COMPUTATIONAL METHODS FOR DETAILING DNA BINDING AFFINITIES AND DIFFERENCES AMONG RELATED PROTEINS

by

Devesh Bhimsaria

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Department of Electrical and Computer Engineering)

at the

UNIVERSITY OF WISCONSIN–MADISON

2016

Date of final oral examination: 09/28/2016

The dissertation is approved by the following members of the Final Oral Committee:
Parameswaran Ramanathan, Professor, Department of Electrical and Computer Engineering
Aseem Z. Ansari, Professor, Department of Biochemistry
Barry D. Vanveen, Professor, Department of Electrical and Computer Engineering
Mikko H. Lipasti, Professor, Department of Electrical and Computer Engineering
Sushmita Roy, Assistant Professor, Department of Biostatistics and Medical Informatics

Dedication

I dedicate my dissertation work to my family—my mother Urmila Bhimsaria, my wife Sakshi Bhimsaria and my daughter Arika Bhimsaria.

Acknowledgments

I would like to thank Dr. Parmeswaran (Parmesh) Ramanathan & Dr. Aseem Ansari for their constant support and guidance, especially for their precious time they gave to me, even on the weekends, hours and hours. Special thanks to Professor Ansari for developing my interest for research and science in me from my undergraduate days. I am also very grateful to Professor Ramanathan for his amazing inputs to my research, for keeping me focused, and providing me freedom to pursue my interest. I want to thank both of them for making me feel comfortable in their labs, with them I never felt away from home. I also thank other Keck Genome Foundry Group members at UW Madison—Professor David Schwartz and Professor Jennifer Reed for guiding me in Biochemistry and Chemical Engineering collaborative Projects. I also thank all the committee members including Dr. Sushmita Roy, Dr. Barry Vanveen and Dr. Mikko Lipasti for their support and guidance.

I thank Jacqui Mendez and Professor Jose A. Rodriguez-Martinez for providing clones for nuclear receptors and HT-SELEX binding data for nuclear receptors respectively. I also thank all the present and past members of Ramanathan lab and Ansari lab for helping me in various projects, specially Prof. Jose A. Rodriguez-Martinez, Graham Erwin, Clayton Carlson and Josh Tietjen for patiently teaching me basics of their field, which helped me a lot in my interdisciplinary projects. I thank all the collaborators and their labs, helping me learn and pursue different projects.

My studies and research all became possible because of countless sacrifices made by my mother Urmila Bhimsaria. I thank her for being my first teacher, my inspiration and ideal. Her overwhelming love gave me strength to pursue my research interest. I also thank my wife Sakshi Bhimsaria for sacrificing her career for my research. Her care for me motivated me to work harder.

I would also like to thank the National Institutes of Health and University of Wisconsin Madison for funding. I thank University of Wisconsin Madison, Department of Biochemistry and Department of Electrical and Computer Engineering for accepting me as a Khorana student earlier and then as a graduate student.

— Devesh Bhimsaria

Contents

List of Tables

List of Figures

Abstract

Transcription factors (TFs) are proteins that bind to specific sites in the genome to control the flow of genetic information from DNA to mRNA. Identifying the preferred binding sites of TFs can help in elucidating gene regulatory networks and in understanding the genetic basis for many diseases [63]. Many experimental platforms and related computational methods are successful in identifying high affinity binding sites for TFs, but leave out medium- and low- binding affinity sites; recent studies show that the medium and low affinity site also play important roles in gene regulation [68]. These methods also fail to capture all the complexities of the protein–DNA binding, especially when the proteins bind in cooperation with other proteins. This thesis develops a novel representation to capture the full range of DNA binding affinity. The thesis also integrates ideas from digital circuit optimization and compressed sensing field to develop computational methods that capture the complexities of protein–DNA binding and elucidate their detailed binding profiles [16, 22].

Proposed algorithm, differential compressed sensing based motif extraction (Diff-COSME) elucidates subtle binding differences that may explain different biological functions of related proteins. The novel differences found between proteins of homeodomain family show that even non-DNA contacting residues of proteins can affect DNA binding; these subtleties are ignored by existing tools [69]. Another proposed algorithm MinSeqFind target proteins exhibiting complex binding specially those which bind DNA as dimers with multiple different orientations. MinSeqFind when applied to thousands of published protein–DNA binding dataset found novel binding sites verified in vivo [32]. MinSeqFind is also used on a newly collected DNA binding data for vitamin-D receptor, thyroid hormone receptors, steroid hormone receptors and other nuclear receptors (NRs). MinSeqFind discovers that several NRs dimerizes with RXRA protein leading to new binding sites. The binding sites identified by MinSeqFind implicates DNA binding of NRs to hundreds of genetic mutations associated with cancer, diabetes, cardiovascular diseases and others. This information can further be used in developing drugs targeting NRs, one of the most drug targeted family of proteins.

Chapter 1

Introduction

Transcription factors (TFs) are proteins that regulate gene expression by binding to specific DNA sequences. Identification of DNA-binding of TFs is critical for understanding how TFs decipher genomic information to regulate gene circuits that control cell function. Different high-throughput experimental methods have been developed to elucidate binding of thousands of TFs [9, 12, 17, 31, 32, 40, 52, 63]. Most of the existing computational tools to analyze high-throughput binding use only high-affinity binding sequences leaving out medium- to low- affinity binding sequences. Thus, these tools do not fully explore complexity of protein–DNA binding [3, 4, 18, 63, 68]. Such tools fail to capture subtle binding differences of related TFs that can lead to different biological functions [20]. Here we introduce MinSeqs, a representation to capture the full range of DNA binding affinity and develop algorithms suited to different complexity of DNA binding to capture binding and differences in binding of related TFs in the form of MinSeqs.

In the following sections, we first discuss the protein–DNA interaction that governs the binding affinity of a protein for different DNA sequences. Next, we briefly introduce few experimental platforms developed to measure DNA bound by protein inside a cellular environment (in vivo) and also when bound in cell-free environment (in vitro). In further sections we explain the position weight matrix (PWM), a widely used method to represent DNA binding specificity; we also describe its limitations in capturing the complexity of DNA binding motifs, especially differences in TFs of same family. Finally, we describe the contributions that we make in addressing these limitations by capturing variety of binding patterns and differences which lead to the discovery of new connections between many TFs and genetic variants associated with diseases.

1.1 Protein–DNA binding

A protein interacts with DNA through hydrogen bonds, van der Waals forces, water mediated bonds and other forces, often to regulate different biological functions of DNA. DNA binding can be either sequence specific or sequence non-specific or combination of both. Two important aspects of such binding are -

1. **Affinity -** Binding affinity of a TF for sequence S is usually defined as the dissociation constant K_d — ratio of off-rate k_{off}(rate of dissociation of TF-S complex) and on-rate k_{on}(rate of formation of the complex) [63].

$$TF + S \underset{k_{on}}{\overset{k_{off}}{\rightleftharpoons}} TF.S \tag{1.1}$$

$$K_d = \frac{k_{off}}{k_{on}} = \frac{[TF][S]}{[TF.S]} \tag{1.2}$$

   The square brackets represent the concentrations of those entities at equilibrium

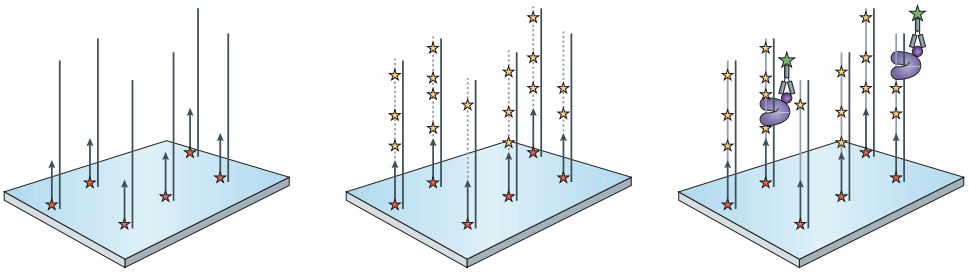
2. **Specificity -** 'Specificity' refers to how well a protein can distinguish between different sequences.

For a given TF, affinity of specific sequences are higher than others, thus when TF search for DNA in genome it prefers the higher affinity sequences.

