaks (proportional to the distance from the maximum KDE score location) to fall within the range: 100-1000bp

The peaks produced during the peakcalling step are further pre-processed using the following steps:

1. *Define the space of putative bound sequence regions* Recent high-throughput sequencing technologies coupled with DNaseI Hypersensitive Sites (DHS) assays have clearly demonstrated that regions of open chromatin tend to be highly enriched in functional DNA elements [19]. Hence, the set of putative regulatory is define to be the DHS peaks assayed in the same experimental conditions and call this the "DNaseI" approach. Ideally, DHS data would be combined with the factor-specific binding evidence (e.g. ChIP-seq) derived from the same cell type. When DHS data is unavailable I propose an alternative strategy—"ensemble" approach—which relies on the assumption that in general ChIP-seq peaks tend to fall within open chromatin regions, irrespective of the specific assay. Hence, the combined set of the

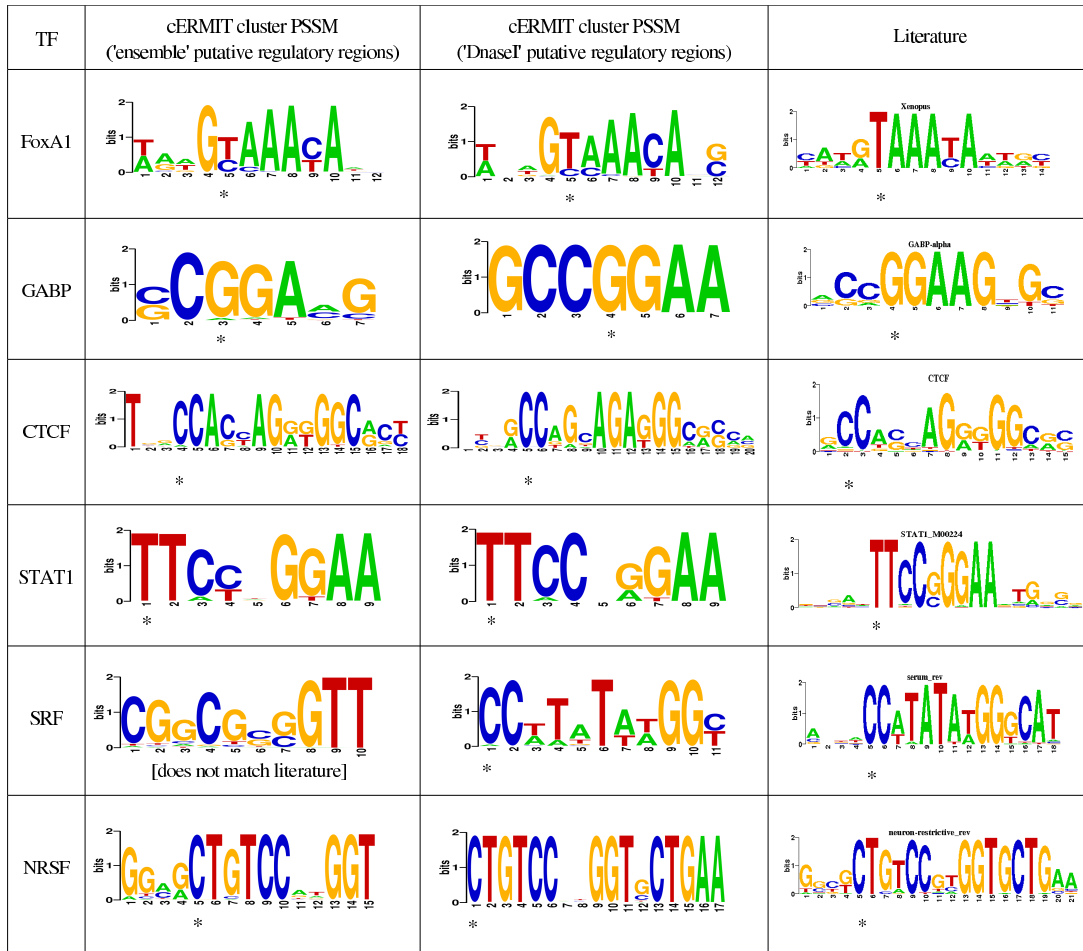


Figure 4.3: cERMIT predictions on human ChIP-seq datasets from [7, 91, 161, 191, 204]

top ChIP-seq peaks from an ensemble of unrelated ChIP-seq datasets would provide a useful proxy to open chromatin.

2. *Assign binding scores based on ChIP-seq data* Each putative regulatory region is assigned the binding score for the corresponding overlapping ChIP-seq peak. If there is no overlapping ChIP-seq peak assign 0. Whenever two putative regulatory regions overlap, merge the two and assign the binding score of the longer of the two original regions.

4.3.2 Analysis Results

I first analyzed six human ChIP-seq datasets on factors STAT1 [161], the insulator binding protein CTCF [7], SRF, GABP [191], FoxA1 [204], and NRSF [91]. Results from the cERMIT analysis are reported in Figure 4.3. I defined the space S of putative regulatory regions based on published DHS data, and contrasted this with an 'ensemble' approach. The latter is suitable for conditions or species for which DHS data is not available, and I took the combined set of high scoring peaks from a panel of ChIP-seq experiments and, after merging overlapping regions, arrived at one final set S used in common for each individual factor. The evidence E was then assigned in factor-specific fashion, based on overlap of the commonly defined regions in S with the factor's ChIP-seq peak regions. The 'ensemble' strategy is effectively an approximation to open chromatin regions, potentially under different experimental conditions depending on the particular ChIP-seq panel used, and provides a reasonable substitute for the DHS data as high scoring ChIP-seq peaks are known to be enriched within DHS sites. The DNaseI approach produced excellent matches to the known literature motif in all 6 datasets in human. The ensemble approach resulted in similar performance, with the exception of SRF factor, which has relatively low enrichment of binding sites in ChIP-seq peak regions compared to the other factors. This seemed to result in too weak a signal to detect based on the whole ensemble of input regions.

The largest single ChIP-seq panel has been published as part of a study of transcription factor binding in mouse embryonic stem cells [25]. I applied cERMIT to twelve datasets from this study: cMyc, nMyc, E2f1, CTCF, Esrrb, Klf4, Nanog, Oct4, Sox2, STAT3, Tcfcp2l1, Zfx. As no DHS data has been published for mouse so far, the ensemble approach was used to define the set of putative regulatory regions. The additional data for the non-sequence-specific factor p300 was also used to define the space of regulatory regions, as its broad repertoire of binding partners should help to define an appropriate target set. Results from the cERMIT analysis are shown in Figure 4.4, which also shows the motifs identified in the original study, using the two popular algorithms Weeder [150] and NestedMICA [44]. In all 12 cases cERMIT recovered a good approximation to the known literature binding specificity. For Zfx, there is no known literature consensus, and in

| TF | cERMIT cluster PSSM (ensemble' putative regulatory regions) | Literature | NestedMICA | Weeder |
|----------|--|------------|------------|--------|
| cMyc | | | | |
| CTCF | | | | |
| E2f1 | | | - | - |
| Esrrb | | | | |
| Klf4 | | | | |
| nMyc | | | | |
| Nanog | | | | |
| Oct4 | | | | |
| Sox2 | | | | |
| STAT3 | | | | |
| Tefcp211 | | | | |
| Zfx | | - | | |

Figure 4.4: cERMIT predictions on mouse ChIP-seq data from [25]

that case cERMIT's prediction agrees with the results reported by the other motif finders. The E2F data set was reportedly noisy, and no motif was reported by the other motif finders; while cERMIT successfully identifies a short GC-rich sequence motif resembling part of the site, it fails to expand to a longer motif matching the longer consensus (e.g. as reported in JASPAR [168, 194]). Finally, in the case of Sox2, cERMIT detected a more precise definition of each binding site than both Weeder and NestedMICA, whose prediction corresponded to motifs spanning sites for both Sox2 and Oct4, which are known to frequently co-occur as a module and co-regulate target genes. This demonstrates a strength of cERMIT as compared to Weeder and NestedMICA; it is able to integrate the quantitative evidence for tens of thousands of putative regulatory regions (35,500 regions for the mouse 'ensemble' set), rather than running on a small set of a few hundred highly scoring regions, in which a co-occurring motif might dominate over the true targets of the assayed factor. This makes the proposed motif discovery pipeline naturally suited to take full advantage of the state-of-the-art high-throughput sequence data.

4.4 Conclusions

I have demonstrated that the cERMIT motif discovery strategy described in Section 3.1 is easily scalable to genome-wide technologies such as ChIP-seq, which provide data for the analysis of a much larger sequence space for putative TF targets. While cERMIT does not require an explicit background model, it detects enriched motifs by virtue of analyzing their occurrence patterns in the complete set of regulatory regions. In higher organisms with a complex non-coding genome, the definition of regulatory regions is a challenging task; however, recent high-throughput approaches to map open chromatin, or factors such as p300 which interact with a range of enhancers, provide a good approximation. In fact, based on the empirical evidence from mouse ChIP-seq data, even a simple joint set of target regions from a panel of different TFs can serve that purpose. Naturally, this will lead to differences in performance if the TFs have a wide range of biological targets. Data on open chromatin under different conditions is expected to increase by efforts of the ENCODE consortium. As the described empirical results show, the definition of putative regulatory regions is

already very good given the current limited data, even though the conditions of DNaseI-chip and ChIP-seq matched for only some experiments.

Together with other recent approaches which utilize available quantitative evidence of regulation, the results reported here demonstrate convincingly that motif finders which make intelligent use of this additional information consistently outperform earlier motif finders. In contrast to the notoriously difficult motif finding problem based on over-representation in sequence alone, the scale-up in genomic experimental techniques, combined with appropriate motif finders, has made great progress on the problem to efficiently decode the regulatory information in complex genomes.

Chapter 5

Applications of cERMIT to Studying the Post-transcriptional Regulation

5.1 Analysis of RBP Regulation Using Transcriptome-wide RNA Cross-linking Data

Here I present a novel methodology for analysis of PAR-CLIP data to generate a transcriptome-wide high-resolution map of RNA-protein interaction sites.

5.2 PAR-CLIP Datasets

The main focus was the analysis of human PAR-CLIP datasets described in [74] which profile the targets of four distinct mRNA-interacting factors. Three of the datasets were generated from immunoprecipitation data of the sequence-specific RBPs Quaking (QKI), Pumilio2 (PUM2), and Insulin-like growth factor 2 binding protein (IGF2BP1). While QKI is a well-studied splicing factor in the nucleus [58], Pumilio2 RBPs are involved in mRNA stability and translation in the cytoplasm [196]. The functions of Pumilio2 are widely studied in a variety of species, and their global RNA targeting properties have been examined across a large phylogeny [62, 63, 59, 99, 132]. IGF2BP1 belongs to a family of genes that are able to regulate translation by their direct binding to target mRNAs.

The fourth dataset consists of pooled libraries assaying members of the Argonaute (AGO) family of RBPs, central components of the RNA-induced silencing complex (RISC) which directs microRNAs (miRNA) to their target transcripts, thereby negatively impacting gene expression [8]. Different from the other RBPs, Argonaute members do not have a specific mRNA recognition site; rather, their targets are specified by the interaction of the miRNA in RISC with partially complementary sequences in the target mRNAs. The seed region of the miRNA is regarded as the important

sequence determinant in target mRNA interactions [111]. AGO cross-linking is currently a popular method to directly identify miRNA targets, but the libraries contain a mixture of all targets of those miRNAs expressed in a particular cellular context.

Initial analysis of the PAR-CLIP data revealed that interaction sites of different proteins exhibit particular patterns of T-to-C conversions, likely reflecting the accessibility of nucleotides in the RNA bound by the protein. Therefore, conversions do not have to include all thymines of a sequence motif equally, and may not even fall directly on top of conserved motifs at the interaction sites. Most notably, miRNA seed matches were observed to be largely devoid of T-to-C conversions, and conversions were predominantly located directly upstream of the 5' end of the seed match, which was also observed by [102].

5.3 Data Pre-processing

The T-to-C conversion event that occurs at the site of RNA-protein crosslinking can be used to identify, with high resolution, where within the read data the actual RBP interaction occurred, and subsequently, which sequence motifs are found at these interaction sites. A recently proposed toolkit, dubbed PARalyzer (PAR-CLIP data analyzer), uses a non-parametric kernel-density estimate classifier to identify the RNA-protein interaction sites from a combination of T-to-C conversions and read density. Next I briefly outline the steps in the PARalyzer analysis which uses as input the full set of aligned reads to the genome and produces as output a set of peak regions with corresponding evidence of RBP cross-linking that is used as input to the motif discovery approach described in detail in Section 5.1.

Genomic Mapping and Peak-calling of PAR-CLIP Data Reads are first aligned to the genome, and those overlapping by at least a single nucleotide are grouped together. To exploit available read data in an effective way, relatively lenient alignment parameters are utilized. Reads are allowed to be as short as 13nt after adapter stripping, and a read may contain up to 2 mismatches restricted to T-to-C conversions (in comparison, the analysis by [74] used a read length of at least

20nt, and allowed for one T-to-C mismatch). Within each read-group, PARalyzer generates two smoothed kernel density estimates, one for T-to-C transitions and one for non-transition events. Nucleotides within the read-groups that maintain a minimum read depth, and where the likelihood of T-to-C conversion is higher than non-conversion, are considered interaction sites.

Initial read-groups are extended either to encompass the full underlying reads that contain a conversion event, or by a generic window size. The choice between these methods is dependent on the cross-linking properties of the analyzed RBP. For example, extending the region by five nucleotides on each side efficiently captures for Pumilio2 binding sites, where crosslinking occurs directly at the motif. In contrast, when assaying the Argonaute protein family in which the miRNA-mRNA interaction site is protected from both digestion and T-to-C conversion events, extending the region based on the underlying reads will include the location of conversion as well as the bound site, i.e. the miRNA seed matches.

For each read-group that contained at least 5 reads, a kernel-density based classifier was utilized to more precisely delineate the region of crosslinking ('signal') versus non-crosslinking ('background'). Class-specific densities were estimated using a Gaussian kernel density estimator with globally fixed precision parameter $\lambda = 3$.

More formally, for a given read-group of length L define $x_{T-to-T}^{(i)}$ and $x_{T-to-C}^{(i)}$ to be the number of observed conversion and non-conversion events, respectively, at an offset i relative to the start, and with a minimum read depth of 5 to be able to estimate conversion frequencies. Let n_{T-to-T} and n_{T-to-C} be the total number of conversion and non-conversion events in the group. For any position $j \in \{1, \dots, L\}$ define:

$$f_{T-to-C}(j) = \sum_{i=1}^L \frac{x_{T-to-C}^{(i)}}{n_{T-to-C}} \frac{1}{\sqrt{2\lambda^2\pi}} e^{-\frac{\|i-j\|^2}{2\lambda^2}} \quad (5.1)$$

$$f_{T-to-T}(j) = \sum_{i=1}^L \frac{x_{T-to-T}^{(i)}}{n_{T-to-T}} \frac{1}{\sqrt{2\lambda^2\pi}} e^{-\frac{\|i-j\|^2}{2\lambda^2}} \quad (5.2)$$

Which, after normalization, produces a non-parametric estimate for the density of conversions

and non-conversions, respectively:

$$k_{T-to-C}(j) = \frac{f_{T-to-C}(j)}{\sum_{j=1}^L f_{T-to-C}(j)} \quad (5.3)$$

$$k_{T-to-T}(j) = \frac{f_{T-to-T}(j)}{\sum_{j=1}^L f_{T-to-T}(j)} \quad (5.4)$$

The nucleotide positions j such that $k_{T-to-C}(j) \geq k_{T-to-T}(j)$ are considered to be interaction sites.

5.4 Analysis of RNA Binding Motifs

The analysis performed in [74] successfully applied standard motif discovery approaches (PhyloGibbs[177], MEME[6]) on the subset of top 100 most highly confident read-groups to predict RNA binding preferences. Such a strategy is well justified in cases where the target-binding motif is of low degeneracy and/or long and hence contains high discriminative signal relative to the background sequence. When this is not the case, a larger set of example sequences with the motif occurrence, with possibly variable binding affinity, can facilitate the search process. When this is not the case, a larger set of example sequences with the motif occurrence, with possibly variable binding affinity, could facilitate the search process. There is currently no method specific to PAR-CLIP data that can identify RBP motifs or miRNA seeds taking into account the entire set of binding evidence. We therefore extended a previous motif finding approach to accurately determine the motifs of sequence-specific RBPs, as well as to identify miRNA targets from Argonaute PAR-CLIP data.

There are two essential components of the motif discovery algorithm: an enrichment function to score evidence of binding for a given sequence motif represented as a k -mer over the alphabet of IUPAC symbols A, C, G, U, W, K, R, Y, S, M, N, and a search strategy that explores the motif space for high-scoring motifs. cERMIT was based on the assumption that evidence was available for an input set of potential regulatory target regions, independent of a specific analyzed factor (e.g., all upstream regions for small genomes such as *S. cerevisiae*, or regions of open chromatin in higher eukaryotes). Here, the regions to be evaluated are based on the same experiments that assayed the

particular protein of interest. We therefore rephrased the original scoring within a classical linear regression framework, allowing for flexible and easily extensible accounting of biases unrelated to protein binding, such as sequence composition or cluster size.

The binding evidence for PARalyzer-generated clusters were reflected in the number of observed T-to-C conversions, which was used (after log₂ transformation) as binding evidence for each sequence cluster. Based on the PAR-CLIP experimental data in [74], the number of observed T-to-C conversions strongly correlated with the total number of reads, which suggested that a very similar motif discovery strategy can also be applied to CLIP-seq datasets [190] by using the (log₂ transformed) number of reads as binding evidences for each cluster.

5.4.1 Sequence-specific RNA Binding Proteins

The motif discovery problem for RNA Binding Proteins (RBPs) such as Pumilio2 [62, 63, 59] and Quaking [58] is closely related to the discovery of DNA-binding preferences of transcription factors (TFs). The proposed approach from Section 3.1 exhibits highly competitive performance in the context of TF binding site discovery [61]. cERMIT differs from most other motif identification tools by making use of the complete quantitative evidence for a genome-wide set of regulatory regions (such as ChIP-chip values for all promoters, or ChIP-seq peaks from high-throughput sequencing). Rather than identifying a motif in a subset of top candidates of arbitrary size, cERMIT ranks all putative target regions based on binding evidence and identifies sequence motifs of flexible length that are highly enriched in targets with high binding evidence. This idea was extended to the setting of RBPs. Here, the regions to be evaluated are based on the same experiments that assayed the particular protein of interest. In order to allow for flexible and easily extensible accounting of biases unrelated to protein binding, I rephrased the original scoring described in 3.2 within a classical linear regression framework, such as sequence composition or cluster size. A useful extension in this setting is allowing for different treatment of genomic regions depending on their annotation. This can be achieved by including the annotation categories as additional covariates in the models, possibly including interaction terms. An alternative, more flexible formulation can be achieved in

the framework of a hierarchical random effects model, with annotation category indicator variables being random intercepts, potentially augmenting the model with additional group-specific covariates.

The motif identification of both QKI and PUM2 datasets were successful in recovering their respective consensus binding motifs [74, 58, 195]. For this analysis, I used read-groups or PARalyzer clusters that contained at least five reads and mapped to a genic region not flagged as a repeat region.

5.4.2 Enrichment Analysis of Argonaute-associated MicroRNAs

For the analysis of AGO dataset, I take advantage of previous studies on the well-established mechanism of miRNA gene regulation [111, 9], which is based on the complementarity of miRNAs to target mRNA transcripts. In particular, I represent each miRNA by a short list of canonical 5 end seeds: 8mer-A1, 8mer-m1, 7mer-A1, 7mer-m1, 7mer-m8, 6mer2-7, 6mer3-8. Instead of performing a *de novo* motif search as in the case of Pumilio2 and Quaking, I can limit the motif search to a pre-specified seed list of known miRNAs, e.g. as defined in miRBase [70]. In cases where additional information on miRNA expression is available, it is possible to further restrict the search to the subset of expressed miRNAs.