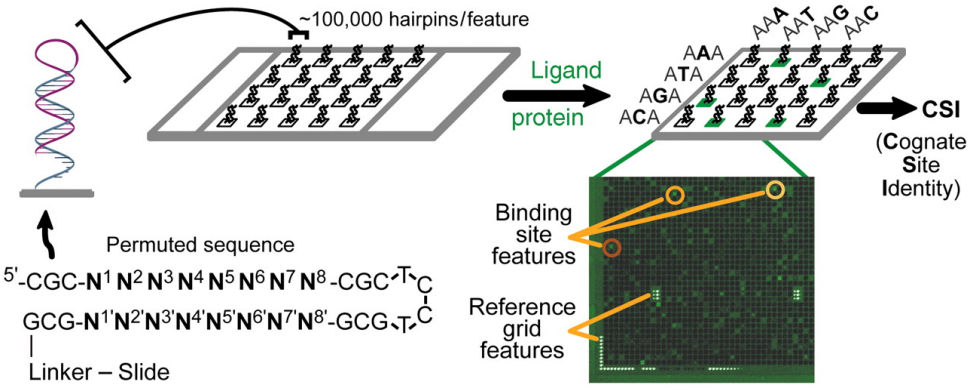
Measuring protein–DNA binding

Many experimental techniques have been developed to identify the DNA binding preferences of proteins. The Electrophoretic mobility shift assay (EMSA) serves as a standard technique to measure protein's binding specificity to a handful of DNA sequences [28,29], but due to the laborious nature, it proves to be impossible to measure binding affinity for thousands of DNA sequences. With the emerging new technologies there has been a considerable advancement in experimental platforms for both in vivo and in vitro binding measurement.

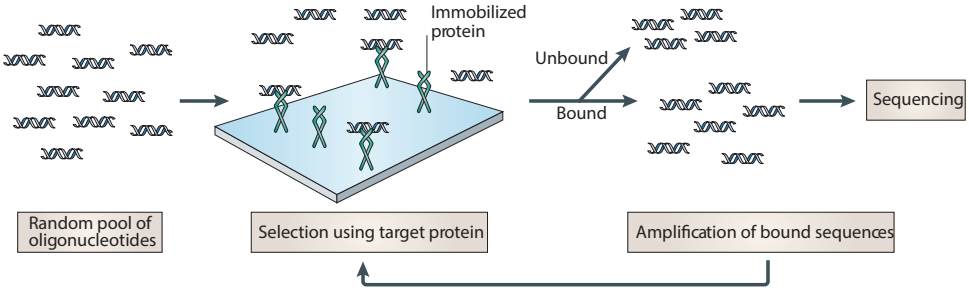
In vivo binding- Experiments performed inside a living cell are termed in vivo. Chromatin immunoprecipitation (ChIP) [47, 60] is used for assaying protein–DNA binding. In ChIP, antibodies corresponding to a protein of interest enrich fragments of genomic DNA to which the protein is bound. These enriched fragments can be analyzed by either high-density microarrays (ChIP-chip) or next-generation sequencing (NGS) (ChIP-Seq — ChIP followed by DNA sequencing) [7,31,38,40,58]. ChIP captures genomic regions bound by protein under the influence of all the cellular components that are result of state and type of cell.



(a) Protein Binding Microarray (PBM) array [9]



(b) Cognate site identification-array (CSI-array) [17]



(c) HT-SELEX (high-throughput systematic evolution of ligands by exponential enrichment)

Figure 1.1: Experimental methods to measure protein–DNA binding in vitro [63]. Array based methods (PBM & CSI) have limited number of different DNA sequences against which protein binding is measured, bound proteins (shown by purple crescents for PBM) is quantified with a fluorescent antibody. Sequencing based methods uses DNA sequences obtained by sequencing DNA fragments bound by a given protein from a random k-mer DNA library (SELEX-Seq or HT-SELEX).

In vitro binding- Experiments performed outside a living cell environment are termed as in vitro. Cognate site identification-array (CSI-array) [17,30,64,66], protein binding microarray (PBM) [9], HT-SELEX (high-throughput systematic evolution of ligands by exponential enrichment) [32,33], SELEX-Seq (SELEX with massively parallel sequencing) [27,45,59] are among the several High-Throughput (HT) in vitro binding methods developed to study individual protein's affinity and specificity for DNA binding. For example, CSI measures binding intensity of a given protein to all 8-mers, i.e. all permutation of A, C, G, T nucleotides of length 10 (more than 1 million sequences) on a microarray (Figure 1.1b). Similarly, PBM measures biding intensity to all 8-mers using 35bp long features (Figure 1.1a). The benefits of using array based methods are a) the user gets to choose the sequences and thus can focus on binding of sequence of interest, and b) the measured output is fluorescent intensity proportional to protein–DNA binding. In sequencing based methods (Figure 1.1c) DNA from various binding experiments can be tagged using DNA barcodes and mixed together to analyze multiple experiments in one sequencing run, making it easier and cheaper to measure binding corresponding to 100s of TFs at a time.

1.2 Analysis of protein–DNA binding

Many computational methods have been developed to study protein–DNA binding data. Most of these methods reduce binding data into a matrix with a small number of variables by displaying it in the format of **position weight matrix (PWM)** motif [57,62], by using only highly preferred sequence for determining specificity, thus leaving out the information about less preferred binding sequences for motif determination [14,17,64]. In the PWM model a score is assigned to each possible base (nucleotide) at each position in the binding site. The total binding score for a sequence is the sum of values corresponding to each nucleotide of that sequence (Figure 1.2). Thus the model can score all possible binding sites for the protein. The 'logo' provides a convenient graphical representation of PWM. PWMs are used for simplicity to visualize and to remove experimental noise. In Figure 1.2 a PWM corresponding to a protein is represented with GCGTGG as the best binding sequence, this PWM model would give best binding score (negative of log normalized score) of 0 to sequence GCGTGG and 0.8 to GCGGGG. Height of each nucleotide in PWM logo represents the information content, taller the nucleotide higher is the protein's sequence specificity at that position. The PWM model makes it easier to distinguish preferred over non-preferred sequences. MEME (Multiple expectation maximization for motif elicitation) [4,5] is extensively used for a lot of the ChIP based experiments. Different computational methods have been developed to analyze specific types of binding data obtained (section A.1). In Table A.1, a few of the best tools to analyze TF binding affinity suited to different

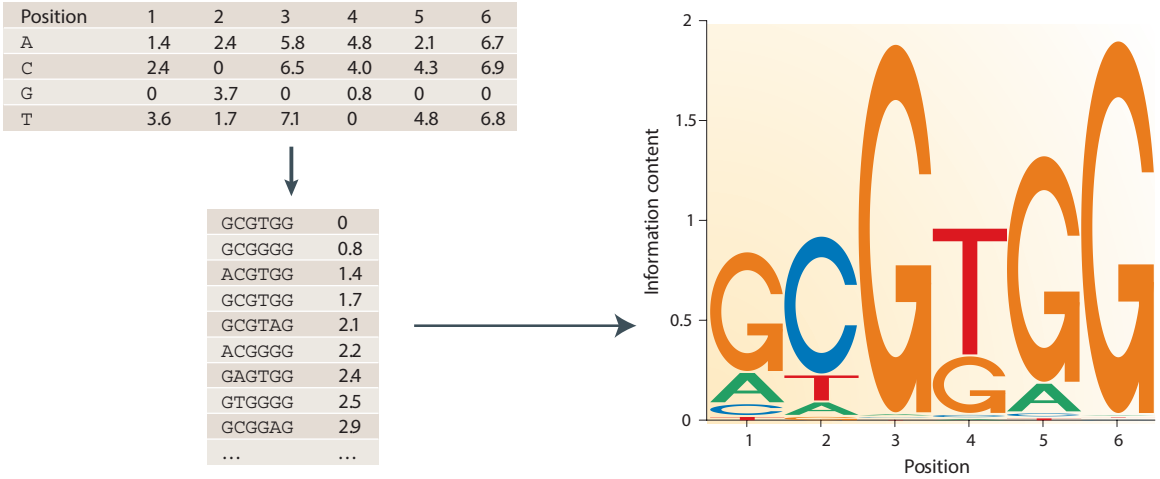


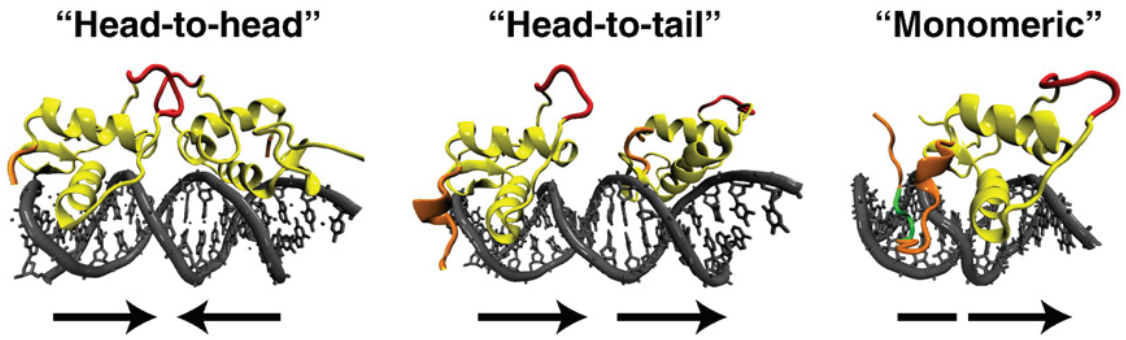
Figure 1.2: Position weight matrix (PWM) logo representation and scoring [63] (Represented -log(Normalized Score))

experiments are presented. A comparison between a few of them for PBM binding data is also presented in Weirauch et al. [68]. In the following subsections we discuss different aspects of protein–DNA binding that are not completely modeled by current computational methods.

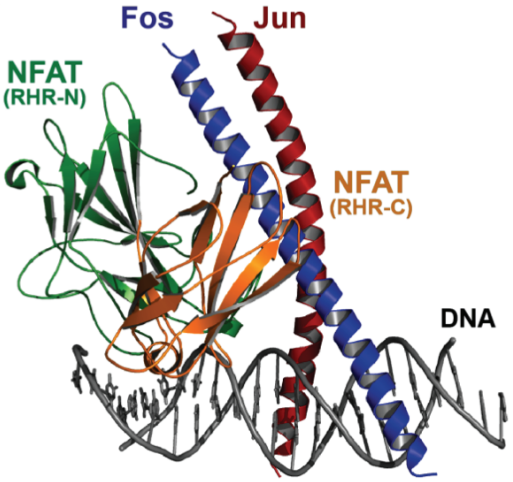
Complexity of protein–DNA binding

TFs can have very complex DNA binding, some bind DNA as monomers[1], homo-dimers, hetero-dimers, and these dimers can have a gap of variable length between the binding (Figure 1.3a). This leads to multiple binding patterns exhibited by one protein. For example, NFAT protein binds DNA – a) just by itself to form NFAT monomer – GGAA, b) as NFAT dimer by dimerizing with another NFAT molecule – binds DNA as GGAA-2gap-TTCC and TTCC-2gap-GGAA, where TTCC is reverse complement of GGAA, c) as a hetero-dimer by dimerizing with complex of two proteins Fos and Jun also known as AP1 – binds DNA as GGAA-gaps-TGACTCA with different preferences of each monomer (Figure 1.3b & 1.3b). A single PWM cannot capture these multiple complex binding patterns. Most of the current computational methods focus on finding a single PWM model or multiple PWM models to fit same best binding motif, leaving out the lower affinity binding patterns (Figure 3.16a). Given the depth of current experimental techniques, just using the best binding sequences is incomplete. There is a need to further explore the complexity of such binding dataset.

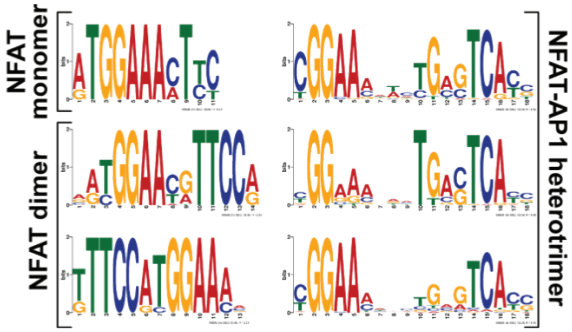
[1]'mer' represents a single protein molecule. Thus if a) single protein molecule binds DNA alone that is called monomer binding, b) if binds DNA with another molecule of same protein known as homo-dimer DNA binding, and c) if binds DNA with other protein molecules then known as hetero-dimer binding



(a) Protein binding to DNA in multiple different ways [1]: Examples of nuclear receptor proteins that bind DNA in multiple orientations.



(b) Structure of NFAT-AP1-DNA complex

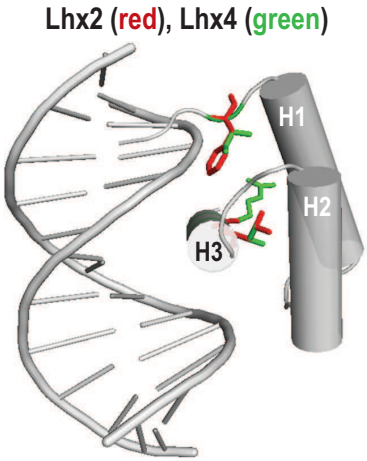


(c) NFAT-preferred DNA binding motifs

Figure 1.3: Complexity of protein–DNA binding

Comparing DNA binding of proteins from the same family

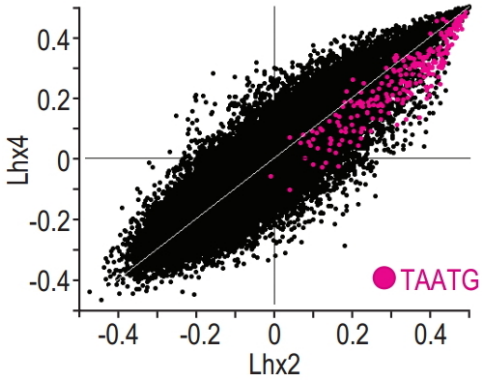
TFs from the same family often have very similar preferences of binding to DNA, although they may significantly differ in functional regulation. For example Lhx2 & Lhx4 protein from homeodomain protein family have similar structure and differs only in their non-DNA contacting residues and they prefer similar DNA sequences, but still both are functionally different, Lhx2 is responsible for blood cell development [51], whereas Lhx4 is involved in pituitary gland development. Figure 1.4a shows the interaction model between DNA and proteins Lhx2 and Lhx4 and Figure 1.4b & 1.4c displays their DNA binding preferences as PWM [9]. Figure 1.4d and 1.4e highlights sequences preferred by Lhx2 and Lhx4 respectively. We were able to find these differences using our method DiSEL (Difference Sequence Specificity Landscape) [11]. Thus a simple PWM model (Figure 1.4b & 1.4c) is not enough to capture such differences.



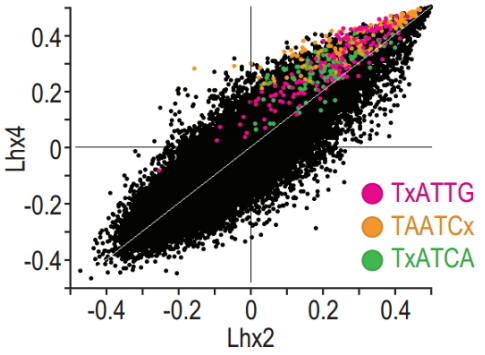
(a) Protein–DNA interaction of Lhx2 & Lhx4: On the left is double helix DNA, which is interacting with protein on right side. Differences are shown in red (Lhx2) and green (Lhx4), rest of the protein sequence is same.

(b) Lhx2 DNA binding PWM logo [8]

(c) Lhx4 DNA binding PWM logo [8]



(d) Scatter plot of DNA binding of Lhx2 Vs Lhx4- showing DNA sequence preferred by Lhx2



(e) Scatter plot of DNA binding of Lhx2 Vs Lhx4- showing DNA sequence preferred by Lhx4 (where x = nucleotide C,G or T)

Figure 1.4: Comparison of Lhx2 & Lhx4 DNA binding [11]

1.3 Contribution

We make the following primary contributions to address the problems discussed above pertaining to protein–DNA binding.