Members from the same miRNA family share canonical seeds, and across families there are also cases where seeds differ by only one nucleotide in a substitution or a shift. This is reflected in the motif enrichment analysis, where such miRNAs receive highly similar scores and are essentially indistinguishable. We therefore post-processed the results to group together miRNAs with highly similar canonical seeds.

For the motif analysis on the combined AGO PAR-CLIP data sets, all human miRNAs available in miRBase v16 were used as input for the restricted motif analysis of the microRNA enrichment analysis tool (mEAT). Despite starting from all known human miRNAs, mEAT automatically ranked the top expressed miRNAs in the cell line on the top of the list of predicted enriched miRNA seed clusters (Table 5.1). Therefore, this enrichment analysis can be used to identify those miRNAs

| cluster # | mirbase | 8-mer | expr rank | microRNA score | p-val | # targets | cumulative #targets |
|------------------|---------------|--------------|-----------|----------------|----------|-----------|---------------------|
| 1 | hsa-mir-16-2 | TGCTGCTA | 22 | 17.93 | 3.50E-20 | 438 | 438(3%) |
| | hsa-mir-15b | TGCTGCTA | 53 | 17.93 | 3.50E-20 | 438 | 438(3%) |
| | hsa-mir-15a | TGCTGCTA | 64 | 17.93 | 3.50E-20 | 438 | 438(3%) |
| | hsa-mir-195 | TGCTGCTA | NA | 17.93 | 3.50E-20 | 438 | 438(3%) |
| | hsa-mir-16-1 | TGCTGCTA | NA | 17.93 | 3.50E-20 | 438 | 438(3%) |
| | hsa-mir-103-2 | ATGCTGCT | 2 | 14.41 | 9.70E-13 | 620 | 620(5%) |
| | hsa-mir-107 | ATGCTGCT | 39 | 14.41 | 9.70E-13 | 620 | 620(5%) |
| | hsa-mir-103-1 | ATGCTGCT | NA | 14.41 | 9.70E-13 | 620 | 620(5%) |
| | hsa-mir-424 | TGCTGCTG | 60 | 12.92 | 1.50E-08 | 632 | 632(5%) |
| | hsa-mir-497 | TGCTGCTG | 133 | 12.92 | 1.50E-08 | 632 | 632(5%) |
| | hsa-mir-646 | AGCTGCTT | NA | 10.5 | 1.10E-06 | 708 | 708(6%) |
| | hsa-mir-503 | CGCTGCTA | 97 | 10.08 | 1.70E-07 | 714 | 714(6%) |
| | 2 | hsa-mir-106b | GCACTTTA | 5 | 17.63 | 8.90E-17 | 455 |
| hsa-mir-20a | | GCACTTTA | 9 | 17.63 | 8.90E-17 | 455 | 1164(9%) |
| hsa-mir-106a | | GCACTTTT | 121 | 15.65 | 1.60E-15 | 565 | 1272(10%) |
| hsa-mir-519c | | TGCACTTT | NA | 14.71 | 7.60E-21 | 689 | 1395(11%) |
| hsa-mir-519c-3p | | TGCACTTT | NA | 14.71 | 7.60E-21 | 689 | 1395(11%) |
| hsa-mir-519a-2 | | TGCACTTT | NA | 14.71 | 7.60E-21 | 689 | 1395(11%) |
| hsa-mir-519b-3p | | TGCACTTT | NA | 14.71 | 7.60E-21 | 689 | 1395(11%) |
| hsa-mir-519a-1 | | TGCACTTT | NA | 14.71 | 7.60E-21 | 689 | 1395(11%) |
| hsa-mir-526bstar | | GCACTTTC | NA | 14.57 | 4.80E-22 | 746 | 1450(12%) |
| hsa-mir-93 | | GCACTTTG | 1 | 12.99 | 1.40E-13 | 790 | 1490(12%) |
| hsa-mir-17 | | GCACTTTG | 10 | 12.99 | 1.40E-13 | 790 | 1490(12%) |
| hsa-mir-20b | | GCACTTTG | NA | 12.99 | 1.40E-13 | 790 | 1490(12%) |
| hsa-mir-519d | | GCACTTTG | NA | 12.99 | 1.40E-13 | 790 | 1490(12%) |
| hsa-mir-520d-3p | | AGCACTTT | NA | 12.15 | 4.20E-11 | 796 | 1496(12%) |
| hsa-mir-520b | | AGCACTTT | NA | 12.15 | 4.20E-11 | 796 | 1496(12%) |
| hsa-mir-520e | | AGCACTTT | NA | 12.15 | 4.20E-11 | 796 | 1496(12%) |
| hsa-mir-372 | | AGCACTTT | NA | 12.15 | 4.20E-11 | 796 | 1496(12%) |
| hsa-mir-520c-3p | | AGCACTTT | NA | 12.15 | 4.20E-11 | 796 | 1496(12%) |
| hsa-mir-520a-3p | | AGCACTTT | NA | 12.15 | 4.20E-11 | 796 | 1496(12%) |
| hsa-mir-3609 | | TCACCTTG | NA | 10.2 | 9.30E-09 | 798 | 1498(12%) |
| 3 | hsa-mir-92a-1 | GTGCAATA | 4 | 13.59 | 4.80E-10 | 223 | 1709(14%) |
| | hsa-mir-32 | GTGCAATA | 95 | 13.59 | 4.80E-10 | 223 | 1709(14%) |
| | hsa-mir-92b | GTGCAATA | 101 | 13.59 | 4.80E-10 | 223 | 1709(14%) |
| | hsa-mir-92a-2 | GTGCAATA | NA | 13.59 | 4.80E-10 | 223 | 1709(14%) |
| | hsa-mir-25 | GTGCAATG | 11 | 11.38 | 2.20E-09 | 239 | 1722(14%) |
| | hsa-mir-363 | GTGCAAIT | 130 | 11.33 | 1.60E-09 | 265 | 1746(14%) |
| | hsa-mir-367 | GTGCAAIT | NA | 11.33 | 1.60E-09 | 265 | 1746(14%) |
| 4 | hsa-mir-454 | TTGCACTA | 108 | 12.04 | 2.30E-04 | 298 | 1904(16%) |
| 5 | hsa-mir-101-2 | GTA CTGTA | 12 | 11.87 | 1.70E-11 | 202 | 2098(17%) |
| | hsa-mir-101-1 | GTA CTGTA | NA | 11.87 | 1.70E-11 | 202 | 2098(17%) |
| | hsa-mir-144 | ATA CTGTA | NA | 9.83 | 8.30E-06 | 260 | 2151(18%) |

Figure 5.1: Top enriched microRNAs based on the Argonaute PAR-CLIP data from [74]

with the strongest impact on mRNA targeting, even in the absence of miRNA expression information. While the initial PAR-CLIP study reported that seed matches could explain about 50% of CCRs, this was based on 6-mer matches to the top 100 expressed individual miRNAs. As the above analysis showed, only the matches of the top ~60 or so miRNAs provide a signal above background. The *de novo* motif analysis here confirms this: The top 5 expressed miRNAs alone can explain ~18% of all targets, but collectively, all 25 significantly enriched seed match families covered only ~30% of the clusters.

5.5 Conclusions

As with many new short-read deep sequencing protocols, the PAR-CLIP approach to elucidate RNA binding sites enables specific opportunities for in-depth analysis and interpretation of genomic data. In addition to mapping sequence-specific RBPs such as PUM2, QKI or IGF2BP1, an anticipated popular application of this protocol will be to study binding by members of the RISC complex, making it possible to identify the joint set of transcriptome-wide miRNA targets under specific conditions. To address the challenges posed by these two scenarios, the motif analysis approach described in Section 3.1 successfully identified binding motifs for sequence-specific RBPs or over-represented miRNA seed-matches.

Chapter 6

Randomized Dimension Reduction and Inference of Population Structure

6.1 Introduction

With the increased availability of large high-dimensional data set the demand for scalable dimension reduction methods based on spectral decomposition of the estimated covariance structure has dramatically increased. In this section I investigate the extension of several such dimension reduction approaches to the novel setting of large number of variables and data samples using a recently introduced randomized algorithm for approximate dimension reduction [162]. Specific approaches I consider in detail are Principal Component Analysis (PCA) and Locality Preserving Projections (LPP) and I also discuss extension to supervised dimension reductions based on generalized eigen-decompositions: Sliced Inverse Regression (SIR) and Localized Sliced Inverse Regression (LSIR). Tests of the runtime and the quality of the approximation provided by the randomized dimension reduction algorithms is preformed on simulated and real SNP microarray data sets.

6.2 Statistical Methods and Algorithms

I consider three common dimension reduction approaches unsupervised dimension reduction [151, 88], localized (non-linear) unsupervised dimension reduction [187, 167, 43, 80, 10], and supervised dimension reduction [116, 36, 115, 79, 66, 65, 112, 143, 185, 35] and discuss algorithmic adaptations that allow for scaling these approaches to massive data sets, tens to hundreds of thousands of observations and millions of variables. The ideas I develop are applicable, in general, to spectral decomposition-based approaches to dimension reduction and integrate recent developments in numerical linear algebra, theoretical computer science and statistics. The main tool I will use to scale these methods are randomized approximate matrix factorization algorithms [72, 162]. There will

be two complementary arguments for randomization. The first argument is the increased computational efficiency resulting from using randomized algorithms. This idea has been well developed in the numerical analysis and theoretical computer science communities. The second argument comes from a statistical perspective and has drawn less attention, the inherent sampling based nature of randomized algorithms provides an algorithmic form of regularization. This can help prevent overfitting and the randomized scalable approximate algorithm can result in better performance than its exact deterministic counterpart due to this regularization.

I will adapt randomized algorithms for matrix factorization to dimension reduction procedures. The appeal of the randomized algorithms is their computational efficiency which is vital in the analysis of large high-dimensional data. In particular, the randomized methods I will use [162, 75] provide flexible control of the degree of accuracy in the factorization and make explicit the tradeoff between computation time and estimation accuracy. In most practical scenarios the data we observe is a random sample with some noise, so it does not make sense to require estimation accuracy of the factorization beyond the level of the noise and the stochasticity of the sample. This perspective is not so prevalent in classical numerics applications and provides an argument for treating the computation time as a regularization parameter in dimension reduction procedure.

Many dimension reduction methods can be stated as a generalized eigendecomposition problem

$$Cv_i = \lambda_i Dv_i,$$

where $\{v_i\}$ are the eigenvectors and $\{\lambda_i\}$ are the eigenvalues. Typically, dimension reduction consists of projecting the data or the observations onto $\{v_i\}_{i=1}^{\ell}$ corresponding to the ℓ eigenvalues greater than zero by some threshold. It is almost always assumed that $\ell \ll p$, where p is the dimension of the observations. The subspace $B = \text{span}(v_1, \dots, v_{\ell})$ will be the result of the dimension reduction methods and the data will be projected onto this subspace. In the case of PCA D is the identity and C is the empirical covariance matrix of the data. The ideas that I will develop are applicable to dimension reduction methods that can be formulated as the above spectral decomposition problem. Some popular representatives include PCA, LPP, SIR, LSIR, Linear Discriminant Analysis (LDA), and Canonical Correlation Analysis (CCA).

To exploit randomized approximation algorithms we will need to assume some structure in the matrices C and D – typically it is assumed that those matrices have low rank. The structure of C and D , depends on the specific dimension reduction method of choice. For this reason, I will focus in detail on two specific dimension reduction methods and apply randomized algorithms to these methods. The first is PCA and is widely used for *unsupervised* dimension reduction – probably the most common dimension reduction tool. The second method I will focus on is LPP [80], also an unsupervised method developed in the context of manifold learning [187, 167, 43, 10]. LPP is expected to outperform PCA when there is local structure in data, for example there are several clusters in the data or the data is concentrated on a manifold. In addition I’ll discuss extensions to two *supervised* dimension reduction methods: SIR [116] and LSIR [200]. Similarly to LPP and PCA, LSIR outperforms SIR when there is local structure in data.

6.2.1 Notation

The following notation and definitions are used repeatedly in the remainder of the current section. For positive integers p and d , $\mathbb{R}^{p \times d}$ stands for the class of all matrices with real entries of dimension $p \times d$, and $\mathbb{S}^{p \times p}$ denotes the sub-class of symmetric positive semi-definite $p \times p$ matrices. For $B \in \mathbb{R}^{p \times d}$, $\text{span}(B)$ denotes the subspace of \mathbb{R}^p spanned by the columns of B . A *basis matrix* for a subspace \mathcal{S} is any full column rank matrix $B \in \mathbb{R}^{p \times d}$ such that $\mathcal{S} = \text{span}(B)$, where $d = \dim(\mathcal{S})$. A *semi-orthogonal* matrix $A \in \mathbb{R}^{p \times d}$ has orthonormal columns, $A^T A = I_p$. For a matrix $\Sigma \in \mathbb{S}^{p \times p}$, the inner product in \mathbb{R}^p defined by $\langle x_1, x_2 \rangle_\Sigma = x_1^T \Sigma x_2^T$ is referred to as the Σ inner product; when $\Sigma = I_p$, this is the usual inner product. A projection relative to the inner product Σ has the matrix representation $P_{B(\Sigma)} = B(B^T \Sigma B)^\dagger B^T \Sigma$, where the projection is onto the $\text{span}(B)$, $B \in \mathbb{R}^{p \times q}$ and † indicates the Moore-Penrose inverse. When $\Sigma = I_p$ I use the abbreviated notation P_B .

Data

Denote the data matrix or *covariates matrix* by $X = (X_1, \dots, X_p)^T \in \mathbb{R}^{n \times p}$, where n is the number of samples and p is the number of variables. If provided, denote the response to be $Y \in \mathbb{R}^n$.

The data and the response are drawn from a joint distribution $(X, Y) \sim \mathcal{P}_{X \times Y}$, which induces the corresponding marginal distributions $X \sim \mathcal{P}_X$ and $Y \sim \mathcal{P}_Y$.

Unless explicitly specified otherwise, assume that both the sample data and the response are centered and scaled to have unit variance, so that $\sum_{i=1}^n Y_i = \sum_{i=1}^n X_{ij} = 0$ and $\sum_{i=1}^n Y_i^2 = \sum_{i=1}^n X_{ij}^2 = 1$ for all $j = 1, \dots, p$. Denote the covariance of X to be $\Sigma = \text{Cov}(X) \in \mathcal{S}^{p \times p}$, and the response variance to be $\sigma_Y = \text{Var}(Y) \in \mathbb{R}^+$. Denote the corresponding sample estimators by $\hat{\Sigma}$ and $\hat{\sigma}_Y$.

Types of Error

Our methodology development will focus on the estimation of a basis for the projective subspace, characterizing linear dimension reductions. Hence, the population quantity of interest will be the span of a *basis matrix* B , where $B \in \mathbb{R}^{p \times d}$, $d \leq p$. Without loss of generality, B is assumed to be a semi-orthogonal matrix and is estimated by two classes of estimators:

- (1) estimate based on exact spectral decomposition: \hat{B} ;
- (2) estimate based on randomization-based spectral decomposition: \tilde{B} .

Of major interest when discussing the inference properties of the proposed procedures are three types of error

- estimation error: $E_e = \|B - \hat{B}\|$
- approximation error: $E_a = \|\hat{B} - \tilde{B}\|$
- total error: $E = \|B - \tilde{B}\| \leq \|B - \hat{B}\| + \|\hat{B} - \tilde{B}\| = E_e + E_a$.

Subspace Distance Metric

The main goal is to minimize the total error E . To judge the quality of the proposed estimators I use the *trace correlation* [201, 206] as the distance metric between subspaces. Let $\mathcal{S}_A = \text{span}(A) \in$

$\mathbb{R}^{s \times d}$ and $\mathcal{S}_B = \text{span}(B) \in \mathbb{R}^{s \times d}$, for any $d, s \in \mathbb{N}^+$. Define the distance between \mathcal{S}_A and \mathcal{S}_B to be

$$D_{tr}(\mathcal{S}_A, \mathcal{S}_B) := 1 - \sqrt{\frac{1}{d} \text{tr}(P_A P_B)}, \quad (6.1)$$

Notice that $0 \leq D_{tr}(\mathcal{S}_A, \mathcal{S}_B) \leq 1$ and $D_{tr}(\mathcal{S}_A, \mathcal{S}_B) = 1$ corresponds to orthogonal subspaces while $D_{tr}(\mathcal{S}_A, \mathcal{S}_B) = 0$ indicates identical subspaces.

6.2.2 Principal Component Analysis

In this section I consider the classic approach for unsupervised dimension reduction based on PCA. I first introduce the problem and a standard numerical approach to solving it, followed by an approximate solution using a randomized algorithm. Next I discuss the inference properties of the exact and the randomized dimension reduction and the potential for applying regularization via the randomized estimator.

PCA is an unsupervised approach widely used in applications where measurements are made on a large number of variables and the objective is to find a set of $k \leq p$ linear combinations of the original p variables

$$\xi_j = b_j^T X, \quad j = 1, 2, \dots, k,$$

which retain as much as possible of the variation in the original data set. Thus the j -th coefficient vector $b_j = (b_{1j}, \dots, b_{pj})$ satisfies the following

- the linear projections ξ_j , $j = 1, 2, \dots, k$ are ordered by decreasing variance: $\text{var}(\xi_1) \geq \dots \geq \text{var}(\xi_k)$
- ξ_i is uncorrelated with ξ_j for all $i \neq j$.

This problem is stably and efficiently solved, for an arbitrary rectangular matrix $X \in \mathbb{R}^{n \times p}$ and all possible values $k \in \{1, \dots, n\}$, by the *Singular Value Decomposition* (SVD), which provides an

algorithmic approach to the estimation of U, S, V such that:

$$X = USV^T, \quad U^T U = V^T V = I_{n \times n}, \quad (6.2)$$

$$S = \text{diag}(l_1, \dots, l_n), \quad l_1 \geq l_2 \geq \dots \geq l_n \geq 0 \quad (6.3)$$

The diagonal entries of S are the singular values, sorted in non-decreasing order, according to the amount of captured variance in the directions defined by the corresponding singular vectors. The first k orthonormal columns of the matrix V (with the k largest singular values), provide the exact solution $B \in \mathbb{R}^{p \times k}$

$$B = \begin{pmatrix} | & & | \\ u_1 & \dots & u_k \\ | & & | \end{pmatrix}. \quad (6.4)$$

A numerically stable and computationally efficient implementation of SVD is provided as part of the industry-standard numerical linear algebra package LAPACK [4] requiring $O(n^2 p)$ time and $O(np)$ of memory. Parallelization of the code, which is based on classic SVD algorithms, is inherently problematic. This prevents the optimal use of, nowadays commonly available, cluster resources with thousands of processors capable of parallel computation. When both the number of variables (p) and the number of data points (n) is large this runtime cost is prohibitively high, which has been one of my major motivations to explore fast alternative solutions which I describe in details in Section 6.2.2.

Randomness and Dimension Reduction

It is often much more efficient to solve a computational problem presented in a high-dimensional space by first transforming it to a lower-dimensional space while preserving its essential structure. This is possible if the support of the sampling distribution of the data is a lower-dimensional (linear) manifold embedded in the high-dimensional ambient space. Random projections provide a very effective tool for constructing such dimension-reduction maps. Most approaches that rely on this ideas can be traced back to the seminal work of Johnson and Lindenstrauss [92], who showed that any set of n points in p -dimensional Euclidean space can be embedded into k -dimensional

Euclidean space, where k is *logarithmic* in n and *independent* of p , preserving all pairwise distances to within an arbitrarily small factor. More formally:

Lemma 6.2.1. (Johnson-Lindenstrauss (1984))

Let $\epsilon \in (0, 1/2)$ and let $x_1, \dots, x_n \in \mathbb{R}^p$ be arbitrary points. Let $k = O(\epsilon^{-2} \log(n)) \in \mathbb{N}$. Then there exists a Lipschitz map $f : \mathbb{R}^p \rightarrow \mathbb{R}^k$ such that

$$(1 - \epsilon) \|x_i - x_j\|_2^2 \leq \|f(x_i) - f(x_j)\|_2^2 \leq (1 + \epsilon) \|x_i - x_j\|_2^2$$

for all $i, j \in \{1, \dots, n\}$.

The original proof of the Lemma was constructive, based on random orthogonal projections. Using probabilistic approaches has allowed for the original proof of Johnson and Lindenstrauss to be greatly simplified and sharpened [55, 89, 40, 2].

It turns out that the provided guarantee regarding the approximate preservation of the pairwise distances is often enough to ensure that a solution, found working in the low dimensional space, is a good approximation to the solution in the original space. There has been rich contributions significantly expanding on these ideas that produce approximation alternatives [45, 169, 72, 118, 17, 162, 75]. Here are a few examples of areas where randomization ideas have contributed in a significant way:

- Efficient methods for factoring the coefficient matrix leads to efficient techniques for solving least squares problems [163].
- Low-dimensional embedding of data under the assumption of manifold structure often reduces to computing low-rank SVD of a matrix derived from the data [28].
- Latent Semantic Indexing is an approach for studying the relationship between text documents and keyword terms, based on the SVD factorization of the document \times term matrix [85].
- Efficient approximations to the Gram matrix have been proposed using the Nystrom Method [46].

Randomized SVD

I first describe an approximate randomized approach for the efficient estimation of a best rank- k approximation of an arbitrary rectangular matrix called *Randomized SVD* [162, 75]. At the core of the approach is the ability to efficiently identify the part of the range of the data that corresponds to the larger variance directions. The task of constructing the low-rank approximation to the given data matrix can be split naturally in two distinct steps which decouple the randomization from the linear algebra.

- **Randomization:** sample high-variance directions in the data using uniformly distributed random orthogonal matrices. In this context, *uniform* is defined in terms of Haar measure, which essentially requires that the distribution not change if multiplied by any freely chosen orthogonal matrix
- **Factorization:** restrict the data to sampled subspace from the Randomization step and apply standard exact factorizations

In essence, the *Randomization* step couples a random projection step with a form of the power iteration method, to enhance the decay of the eigenspectrum, while leaving the singular vectors unchanged. The *Factorization* step rotates the orthogonal basis constructed during the *Randomization* step to the canonical eigenvector basis and recovers estimates of the *variances* (eigenvalues) for the top eigenvectors. Next I provide a more detailed description of the above steps.

Randomization: Sample the Range of the Data Sample linear combination of *all* directions of the column space of the data with weights proportional to the corresponding empirical eigenvalues raised to a power $t \in \mathbb{N}$ (fixed parameter). Specifically, given a random projection matrix $\Omega_{n \times l}$, e.g. $\Omega_{ij} \sim N(0, 1), l \ll n$,

$$\underbrace{F^{(t)}}_{n \times l} = (X X^T)^t \Omega.$$

As a result the columns of $F^{(t)}$ contain a random sample from $\text{span}(X)$. Additional insight can be gained by considering the SVD representation of the data matrix:

$$\begin{aligned} F^{(t)} &= US^{2t}V^T\Omega \\ &= US^{2t}\Omega^*. \end{aligned}$$

Notice that $\Omega^* := V^T\Omega \Rightarrow \Omega_{ij}^* \sim N(0, 1) \forall i, j$ (other random structures work equally well in practice, e.g. $\Omega_{ij} \sim \text{Uniform}(0, 1)$, but the theoretical results are easier to derive for the Gaussian case). Hence, $F_j^{(t)} \sim N(0, US^{4i}U^T)$, where $F_j^{(t)}$ denotes the j -th column of $F^{(t)}$. Another interpretation of the construction of $F_j^{(t)}$ is that after scaling according to their corresponding eigenvalues (*importance weights*), the columns of U are sampled *uniformly* using the unstructured Gaussian random projection Ω^* to produce a linear combination of the left eigenvectors of X . Next, in a similar fashion to the (blocked) Lanczos approach (Chapter 9, in [67]) the matrix R_t is constructed by concatenating all $F^{(i)}$ for $i = 1, \dots, t$:

$$\begin{aligned} R_t &\equiv (F^{(1)} \mid F^{(2)} \mid \dots \mid F^{(t)}) \\ &= (US^2\Omega^* \mid US^4\Omega^* \mid \dots \mid US^{2t}\Omega^*). \end{aligned}$$

Each column of R_t corresponds to a random linear combination of the sample eigenvectors U , weighted by *powers* of the corresponding eigenvalues in S . Blocks $F^{(i)}$ that correspond to larger values of i contain columns with increased weight for the higher variance relative to the lower variance directions, while lower values of i result in more *uniform* weights, hence allowing for larger influence of lower variance directions.

Factorization: Exact SVD in Lower-Dimensional Space This step relies on mature technology, utilizing well-established deterministic linear algebra routines to project the data onto the inferred subspace during the *Randomization* step. First, the orthonormal basis of the sampled subspace is constructed using a pivoted QR or SVD factorization. Then an exact SVD is performed in the lower-dimensional space. Both the pivoted QR and the SVD have efficient implementation

for dense matrices in the numerical linear algebra package LAPACK [4]. See Algorithm (1) for a detailed description of the proposed Randomized SVD approach [162].

Randomized SVD 1

input

X : data matrix, $n \times p$

k : number of required top variance directions

t : number of power iterations

output

$\hat{\Sigma}$: singular values (top k)

\hat{U} : left eigenvectors ($n \times k$)

\hat{V} : right eigenvectors ($p \times k$)

Stage 1: Find approximate orthonormal basis for the range of X

1. Set $l = k + 12$
2. Generate $\Omega(n \times l)$ s.t. $\Omega_{ij} \sim N(0, 1) \quad \forall i, j$
3. Construct
$$R^{(0)} = XX^T\Omega$$
$$R^{(j)} = XX^TR^{(j-1)} \text{ for } j = 1, \dots, t$$
4. Factorize $R = (R^{(0)} | R^{(1)} | \dots | R^{(t)}) = QS, \quad Q^TQ = I$

Stage 2: Project data onto the orthonormal basis Q and do SVD

1. Construct $B = XX^TQ$
 2. Factorize $B = U\Sigma W^T, \quad \Sigma = \text{diag}(\sigma_1, \dots, \sigma_{(t+1)l}), \quad U^TU = W^TW = I$
 3. Set
$$\hat{U} = \begin{pmatrix} | & & | \\ u_1 & \dots & u_k \\ | & & | \end{pmatrix}$$
$$\hat{V} = X^T\hat{U}$$
$$\hat{\Sigma} = \text{diag}(\sqrt{\sigma_1}, \dots, \sqrt{\sigma_k})$$
-

Runtime Analysis The main runtime bottleneck is the *Randomization* step of the algorithm, which in addition turns out to be of interest for reasons of statistical inference. The asymptotic runtimes for both steps are as follows:

- Randomization: $O(tlnp)$, where $l \approx k$, which is the rank of the required approximation
- Factorization: $O(lnp + t^2l^2n)$

The overall asymptotic runtime is $O(tlnp + t^2l^2n)$, with small values of t and l relative to $n \ll p$. The runtime in both steps is dominated by the multiplication by the data matrix. Hence, if the data matrix is sparse or amenable to fast multiplication, then the runtime can be further reduced. The computational efficiency of the matrix multiplication could be improved, even in the case of dense data matrix, by using ideas from [75, 118] who suggest the use of variants of the Fourier transform (FFT) to introduce “structured” randomness that allows for more efficient multiplication than using dense matrix containing Gaussian noise. Another relatively simple, yet potentially very effective speed-up could be implemented by means of parallelizing the matrix multiplication, which can be performed in a distributed fashion.

Error Bound The Randomized SVD algorithm introduced in the previous section provides not only a highly computationally efficient solution but also achieves very high accuracy [162, 75]. Despite the fact that the algorithm fails to produce good approximation with a certain probability, this probability of failure can be easily controlled to be negligibly small. In addition the algorithm is not sensitive to the random number generator that is used. Hence, for all practical purposes it can be treated as a computational device that efficiently computes a good deterministic approximation to the exact sample estimate. On the other hand the structural randomness in Randomized SVD could be exploited to serve as a regularization control with potential utility in inference settings which is discussed further in Section 6.2.2.

The authors in [162] proved a strong relative error bound. In particular, for a given $n \times p$ matrix X and a pre-specified rank of the required approximation k , the algorithm produces orthonormal \hat{U} , \hat{V} , and diagonal $\hat{\Sigma}$ with non-negative entries, such that:

$$\|X - \hat{U}\hat{\Sigma}\hat{V}^T\|_2 \leq Cn^{1/(4t)}\sigma_{k+1}$$

with very high probability (typically $1 - 10^{-15}$, independent of X), where $\|\cdot\|_2$ denotes the spectral norm and C is a constant independent of X . Numerical experiments show [162] that the theoretical bound holds in practice with greatly reduced scaling constants as compared to the theoretical prediction, independent of the matrix structure.

Inference Properties

In this section I study the implications of introducing a random projection step in the inference of the dimension reduction subspace based on the PCA objective function and provide arguments that, in addition to the attractive computational properties, the Randomized SVD approach opens up opportunities for imposing implicit regularization by the introduction of small amount of “random noise”.

Implicit Regularization Control The key parameter that controls both the computational efficiency and the approximation accuracy of the randomized estimator described in Algorithm 1 is t

since

$$E_a^{(t)} = \|\hat{B} - \tilde{B}^{(t)}\|,$$

with $\|\hat{B} - \tilde{B}^{(t)}\| \downarrow 0$ as t increases. As noted in Section 6.2.1

$$\|B - \tilde{B}^{(t)}\| \leq \|B - \hat{B}\| + \|\hat{B} - \tilde{B}^{(t)}\| \quad (6.5)$$

$$E^{(t)} \leq E_e + E_a^{(t)} \quad (6.6)$$

The tightest upper bound is achieved for $E_a^{(t)} = 0$, yet we have no control over E_e , so for any finite sample size the order of the right hand side can be no smaller than $O(E_e)$. This suggests that t needs to be large enough to ensure that $E_a^{(t)} = O(E_e)$ and any further reduction of $E_a^{(t)}$ would be unnecessary.

When the sample size is small relative to the variable dimension the resulting estimator \hat{B} tends to have high variance and the bound described in Equation (6.6) could be quite loose for large values of t . On the other hand, small values of t could have beneficial regularization effect on the subspace estimate \tilde{B} , as lower sample variance directions are allowed to have higher impact. This could be a useful feature in the context of an optimality criterion which depends on a (possibly unknown) subspace spanned by a subset of the top sample eigenvectors, hence allowing for the possibility of preferring lower variance directions over higher variance directions. One such example is when the subspace estimate is used for prediction. In that case data-adaptive estimation of t can be used to select the optimal value for t that would allow the incorporation of lower variance directions that have predictive power.

In addition, if the data has approximately low-rank structure with all signal concentrated in the top few eigenvectors, the low-rank factorization of the sample covariance produces a good approximation when the level of noise is small. Hence, it would be desirable to have peaked distribution on the sampling weights of the variance directions with little or no regularization being necessary (t small). In this setting the noise eigenvalues decay very fast with t which results in a very efficient computation.

In the presence of substantial added noise, with most of the signal concentrated in the top

few eigenvectors, yet still having some signal in the lower variance directions, the lower variance directions may contain useful information. Hence it would be helpful to allow for non-negligible probability of sampling from these directions. Large values of t may not necessarily achieve that, as they would result in improved estimation accuracy of the space spanned by the top few sample eigenvectors at the expense of lower variance directions. In this case *intermediate* or even *small* values of t are more appropriate.

In both cases of small and substantial noise a data dependent approach to selecting t that is data dependent is useful. One such approach, based on the idea of cross-validation is presented next.

Selecting the Regularization Parameter In the unsupervised setting of PCA and LPP it is possible to use the reconstruction accuracy of the projection of each $\{X_i\}_{i=1}^n$ onto the dimension reduction subspace $\tilde{B}^{(t)}$ as the optimization criterion, which in the population corresponds to

$$t^* = \arg \min_{t \in \{1, \dots, t_{max}\}} \mathbb{E}_X \|(I - P_{\tilde{B}^{(t)}})X\|_2^2.$$

Leave-one-out cross-validation (LOO-CV) provides an approximately unbiased estimate of this error based on the random sample $\{X_i\}_{i=1}^n$

$$\text{CV}_{\text{loo}}^{(t)} = \frac{1}{n} \sum_{i=1}^n \|(I - P_{\tilde{B}_{(-i)}^{(t)}})X_i\|_2^2 = \frac{1}{n} \sum_{i=1}^n \|(I - \tilde{B}_{(-i)}^{(t)} \tilde{B}_{(-i)}^{(t)T})X_i\|_2^2,$$

where $\tilde{B}_{(-i)}^{(t)} \in \mathbb{R}^{p \times k}$ is the orthogonal basis matrix for the dimension reduction subspace inferred using $\{X_{(-i)}\}$, the full data set excluding the i -th sample. An equivalent form for the LOO-CV is

$$\text{CV}_{\text{loo}}^{(t)} = \frac{1}{n} \sum_{i=1}^n \left(\frac{X_i - \hat{X}_i^{(t)}}{1 - (H_{(-i)}^{(t)})_{ii}} \right)^2,$$

where $H_{(-i)}^{(t)} = \tilde{B}_{(-i)}^{(t)} \tilde{B}_{(-i)}^{(t)T}$ is the *hat* matrix and $\hat{X}_i^{(t)} = H_{(-i)}^{(t)} X_i$. It is necessary to optimize over the full range of allowable values for t to arrive at the final estimate

$$\hat{t}_{\text{loo}}^* = \arg \min_{t \in \{1, \dots, t_{max}\}} \text{CV}_{\text{loo}}^{(t)}.$$

A more computationally efficient estimate that has lower variance but higher bias than LOO-CV is provided by the c -fold cross-validation (e.g. $c = 5$ or 10). To simplify notation let $n = a \times c$ and denote the c -fold cross-validation error estimate for a pre-specified value of the parameter t to be $\text{CV}_{c\text{-fold}}^{(t)}$. Then

$$\text{CV}_{c\text{-fold}}^{(t)} = \frac{1}{a} \sum_{i=1}^a \frac{1}{c} \sum_{j=1}^c \|(I - P_{\tilde{B}_{(-i)}^{(t)}})X_{(i)}^j\|_2^2 = \frac{1}{n} \sum_{i=1}^a \sum_{j=1}^c \|(I - \tilde{B}_{(-i)}^{(t)} \tilde{B}_{(-i)}^{(t)T})X_{(i)}^j\|_2^2,$$

where $\tilde{B}_{(-i)}^{(t)} \in \mathbb{R}^{p \times k}$ is the orthogonal basis matrix for the dimension reduction subspace inferred using the full data set excluding the i -th subset of size a , $X_{(i)}$, where the input data $\{X_i\}_{i=1}^n$ is randomly partitioned into c equally sized disjoint subsets $\{X_{(i)}\}_{i=1}^c$ and $X_{(i)}^j$ for $j = 1, \dots, a$, denotes the j -th sample point within subset i . It is necessary to optimize over the full range of allowable values for t to arrive at the final estimate

$$\hat{t}_{c\text{-fold}}^* = \arg \min_{t \in \{1, \dots, t_{max}\}} \text{CV}_{c\text{-fold}}^{(t)}.$$

Runtime Analysis The overall asymptotic runtime for the c -fold cross-validation procedure scales linearly with c for a fixed value of the regularization parameter t . In more detail, for a fixed value of $t \in \{1, \dots, t_{max}\}$ there are two major computational steps:

- estimate projective subspace $\tilde{B}_{(-i)}^{(t)}$, for $i = 1, \dots, c$:

$$O(c \times [tl(n-a)p + t^2l^2(n-a)])$$

- project data and construct error estimate: $O(lnp)$

Hence, the overall asymptotic runtime is dominated by the projective subspace estimation and adds up to $O(\sum_{t=1}^{t_{max}} c \times [tl(n-a)p + t^2l^2(n-a)])$. For small values of t_{max} this is on the order of a single iteration of the Randomized SVD algorithm (see 6.2.4).

Bounds on Estimators Exact PCA produces an optimal subspace estimate in terms of average distance of the sample data from their projections onto the subspace of the pre-specified dimension

k . The quality of the estimate has been studied from the point of view of *reconstruction error* [175] and *subspace estimation error* [207]. In the latter case the authors derived an upper bound on the error that depends only on the gap between the k -th and $k+1$ -st eigenvalues. They considered the general case when the data $(X_1, \dots, X_n) \in \mathcal{X}$ is mapped to a Reproducing Kernel Hilbert Space (RKHS) \mathcal{H}_κ with kernel function κ through the feature map $\phi(x)$. In that case, the objective of PCA remains unchanged: recover the k dimensional subspace B_k , such that the projection of $\phi(x)$ onto B_k has maximum averaged square norm. I focus on the case of $\mathcal{H} = \mathbb{R}^p$, for some $p \in \mathbb{N}^+$ and the feature map is the identity.

For a general kernel the follow holds. Denote the covariance operator of variable $\phi(x)$ by K and its empirical version by K_n . Then, assuming a bounded kernel $\sup_{x \in \mathcal{X}} \kappa(x, x) < M$ for some $M < \infty$, the following holds:

Theorem 6.2.2. (Zwald-Blanchard) [207]

Let B_d and \hat{B}_d be the subspaces spanned by the first d eigenvectors of the covariance operator K , resp. K_n . Denoting $\lambda_1 > \lambda_2 > \dots$ the eigenvalues of K , if $k > 0$ is such that $\lambda_k > 0$, put $\delta_k = \frac{1}{2}(\lambda_k - \lambda_{k+1})$ and

$$B_k = \frac{2M}{\delta_k} \left(1 + \sqrt{\frac{\xi}{2}} \right)$$

Then, provided that $n \geq B_k^2$, the following holds with probability at least $1 - \epsilon^{-\xi}$

$$\|P_{S_k} - P_{\hat{S}_k}\| \leq \frac{B_k}{\sqrt{n}}$$

this entails in particular

$$\hat{S}_k \subset \{g + h, g \in S_k, h \in S_k^\perp, \|h\|_{\mathcal{H}_\kappa} \leq B_k n^{-\frac{1}{2}} \|g\|_{\mathcal{H}_\kappa}\} \quad (6.7)$$

where $S_k = \text{span}(\phi_1, \dots, \phi_k)$ and $\hat{S}_k = \text{span}(\hat{\phi}_1, \dots, \hat{\phi}_k)$ are the top k eigenvectors of K , resp. K_n .

Note that (6.7) has a geometric interpretation: the tangent between any vector in \hat{S}_k and its projection onto S_k is bounded by B_k/\sqrt{n} .

For PCA it is also possible to bound the difference between the empirical covariance and the population covariance by the following analysis. Let $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i$ and

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n (X_i - \hat{\mu})(X_i - \hat{\mu})^\top.$$

In the following, I adopt the notation $u \otimes v = uv^\top$. Then

$$\hat{\Sigma} = \frac{1}{n} (X_i - \hat{\mu}) \otimes (X_i - \hat{\mu})$$

and

$$\Sigma = \mathbb{E}[(X - \mu) \otimes (X - \mu)].$$

For a matrix A , $\|A\|$ represents the Hilbert-Schmidt norm and $\|A\|_{op}$ represents the operator norm. We will need the following fact:

$$\|u \otimes v\| = \|u\| \|v\|$$

Theorem 6.2.3. *Assume $\mu = \mathbb{E}[X]$ exists and $M_4 = \mathbb{E}[\|X - \mu\|^4] < \infty$. Then in probability*

$$\mathbb{E}[\|\hat{\Sigma} - \Sigma\|] \leq \frac{M_2 + \sqrt{M_4 - \|\Sigma\|^2}}{\sqrt{n}} \leq \frac{2\sqrt{M_4}}{\sqrt{n}}$$

where $M_2 = \mathbb{E}[\|X - \mu\|^2]$.