1. **Introducing MinSeqs –** Inspired by the notion of minterms and implicants and the associated methods in digital circuit optimization, we develop a novel representation to characterize the binding preference of a TF using a concept called Weighted Minterm Sequences (or MinSeqs). The set of MinSeqs collectively capture the detailed binding profile of a TF from the information dense experimental data. Various complex binding patterns with protein partners are captured in the set of MinSeqs. Different weights assigned to MinSeqs enable them to capture not only the best affinity sites but also medium and low affinity sites, giving a full binding profile [11, 14, 17, 23, 24, 54, 64].

2. **COSME to capture protein–DNA binding –** By leveraging recent advances in compressed sensing (CS) section of signal processing, we develop a method to extract MinSeqs from microarray binding data, we term the method as compressed sensing based motif extraction (COSME). COSME captures various high, medium and low affinity binding sites in form of MinSeq set, which can then be used to predict binding in genomic sites. COSME gives a better characterization of DNA binding when compared to existing computational tools [3, 68].

3. **Diff-COSME to explore binding differences in related proteins –** To elucidate binding differences among related proteins we develop differential-COSME (Diff-COSME). By capturing the complete profile of binding, Diff-COSME is able to discover subtle differences in binding of similar proteins missed by other computational methods. These subtle binding differences can lead to a change in the gene network of the cell leading to phenotypic variations or diseases in humans.

4. **MinSeqFind to capture complex DNA binding motifs with partners–** We introduce MinSeqFind, a novel method to capture complex DNA binding motifs from HT-SELEX binding experiment. MinSeqFind explores longer binding patterns, with and without gaps, and capture multiple partners of proteins in the form of MinSeqs. MinSeqFind not only captures such complex motifs in the form of MinSeqs, but it also filters out many platform related biases like bias introduced due to the primer binding region and PCR bias, that were previously not addressed. When applied to published datasets, MinSeqFind captures known as well as novel secondary binding motifs missed by state of the art computational methods [2].

5. **Effect of protein partners and ligands on nuclear receptor's DNA binding–** The nuclear receptor (NRs) family of TFs is one of the most common classes of drug

targets (13% of all FDA-approved drugs), which has been studied extensively for decades. In the presence of ligands (drugs) or protein partners, a given NR activates or represses transcription of certain genes. Disruption in NR function has been implicated in several diseases, including allergies, asthma, obesity, several forms of cancer and diabetes [25]. To elucidate the disruption in DNA binding of NRs due to a protein partner and ligands, we measure their DNA binding with and without ligands and partner proteins. From HT-SELEX binding measurements of NRs, MinSeqFind captures many known binding patterns and at the same time reveals novel binding patterns. MinSeqFind captures variation in binding due to drugs or ligands, also finds multiple binding patterns of RXRA protein with different binding protein partners. MinSeqFind also identifies genetic variants that alter NR binding. In fact, we find more than 300 genetic variants associated with disease that alter NR binding. These findings have direct implications on mechanisms of diseases like cancer, diabetes and cardiovascular diseases.

MinSeqs serves as a representation tool to capture the binding patterns of various TFs as well as a tool to differentiate binding of TFs having subtle variation in their binding. Such newly found patterns and differences can be used to assemble new gene networks. In chapter 2 we first introduce MinSeqs and explain the COSME and Diff-COSME algorithm and explore differences found by Diff-COSME. In chapter 3 we extend our approach to find complex DNA binding motifs using MinSeqs for HT-SELEX data for nuclear receptor proteins. In the last chapter we discuss the possibility of extending MinSeqs to in vivo binding data and using shape features for MinSeqs.

Chapter 2

COSME & Diff-COSME to Elucidate Protein–DNA Binding

TFs have one or more DNA binding domain (DBD) which contains a structural unit that recognizes DNA. Different parts of the DBD interacts with only specific nucleotides making DNA binding as sequence specific. The DBD acts as an independent unit and binds DNA and mostly one part of DBD get affected by DNA binding of other part and thus their binding is dependent on each other. Until a decade ago due to lack of high-throughput DNA binding platforms, it wasn't possible to study all such dependency and thus, the PWM served as a great tool to capture DNA binding where each nucleotide binding to DBD of protein is independent of each other [61, 63]. With the advent of many high-throughput DNA binding methods, we don't have to limit ourselves and we can explore nucleotide dependencies as well. We develop weighted Minterm Sequences or MinSeqs as a representation tool to capture such nucleotide dependencies, inspired by the notion of minterms, implicants and the associated methods in digital circuit optimization.

To extract MinSeqs from microarray DNA binding data, we use compressed sensing (CS) – a recently developed signal processing technique [16] for efficiently acquiring signal by finding solutions to underdetermined linear system. CS is a fast growing field and many applications of CS are being implemented in diverse disciplines. We formulate our problem of capturing DNA binding of TF into an underdetermined linear system by defining a linear model of binding. We further exploit CS algorithms like compressive sampling matching pursuit (CoSaMP) for sparse solution of MinSeqs. The solution provided here is applicable not just for comparing TF binding, but also finding new information regarding a given TF as well. We develop the COSME algorithm that focuses on fully characterizing DNA binding of TFs using MinSeqs and Diff-COSME algorithm that finds differences in the binding preferences of TFs belonging to the same family.

2.1 MinSeqs to represent protein–DNA binding

A MinSeq is a k-mer comprised of sequence of A, C, G, T, and N, where N represents any nucleotide A, C, G, or T. A gap is considered in the form of N as many proteins like to bind DNA with gaps in middle. For example, the sequence ACGNTGA is a 7-mer-Minseq with ACG followed by single nucleotide gap (N), which is followed by TGA. Each MinSeq has an associated binding intensity or weight to it, which is related to the affinity of the TF for that sequence. Thus, a set of MinSeqs captures sequences of different affinity without leaving out any medium- or low- affinity sequences.

2.2 Modeling TF-DNA binding

We model the binding preference of a TF as follows. Let us consider there are T TF to be compared, which has similar PWM binding motif. Let y_t denote the $M_t \times 1$ vector of probe intensities obtained from a microarray after normalization of array based defects for t^{th} TF. Thus, M_t is the number of probes in the microarray corresponding to t^{th} TF. In the case of PBM data set, M_t is approximately 40,000.

Let B denote background for scale shifting and S denote the set of all variables needed to characterize the probe intensities, such a set S is termed as MinSeq set or set of MinSeqs. 'MinSeq' term is inspired by the notion of 'minterms' from digital circuit optimization. For instance, for the TFs in the PBM data set, S may include all possible k-mer MinSeqs for $k = 4, 5, \ldots, 10$. Many TFs bind as dimers and their binding preferences may contain gaps, e.g., 4-mer, followed by two gaps, followed by another 4-mer, thus considering in the set S all possible sequences of length of 4 to 10 including a certain number of gaps in the middle (gapped k-mer MinSeqs). Thus, the cardinality of S may easily exceed 10 million. Let H_t denote a sparse "mapping matrix" representing the relationship between elements in S and probes. More formally, let $(h_t)_{ij}$ equal the number of times element $j \in S$ occurs in i^{th} probe of microarray for t^{th} TF.

The mapping matrix H will have rows equal to the number of probes, and the number of columns corresponds to the possible MinSeqs in S. Thus, the column entries corresponding to each 4-mer MinSeqs will be: AAAA, AAAC, ...,TTTT and similarly 5-mer, ..., 10mer. For example, in Figure 2.1 consider a 25mer probe AATGACATGACTGACATAAAAAACG. The mapping matrix will have non-zero entries for columns corresponding to 4-mer MinSeqs AATG, ATGA, ..., AACG rest will be 0. Similar counting is done for all 5-mer MinSeqs and so on as represented. If single gapped 4-mer MinSeqs are also considered in set S, then there will be non-zero entries corresponding to ANTGA, AANGA, AATNA (derived from 1st 5-mer AATGA) and ANGAC, ATNAC, ATGNC (derived from 2nd 5-mer ATGAC) and so on. Similar mapping is done for all the probes.