Proof. It is easy to check that

$$\hat{\Sigma} - \Sigma = (\hat{\mu} - \mu) \otimes (\hat{\mu} - \mu) + \left[\frac{1}{n} \sum_{i=1}^n (X_i - \mu) \otimes (X_i - \mu) - \Sigma \right].$$

Direct calculation gives

$$\mathbb{E}[\|\hat{\mu} - \mu\|^2] = \frac{\mathbb{E}[\|X - \mu\|^2]}{n} = \frac{M_2}{n} \leq \frac{\sqrt{M_4}}{\sqrt{n}}.$$

Hence

$$\mathbb{E}[\|(\hat{\mu} - \mu) \otimes (\hat{\mu} - \mu)\|] = \mathbb{E}[\|\hat{\mu} - \mu\|^2] \leq \frac{M_2}{\sqrt{n}}.$$

By definition $\Sigma = \mathbb{E}[(X_i - \mu) \otimes (X_i - \mu)]$, hence

$$\begin{aligned} \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n (X_i - \mu) \otimes (X_i - \mu) - \Sigma \right\|^2 \right] &= \frac{1}{n} (\mathbb{E}[\|(X - \mu) \otimes (X - \mu)\|^2] - \|\Sigma\|^2) \\ &= \frac{1}{n} (\mathbb{E}[\|X - \mu\|^4] - \|\Sigma\|^2). \end{aligned}$$

The desired estimate follows from the triangle inequality and the two estimates above. \square

This theorem tells us that the upper bound of the estimate $\hat{\Sigma}$ for Σ depends only on the moments of the random variable X , but not on the dimensionality. Hence, if X has a low dimensional structure, this theorem ensures that the error bound on the estimation error depends on the intrinsic, rather than the extrinsic dimensionality of the data. To be more precise, suppose $X \in \mathbb{R}^d \subset \mathbb{R}^p$, hence $\text{rank}(\Sigma) = d$. Let (σ_i, v_i) be the eigenvalues and the corresponding eigenvectors. Note, only the top d eigenvalues are nonzero. Let $\eta_i = (X - \mu)^T v_i$ be the projection of $X - \mu$ onto the direction v_i . Then η_i has mean zero, variance σ_i , and are uncorrelated. The SVD decomposition $X - \mu = \sum_{i=1}^n \sqrt{\sigma_i} \eta_i v_i$, provides simple estimates for the quantities in the upper bound in Theorem 6.2.3:

$$\begin{aligned} M_2 &= \mathbb{E}[\|X - \mu\|^2] = \sum_{i=1}^p \mathbb{E}[\eta_i^2] = \sum_{i=1}^n \sigma_i = \sum_{i=1}^d \sigma_i \\ M_4 &= \mathbb{E}[\|X - \mu\|^4] = \mathbb{E} \left[\left(\sum_{i=1}^n \eta_i^2 \right)^2 \right] = \mathbb{E} \left[\left(\sum_{i=1}^d \eta_i^2 \right)^2 \right] \\ \|\Sigma\|^2 &= \text{Tr}(\Sigma^2) = \sum_{i=1}^n \sigma_i^2 = \sum_{i=1}^d \sigma_i^2 \end{aligned}$$

All these quantities depends on d , not on p . Furthermore, the upper bound can be independent of p even when X do not have low dimensional structure, provided that their *projections* onto the PC directions all have “light” tails to ensuring that the total moments are absolutely bounded.

By combining the above analysis with matrix perturbation theory, it follows that

$$\|\hat{B} - B\| \leq \frac{\kappa}{\sqrt{n}}$$

with κ depending on the moments of X . In the case of interest for us when X has a d dimensional low-rank structure, κ depends on d , not on p .

6.2.3 Dimension Reduction Via Graph Embeddings

Dimension reduction based on graph embeddings seek to map the original data points to a lower dimensional set of points while preserving neighborhood relationships. The theoretical assumptions underlying these methods are that the data lies on a smooth manifold embedded in high-dimensional ambient space. This manifold is unknown and needs to be inferred from the data. Given enough observations the manifold can be reasonably represented as a $G = (E, V)$ [27] where the vertices $\{v_1, \dots, v_n\}$ correspond to the n observations $\{x_1, \dots, x_n\}$ and the edges corresponds to which points are close to each other. For example this neighborhood relationship can be encoded in a sparse adjacency matrix W , which is assumed to be symmetric. Given this adjacency matrix the Laplacian eigenmaps (LE) algorithm [10] embeds the data into a low dimensional space preserving local relationships between points.

Given the adjacency or association matrix the Graph Laplacian is constructed $L = D - W$, where D is a diagonal matrix with $D_{ii} = \sum_j W_{ij}$. A spectral decomposition of L

$$Lv_i = \lambda_i v_i$$

results in eigenvalues $\lambda_1 = 0 \leq \lambda_2 \leq \dots \leq \lambda_n$ with $v_1 = \mathbf{1}$. Projecting the matrix L onto the ℓ eigenvectors corresponding to the smallest k eigenvalues greater than zero embeds the n points into a ℓ dimensional space. Under certain conditions [11], the Graph Laplacian converges to the Laplace-Beltrami operator on the underlying manifold. This provides a theoretical motivation for the embedding. This embedding needs to be recomputed when a new data point is introduced and typically will not be a linear projection of the data. For computational reasons it would be advantageous to have a linear projection that can be applied to new data points without having to recompute the spectral decomposition of the graph Laplacian. The goal of Locality Preserving Projections [80] is to provide a linear approximation to the non-linear embedding of Laplacian eigenmaps (LE) [10].

The dimension reduction procedure starts by specifying the dimension of the transformed space to be $\ell < n$. Let the parameter defining the neighborhood size be k . Locality Preserving Projections

(LPP) [80] results in a generalized eigendecomposition problem. Assume that the k -NN adjacency matrix $W = J_k^T J_k$, where

$$J_k = \begin{pmatrix} \delta_{1 \sim 1} & \delta_{1 \sim 2} & \cdots & \delta_{1 \sim n} \\ \vdots & \ddots & \ddots & \vdots \\ \delta_{n \sim 1} & \cdots & \delta_{n \sim n-1} & \delta_{n \sim n} \end{pmatrix}_{n \times n}$$

$$\delta_{r \sim s} = \begin{cases} \exp\left\{-\frac{\|\tilde{X}_{s_1} - \tilde{X}_{s_2}\|}{b}\right\} & \text{if sample } s_1 \text{ is among the } r\text{-NN of sample } s_2 \text{ or vice versa} \\ 0 & \text{otherwise} \end{cases}$$

$$\implies (\forall r) \sum_{j=1}^n \delta_{r \sim j} = k$$

The generalized eigendecomposition problem reduces to

$$\begin{aligned} X^T L X e &= \lambda X^T D X e & (6.8) \\ X^T (D - W) X e &= \lambda X^T D X e \\ X^T W X e &= (1 - \lambda) X^T D X e \\ X^T J_k^T J_k X e &= (1 - \lambda) X^T D X e \end{aligned}$$

The column vectors that are the solutions to equation (6.8) are the required embedding directions $\{e_j\}$, ordered according to their generalized eigenvalues $0 = \lambda_0 < \lambda_1 < \dots < \lambda_{\ell-1}$. Hence the neighborhood-preserving optimal embedding according to the LPP criterion is:

$$x_i \rightarrow A^T y_i, \quad \text{where } A = (e_0, e_1, \dots, e_{\ell-1}).$$

LPP is obtained from a graph embedding, using a nearest neighbor graph trying to capture *local* manifold structure. If a complete graph is used, e.g. Euclidean inner products between data points, then a very similar result to PCA is produced, emphasizing the *global* structure of the data.

$$\begin{aligned} X^T L X e &= \lambda X^T D X e & (6.9) \\ X^T W X e &= (1 - \lambda) X^T D X e \\ X^T X X^T X e &= (1 - \lambda) X^T D X e. \end{aligned}$$

Since the diagonal matrix D is close to identity matrix, $X^TDX \approx X^TX$ and the minimum eigenvalues of equation (6.9) correspond to the maximum eigenvalues of

$$\begin{aligned}X^TXX^T Xe &= \lambda X^T Xe \\X^T Xe &= \lambda e,\end{aligned}$$

which is the optimization problem that is solved by PCA. Hence, LPP with a complete inner product graph is similar to PCA. The only difference is that the matrix D is used to measure the local density around each data point (by its degree on the neighborhood graph) while PCA treats all point equally.

Randomized Algorithm for LPP

In this section I describe in more detail the incorporation of the Randomized dimension reduction ideas as part of the Locality Preserving Projections algorithm.

Locality Preserving Projections 2

input X : data matrix ($n \times p$)

r : number of nearest neighbors in the affinity graph

b : bandwidth for the Gaussian kernel density estimator

k : number of required embedding directions

t : number of power iterations for Randomized SVD

output \hat{V}_{lpp} : embedding directions ($p \times k$)

Stage 1: Project data onto the top k eigenvectors of X from Randomized SVD

1. Estimate $[\tilde{U}, \tilde{S}, \tilde{V}] = \text{RandomizedSVD}(X, k, t)$

2. Project $\tilde{X} = X\tilde{V}$ ($n \times k$)

3. Set

$$J_r = \begin{pmatrix} \delta_{1 \sim 1} & \delta_{1 \sim 2} & \cdots & \delta_{1 \sim n} \\ \vdots & \ddots & \ddots & \vdots \\ \delta_{n \sim 1} & \cdots & \delta_{n \sim n-1} & \delta_{n \sim n} \end{pmatrix}_{n \times n} \quad \text{[use sparse representation, } r \times n \text{ space]}$$

$$\delta_{s_1 \sim s_2} = \begin{cases} \exp\left\{-\frac{\|\tilde{X}_{s_1} - \tilde{X}_{s_2}\|}{b}\right\} & \text{if } s_1 \text{ is among the } r\text{-NN of } s_2 \text{ or vice versa} \\ 0 & \text{otherwise} \end{cases}$$

$$D = \text{diag}\left(\sum_{j=1}^n \delta_{1 \sim j}, \dots, \sum_{j=1}^n \delta_{n \sim j}\right)$$

4. Construct

$$L = \tilde{X}^T J_r^T$$

$$\Sigma = LL^T \quad \Gamma = \tilde{X}^T D \tilde{X}$$

Stage 2: Solve Generalized Eigendecompositon

1. Solve (Cholesky) $\Sigma v_i = \lambda_i \Gamma v_i$, for $i = 1, \dots, k$

2. Set $\hat{V}_{lpp} = (v_1 | \dots | v_k)$

Runtime Analysis The main runtime bottleneck of Randomized LPP is Stage 1, where the Randomized SVD estimates are used to project the data. Hence, the overall runtime is $O(tlnp + t^2l^2n)$ (see 6.2.4). When c -fold cross-validation is used to estimate an optimal value for t the runtime becomes $O(\sum_{t=1}^{t_{max}} c \times [tl(n-a)p + t^2l^2(n-a)])$ (see 6.2.2).

6.2.4 Supervised Dimension Reduction

In this section we assume the data matrix to be $X \in \mathbb{R}^{p \times n}$, and to contain samples as rows and variables as columns. The response is univariate $Y \in \mathbb{R}$. Dimension reduction is typically only the first step in the data analysis which often also includes the steps of data visualization and regression. If the focus is on the regression, one typically seeks a low dimensional projection or embedding $R(X) \in \mathbb{R}^d$, $d \ll p$, for which a simple predictive model can be used to predict a future response $Y \in \mathbb{R}$

$$Y = f(R(X)) + \varepsilon,$$

here ε corresponds to noise or error. A natural criterion that the embedding or projection should satisfy is $\mathbb{E}[Y | X] = \mathbb{E}[Y | R(X)]$. Replacing X by $R(X)$ is termed *sufficient dimension reduction* whenever $R(X)$ retains all the relevant information for predicting Y . $R(X)$ is minimal sufficient if any other sufficient reduction $T(X)$ is a function of R [35]. The idea of sufficiency in dimension reduction was clearly captured in [35] which develops the following definition following:

Definition 6.2.4. A reduction $R: \mathbb{R}^p \rightarrow \mathbb{R}^q$, $q < p$, is sufficient if it satisfies one of the following statements:

1. *Inverse reduction*, $X | Y, R(X) \stackrel{d}{=} X | R(X)$, ($Z \stackrel{d}{=} U$ denotes equivalence in distribution between Z and U);
2. *Forward reduction*, $Y | X \stackrel{d}{=} Y | R(X)$;
3. *Joint reduction*, $X \perp\!\!\!\perp Y | R(X)$.

If (X, Y) have a joint distribution then the above conditions are equivalent and using $R(X)$ for prediction has a strong motivation, $\mathbb{E}[Y | X] = \mathbb{E}[Y | R(X)]$. The population quantity that

uniquely identifies a particular sufficient dimension reduction is the subspace $S_{Y|X} \subseteq \text{colspace}(X) \subseteq \mathbb{R}^p$, which captures all the information in the data relevant to prediction. Any basis $\eta = (\eta_1, \dots, \eta_d) \in \mathbb{R}^{p \times d}$ of $S_{Y|X}$ can be used to define the sufficient dimension reductions $R(X) = \eta^T X$. A parsimonious target of supervised dimension reduction is often taken to be the intersection of all dimension reduction subspaces—the central subspace, which from now on I denote as $S_{Y|X}$. Under mild conditions [33, 34] $S_{Y|X}$ exists and is a dimension reduction space (for more details see Section 6.2.4).

Sufficient Dimension Reduction in Inverse Linear Regression

We now introduce two algorithms, Sliced Inverse Regression (SIR) [116] and Localized Sliced Inverse Regression (LSIR) [200], for estimating the the sufficient dimension reductions $R(X)$. Both algorithms fit within the inverse regression framework and are based on a semi-parametric statistical model

$$Y | X = f(G^T X, \varepsilon), \quad G = (g_1, \dots, g_d) \in \mathbb{R}^{p \times d}, \quad \mathbb{E}[\varepsilon | X] = 0, \quad (6.10)$$

where f is a density function and g_i are the d orthonormal bases of the dimension reduction subspace $S_{Y|X}$ (the column space of $G = S_{Y|X}$) that contains the predictive information on Y . The new d variates $Z = G^T X$, can replace X without any loss of information about the regression as $Y \perp\!\!\!\perp X | Z$. Notice, an important distinction from PCA is that the objective explicitly includes the response Y as part of the optimality criterion.

Sliced Inverse Regression Sliced Inverse Regression (SIR) is a dimension reduction approach introduced by [116] which is based on the semi-parametric model in (6.10). The statistical quantities underlying SIR are the inverse regression function, the covariance of the inverse regression function, and the marginal covariance of the covariates

$$\eta_Y = \mathbb{E}_X[X | Y], \quad \Gamma = \text{cov}(\eta_Y), \quad \Sigma = \text{cov}(X)$$

with the central statistical assumption that

$$\text{span}(\mathbb{E}_X[X | Y] - \mathbb{E}_X[X]) \in \text{span}(\Sigma G) \quad (6.11)$$

where $\text{span}(G) = S_{Y|X}$ is the dimension reduction subspace. One setting where the above assumption is satisfied is that $\mathbb{E}[X | \eta_Y^T X]$ is a linear function of X , the *linearity condition* [32]. The basis G for the dimension reduction subspace can be computed from the following generalized eigendecomposition problem

$$\Gamma g_i = \lambda_i \Sigma g_i, \quad \text{for } i = 1, \dots, d \quad (6.12)$$

and $G = \{(g_1, \dots, g_d) | \lambda_1, \dots, \lambda_d > 0\}$, the eigenvectors corresponding to non-zero eigenvalues. Given observations $\{(x_i, y_i)\}_{i=1}^n$ an empirical estimate \hat{G} is computed via the following algorithm [116]:

- (1) Compute an empirical estimate of Σ ,

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})(x_i - \hat{\mu})^T$$

where $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i$ is the sample mean.

- (2) Divide the samples into H groups (or slices) S_1, \dots, S_H according to the value of y . Compute an empirical estimate of Γ ,

$$\hat{\Gamma} = \sum_{h=1}^H \frac{n_h}{n} (\hat{\mu}_h - \hat{\mu})(\hat{\mu}_h - \hat{\mu})^T, \quad \hat{\mu}_h = \frac{1}{n_h} \sum_{j \in S_h} x_j$$

with $\hat{\mu}_h$ as the sample mean for group h and n_h the group size.

- (3) Estimate the d.r. directions β by solving the generalized eigendecomposition problem

$$\hat{\Gamma} \hat{g}_i = \hat{\lambda}_i \hat{\Sigma} \hat{g}_i. \quad (6.13)$$

- (4) Set $\hat{G} = \{(\hat{g}_1, \dots, \hat{g}_d) | \hat{\lambda}_1, \dots, \hat{\lambda}_d > \delta\}$ where $\delta > 0$ is a small threshold.

Localized Sliced Inverse Regression Problems with SIR arise when the statistical assumption in (6.11) is not satisfied, that is the predictive structure in the data is nonlinear. This can be

due to the data being concentrated near a manifold or there being clusters in the data. SIR can be adapted to address this setting by taking into account local structure of the explanatory variables conditioned on the response variable. The idea behind Localized Sliced Inverse Regression (LSIR) [200] is based on the observation in manifold learning that Euclidean structure around a data point in \mathbb{R}^p is only useful locally. Therefore, when observations in a slice are far apart in the ambient space, computing a global average μ_h for a slice is not meaningful. Instead it is more appropriate to consider local averages. The one difference between SIR and LSIR is how the empirical estimate of Γ is computed. The following is the algorithm for LSIR:

- (1) Compute $\hat{\Sigma}$ as in SIR.
- (2) Slice the samples into H groups as in SIR. For each sample (x_i, y_i) compute

$$\hat{\mu}_{i,\text{loc}} = \frac{1}{k} \sum_{j \in s_i} x_j,$$

where

$$s_i = \{j : x_j \text{ belongs to the } k\text{-nearest neighbors of } x_i \text{ in } S_h\},$$

and h indexes the group S_h to which i belongs. Then a localized version of Γ is computed

$$\hat{\Gamma}_{\text{loc}} = \frac{1}{n} \sum_{i=1}^n (\hat{\mu}_{i,\text{loc}} - \hat{\mu})(\hat{\mu}_{i,\text{loc}} - \hat{\mu})^T.$$

- (3) Solve the generalized eigendecomposition problem

$$\hat{\Gamma}_{\text{loc}} \hat{g}_i = \hat{\lambda}_i \hat{\Sigma} \hat{g}_i. \tag{6.14}$$

Exact Solutions

For both SIR and LSIR the most straightforward solution to the generalized eigendecomposition in equations (6.13) and (6.14) is based on the Cholesky decomposition of $\hat{\Sigma}$ and typically has runtime complexity $O(np^2)$. We provide a procedure based on the SVD of the data matrix X to provide

exact solutions to equations (6.13) and (6.14). One motivation for this procedure is that extensions to very large data sets are feasible by randomized algorithms.

The algorithm consists of a few steps:

- (1) Sort the samples in decreasing value of the response, partitioning them into H (approximately) equally-sized slices. Denote the number of samples in each slice i to be n_i .
- (2a) For SIR construct the following matrix

$$J_{\text{sir}} = \begin{pmatrix} J_1 & & 0 \\ & \ddots & \\ 0 & & J_H \end{pmatrix}_{n \times n}, \quad J_i = \begin{pmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \dots & 1 \end{pmatrix}_{n_i \times n_i},$$

and

$$\hat{\Gamma} = X^T J_{\text{sir}}^T J_{\text{sir}} X.$$

- (2b) For LSIR construct the following matrix

$$J_{(\text{lsir})} = \begin{pmatrix} J_1 & & 0 \\ & \ddots & \\ 0 & & J_H \end{pmatrix}_{n \times n}, \quad J_i = \frac{1}{\sqrt{k}} \begin{pmatrix} 1 & \delta_{1 \sim 2} & \dots & \delta_{1 \sim n_i} \\ \vdots & \ddots & \ddots & \vdots \\ \delta_{n_i \sim 1} & \dots & \delta_{n_i \sim n_i - 1} & 1 \end{pmatrix}_{n_i \times n_i},$$

where k is the number of nearest neighbors and in each block $\delta_{r \sim s} = 1$ if observation r is among the k -nearest neighbors of s and is 0 otherwise. The following equivalence holds

$$\hat{\Gamma}_{\text{loc}} = X^T J_{(\text{lsir})}^T J_{(\text{lsir})} X.$$

The above k-NN computation requires distances between points. We use the SVD of the data matrix, $X = USV^T$, to compute these distances by computing the distances in the eigenvector coordinates of the marginal covariance. Each sample in this coordinate system is a row in the matrix $\tilde{X} = XV = US$ and the distance between the i -th and j -th sample is $\|\tilde{X}_i - \tilde{X}_j\|_2$ where \tilde{X}_i and \tilde{X}_j correspond to the i -th and j -th rows.

(3) The solution for (L)SIR based on the SVD decomposition of X reduces to

$$\begin{aligned}
\hat{\Gamma}_{(\text{loc})} e &= \lambda X^T X e \\
X^T J_{(1)\text{sir}}^T J_{(1)\text{sir}} X e &= \lambda X^T X e \\
V(US)^T J_{(1)\text{sir}}^T J_{(1)\text{sir}} (US) V^T e &= \lambda V(US)^T (US) V^T e \\
V \tilde{X}^T J_{(1)\text{sir}}^T J_{(1)\text{sir}} \tilde{X} V^T e &= \lambda V \tilde{X}^T \tilde{X} V^T e \\
\Rightarrow \tilde{X}^T J_{(1)\text{sir}}^T J_{(1)\text{sir}} \tilde{X} f &= \lambda \tilde{X}^T \tilde{X} f
\end{aligned}$$

where $f \equiv V^T e$. When $n < p$ the system $V^T e = f$ is underdetermined so we select the solution with minimal norm, which is provided by the set $\{e_j = V f_j, j = 1, \dots, n-1\}$, ordered according to the corresponding generalized eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{n-1}$.

Supervised Locality Preserving Projections LPP can be adapted to be a supervised dimension reduction methodology by replacing $J_{(1)\text{sir}}$ with J_{lpp} in equation (6.15). This results in re-defining the notion of “local neighborhoods” to only the neighbors in covariate space that have similar responses (i.e. fall within the same “slice”). A similar approach was proposed in [81], in which the authors suggest the use of a block diagonal structure for the adjacency matrix based on provided class labels in the context of classification.

Approximate Randomized Solutions

The exact solution presented in the previous section takes advantage of the matrix arguments in the generalized eigendecomposition problems defined in equations (6.13) and (6.14). The exact solution is also based on the SVD of X . The appeal of this formulation in extensions to very large data is the exact SVD of X can be replaced with the randomized algorithm described in 6.2.2 to estimate the top r eigenvectors and eigenvalues of $\hat{\Sigma}$ where r is large enough so that the span of the eigenvectors contains the effective dimension reduction subspace. The regularization parameter t in the randomized algorithm is assumed fixed.

The randomized solution follows the same steps described in Section 6.2.4, replacing the exact

SVD factorization of the data matrix with the approximate solution provided by Randomized SVD.

In more detail, the Randomized (L)SIR steps are as follows,

(1a) Estimate Randomized SVD as described in Section 6.2.2 to produce a rank- r approximation to the data matrix:

$$X \approx \tilde{U} \tilde{S} \tilde{V}^T.$$

(1b) Construct derived variates

$$\tilde{X} = X \tilde{V} \quad (n \times r).$$

(2a) For SIR construct the J_{sir} matrix

(2b) For LSIR construct the J_{lsir} matrix,

where the k -NN matrix (k fixed) is constructed using the derived variates \tilde{X} . The distances between points are computed using the approximate eigenvector coordinates of the marginal covariance.

Each sample in this coordinate system is a row in the matrix $\tilde{X} = \tilde{U} \tilde{S}$ and the distance between the i -th and j -th sample is $\|\tilde{X}_i - \tilde{X}_j\|_2$ where \tilde{X}_i and \tilde{X}_j correspond to the i -th and j -th rows.

(3) The solution for (L)SIR based on the Randomized SVD decomposition of X reduces to

$$\tilde{X}^T J_{(l)\text{sir}}^T J_{(l)\text{sir}} \tilde{X} f = \lambda \tilde{X}^T \tilde{X} f,$$

where $f \equiv V^T e$. The system $V^T e = f$ is underdetermined as $V \in \mathbb{R}^{p \times r}$ and $r \ll p$, but we work in basis spanned by the estimated top r eigenvectors of the marginal covariance matrix, hence the unique solution that falls in that subspace is provided by the set $\{e_j = V f_j, j = 1, \dots, n-1\}$, ordered according to the corresponding generalized eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{n-1}$.

Runtime Analysis The main runtime bottleneck in Randomized (L)SIR is calculating the Randomized SVD estimates which are used to project the data. Following that step is the solution to a generalized eigendecomposition problem of dimension r , which takes $O(r^3)$ time. As $r \ll n < p$ this is computationally feasible for large data sets. Hence, the overall asymptotic runtime is the same as that for Randomized SVD: $O(t \ln p + t^2 l^2 n)$ (see 6.2.4). When c -fold cross-validation is used to estimate an optimal value for t the runtime becomes $O(\sum_{t=1}^{t_{\text{max}}} c \times [t l (n-a) p + t^2 l^2 (n-a)])$ (see 6.2.2).

Inference Properties

In this section I describe some known population results in the classic case of SIR and a closely related likelihood based formulation for sufficient dimension reduction.

Model-free Framework for Inverse Regression The population quantity of interest in inverse regression approaches to dimension reduction is the central subspace $S_{Y|X} = \text{span}(g_1, \dots, g_d)$, as any orthonormal basis for $S_{Y|X}$ can be used to construct the orthogonal projection matrix $P_{S_{Y|X}}$. An actual model linking the linearly transformed covariates to the response need not be specified for the estimation of g_1, \dots, g_d .

If the additional *linearity assumption* is satisfied

$$\mathbb{E}[X | P_{S_{Y|X}}] = P_{S_{X|Y}}X, \quad (6.15)$$

then the centered inverse regression is confined to $S_{Y|X}$, $\text{span}(\mathbb{E}_X[X | Y] - \mathbb{E}_X[X]) \subseteq S_{Y|X}$ and $S_{\mathbb{E}_X[Y|X]} \subseteq S_{Y|X}$. An interesting observation was that as $n, p \rightarrow \infty$ linear combinations of the covariates are approximately normally distributed [76], this ensures that the *linearity condition* is asymptotically satisfied. The practical implication of this result is yet to be clearly understood. If, in addition, we assume that

$$Y \perp\!\!\!\perp X | P_{S_{\mathbb{E}[X|Y]}}X, \quad (6.16)$$

then it was shown that $S_{\mathbb{E}[Y|X]} = S_{Y|X}$ [26]. If condition (6.16) is not satisfied then [116] and [36] suggested that we can consider higher conditional moments of $X | Y$ or consider the "central kth-moment subspace"(CKMS) [203]. The localization in LSIR also addresses this situation.

Model-based Framework for Inverse Regression Principal Fitted Components (PFC) is a general parametric framework for inverse regression proposed in [35] and extended in [31, 3], which is based on a multivariate normal model for the conditional distribution of $X|Y$, allowing for a variety of covariance structures. Let $X_y \in \mathbb{R}^p$ denote the random vector with conditional distribution $X|(Y = y)$, $E[Y] = \mu_y$, and $E[X] = \mu$. In addition let $\Gamma \in \mathbb{R}^{p \times d}$ be a basis for

the span($\mu_y - \mu$). Then, assuming a normal conditional model, the authors in [35] postulate the following model

$$X_y = \mu + \Gamma \nu_y + \sigma \epsilon, \quad (6.17)$$

where $\epsilon \sim N(0, \Delta)$, $\Delta \in \mathbb{S}^{p \times p}$ and ϵ is independent of Y . The response Y is incorporated as part of the conditional mean through $\nu_y \in \mathbb{R}^d$. In particular, $\nu_y = (\Gamma^T \Gamma)^{-1} \Gamma^T (\mu_y - \mu) = \Gamma^T (\mu_y - \mu)$, provides the coordinates for the centered conditional mean vector $\mu_y - \mu$ in the basis defined by the semi-orthogonal matrix Γ . Under this model $R(X) = \Gamma^T \Delta^{-1} X$ is a minimal sufficient reduction and hence the central subspace is characterized by $S_{Y|X} = \Delta^{-1} \Gamma = \{\Delta^{-1} z : z \in S_\Gamma\}$. This is the same target *population* quantity estimated by the SIR algorithm, which implies that in the class of *normal inverse* models (6.17), *moment-free* (SIR) and *model-based* (PFC) inverse dimension reduction coincide in the population. The above parametric framework was further extended in a Bayesian setting by [126].

In the special case of *isotropic* conditional covariance $\Delta = \sigma^2 I$, the subspace for the sufficient dimension reduction is the span of the top eigenvectors of the marginal covariance $\Sigma = Cov(X)$, which are estimated, using maximum likelihood, by the top eigenvectors of the sample covariance matrix. This provides a parametric model justifying the use of Principal Components as a dimension reduction approach for regression, that assumes a very general mean structure and quite restrictive conditional covariance structure.

Bounds on Estimators Similar to the case of PCA we can bound the deviation between the estimated dimension reduction subspace and the population dimension subspace for SIR. Theorem 6.2.3 can be applied to both the sample covariance matrix as well as the estimate of the covariance of the inverse regression resulting in a bound of the form

$$\|\hat{B} - B\| \leq \frac{\kappa}{\sqrt{n}},$$

where κ depends on the moments of X and $X | Y = y$. In this case the upper bound also turns out to depend on M_4 .

We do not yet have an analysis for the case of LSIR or supervised LPP.

6.3 Results

In this section I describe the application of the algorithms described in Section 6.2.2 and Section 6.2.3 in the context of a simulated Wishart and low-rank covariance structure and a SNP microarray experimental data from Wellcome Trust Case Control Consortium (WTCCC) [1] and a genotyping study of European Populations from POPRES [138] reported in [146].

I first investigate the *runtime* properties of the randomized procedures as compared to the exact sample estimates, followed up a study of the effect of the regularization parameter on the estimation accuracy of the dimension reduction subspace.

6.3.1 Simulation

The following section contains results from empirical runs of the iterative approximation scheme with comparisons to an exact SVD approach implemented as part of the LAPACK package (DGESVD routine). The agreement between the estimates of the subspace spanned by the top eigenvectors and the true subspace is measured using trace correlation [201, 206] as described in Section 6.2.1.

Random Wishart Covariance Structure

The input data was a randomly generated rectangular matrix of dimension $m \times n$ with independent $N(0, 1)$ entries. Our main goal in this section is to examine the ability of the iterative scheme to produce fast and accurate approximation to the span of the top few eigenvectors of the sample covariance matrix. Notice that this is the worst-case scenario for the Randomized algorithm whereby all population eigenvalues are the same i.e. no eigenvalue decay. Hence, the observed performance gains over the exact method can be expected to be much more substantial in similar size data set that contains *low rank structure*, and small values for the parameter t would be sufficient to achieve excellent approximation accuracy (see Section 6.3.1). We are going to illustrate the relative improvement in terms of runtime over the exact SVD method in two distinct dimension size regimes.

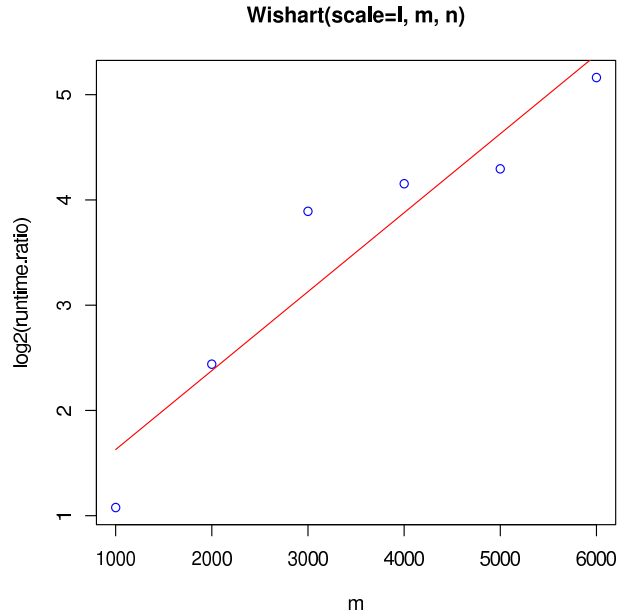


Figure 6.1: Similar matrix dimensions result in increased runtime gains over exact PCA

| m | runtime(exact) sec(s) | $\frac{\text{exact}}{\text{approx}}$ | $\log_2 \frac{\text{exact}}{\text{approx}}$ |
|------|-----------------------|--------------------------------------|---|
| 1000 | 38 | 2.11 | 1.08 |
| 2000 | 217 | 5.43 | 2.44 |
| 3000 | 1172 | 14.84 | 3.89 |
| 4000 | 1887 | 17.80 | 4.15 |
| 5000 | 2828 | 19.64 | 4.30 |
| 6000 | 6983 | 35.81 | 5.16 |

Table 6.1: n=9000, varying m

Decreasing Difference Between the Data Dimensions Figure 6.1 (based on Table 6.1) illustrates the runtime behavior of the iterative scheme relative to the exact SVD as the *difference between the data dimensions decreases*. In this scenario Randomized PCA show exponential runtime gains. This is expected to be the regime of greatest interest, in which an abundance of high dimensional samples are to be analyzed in the context of very high-dimensional genotype data for each individual sample.

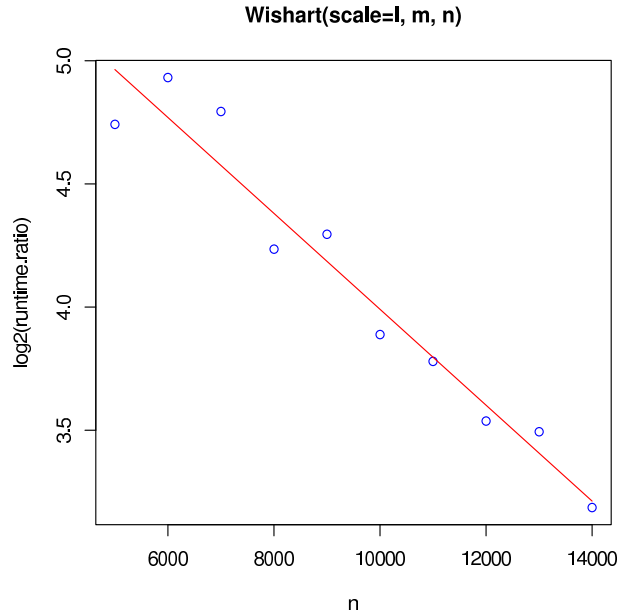


Figure 6.2: Different matrix dimensions result in decreased runtime gains over exact PCA

| n | runtime(exact) sec(s) | $\frac{\text{exact}}{\text{approx}}$ | $\log_2 \frac{\text{exact}}{\text{approx}}$ |
|-------|-----------------------|--------------------------------------|---|
| 5000 | 1873 | 26.76 | 4.74 |
| 6000 | 2533 | 30.52 | 4.93 |
| 7000 | 3134 | 27.73 | 4.79 |
| 8000 | 2637 | 18.84 | 4.24 |
| 9000 | 2828 | 19.64 | 4.30 |
| 10000 | 2873 | 14.81 | 3.89 |
| 11000 | 2966 | 13.73 | 3.78 |
| 12000 | 3088 | 11.61 | 3.54 |
| 13000 | 3222 | 11.27 | 3.49 |
| 14000 | 3268 | 9.10 | 3.19 |

Table 6.2: m=5000, varying n

Increasing Difference Between the Data Dimensions In many high dimensional datasets the number of samples is much smaller than the number of measured features, as it is often infeasible to collect a large amount of data. Figure 6.2 (based on Table 6.2) illustrates how Randomized

PCA runtime changes as we approach such a scenario in which the *difference between the data dimensions increases*. In this case the relative runtime gains of Randomized PCA diminish exponentially fast. Hence, unless runtime gain of a factor of 2 or 3 is essential an exact SVD approach would be recommended.

Low Rank Covariance Structure

When the matrix is of low rank, usually a single iteration of the approximate scheme produces a very accurate approximation to the top eigenvectors and eigenvalues. We expect this to be (approximately) the case in many real-world datasets as illustrated in the next section with large genetic data sets. Figure 6.3 illustrates this behavior of Randomized PCA on a spiked covariance matrix in which top 5 eigenvalues contain signal and all other contain only noise. The simulation experiments were replicated 1000 times.

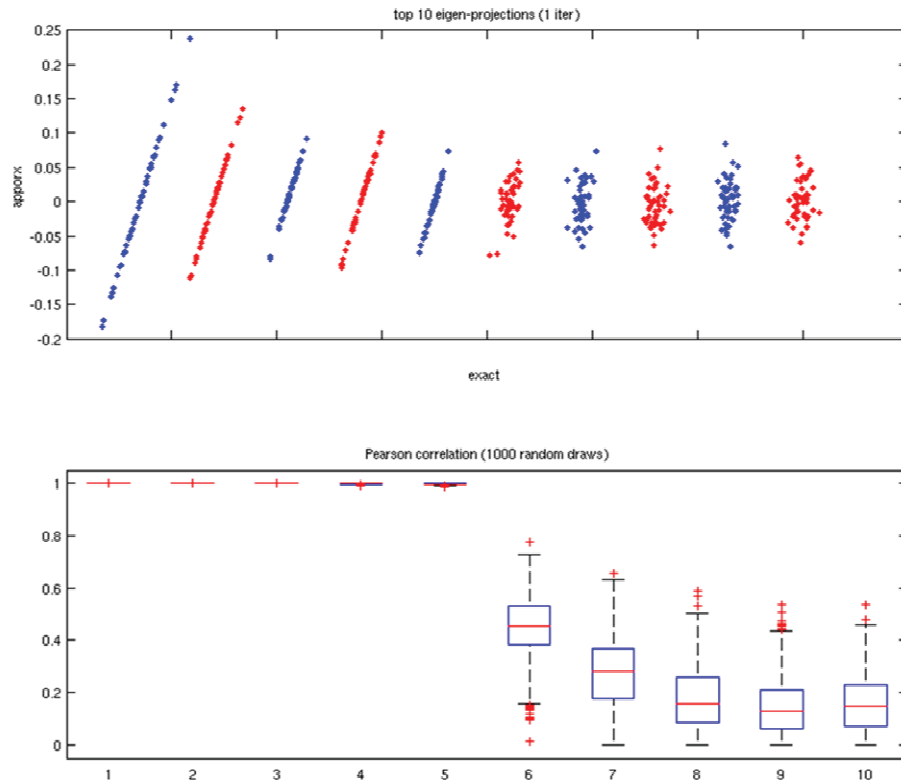


Figure 6.3: Application of Randomized PCA to a spiked Wishart covariance structure

6.3.2 SNP Data

In this section I apply Randomized PCA to some moderate size genetic data sets, evaluating its ability to provide a good approximation to the dimension reduction subspace spanned by the top few eigenvectors as compared to the exact methods.

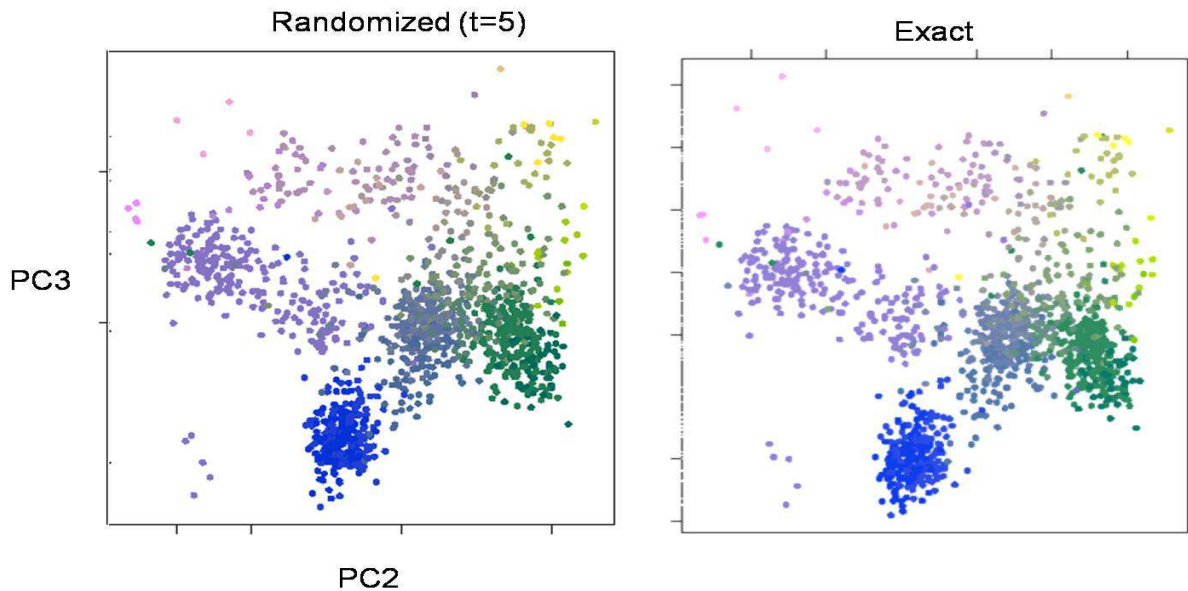


Figure 6.4: Estimated axes of variation (PC2, PC3) based on data from Novembre *et al.* [146].