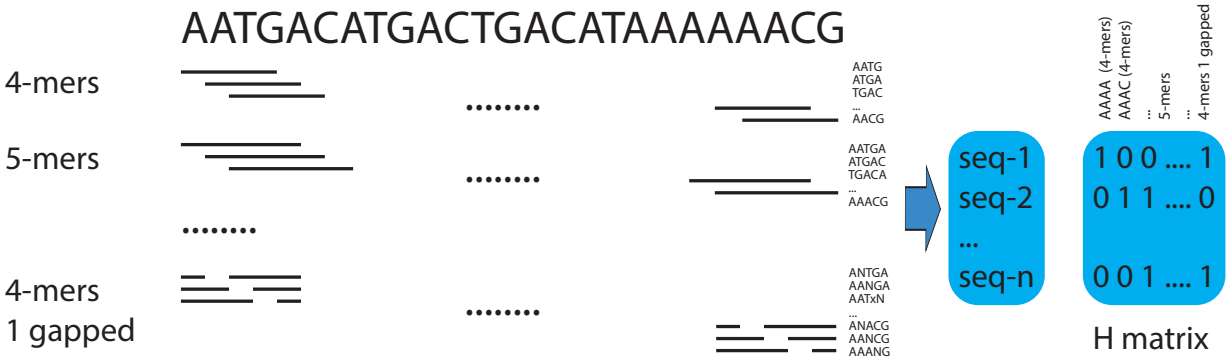


Figure 2.1: Construction of design matrix H - h_{ij} = number of times MinSeq $j \in S$ occurs in i^{th} probe of microarray

Let vector x_t denote the $N \times 1$ vector of estimated binding preferences of t^{th} TF with the i^{th} element as MinSeq x_{it} representing binding preference for i^{th} MinSeq and where $N = |S|$. We model the relationship between x_t and y_t as an underdetermined linear system of the form -

$$y_t = H_t x_t + \eta_t \tag{2.1}$$

where, η_t is the estimation error.

For T TFs to be compared there are some binding preferences common to all the TFs and some binding preferences that are characteristic of an individual TF. Thus, after scaling and shifting y_t to same scale, we divide x_t into c and d_t. Where c is binding preference common to all TFs and d_t represents individual binding preference corresponding to t^{th} TF.

$$\begin{aligned}
y_1 &= H_1 x_1 + \eta_1 = H_1 c + H_1 d_1 + \eta_1 \\
y_2 &= H_2 x_2 + \eta_2 = H_2 c + H_2 d_2 + \eta_2 \\
&\vdots \\
y_t &= H_t x_t + \eta_t = H_t c + H_t d_t + \eta_t \\
&\vdots \\
y_T &= H_T x_T + \eta_T = H_T c + H_T d_T + \eta_T
\end{aligned} \tag{2.2}$$

which can be represented in a combined form as-

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_t \\ \vdots \\ y_T \end{bmatrix} = \begin{bmatrix} H_1 \\ H_2 \\ \vdots \\ H_t \\ \vdots \\ H_T \end{bmatrix} c + \begin{bmatrix} H_1 \\ 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \end{bmatrix} d_1 + \cdots + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ H_t \\ \vdots \\ 0 \end{bmatrix} d_t + \cdots + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \vdots \\ H_T \end{bmatrix} d_T + \begin{bmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_t \\ \vdots \\ \eta_T \end{bmatrix} \tag{2.3}$$

$$\mathbf{y} = \mathbf{H}c + \breve{\mathbf{H}}_1 d_1 + \cdots + \breve{\mathbf{H}}_t d_t + \cdots + \breve{\mathbf{H}}_T d_T + \boldsymbol{\eta} \tag{2.4}$$

where $\mathbf{y}$, $\mathbf{H}$ and $\boldsymbol{\eta}$ are concatenated matrix of y_t, H_t and η_t $\forall t$ respectively. $\breve{\mathbf{H}}_t$ is a matrix corresponding to H_t such that rest all entries are zero. Equation 2.4 is the first of its kind in the literature to capture the similarity and differences between protein–DNA binding.

2.3 Problem statement & strategy

Problem Statement

In equation 2.4 given $\mathbf{y}$, $\mathbf{H}$ and $\breve{\mathbf{H}}_t$ $\forall t$, find the values c, d_t $\forall t$, minimizing the error $\boldsymbol{\eta}$.

Compressed-MinSeqs to capture protein–DNA binding

c and d_t comes from set S, which spans to millions of MinSeqs, but we don't have so many equations (which is equal to the number of probes), thus it is an underdetermined system. Thus we need to compress our search space from S to a small number of MinSeqs. We use knowledge of protein–DNA binding and algorithms from compressed sensing field for such compression. We call the binding variables thus obtained **compressed-MinSeqs**.

Solution strategy

Common part c can be extracted using different existing PWM based methods (like BEEML, MatrixRECUCE) or k-mer based methods (like Annala et al. [3]). PWM is based on the assumption that protein–DNA binding occurs in a way that each nucleotide position independently contributes to the binding [63,65], but there are many cases that are not captured by PWMs [17,33]. Thus PWMs can't fully capture the common binding, but PWMs are very easy to comprehend. k-mer based methods assume complete dependence between each nucleotide position and are of the format k-mer and corresponding intensity and need many such MinSeqs (of the order of thousands) to capture binding, but they limit their search space to top few binding sequences only and also difficult to comprehend. We chose to use PWM to capture the position-independent component, and the rest MinSeqs from set S (MinSeqs). S has millions of MinSeqs (much more that other k-mer based methods) and thus becomes underdetermined system, we use compressed sensing to acquire a small subset of S which is significant for binding. In order to apply CS we need to scale MinSeqs. Together PWM and MinSeqs define C, which are easy to comprehend.

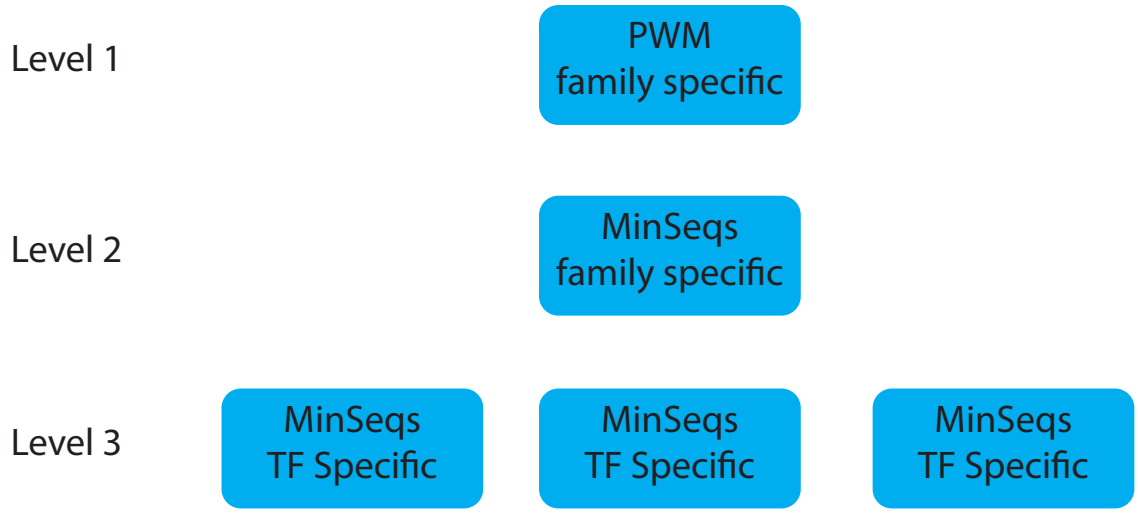


Figure 2.2: Breakdown of PWMs and MinSeqs for TFs having similar DNA binding

Currently there is no existing method to extract individual preferences d_t. We use compressed sensing to capture subset of S unique to t^{th} TF, we assume that most of the position-independent components are captured by the common PWM.