European Data Set Figure 6.4 illustrates the ability of the Randomized SVD to recover a good approximation to the top 4 PCs in [146]. The genotype data is collected from European individuals as part of the POPRES [138] project. The DNA of a total of $\sim 3,000$ European individuals was assayed at $\sim 500,000$ locations. After data pre-processing 1,387 individual samples and $\sim 200,000$ SNPs were used in the downstream analysis. The different colors on Figure 6.4 represent different geographic origin of the individuals projected onto the 2nd and 3th principal component (PC). Good separation is evident among the geographic subpopulation. The value of t used was 5 and it was sufficient to produce empirical correlations of the exact sample PCs with the approximate estimates as follows (starting from PC2): 0.999, 0.996, 0.930, 0.523, etc.

WTCCC Crohn’s Disease Genotype Data In this section I compare the performance of Randomized SVD to the exact SVD on a subset of the Crohn’s disease SNP data set from the Wellcome Trust Case Control Consortium [1]. $\sim 5,000$ individuals were assayed at $\sim 500,000$ locations. After data pre-processing $\sim 5,000$ individual samples and $\sim 200,000$ SNPs were used in the downstream analysis. Table 6.3 shows results from running Randomized SVD on a subset of the data from chr 17,19,21, and 22.

To study the properties of the Randomized SVD approach I used two different stopping rules. Table 6.3 summarizes the results from a single iteration of Randomized SVD while Table 6.4 contains the results from running the iterative scheme until high subspace accuracy has been achieved (distance of 1.0×10^{-3} or less has been achieved, that typically corresponds to 2-3 d.p. accuracy of the eigenvalue estimates for the sample covariance matrix). The dimension of the target subspace was fixed to be 10. In both scenarios the results illustrate that Randomized SVD provides a highly scalable approach to inference of the axes of genetic variation that correspond to the top few Principal Components of the sample covariance matrix of the SNP data matrix.

| data | runtime(exact) | runtime(approx) | subspace dist. |
|-----------------------------------|---------------------|-----------------|----------------|
| 4,686×6,041 (chr 19) | 37 min(s) 34 sec(s) | 6 sec(s) | 0.01 |
| 4,686×11,943 (chr 19, 22) | 48 min(s) 1 sec(s) | 13 sec(s) | 0.05 |
| 4,686×18,715 (chr 19, 21, 22) | 72 min(s) 8 sec(s) | 38 sec(s) | 0.11 |
| 4,686×29,406 (chr 17, 19, 21, 22) | 72 min(s) 29 sec(s) | 31 sec(s) | 0.09 |

Table 6.3: Crohns Disease; 200K SNP array; 5,000 individuals. Single iteration of Randomized SVD

| data | runtime(exact) | $\frac{\text{exact}}{\text{approx}}$ | $\log_2 \frac{\text{exact}}{\text{approx}}$ | subspace dist. |
|--------------------------------|---------------------|--------------------------------------|---|----------------------|
| 4,686×6,041 (chr 19) | 34 min(s) 40 sec(s) | 189.09 | 7.56 | 3.9×10^{-7} |
| 4,686×11,943 (chr 19,22) | 35 min(s) 46 sec(s) | 79.48 | 6.31 | 3.3×10^{-7} |
| 4,686×18,715 (chr 19,21,22) | 45 min(s) 29 sec(s) | 62.02 | 5.95 | 1.9×10^{-7} |
| 4,686×29,406 (chr 17,19,21,22) | 61 min(s) 34 sec(s) | 53.54 | 5.74 | 3.2×10^{-7} |

Table 6.4: Required minimum distance to the space spanned by the exact top eigenvectors:
 1.0×10^{-3}

Chapter 7

Conclusions

7.1 Direct Evidence of Regulation Improves *De Novo* Motif Discovery Predictions

In this section I briefly summarize some of the general conclusions and propose future directions based on the proposed approach for leveraging high-throughput binding evidence to achieve improved models for the binding preferences of RNA and DNA binding factors as well as their functional occurrences in the genome.

7.1.1 Genome-wide Binding Evidence

Motif finding with an objective function based on genome-wide evidence of regulation provides a flexible and successful framework to integrate sequence data with high-throughput binding or expression information. In particular, I presented a flexible and successful pipeline to analyze regulatory information resulting from applications of next-generation sequencing technology. I have also demonstrated the usefulness to integrate high-level information on the genome wide set of regulatory regions (such as defined by DNaseI hypersensitive sites), with quantitative data on the genome-wide affinity of individual regulatory factors.

For the case of TF binding site discovery in ChIP-chip data cERMIT was compared to existing state-of-the art approaches and showed very competitive performance. In the case of Chip-seq the method successfully recovered the known binding preferences of the factor under study, with much improved quality possibly due to the larger amount of binding data as well as the much higher resolution provided by the sequencing-based experimental assays.

Binding Site Representation

Using IUPAC consensus motifs as described in Section 3.6 occasionally results in underestimating the motif degeneracy, as I build a PSSM description based on the consensus sequence in the predicted set of bound genes. Together with the objective function currently used by cERMIT described in Section and 3.4.2, targeting oligomer most strongly associated with the evidence provided, this means that reported motif predictions should not be considered as quantitative models of actual binding affinity, but rather as the core of a functional motif. In addition, similarly to others before us [41], the motif search implemented by cERMIT is based on the assumption that the experimental setup ensures sufficient concentration of the factor in order for it not to be a limiting step in the sequence binding reactions. This allows us currently to approximate the inherently stochastic DNA-TF interaction by modeling it as a binary event.

7.1.2 Transcriptome-wide Binding Evidence

The different, and in many cases unknown, cross-linking properties for RBPs presents a challenge for all CLIP protocols, and requires small adjustments as to how to call and expand read clusters to ensure the inclusion of the binding site to be identified in the downstream motif analysis. In instances of newly studied proteins, for which the motif or conversion pattern are not known, e.g. the recently analyzed HuR protein [134], it is thus best to use PARalyzer with the extend-by-read option in combination with the output of motif finding to determine if significant top-scoring motifs tend to have specific locations of high conversion. If it is the case, as it is e.g. for PUM2, that there is at least one location of high conversion, then a tighter cluster extension can be used to reduce the size of the interaction map.

In addition to the RBP-specific sequence affinity preferences, the RBP-RNA interaction has been shown to be influenced by the secondary structure of the targeted RNA sequence and has been successfully exploited in previous work on RBP motif discovery [83, 117, 93]. Incorporating information on the RBP structural preferences into the motif analysis proposed in the current work could be implemented by means of a prior distribution on the binding evidence for individual sequence

regions inferred by PARalyzer, biasing the motif discovery towards high-scoring sequence patterns that contain favorable sequence context for RBP binding. This could help filter out non-specific interactions with highly abundant mRNAs. In the context of AGO-mediated regulation, a prior based on the predicted microRNA-mRNA duplex stability could be used in a similar fashion.

Due to the use of 4SU nucleoside analogue in the original PAR-CLIP protocol, the U content of an actual binding site and its vicinity will obviously impact the identification of RBP binding sites. If a recognition site does not contain any uridines, precise delineation using this approach is compromised; on the other hand, many U residues may either cause problems with alignment due to the potential of many mismatches, and/or to spread out the signal over multiple positions. The current investigations of additional amenable photoactivatable nucleosides [106], complemented by the use of different digestion enzymes [101], are expected to reduce potential biases, and can easily be specified in PARalyzer. As such, the RBP motif analysis pipeline provides a standardized solution for the analysis of RBP binding sites via PAR-CLIP, for subsequent motif finding for sequence-specific RBPs, and for the elucidation of post-transcriptional regulatory mechanisms and networks.

7.1.3 Future Directions

The presented motif discovery approach in high-throughput binding data from Section 3.1 provides a useful set of ideas but there is scope for improvement, both in terms of a more flexible motif description as well as on the motif search strategy, by means of a stochastic search in place of the greedy approach. I expect that this will allow us to pick up more degenerate signals and to provide more quantitative models of the recovered functional sites. Instead of merely defining genome partitions by the presence of motifs, a probabilistic framework based on a joint likelihood as well as a formal model of uncertainty would also allow for the simultaneous inference of a motif model and the most probable set of target genes. Different types of sampling moves can be also incorporated that will enhance cERMIT's ability to explore the motif search space. This may allow us to better capture motifs with two half-sites separated by highly degenerate spacer regions, or combinations

of two or more motifs. To that end partitions can be defined to be based on high scoring motifs that co-occur in regulatory sequences. Ultimately, the main goal in motif discovery is to approach the harder problem of detecting combinatorial interactions of different factors that distinguish between biological states, be it between different tissues, specific developmental stages, or normal vs. cancer conditions.

7.2 Extensions to the Framework for Inference of Population Structure

In this section I briefly outline some possible extensions that would allow to take full advantage of the proposed computational framework for approximate estimation of the optimal dimension reduction subspace relevant to large genetic data sets.

7.2.1 Statistical Inference of the Dimensionality of the Population Structure

Randomized SVD was demonstrated to produce highly accurate estimates of the subspace spanned by the top few eigenvectors of the sample covariance estimates. This subspace could be very useful for visual inspection of the projections onto the top 2-3 eigenvectors to identify natural groups and individual patterns. In addition, random matrix theory has been successfully adapted to the problem of formal inference of population structure using the full SVD of the (normalized) genotype matrix [149]. Even though, typically, the top few eigenvectors capture the population structure present in the data, testing exactly how many eigenvectors are needed currently relies on estimates of all sample eigenvalues. Hence, new statistical theory needs to be developed to address the setting of large number of genotypes (SNPs) as well as large number of genotyped individuals (samples) where estimation of the full SVD is not feasible.

One promising direction to explore would be to take advantage of the knowledge that in structure in the genetic data from real world populations tends to be of small dimension, hence most of the population eigenvalues of the covariance matrix should correspond to noise directions and are

approximately equal. Combining this idea with accurate estimates of the top few sample eigenvalues e.g. using Randomized SVD, could provide the necessary information to construct a suitable statistic to test for the dimensionality of the structure present in the population under study.

7.2.2 Theoretical Results for Localized Sliced Inverse Regression

For the case of PCA and SIR Theorem 6.2.2 and Theorem 6.2.3 provide a bound on the deviation between the estimated dimension reduction subspace and the population dimension subspace was derived. It would be desirable to prove a similar result for Localized Sliced Inverse Regression (LSIR) and supervised LPP. This is problematic as the population object itself that LSIR estimates is difficult to define and analyze.

Appendix A

Website With Software and Data Sets

A.1 TF Binding Motif Discovery

<http://tools.genome.duke.edu/generegulation/transcription/cERMIT/>

A.2 PAR-CLIP Motif Analysis

<http://www.genome.duke.edu/labs/ohler/research/mEAT/index.php>

A.3 Randomized Eigendecomposition Analysis

<http://stat.duke.edu/sayan/eigen.htm>

Bibliography

- [1] Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, 2007.
- [2] Dimitris Achlioptas. Database-friendly random projections. In *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, PODS '01, pages 274–281, New York, NY, USA, 2001. ACM.
- [3] Kofi P. Adragani and R. Dennis Cook. Sufficient dimension reduction and prediction in regression. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906):4385–4405, 2009.
- [4] E. Anderson, Z. Bai, J. Dongarra, A. Greenbaum, A. McKenney, J. Du Croz, S. Hammerling, J. Demmel, C. Bischof, and D. Sorensen. Lapack: a portable linear algebra library for high-performance computers. In *Proceedings of the 1990 ACM/IEEE conference on Supercomputing*, Supercomputing '90, pages 2–11, Los Alamitos, CA, USA, 1990. IEEE Computer Society Press.
- [5] Daehyun Baek, Judit Villn, Chanseok Shin, Fernando D. Camargo, Steven P. Gygi, and David P. Bartel. The impact of micrnas on protein output. *Nature*, 455(7209):64–71, 2008.
- [6] Timothy L. Bailey and Charles Elkan. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. In *Machine Learning*, pages 51–80, 1995.
- [7] Artem Barski, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Dustin E. Schones, Zhibin Wang, Gang Wei, Iouri Chepelev, and Keji Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823 – 837, 2007.
- [8] David P. Bartel. Micrnas: Genomics, biogenesis, mechanism, and function. *Cell*, 116(2):281 – 297, 2004.
- [9] David P. Bartel. Micrnas: Target recognition and regulatory functions. *Cell*, 136(2):215 – 233, 2009.
- [10] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.
- [11] Mikhail Belkin and Partha Niyogi. Towards a theoretical foundation for laplacian-based manifold methods. 2005.
- [12] Adam C. Bell, Adam G. West, and Gary Felsenfeld. The Protein CTCF Is Required for the Enhancer Blocking Activity of Vertebrate Insulators. *Cell*, 98(3):387–396, August 1999.
- [13] Michael F Berger, Anthony A Philippakis, Aaron M Qureshi, Fangxue S He, Preston W Estep, and Martha L Bulyk. Compact, universal dna microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotech*, 24(11):1429–1435, 2006.

- [14] Mathieu Blanchette, Benno Schwikowski, and Martin Tompa. Algorithms for phylogenetic footprinting. *Journal of Computational Biology*, 9(2):211–223, 2002.
- [15] Mathieu Blanchette and Martin Tompa. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Research*, 12(5):739–748, 2002.
- [16] Juan S. Bonifacino, Esteban C. Dell’Angelica, and Timothy A. Springer. *Immunoprecipitation*. John Wiley & Sons, Inc., 2001.
- [17] Christos Boutsidis, Michael W. Mahoney, and Petros Drineas. An improved approximation algorithm for the column subset selection problem. In *Proceedings of the twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA ’09, pages 968–977. Society for Industrial and Applied Mathematics, 2009.
- [18] Paul L. Boutz, Peter Stoilov, Qin Li, Chia-Ho Lin, Geetanjali Chawla, Kristin Ostrow, Lily Shiue, Manuel Ares, and Douglas L. Black. A post-transcriptional regulatory switch in polypyrimidine tract-binding proteins reprograms alternative splicing in developing neurons. *Genes & Development*, 21(13):1636–1652, 2007.
- [19] Alan P. Boyle, Sean Davis, Hennady P. Shulha, Paul Meltzer, Elliott H. Margulies, Zhiping Weng, Terrence S. Furey, and Gregory E. Crawford. High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 132(2):311 – 322, 2008.
- [20] Alan P. Boyle, Justin Guinney, Gregory E. Crawford, and Terrence S. Furey. F-seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics*, 24(21):2537–2538, 2008.
- [21] M. Buck and J. Lieb. ChIP-chip: Considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, 83:349–360, 2004.
- [22] Natascha Bushati and Stephen M. Cohen. microRNA functions. *Annual review of cell and developmental biology*, 23(1):175–205, May 2007.
- [23] Harmen J. Bussemaker, Barrett C. Foat, and Lucas D. Ward. Predictive modeling of genome-wide mrna expression: From modules to molecules. *Annual Review of Biophysics and Biomolecular Structure*, 36(1):329–347, 2007.
- [24] Harmen J. Bussemaker, Hao Li, and Eric D. Siggia. Regulatory element detection using correlation with expression. *Nat Genet*, 27(2):167–174.
- [25] Xi Chen, Han Xu, Ping Yuan, Huss Mikael Fang, Fang, Vinsensius B. Vega, Eleanor Wong, Weiwei Orlov, Yuriy L. Zhang, Jianming Jiang, Yuin-Han Loh, Hock Chuan Yeo, Zhen Xuan Yeo, Vipin Narang, Kunde Ramamoorthy Govindarajan, Bernard Leong, Atif Shahab, Yijun Ruan, Guillaume Bourque, Wing-Kin Sung, Neil D. Clarke, Chia-Lin Wei, and Huck-Hui Ng. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, 133(6):1106–1117, 2008.
- [26] Francesca Chiaromonte, R. Dennis Cook, and Bing Li. Sufficient dimension reduction in regressions with categorical predictors. *The Annals of Statistics*, 30(2):475–497.

- [27] F R K Chung. *Spectral Graph Theory*, volume 92. American Mathematical Society, 1997.
- [28] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences of the United States of America*, 102(21):7426–7431, 2005.
- [29] ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146):799–816, June 2007.
- [30] Human Genome Sequencing Consortium International. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, 2004.
- [31] R. D. Cook and L. Forzani. Principal fitted components for dimension reduction in regression. *Statistical Science*, 23:485–501, 2008.
- [32] R. Dennis Cook and Liqiang Ni. Sufficient Dimension Reduction via Inverse Regression: A Minimum Discrepancy Approach. *Journal of the American Statistical Association*, 100(470):410–428, June 2005.
- [33] R.D. Cook. Graphics for regressions with a binary response. *J. Amer. Statist. Assoc.*, 91:983–92, 1996.
- [34] R.D. Cook. *Regression Graphics: Ideas for Studying Regressions Through Graphics*. Wiley, 1998.
- [35] R.D. Cook. Fisher lecture: Dimension reduction in regression. *Statistical Science*, 22(1):1–26, 2007.
- [36] R.D. Cook and S. Weisberg. Discussion of li (1991). *J. Amer. Statist. Assoc.*, 86:328–332, 1991.
- [37] David Corcoran*, Stoyan Georgiev*, Neel Mukherjee, Eva Gottwein, Rebecca Skalsky, Jack D. Keene, and Uwe Ohler. Definition of regulatory rna binding sites from par-clip massive short-read sequence data. *Genome Biology in review*, 2011.
- [38] Gloria Coruzzi, Rodrigo Gutierrez, Erich Grotewold, and Nathan Springer. *The Plant Genome: Decoding the Transcriptional Hardwiring*, pages 196–228. Wiley-Blackwell, 2009.
- [39] Gregory E. Crawford, Ingeborg E. Holt, James Whittle, Bryn D. Webb, Denise Tai, Sean Davis, Elliott H. Margulies, YiDong Chen, John A. Bernat, David Ginsburg, Daixing Zhou, Shujun Luo, Thomas J. Vasicek, Mark J. Daly, Tyra G. Wolfsberg, and Francis S. Collins. Genome-wide mapping of dnase hypersensitive sites using massively parallel signature sequencing (mpss). *Genome Research*, 16(1):123–131, 2006.
- [40] Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of johnson and lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003.

- [41] Marko Djordjevic, Anirvan M. Sengupta, and Boris I. Shraiman. A biophysical approach to transcription factor binding site discovery. *Genome Research*, 13(11):2381–2390, 2003.
- [42] Lori E. Dodd, Srikumar Sengupta, I-How Chen, Johan A. den Boon, Yu-Juen Cheng, William Westra, Michael A. Newton, Beth F. Mittl, Lisa McShane, Chien-Jen Chen, Paul Ahlquist, and Allan Hildesheim. Genes involved in dna repair and nitrosamine metabolism and those located on chromosome 14q32 are dysregulated in nasopharyngeal carcinoma. *Cancer Epidemiology Biomarkers and Prevention*, 15(11):2216–2225, 2006.
- [43] D. Donoho and C. Grimes. Hessian eigenmaps: new locally linear embedding techniques for highdimensional data. *PNAS*, 100:5591–5596, 2003.
- [44] Thomas A. Down and Tim J. P. Hubbard. Nestedmica: sensitive inference of over-represented motifs in nucleic acid sequence. *Nucleic Acids Research*, 33(5):1445–1453, 2005.
- [45] Petros Drineas, Ravi Kannan, and Michael W. Mahoney. Fast monte carlo algorithms for matrices ii: Computing a low-rank approximation to a matrix. *SIAM J. Comput.*, 36:158–183, July 2006.
- [46] Petros Drineas and Michael W. Mahoney. On the nystrom method for approximating a gram matrix for improved kernel-based learning. *J. Mach. Learn. Res.*, 6:2153–2175, December 2005.
- [47] Eran Eden, Doron Lipson, Sivan Yogev, and Zohar Yakhini. Discovering motifs in ranked lists of dna sequences. *PLoS Comput Biol*, 3(3):e39, 03 2007.
- [48] Olivier Elemento, Noam Slonim, and Saeed Tavazoie. A universal framework for regulatory element discovery across all genomes and data types. *Molecular Cell*, 28(2):337 – 350, 2007.
- [49] Olivier Elemento and Saeed Tavazoie. Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach. *Genome Biology*, 6(2):R18, 2005.
- [50] Ana Eulalio, Jan Rehwinkel, Mona Stricker, Eric Huntzinger, Schu-Fee Yang, Tobias Doerks, Silke Dorner, Peer Bork, Michael Boutros, and Elisa Izaurralde. Target-specific requirements for enhancers of decapping in mirna-mediated gene silencing. *Genes & Development*, 21(20):2558–2570, 2007.
- [51] Gary Felsenfeld and Mark Groudine. Controlling the double helix. *Nature*, 421(6921):448–453, 2003.
- [52] Robert D. Finn, Jaina Mistry, John Tate, Penny Coghill, Andreas Heger, Joanne E. Pollington, O. Luke Gavin, Prasad Gunasekaran, Goran Ceric, Kristoffer Forslund, Liisa Holm, Erik L. L. Sonnhammer, Sean R. Eddy, and Alex Bateman. The pfam protein families database. *Nucleic Acids Research*, 38(suppl 1):D211–D222, 2010.

- [53] Barrett C. Foat, S. Sean Houshmandi, Wendy M. Olivas, and Harmen J. Bussemaker. Profiling condition-specific, genome-wide regulation of mRNA stability in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, 102(49):17675–17680, 2005.
- [54] Barrett C. Foat, Alexandre V. Morozov, and Harmen J. Bussemaker. Statistical mechanical modeling of genome-wide transcription factor occupancy data by matrixreduce. *Bioinformatics*, 22(14):e141–e149, 2006.
- [55] P. Frankl and H. Maehara. The johnson-lindenstrauss lemma and the sphericity of some graphs. *J. Comb. Theory Ser. A*, 44:355–362, June 1987.
- [56] Matthew L Freedman, David Reich, Kathryn L Penney, Gavin J McDonald, Andre A Mignault, Nick Patterson, Stacey B Gabriel, Eric J Topol, Jordan W Smoller, Carlos N Pato, Michele T Pato, Tracey L Petryshen, Laurence N Kolonel, Eric S Lander, Pamela Sklar, Brian Henderson, Joel N Hirschhorn, and David Altshuler. Assessing the impact of population stratification on genetic association studies. *Nat Genet*, 36(4):388–393, 2004.
- [57] Robin C. Friedman, Kyle Kai-How Farh, Christopher B. Burge, and David P. Bartel. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research*, 19(1):92–105, 2009.
- [58] Andre Galarneau and Stephane Richard. Target RNA motif and target mRNAs of the quaking star protein. *Nat Struct Mol Biol*, 12(8):691–698, 2005.
- [59] Alessia Galgano, Michael Forrer, Lukasz Jaskiewicz, Alexander Kanitz, Mihaela Zavolan, and Andr P. Gerber. Comparative analysis of mRNA targets for human PUF-family proteins suggests extensive interaction with the miRNA regulatory system. *PLoS ONE*, 3(9):e3164, 09 2008.
- [60] Miklos Gaszner and Gary Felsenfeld. Insulators: exploiting transcriptional and epigenetic mechanisms. *Nat Rev Genet*, 7(9):703–713, 2006.
- [61] Stoyan Georgiev, Alan Boyle, Karthik Jayasurya, Xuan Ding, Sayan Mukherjee, and Uwe Ohler. Evidence-ranked motif identification. *Genome Biology*, 11(2):R19, 2010.
- [62] André P. Gerber, Daniel Herschlag, and Patrick O. Brown. Extensive association of functionally and cytotopically related mRNAs with PUF family RNA-binding proteins in yeast. *PLoS biology*, 2(3), March 2004.
- [63] Andr P. Gerber, Stefan Luschnig, Mark A. Krasnow, Patrick O. Brown, and Daniel Herschlag. Genome-wide identification of mRNAs associated with the translational regulator pumilio in *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences of the United States of America*, 103(12):4487–4492, 2006.
- [64] Antonio J. Giraldez, Yuichiro Mishima, Jason Rihel, Russell J. Grocock, Stijn Van Dongen, Kunio Inoue, Anton J. Enright, and Alexander F. Schier. Zebrafish mir-430 promotes deadenylation and clearance of maternal mRNAs. *Science*, 312(5770):75–79, 2006.

- [65] Amir Globerson and Sam Roweis. Metric learning by collapsing classes. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 451–458. MIT Press, Cambridge, MA, 2006.
- [66] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov. Neighbourhood component analysis. In *Advances in Neural Information Processing Systems 17*, pages 513–520, 2005.
- [67] Gene H. Golub and Charles F. Van Loan. *Matrix computations*. John Hopkins University Press, Baltimore, MD, 3rd edition, 1996.
- [68] Raluca Gordân and Alexander J Hartemink.
- [69] Raluca Gordân, Leelavati Narlikar, and Alexander J. Hartemink. A fast, alignment-free, conservation-based method for transcription factor binding site discovery. In *Proceedings of the 12th annual international conference on Research in computational molecular biology, RECOMB’08*, pages 98–111, Berlin, Heidelberg, 2008. Springer-Verlag.
- [70] Sam Griffiths-Jones, Russell J. Grocock, Stijn van Dongen, Alex Bateman, and Anton J. Enright. mirbase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Research*, 34(suppl 1):D140–D144, 2005.
- [71] D S Gross and W T Garrard. Nuclease hypersensitive sites in chromatin. *Annual Review of Biochemistry*, 57(1):159–197, 1988.
- [72] Per gunnar Martinsson, Vladimir Rokhlin, and Mark Tygert. A randomized algorithm for the approximation of matrices. 2006.
- [73] Huili Guo, Nicholas T. Ingolia, Jonathan S. Weissman, and David P. Bartel. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature*, 466(7308):835–840, 2010.
- [74] Markus Hafner, Markus Landthaler, Lukas Burger, Mohsen Khorshid, Jean Hausser, Philipp Berninger, Andrea Rothballer, Manuel Ascano Jr., Anna-Carina Jungkamp, Mathias Munschauer, Alexander Ulrich, Greg S. Wardle, Scott Dewell, Mihaela Zavolan, and Thomas Tuschl. Transcriptome-wide identification of rna-binding protein and microRNA target sites by par-clip. *Cell*, 141(1):129 – 141, 2010.
- [75] N. Halko, P.-G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *ArXiv e-prints*, September 2009.
- [76] Peter Hall and Ker-Chau Li. On almost linearity of low dimensional projections from high dimensional data. *The Annals of Statistics*, 21(2):867–889.
- [77] C. Harbison, D. Gordon, T. Lee, N. Rinaldi, K. Macisaac, T. Danford, N. Hannett, J. Tagne, D. Reynolds, J. Yoo, E. Jennings, J. Zeitlinger, D. Pokholok, M. Kellis, P. Rolfe, K. Takusagawa, E. Lander, D. Gifford, E. Fraenkel, and R. Young. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431:99–104, 2004.

- [78] Christopher T. Harbison, D. Benjamin Gordon, Tong Ihn Lee, Nicola J. Rinaldi, Kenzie D. Macisaac, Timothy W. Danford, Nancy M. Hannett, Jean-Bosco Tagne, David B. Reynolds, Jane Yoo, Ezra G. Jennings, Julia Zeitlinger, Dmitry K. Pokholok, Manolis Kellis, P. Alex Rolfe, Ken T. Takusagawa, Eric S. Lander, David K. Gifford, Ernest Fraenkel, and Richard A. Young. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431(7004):99–104, 2004.
- [79] T. Hastie and R. Tibshirani. Discriminant adaptive nearest neighbor classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(6):607–616, 1996.
- [80] Xiaofei He and Partha Niyogi. Locality preserving projections. In *NIPS*, 2003.
- [81] Xiaofei He, Shuicheng Yan, Yuxiao Hu, Partha Niyogi, and Hong jiang Zhang. Face recognition using laplacianfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:328–340, 2005.
- [82] Jacques van Helden, Alma. F. Rios, and Julio Collado-Vides. Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Research*, 28(8):1808–1818, 2000.
- [83] Michael Hiller, Rainer Pudimat, Anke Busch, and Rolf Backofen. Using rna secondary structures to guide sequence motif finding towards single-stranded regions. *Nucleic Acids Research*, 34(17):e117, 2006.
- [84] Brad G Hoffman and Steven J M Jones. Genome-wide identification of dna-protein interactions using chromatin immunoprecipitation coupled with flow cell sequencing. *J Endocrinol*, 201(1):1–13, 2009.
- [85] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '99*, pages 50–57, New York, NY, USA, 1999. ACM.
- [86] C Holdridge and D Dorsett. Repression of hsp70 heat shock gene transcription by the suppressor of hairy-wing protein of drosophila melanogaster. *Mol. Cell. Biol.*, 11(4):1894–1900, 1991.
- [87] Roger A. Hoskins, Jane M. Landolin, James B. Brown, Jeremy E. Sandler, Hazuki Takahashi, Timo Lassmann, Charles Yu, Benjamin W. Booth, Dayu Zhang, Kenneth H. Wan, Li Yang, Nathan Boley, Justen Andrews, Thomas C. Kaufman, Brenton R. Graveley, Peter J. Bickel, Piero Carninci, Joseph W. Carlson, and Susan E. Celniker. Genome-wide analysis of promoter architecture in *Drosophila melanogaster*. *Genome Research*, 21(2):182–192, February 2011.
- [88] H. Hotelling. Analysis of a complex of statistical variables in principal components. *Journal of Educational Psychology*, 24:417–441, 1933.
- [89] Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing, STOC '98*, pages 604–613, New York, NY, USA, 1998. ACM.

- [90] V. Iyer, C. Horak, C. Scafe, D. Botstein, M. Snyder, and P. Brown. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature*, 409:533–538, 2001.
- [91] David S. Johnson, Ali Mortazavi, Richard M. Myers, and Barbara Wold. Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316(5830):1497–1502, 2007.
- [92] William Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. In *Conference in modern analysis and probability (New Haven, Conn., 1982)*, volume 26 of *Contemporary Mathematics*, pages 189–206. American Mathematical Society, 1984.
- [93] Hilal Kazan, Debashish Ray, Esther T. Chan, Timothy R. Hughes, and Quaid Morris. Rna-context: A new method for learning the sequence and structure binding preferences of rna-binding proteins. *PLoS Comput Biol*, 6(7):e1000832, 07 2010.
- [94] Jack D. Keene. Rna regulons: coordination of post-transcriptional events. *Nat Rev Genet*, 8(7):533–543, 2007.
- [95] Jack D Keene, Jordan M Komisarow, and Matthew B Friedersdorf. Rip-chip: the isolation and identification of mrnas, micromnas and protein components of ribonucleoprotein complexes from cell extracts. *Nat. Protocols*, 1(1):302–307, 2006.
- [96] M A Keene, V Corces, K Lowenhaupt, and S C Elgin. Dnase i hypersensitive sites in drosophila chromatin occur at the 5' ends of regions of transcription. *Proceedings of the National Academy of Sciences*, 78(1):143–146, 1981.
- [97] Manolis Kellis, Nick Patterson, Bruce Birren, Bonnie Berger, and Eric S. Lander. Methods in comparative genomics: Genome correspondence, gene identification and regulatory motif discovery. *Journal of Computational Biology*, 11(2-3):319–355, 2004.
- [98] Manolis Kellis, Nick Patterson, Matthew Endrizzi, Bruce Birren, and Eric S. Lander. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423(6937):241–254, 2003.
- [99] Aaron M. Kershner and Judith Kimble. Genome-wide analysis of mrna targets for caenorhabditis elegans fbf, a conserved stem cell regulator. *Proceedings of the National Academy of Sciences*.
- [100] Tae Hoon Kim, Leah O. Barrera, Ming Zheng, Chunxu Qu, Michael A. Singer, Todd A. Richmond, Yingnian Wu, Roland D. Green, and Bing Ren. A high-resolution map of active promoters in the human genome. *Nature*, 436(7052):876–880, 2005.
- [101] Shivendra Kishore, Lukasz Jaskiewicz, Lukas Burger, Jean Hausser, Mohsen Khorshid, and Mihaela Zavolan. A quantitative analysis of clip methods for identifying binding sites of rna-binding proteins. *Nat Meth*, 8(7):559–564, 2011.
- [102] Shivendra Kishore, Sandra Luber, and Mihaela Zavolan. Deciphering the role of rna-binding proteins in the post-transcriptional control of gene expression. *Briefings in Functional Genomics*, 9(5-6):391–404, 2010.

- [103] Ana Kozomara and Sam Griffiths-Jones. mirbase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Research*, 39(suppl 1):D152–D157, 2011.
- [104] D.S. Latchman. *Eukaryotic transcription factors*. Elsevier/Academic Press, 2008.
- [105] CE Lawrence, SF Altschul, MS Boguski, JS Liu, AF Neuwald, and JC Wootton. Detecting subtle sequence signals: a gibbs sampling strategy for multiple alignment. *Science*, 262(5131):208–214, 1993.
- [106] Svetlana Lebedeva, Marvin Jens, Kathrin Theil, Björn Schwanhüsser, Matthias Selbach, Markus Landthaler, and Nikolaus Rajewsky. Transcriptome-wide analysis of regulatory interactions of the rna-binding protein hur. *Molecular Cell*, In Press, Corrected Proof:–, 2011.
- [107] Rosalind C. Lee, Rhonda L. Feinbaum, and Victor Ambros. The *c. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell*, 75(5):843 – 854, 1993.
- [108] T. Lee, N. Rinaldi, F. Robert, D. Odom, Z. Bar-Joseph, G. Gerber, N. Hannett, C. Harbison, C. Thompson, I. Simon, J. Zeitlinger, E. Jennings, H. Murray, D. Gordon, B. Ren, J. Wyrick, J. Tagne, T. Volkert, E. Fraenkel, D. Gifford, and R. Young. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, 298:799–804, 2002.
- [109] Boris Lenhard, Albin Sandelin, Luis Mendoza, Par Engström, Niclas Jareborg, and Wyeth Wasserman. Identification of conserved regulatory elements by comparative genome analysis. *Journal of Biology*, 2(2):13, 2003.
- [110] Michael Levine and Robert Tjian. Transcription regulation and animal diversity. *Nature*, 424(6945):147–151, 2003.
- [111] Benjamin P. Lewis, Christopher B. Burge, and David P. Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120(1):15 – 20, 2005.
- [112] Bing Li, Hongyuan Zha, and Francesca Chiaromonte. Contour regression: A general approach to dimension reduction. *The Annals of Statistics*, 33(4):1580–1616, 2005.
- [113] Heng Li, Jue Ruan, and Richard Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, 18(11):1851–1858, 2008.
- [114] Jun Z. Li, Devin M. Absher, Hua Tang, Audrey M. Southwick, Amanda M. Casto, Sohini Ramachandran, Howard M. Cann, Gregory S. Barsh, Marcus Feldman, Luigi L. Cavalli-Sforza, and Richard M. Myers. Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, 319(5866):1100–1104, 2008.
- [115] K. C. Li. On principal hessian directions for data visualization and dimension reduction: another application of Stein’s lemma. *J. Amer. Statist. Assoc.*, 87:1025–1039, 1992.
- [116] K.C. Li. Sliced inverse regression for dimension reduction (with discussion). *J. Amer. Statist. Assoc.*, 86:316–342, 1991.

- [117] Xiao Li, Gerald Quon, Howard D. Lipshitz, and Quaid Morris. Predicting in vivo binding sites of rna-binding proteins using mrna secondary structure. *RNA*, 16(6):1096–1107, 2010.
- [118] Edo Liberty, Franco Woolfe, Per-Gunnar Martinsson, Vladimir Rokhlin, and Mark Tygert. Randomized algorithms for the low-rank approximation of matrices. *Proceedings of the National Academy of Sciences*, 104(51):20167–20172, 2007.
- [119] Chaim Linhart, Yonit Halperin, and Ron Shamir. Transcription factor and microrna motif discovery: The amadeus platform and a compendium of metazoan target sets. *Genome Research*, 18(7):1180–1189, 2008.
- [120] X. Liu, D. L. Brutlag, and J. S. Liu. Bioprospector: Discovering conserved dna motifs in upstream regulatory regions of co-expressed genes. In *Pac. Symp. Biocomput*, pages 127–138, 2001.
- [121] X. Shirley Liu, Douglas L. Brutlag, and Jun S. Liu. An algorithm for finding protein-dna binding sites with. *Nat Biotech*, 20(8):835–839, 2002.
- [122] Bradley M. Lunde, Claire Moore, and Gabriele Varani. Rna-binding proteins: modular design for efficient function. *Nat Rev Mol Cell Biol*, 8(6):479–490, 2007.
- [123] Kenzie MacIsaac, Ting Wang, D Benjamin Gordon, David Gifford, Gary Stormo, and Ernest Fraenkel. An improved map of conserved regulatory sites for *saccharomyces cerevisiae*. *BMC Bioinformatics*, 7(1):113, 2006.
- [124] Udi Manber and Gene Myers. Suffix Arrays: A New Method for On-Line String Searches. In *Proceedings of the first annual ACM-SIAM symposium on Discrete algorithms, SODA '90*, pages 319–327, Philadelphia, PA, USA, 1990. Society for Industrial and Applied Mathematics.
- [125] Giovanni Manzini and Paolo Ferragina. Engineering a lightweight suffix array construction algorithm. *Algorithmica*, 40:33–50, 2004.
- [126] Kai Mao, Feng Liang, and Sayan Mukherjee. Supervised dimension reduction using bayesian mixture modeling. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010.
- [127] Patricia A Maroney, Yang Yu, Jesse Fisher, and Timothy W Nilsen. Evidence that micrornas are associated with translating messenger rnas in human cells. *Nat Struct Mol Biol*, 13(12):1102–1107, 2006.
- [128] James D. McGhee, William I. Wood, Maureen Dolan, James Douglas Engel, and Gary Felsenfeld. A 200 base pair region at the 5' end of the chicken adult [beta]-globin gene is accessible to nuclease digestion. *Cell*, 27(1, Part 2):45 – 55, 1981.
- [129] P Menozzi, A Piazza, and L Cavalli-Sforza. Synthetic maps of human gene frequencies in europeans. *Science*, 201(4358):786–792, 1978.