First we solve for c by considering d_t $\forall t$ equal to zero from equation 2.4 and then solve for d_t by using independent equation 2.2. Figure 2.2 displays the intended output from the method. It shows breakdown for PWM and MinSeqs for TFs of same family having similar DNA binding preferences. First two levels c - PWM and family specific MinSeqs contain information about binding which is same for all the TFs, third level d_t contains information which is unique to each TF. This way we'll have representation of what is common and what is different in their DNA binding preferences. To find characteristics of the individual TF we leave out the 3rd level d_t.

Estimating DNA binding of TF for new set of sequences

Our aim here is to first solve for x_t in equation 2.1 to get estimate $\hat{x}_t = \hat{c} + \hat{d}_t$, and using that, binding for a any new set of sequences can be predicted for TF t as follows-

$$\begin{aligned} \tilde{y}_t &= \tilde{H}_t \hat{x}_t \\ &= \tilde{H}_t \hat{c} + \tilde{H}_t \hat{d}_t \end{aligned} \tag{2.5}$$

Where $\tilde{y}_t$ is the predicted intensity and $\tilde{H}_t$ is mapping matrix for the new set of sequences. Thus, once a good estimate for $\hat{x}_t$ is made, that information is sufficient for a TF t to get binding to any new sequence. Until recently, PWMs were only used for this estimation, which failed to identify differences between related TFs.

2.4 Common binding

To solve for c, we first force $d_t\ \forall t$ to 0 in equation 2.4, which reduces then to-

$$\mathbf{y} = \mathbf{H}c + \boldsymbol{\eta} \tag{2.6}$$

Since we are calculating c ignoring effect of d_t, for a set of TFs above equation will give us an initial estimate for c, which has to be improved iteratively with d_t, which will be discussed in next section. But if the goal is to characterize only single TF binding and not to compare (means no d_t), in that equation 2.6 will provide a complete characterization of binding.

Further c is divided into b and n to capture position independent binding by PWM and k-mer dependent part respectively.

$$\mathbf{y} = \mathbf{H}b + \mathbf{H}n + \boldsymbol{\eta} \tag{2.7}$$

PWM solution

Considering only PWM part, equation 2.7 reduces to

$$\mathbf{y} = \mathbf{H}b + \boldsymbol{\eta} \tag{2.8}$$

If we consider PWM of length 8 (for PBM) then 8-mer MinSeqs in b contains non zeros entries (in addition to background); rest are all zero (4-mer, 5mer etc, and all gapped k-mers). Separating background B and PWM 8-mer from b (rest are zeros) and by matrix operations above equation reduces to summation of binding affinity due to all 8-mers in a particular probe as-

$$\mathbf{y}_i = \sum_{k=1}^{L_s - L_p + 1} I(probe(k : k + L_p - 1)) + B + \boldsymbol{\eta}_i \tag{2.9}$$

where $\mathbf{y}_i$ is measured binding intensity of probe sequence i, L_s is length of probe sequence, L_p is length of PWM matrix (in this case it is 8) and $I(probe(k : k + L_p - 1))$ is PWM binding affinity for an 8mer sequence from k-th position to $k + L_p - 1$. B corresponds to shift for background normalization. PWM assumes each nucleotide position has independent binding effect thus $I(probe(k : k + L_p - 1))$ can be calculated by following formula (similar to Figure 1.2) [26].

$$I(probe(k : k + L_p - 1)) = M \prod_{l=1}^{L_p} w_{l,probe(k+j-1)} \tag{2.10}$$

where M is scaling factor. A reference sequence S_{ref} is used for above PWM model, which is taken as the best median 8-mer then, $w_{l,b}$ is the multiplicative change corresponding to free energy change due to mutation at l-th position to nucleotide base b. Thus if base b is same as in S_{ref}, $w_{l,b}$ will be equal to 1, and if whole 8-mer is S_{ref} then $I(S_{ref}) = M$. Thus the scaling factor M corresponds to binding intensity of S_{ref} (after background shift B). So the PWM model can be represented using $3L_p$ variables (given S_{ref}), since at each nucleotide position $w_{l,b}$ corresponding to S_{ref} is 1, rest 3 has to be computed.

Thus equation 2.9 can be rewritten as-

$$\mathbf{y}_i = M \sum_{k=1}^{L_s-L_p+1} \prod_{l=1}^{L_p} w_{l,probe(k+j-1)} + B + \boldsymbol{\eta}_i \tag{2.11}$$

Now we can use the least square fit (MatrixREDUCE [26]) to above equation[1] to get M, F and $w_{l,b}$ which provides solution to PWM part of equation 2.7. Other methods such as BEEML-PBM [72] and DeepBind [2] can also be used for PWM solution for data from microarrays.

The PWM is widely accepted as a standard to represent binding pattern as it captures nucleotide position independent binding in a visually-intuitive display [57]. But as stated earlier the PWM fails to capture many details, and thus it is not sufficient for analyzing data from high-throughput methods. Thus, next we capture nucleotide dependent binding.

COSME to capture protein–DNA binding

Candes and Tao [15] [16] proved that given the knowledge about a signal's sparsity, the signal may be reconstructed with fewer samples than the Nyquist-Shannon theorem requires. This idea is the basis of compressed sensing and we use this for solving the second part of equation 2.7-

$$\mathbf{y} = \mathbf{H}\eta + \boldsymbol{\eta} \tag{2.12}$$

Equation 2.12 is a linear underdetermined system, with H and η being sparse. Here we assume sparsity of η using knowledge of protein–DNA binding.

Considering our signal to be sparse, we search for the sparsest signal η which yields $\mathbf{y}$ within a given error limit θ-

$$\hat{\eta} = \arg\,|_\eta \, \|\eta\|_0 \text{ s.t. } \|\mathbf{y} - \mathbf{H}\eta\|_2^2 \leqslant \theta \tag{2.13}$$

where $\|\eta\|_0$ is l_0 norm of vector η and defined as the number of nonzero entries in η.

[1] Maximum binding intensity of sequence or it's reverse complement is used for PWM prediction for equation 2.10 & 2.11.

Note that $\hat{n}$ is a subset of S which needs to be determined. The solution to equation 2.13 is L_0 norm minimization, which leads to NP-hard problem. Since the variable space is large, that is not computationally feasible.

Effective solutions for this formulation can be obtained from recent research results in the area of compressed sensing. Compressive sampling matching pursuit (CoSaMP) [41] [22], which is an algorithm, came from CS and provides an approximate solution to the problem. Since n contains information about different length MinSeqs (with and without gaps), thus there is some inherent dependency (non-orthogonal vectors) like TGAC, TGACA, TGACAT, will be affected by each other. We chose $2^{(\text{Length of kmer})}$ weight to counter that, which corresponds to minimum mean square error (discussed in next chapter) and n is represented as $n = Dp$, where D is a diagonal matrix whose entries are $2^{(\text{Length of kmer})}$ corresponding to each k-mer. Thus equation 2.12 becomes-

$$\begin{aligned} \mathbf{y} &= \mathbf{H}Dp + \boldsymbol{\eta} \\ &= \mathbf{C}p + \boldsymbol{\eta} \end{aligned} \tag{2.14}$$

where $\mathbf{C} = \mathbf{H}D$ and $c_{i,j}$ = number of times j-th k-mer found in i-th probe$*2^{(\text{Length of kmer})}$.

Since the MinSeqs are non-orthogonal, whenever one MinSeq is added or taken away from the set it affects on all other MinSeqs. Thus we add one more step of recalculation after pruning in CoSaMP to get modified CoSaMP, which is used to solve equation 2.14.