- [130] Melissa J. Moore. From birth to death: The complex lives of eukaryotic mRNAs. *Science*, 309(5740):1514–1518, 2005.
- [131] Vamsi K Mootha, Cecilia M Lindgren, Karl-Fredrik Eriksson, Aravind Subramanian, Smita Sihag, Joseph Lehar, Pere Puigserver, Emma Carlsson, and Ridderstr. Pgc-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, 34:267 – 73, 2003/07// 2003.
- [132] Adam R. Morris, Neelanjan Mukherjee, and Jack D. Keene. Ribonomic analysis of human pum1 reveals cis-trans conservation across species despite evolution of diverse mRNA target sets. *Mol. Cell. Biol.*, 28(12):4093–4103, 2008.
- [133] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat Meth*, 5(7):621–628, 2008.
- [134] Neelanjan Mukherjee, David L. Corcoran, Jeffrey D. Nusbaum, David W. Reid, Stoyan Georgiev, Markus Hafner, Manuel Ascano Jr., Thomas Tuschl, Uwe Ohler, and Jack D. Keene. Integrative regulatory mapping indicates that the RNA-binding protein HUR couples pre-mRNA processing and mRNA stability. *Molecular Cell*, In Press, Corrected Proof:–, 2011.
- [135] Sonali Mukherjee, Michael F Berger, Ghil Jona, Xun S Wang, Dale Muzzey, Michael Snyder, Richard A Young, and Martha L Bulyk. Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat Genet*, 36(12):1331–1339, 2004.
- [136] Leelavati Narlikar, Raluca Gordn, and Alexander J Hartemink. A nucleosome-guided map of transcription factor binding sites in yeast. *PLoS Comput Biol*, 3(11):e215, 11 2007.
- [137] Leelavati Narlikar and Alexander J. Hartemink. Sequence features of DNA binding sites reveal structural class of associated transcription factor. *Bioinformatics*, 22(2):157–163, 2006.
- [138] Matthew R. Nelson, Katarzyna Bryc, Karen S. King, Amit Indap, Adam R. Boyko, John Novembre, Linda P. Briley, Yuka Maruyama, Dawn M. Waterworth, Grard Waeber, Peter Vollenweider, Jorge R. Oksenberg, Stephen L. Hauser, Heide A. Stirnadel, Jaspal S. Kooner, John C. Chambers, Brendan Jones, Vincent Mooser, Carlos D. Bustamante, Allen D. Roses, Daniel K. Burns, Margaret G. Ehm, and Eric H. Lai. The population reference sample, POPRES: A resource for population, disease, and pharmacological genetics research. *The American Journal of Human Genetics*, 83(3):347 – 358, 2008.
- [139] Shane Neph and Martin Tompa. Microfootprinter: a tool for phylogenetic footprinting in prokaryotic genomes. *Nucleic Acids Research*, 34(suppl 2):W366–W368, 1 July 2006.
- [140] Michael A. Newton, Fernando A. Quintana, Srikumar den Boon, Johan A. Sengupta, and Paul Ahlquist. Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. *Ann. Appl. Stat.*, 1(1):85–106, 2007.
- [141] Ting Ni, David L Corcoran, Elizabeth A Rach, Shen Song, Eric P Spana, Yuan Gao, Uwe Ohler, and Jun Zhu. A paired-end sequencing strategy to map the complex landscape of transcription initiation. *Nat Meth*, 7(7):521–527, 2010.

- [142] Rasmus Nielsen, Ines Hellmann, Melissa Hubisz, Carlos Bustamante, and Andrew G. Clark. Recent and ongoing selection in the human genome. *Nat Rev Genet*, 8(11):857–868, 2007.
- [143] J. Nilsson, F. Sha, and M.I. Jordan. Regression on manifolds using kernel dimension reduction. In *Proceedings of the 24th International Conference on Machine Learning*, 2007.
- [144] Marcelo A. Nobrega, Ivan Ovcharenko, Veena Afzal, and Edward M. Rubin. Scanning human gene deserts for long-range enhancers. *Science*, 302(5644):413, 2003.
- [145] Stephanie Nottrott, Martin J Simard, and Joel D Richter. Human let-7a mirna blocks protein production on actively translating polyribosomes. *Nat Struct Mol Biol*, 13(12):1108–1114, 2006.
- [146] John Novembre, Toby Johnson, Katarzyna Bryc, Zoltan Kutalik, Adam R. Boyko, Adam Auton, Amit Indap, Karen S. King, Sven Bergmann, Matthew R. Nelson, Matthew Stephens, and Carlos D. Bustamante. Genes mirror geography within europe. *Nature*, 456(7218):98–101, 2008.
- [147] S Ogbourne and T M Antalis. Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes. *Biochem. J.*, 331(1):1–14, 1998.
- [148] Uwe Ohler and Heinrich Niemann. Identification and analysis of eukaryotic promoters: recent computational approaches. *Trends in Genetics*, 17(2):56 – 60, 2001.
- [149] Nick Patterson, Alkes L Price, and David Reich. Population structure and eigenanalysis. *PLoS Genet*, 2(12):e190, 12 2006.
- [150] Giulio Pavesi, Giancarlo Mauri, and Graziano Pesole. An algorithm for finding signals of unknown length in dna sequences. *Bioinformatics*, 17(suppl 1):S207–S214, 2001.
- [151] Karl Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559–572, 1901.
- [152] Christian P. Petersen, Marie-Eve Bordeleau, Jerry Pelletier, and Phillip A. Sharp. Short rnas repress translation after initiation in mammalian cells. *Molecular Cell*, 21(4):533 – 542, 2006.
- [153] Alkes Price, Sriram Ramabhadran, and Pavel A. Pevzner. Finding subtle motifs by branching from sample strings. *Bioinformatics*, 19(suppl 2):ii149–ii155, 2003.
- [154] J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959, 2000.
- [155] Yuan Qi, Alex Rolfe, Kenzie D MacIsaac, Georg K Gerber, Dmitry Pokholok, Julia Zeitlinger, Timothy Danford, Robin D Dowell, Ernest Fraenkel, Tommi S Jaakkola, Richard A Young, and David K Gifford. High-resolution computational models of genome binding events. *Nat Biotech*, 24(8):963–970, 2006.

- [156] Elizabeth Rach, Hsiang-Yu Yuan, William Majoros, Pavel Tomancak, and Uwe Ohler. Motif composition, conservation and condition-specificity of single and alternative transcription start sites in the drosophila genome. *Genome Biology*, 10(7):R73, 2009.
- [157] Jan Rehwinkel, Pavel Natalin, Alexander Stark, Julius Brennecke, Stephen M. Cohen, and Elisa Izaurralde. Genome-wide analysis of mRNAs regulated by Droscha and Argonaute proteins in *Drosophila melanogaster*. *Mol. Cell. Biol.*, 26(8):2965–2975, 2006.
- [158] Brenda J. Reinhart, Frank J. Slack, Michael Basson, Amy E. Pasquinelli, Jill C. Bettinger, Ann E. Rougvie, H. Robert Horvitz, and Gary Ruvkun. The 21-nucleotide let-7 RNA regulates developmental timing in *Drosophila*. *Nature*, 403(6772):901–906, 2000.
- [159] B. Ren, F. Robert, J. Wyrick, O. Aparicio, and E. Jennings. Genome-wide location and function of DNA binding proteins. *Science*, 290:2306–2309, 2000.
- [160] B. Ren, F. Robert, J.J. Wyrick, R.O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T.L. Volkert, C.J. Wilson, S.P. Bell, and R.A. Young. Genome-wide location and function of DNA binding proteins. *Science*, 290:2306–2309, December 2000.
- [161] Gordon Robertson, Martin Hirst, Matthew Bainbridge, Misha Bilenky, Yongjun Zhao, Thomas Zeng, Ghia Euskirchen, Bridget Bernier, Richard Varhol, Allen Delaney, Nina Thiessen, Obi L Griffith, Ann He, Marco Marra, Michael Snyder, and Steven Jones. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Meth*, 4(8):651–657, 2007.
- [162] V. Rokhlin, A. Szlam, and M. Tygert. A randomized algorithm for principal component analysis. *ArXiv e-prints*, September 2008.
- [163] Vladimir Rokhlin and Mark Tygert. A fast randomized algorithm for overdetermined linear least-squares regression. *Proceedings of the National Academy of Sciences*, 105(36):13212–13217, 2008.
- [164] Ian M. Rosenberg. *Protein Analysis and Purification: Benchtop Techniques*. 2004.
- [165] Noah A. Rosenberg, Jonathan K. Pritchard, James L. Weber, Howard M. Cann, Kenneth K. Kidd, Lev A. Zhivotovsky, and Marcus W. Feldman. Genetic structure of human populations. *Science*, 298(5602):2381–2385, 2002.
- [166] Frederick P. Roth, Jason D. Hughes, Preston W. Estep, and George M. Church. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat Biotech*, 16(10):939–945, 1998.
- [167] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.
- [168] Albin Sandelin, Wynand Alkema, Pr Engström, Wyeth W. Wasserman, and Boris Lenhard. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research*, 32(suppl 1):D91–D94, 2004.

- [169] Tamas Sarlos. Improved approximation algorithms for large matrices via random projections. In *Foundations of Computer Science, 2006. FOCS '06. 47th Annual IEEE Symposium on*, pages 143–152, oct. 2006.
- [170] Eric W. Sayers, Tanya Barrett, Dennis A. Benson, Evan Bolton, Stephen H. Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M. Church, Michael DiCuccio, Scott Federhen, Michael Feolo, Ian M. Fingerman, Lewis Y. Geer, Wolfgang Helmberg, Yuri Kapustin, David Landsman, David J. Lipman, Zhiyong Lu, Thomas L. Madden, Tom Madej, Donna R. Maglott, Aron Marchler-Bauer, Vadim Miller, Ilene Mizrahi, James Ostell, Anna Panchenko, Lon Phan, Kim D. Pruitt, Gregory D. Schuler, Edwin Sequeira, Stephen T. Sherry, Martin Shumway, Karl Sirotkin, Douglas Slotta, Alexandre Souvorov, Grigory Starchenko, Tatiana A. Tatusova, Lukas Wagner, Yanli Wang, W. John Wilbur, Eugene Yaschenko, and Jian Ye. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 39(suppl 1):D38–D51, 2011.
- [171] Daniela Schmitter, Jody Filkowski, Alain Sewer, Ramesh S. Pillai, Edward J. Oakeley, Mihaela Zavolan, Petr Svoboda, and Witold Filipowicz. Effects of dicer and argonaute down-regulation on mrna levels in human hek293 cells. *Nucleic Acids Research*, 34(17):4801–4815, 2006.
- [172] Kathy Seggerson, Lingjuan Tang, and Eric G. Moss. Two genetic circuits repress the *caenorhabditis elegans* heterochronic gene *lin-28* after translation initiation. *Developmental Biology*, 243(2):215 – 225, 2002.
- [173] Thomas Sellke, M. J. Bayarri, and James O. Berger. Calibration of p values for testing precise null hypotheses. *The American Statistician*, 55(1):62–71.
- [174] Chantelle F. Sephton, Can Cenik, Alper Kucukural, Eric B. Dammer, Basar Cenik, YuHong Han, Colleen M. Dewey, Frederick P. Roth, Joachim Herz, Junmin Peng, Melissa J. Moore, and Gang Yu. Identification of neuronal rna targets of tdp-43-containing ribonucleoprotein complexes. *Journal of Biological Chemistry*, 286(2):1204–1215, 2011.
- [175] J. Shawe-Taylor, C.K.I. Williams, N. Cristianini, and J. Kandola. On the eigenspectrum of the gram matrix and the generalization error of kernel-pca. *Information Theory, IEEE Transactions on*, 51(7):2510 – 2522, july 2005.
- [176] S. T. Sherry, M.-H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. dbsnp: the ncbi database of genetic variation. *Nucleic Acids Research*, 29(1):308–311, 2001.
- [177] Rahul Siddharthan, Eric D Siggia, and Erik van Nimwegen. Phylogibbs: A gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput Biol*, 1(7):e67, 12 2005.
- [178] Adam Siepel, Gill Bejerano, Jakob S. Pedersen, Angie S. Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram Clawson, John Spieth, LaDeana W. Hillier, Stephen Richards, George M. Weinstock, Richard K. Wilson, Richard A. Gibbs, W. James Kent, Webb Miller, and David Haussler. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, 15(8):1034–1050, 2005.

- [179] Saurabh Sinha and Martin Tompa. Discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Research*, 30(24):5549–5560, 2002.
- [180] Robert Sladek, Ghislain Rocheleau, Johan Rung, Christian Dina, Lishuang Shen, David Serre, Philippe Boutin, Daniel Vincent, Alexandre Belisle, Samy Hadjadj, Beverley Balkau, Barbara Heude, Guillaume Charpentier, Thomas J. Hudson, Alexandre Montpetit, Alexey V. Pshezhetsky, Marc Prentki, Barry I. Posner, David J. Balding, David Meyre, Constantin Polychronakos, and Philippe Froguel. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, 445(7130):881–885, 2007.
- [181] Stephen T. Smale. Core promoters: active contributors to combinatorial gene regulation. *Genes & Development*, 15(19):2503–2508, 2001.
- [182] Stephen T. Smale and James T. Kadonaga. The RNA polymerase II core promoter. *Annual review of biochemistry*, 72(1):449–479, 2003.
- [183] Peter Stoilov, Rosette Daoud, Oliver Nayler, and Stefan Stamm. Human tra2-beta1 autoregulates its protein concentration by influencing alternative splicing of its pre-mrna. *Human Molecular Genetics*, 13(5):509–524, 2004.
- [184] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, 2005.
- [185] M. Sugiyama. Dimension reduction of multimodal labeled data by local fisher discriminant analysis. *Journal of Machine Learning Research*, 8:1027–1061, 2007.
- [186] Danilo A. Tagle, Ben F. Koop, Morris Goodman, Jerry L. Slightom, David L. Hess, and Richard T. Jones. Embryonic epsilon and gamma globin genes of a prosimian primate (*galago crassicaudatus*) : Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *Journal of Molecular Biology*, 203(2):439 – 455, 1988.
- [187] J. Tenenbaum, V. de Silva, and J. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.
- [188] Martin Tompa, Nan Li, Timothy L Bailey, George M Church, Bart De Moor, Eleazar Eskin, Alexander V Favorov, Martin C Frith, Yutao Fu, W James Kent, Vsevolod J Makeev, Andrei A Mironov, William Stafford Noble, Giulio Pavesi, Graziano Pesole, Mireille Regnier, Nicolas Simonis, Saurabh Sinha, Gert Thijs, Jacques van Helden, Mathias Vandenberg, Zhiping Weng, Christopher Workman, Chun Ye, and Zhou Zhu. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotech*, 23(1):137–144, 2005.
- [189] Nathan D. Trinklein, Shelley J. Force Aldred, Alok J. Saldanha, and Richard M. Myers. Identification and functional analysis of human transcriptional promoters. *Genome Research*, 13(2):308–312, 2003.

- [190] Jernej Ule, Kirk B. Jensen, Matteo Ruggiu, Aldo Mele, Aljaz Ule, and Robert B. Darnell. Clip identifies nova-regulated mna networks in the brain. *Science*, 302(5648):1212–1215, 2003.
- [191] Anton Valouev, David S Johnson, Andreas Sundquist, Catherine Medina, Elizabeth Anton, Serafim Batzoglou, Richard M Myers, and Arend Sidow. Genome-wide analysis of transcription factor binding sites based on chip-seq data. *Nat Meth*, 5(9):829–834, 2008.
- [192] J. van Helden, B. Andr, and J. Collado-Vides. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *Journal of Molecular Biology*, 281(5):827 – 842, 1998.
- [193] Axel Visel, Matthew J. Blow, Zirong Li, Tao Zhang, Jennifer A. Akiyama, Amy Holt, Ingrid Plajzer-Frick, Malak Shoukry, Crystal Wright, Feng Chen, Veena Afzal, Bing Ren, Edward M. Rubin, and Len A. Pennacchio. Chip-seq accurately predicts tissue-specific activity of enhancers. *Nature*, 457(7231):854–858, 2009.
- [194] Dominique Vlieghe, Albin Sandelin, Pieter J. De Bleser, Kris Vleminckx, Wyeth W. Wasserman, Frans van Roy, and Boris Lenhard. A new generation of jaspar, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Research*, 34(suppl 1):D95–D97.
- [195] E K White, T Moore-Jarrett, and H E Ruley. Pum2, a novel murine puf protein, and its consensus mna-binding site. *RNA*, 7(12):1855–1866, 2001.
- [196] Marvin Wickens, David S. Bernstein, Judith Kimble, and Roy Parker. A puf family portrait: 3'utr regulation as a way of life. *Trends in Genetics*, 18(3):150 – 157, 2002.
- [197] Bruce Wightman, Ilho Ha, and Gary Ruvkun. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in *c. elegans*. *Cell*, 75(5):855 – 862, 1993.
- [198] Carl Wu. The 5[prime] ends of drosophila heat shock genes in chromatin are hypersensitive to dnase i. *Nature*, 286(5776):854–860, 1980.
- [199] Ligang Wu, Jihua Fan, and Joel G. Belasco. Micrnas direct rapid deadenylation of mrna. *Proceedings of the National Academy of Sciences of the United States of America*, 103(11):4034–4039, 2006.
- [200] Qiang Wu, Feng Liang, and Sayan Mukherjee. Localized sliced inverse regression. *Journal of Computational and Graphical Statistics*, 19(4):843–860, 2010.
- [201] Zhishen Ye and Robert E. Weiss. Using the bootstrap to select one of a new class of dimension reduction methods. *Journal of the American Statistical Association*, 98(464):pp. 968–979, 2003.
- [202] Alper Yilmaz, Milton Y. Nishiyama, Bernardo Garcia Fuentes, Glauca Mendes Souza, Daniel Janies, John Gray, and Erich Grotewold. Grassius: A platform for comparative regulatory genomics across the grasses. *Plant Physiology*, 149(1):171–180, 2009.

- [203] Xiangrong Yin and R. Dennis Cook. Dimension reduction for the conditional k th moment in regression. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 64(2):159–175.
- [204] Yong Zhang, Tao Liu, Clifford Meyer, Jerome Eeckhoute, David Johnson, Bradley Bernstein, Chad Nusbaum, Richard Myers, Myles Brown, Wei Li, and X Shirley Liu. Model-based analysis of chip-seq (macs). *Genome Biology*, 9(9):R137, 2008.
- [205] Jing Zhao, Toshiro K. Ohsumi, Johnny T. Kung, Yuya Ogawa, Daniel J. Grau, Kavitha Sarma, Ji Joon Song, Robert E. Kingston, Mark Borowsky, and Jeannie T. Lee. Genome-wide identification of polycomb-associated rnas by rip-seq. *Molecular cell*, 40(6):939–953, 2010.
- [206] Yu Zhu and Peng Zeng. Fourier methods for estimating the central subspace and the central mean subspace in regression. *Journal of the American Statistical Association*, 101(476):1638–1651, 2006.
- [207] Laurent Zwald and Gilles Blanchard. On the convergence of eigenspaces in kernel principal component analysis. In *NIPS*, 2005.

Biography

Stoyan Georgiev was born on February 15, 1980 in Sofia, Bulgaria. He grew up in Sofia, where he graduated from Second English Language High School in 1999. In 1998 received A levels in Mathematics, Further Mathematics and Computing from UK as a student at Rossall School. After two years of undergraduate study in the Computer Science program at Sofia University, he enrolled in the Computer Science and Mathematics program at the University of Bridgeport (Connecticut, USA) where he received a Bachelor's of Science degree in Computer Science and Mathematics (minor in Computer Engineering). In 2005 he joined the Ph.D. program in Computational Biology and Bioinformatics at Duke University.

Refereed Journal and Conference Publications

- Georgiev S, Boyle AP, Jayasurya K, Ding X, Mukherjee S and Ohler U. **Evidence-ranked motif identification.** *Genome Biology*. 2010. **11**:R19
- Mukherjee N, Corcoran DL, Nusbaum JD, Reid DW, Georgiev S, Hafner M, Ascano M Jr, Tuschl T, Ohler U and Keene JD. **Integrative Regulatory Mapping Indicates that the RNA-Binding Protein HuR Couples Pre-mRNA Processing and mRNA Stability.** *Molecular Cell*. 2011.
- Corcoran DL*, Georgiev S*, Mukherjee N, Gottwein E, Skalsky RL, Keene JD, Ohler U. **Definition of RNA binding sites from PAR-CLIP short-read sequence data.** (*accepted Genome Biology*) [* co-primary authors]
- Petricka J, Schauer M, Megraw M, Breakfield N, Thompson J, Georgiev S, Soderblom E, Ohler U, Moseley M, Grossniklaus U, and Benfey P **The Protein Expression Landscape of the Arabidopsis Root.** *submitted Molecular Systems Biology*