Algorithm 1 Modified CoSaMP - Compressed sensing based motif extraction (COSME)

Input: matrix $\mathbf{H}$, diagonal matrix D, measurement vector $\mathbf{y}$, sparsity K, max iteration i_{max}
Output: K-sparse approximation $\hat{p}$ to true signal p
  Initialize: $\hat{p}_0 = 0, \mathbf{r} = \mathbf{y}, i = 0, \mathbf{C} = \mathbf{H}D$
  while $(i < i_{max}$ & $supp(\hat{p}_{i-1}) \neq supp(\hat{p}_i))$ **do**
    $i \leftarrow i + 1$
    $e \leftarrow \mathbf{C}^T\mathbf{r}$ {form residual signal estimate}
    $\Omega \leftarrow supp(\tau(e, 2K))$ {prune residual}
    $T \leftarrow \Omega \cup supp(\hat{p}_{i-1})$ {merge supports}
    $b|_T \leftarrow \mathbf{C}_T^{\dagger}\mathbf{y}, b|_{T^C} \leftarrow 0$ {form signal estimate}
    $\hat{p}_{temp} \leftarrow \tau(b, K)$ {prune signal using model}
    $T_2 \leftarrow supp(\hat{p}_{temp})$ {get support of pruned signal}
    $\hat{p}_i \leftarrow \mathbf{C}_{T_2}^{\dagger}\mathbf{y}, \hat{p}_i|_{T_2^C} \leftarrow 0$ {reestimate signal}
    $\mathbf{r} \leftarrow \mathbf{y} - \mathbf{C}\hat{p}_i$ {update measurement residual}
  end while
  return $\hat{n} \leftarrow \hat{p}_i/D$

The proposed algorithm 1 (COSME) is for computing sparse non-orthogonal variables (MinSeqs) for our system, where $\tau(e, K)$ denotes a thresholding operator on e that sets all but the K entries of e with the largest magnitudes to zero, and $b|_T$ denotes the

restriction of b to the entries indexed by T. Here K is number of non-zero entries used as input.

The algorithm starts with null set $\hat{p}_0$ and $\mathbf{y}$ as residual signal. At every iteration new 2K MinSeqs are selected by taking top 2K rows of $\mathbf{C}^T\mathbf{r}$. These 2K MinSeqs are merged with any MinSeqs carried from previous iteration, this set is estimated using $\mathbf{y}$ intensity and $\mathbf{C}$ matrix using least square solution. Top K MinSeqs from those according to magnitude are picked then re-estimated and kept for next iteration and residual signal is calculated. Algorithm is terminated after maximum iteration is reached or consecutive two iterations result in same support (set of MinSeqs) and the final set of MinSeqs (compressed-MinSeqs) is used as estimated output. Here the set of indices corresponding to the nonzero entries is denoted by support of θ i.e. $\text{supp}(\theta)$.

The benefit of such a solution space is that it is not restricted to 2 or 3 related sequences as most of the PWM methods are, so it can capture completely different sequences and sequences of variable length. Also it does not post any restriction based on absolute intensity of single probe. By including the MinSeqs with gap, it covers further more binding patterns. If there are more sequences which one wants to characterize, can add to variable space accordingly. A lot of TFs bind DNA in a manner such that each nucleotide position contribute independently to binding. Such binding is captured by PWMs in lesser variables. COSME alone will take lot of MinSeqs (K) to characterize such binding. So the combination of both PWM and MinSeqs is needed, which is provided in the next sub-section.

Hybrid of PWM & COSME to capture binding

In this section we present a hybrid model, which can capture independent nucleotide binding using PWM - b and dependent ones using COSME - n. Algorithm 2 is the proposed algorithm for estimating PWM and K MinSeqs iteratively. It starts with estimating PWM from $\mathbf{y}$ (to get b) and get residual signal $\mathbf{r}_1$. Then fits K MinSeqs to $\mathbf{r}_1$ using algorithm 1 and estimates $\mathbf{r}_2$ using these K MinSeqs, which is then used for next iteration to re-estimate PWM. The algorithm terminates when maximum iteration reached or correlation between consecutive calculated b is above a threshold. The output from the algorithm also includes M, B since that provides a scaling and shift between PWM and the variable values n. The proposed algorithm does not provide any mathematical guarantee to converge, but according to our understanding of protein–DNA binding behavior and different test data, it is expected to converge.

The proposed changes are that – some other PWM estimation method can be used to replace the one used in this algorithm. Also if there are multiple binding sites, accordingly more PWMs can be estimated instead of a single PWM. This combined solution is ideal for capturing the benefits of both models.

Algorithm 2 PWM + CoSaMP

Input: matrix $\mathbf{H}$, diagonal matrix D, probe intensities $\mathbf{y}$, sparsity K, max iteration for combined algorithm i_{max}, max iteration for COSME j_{max}, maximum PWM difference θ

Output: PWM + K-sparse approximation $\hat{p}$

Initialize: $\mathbf{r}_2 = \mathbf{y}, i = 0$
while true **do**
  $i \leftarrow i + 1$
  {PWM,M,B}$\leftarrow$ PWM_calculation(r_2) {PWM estimate from residual signal}
  $\hat{b_i} \leftarrow$ b_calculation(PWM,M,B)
  $\mathbf{r}_1 \leftarrow \mathbf{y} - \mathbf{H}\hat{b_i}$ {residual from i-th PWM}
  $\hat{n}_i \leftarrow$ COSME$(\mathbf{H}, D, \mathbf{r}_1, K, j_{max})$ {K sparse solution from algorithm 1 }
  $\mathbf{r}_2 \leftarrow \mathbf{y} - \mathbf{H}\hat{n}_i$ {residual from i-th iteration of K MinSeqs}
  if correlation$(\hat{b_i}, \hat{b_{i-1}}) > \theta$ or $i >= i_{max}$ **then**
    break
  end if
end while
return $\hat{n} \leftarrow \hat{n}_i$ & {PWM,M,B}

2.5 Diff-COSME to prune out differential binding

Many TFs belong to the same protein family, and have similar but not the same binding preferences. Characterizing the differences in the binding preferences of TFs in the same family is critical for understanding the differences in their function and their role in genetic regulation. However, as mentioned earlier, none of the existing computational technique for characterizing the DNA binding preferences of TFs focus on differences in the binding of similar TFs. Instead, they mostly focus on the characterizing the primary binding preferences, which in turn, may be identical for several TFs in the same family. Thus, we develop a novel method to extract binding differences in related TFs.

Here we discuss estimation of common PWM and MinSeqs for multiple TFs–DNA binding data. In this section differences d_t unique to t^{th} TF is calculated by extending algorithm 2. The proposed algorithm Diff-COSME (Differential-COSME or Difference by COSME) is described in Figure 2.3 as a flowchart to find differences between DNA binding of similar T TFs. Here output is one PWM and K_1 MinSeqs $\hat{n}$ common to all T TFs and K_2 MinSeqs $\hat{d}_t$ corresponding to t^{th} TF. Thus inputs to algorithm are – matrix $\mathbf{H}$, diagonal matrix D, array intensity $\mathbf{y}$, sparsity K_1 & K_2 corresponding to MinSeqs common to T TFs and individuals respectively, max iteration for combined algorithm i_{max}, max iteration for COSME j_{max1} & j_{max2}, minimum correlation θ to exit iteration.

Algorithm Diff-COSME starts with scaling by determining initial PWM $\mathbf{y}$, by maximizing the sum of correlation corresponding to each protein–DNA binding data. This is done so without using shift and scale information. Shift and scale for each array is calculated using initial PWM fit to each array data and thus shift and scale to bring all

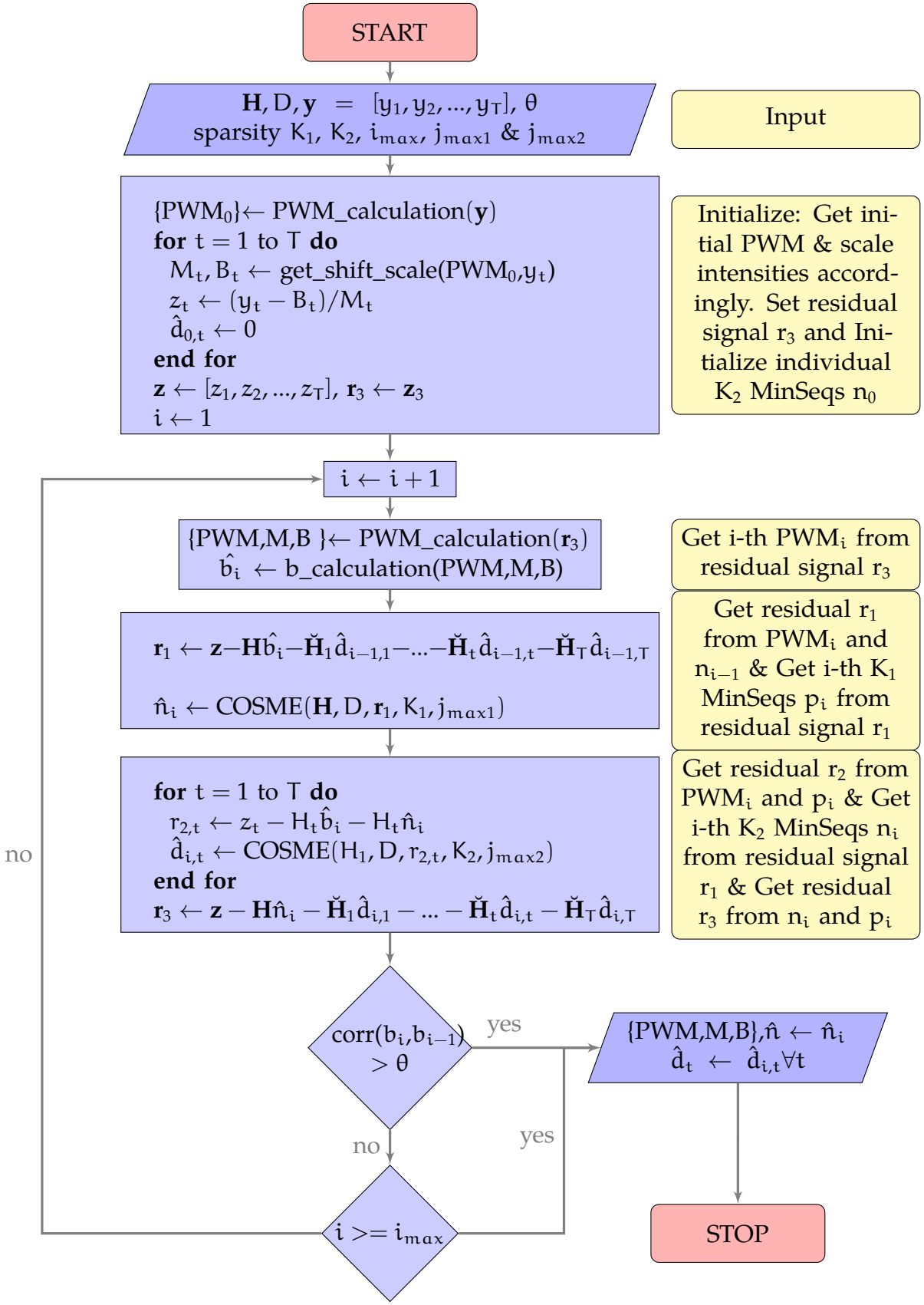


Figure 2.3: Finding differences in DNA binding of related TFs using Diff-COSME

of them to same distribution to get z_t. Using z_t iteratively, first PWM, then common K_1 MinSeqs and then for each individual K_2 MinSeqs are estimated in a round robin fashion using the residual signal from last two signals. 1) While estimating PWM, residual signal is calculated by subtracting common & individual MinSeqs estimate from $\mathbf{z}$ ($\mathbf{r}_3 \leftarrow \mathbf{z} - \mathbf{H}\hat{n}_i - \check{\mathbf{H}}_1\hat{d}_{i,1} - ... - \check{\mathbf{H}}_t\hat{d}_{i,t} - \check{\mathbf{H}}_T\hat{d}_{i,T}$), 2) to estimate K_1 common MinSeqs, residual signal is calculated by subtracting individual MinSeqs and PWM estimate from $\mathbf{z}$ ($\mathbf{r}_1 \leftarrow \mathbf{z} - \mathbf{H}\hat{b_i} - \check{\mathbf{H}}_1\hat{d}_{i-1,1} - ... - \check{\mathbf{H}}_t\hat{d}_{i-1,t} - \check{\mathbf{H}}_T\hat{d}_{i-1,T}$), 3) to estimate K_2 individual MinSeqs, residual signal is calculated for each TF by subtracting common MinSeqs and PWM estimate from $\mathbf{z}_t$($r_{2,t} \leftarrow z_t - H_t\hat{b}_i - H_t\hat{n}_i$). Here $\hat{d}_{i,l}$ and $\hat{n}_i$ are i-th iteration estimate of individual and common MinSeqs respectively. Algorithm terminates when maximum iteration reached or correlation between two consecutive b is more than a threshold. Note that output MinSeqs are scaled to M_t because of initial scaling.

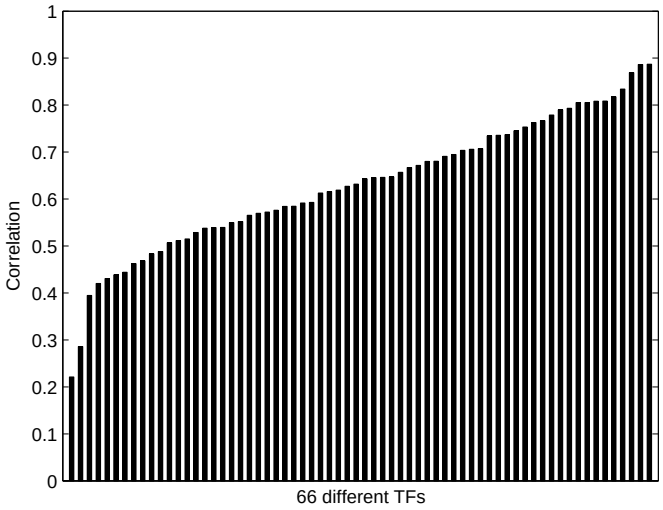
There are 3 outputs corresponding to each protein in this case. PWM and $\hat{n}$ are common to all which captures nucleotide independent and dependent binding which are similar in all the proteins. And another set of MinSeqs $\hat{d}_t$ corresponding to each protein captures dependent binding for individual ones.

2.6 Results

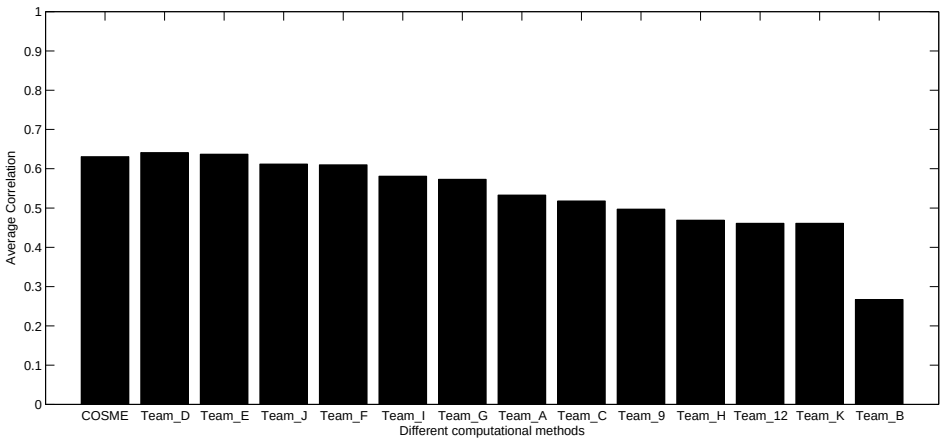
We use MinSeqs extraction using COSME and Diff-COSME for DNA binding of different proteins measured using protein binding microarray. First we apply COSME on data from the challenge posted by DREAM (Dialogue for Reverse Engineering Assessments and Methods) group [68] for PBM to compare performance to existing methods and then Diff-COSME to find differences in proteins exhibiting similar binding profile. MinSeqs not only found differences in similar proteins, but also in different variants of same proteins.

DNA binding of proteins measured by PBM array

We consider PBM data from an online challenge DREAM [68]. In the challenge, there were 2 set of arrays HK and ME for which PBM measurements were done. Both HK and ME contains all 10mers on array following de-bruijn pattern, but with different arrangements. For 20 TFs binding data was provided for both the arrays and for 66 TFs binding data for only one array was provided and for the other it was to be estimated, results from the estimations was used to evaluate different computational algorithms. Normalized array data [3, 68] was used for COSME algorithm 1 to get binding profile of all 66 TFs in the form of MinSeqs. Correlation between estimated and measured binding intensity of 66 TFs shows COSME captures binding information comparative to existing computational tools [68] (Figure 2.4).



(a) Pearson correlation between estimated and measured binding intensity of 66 different TFs



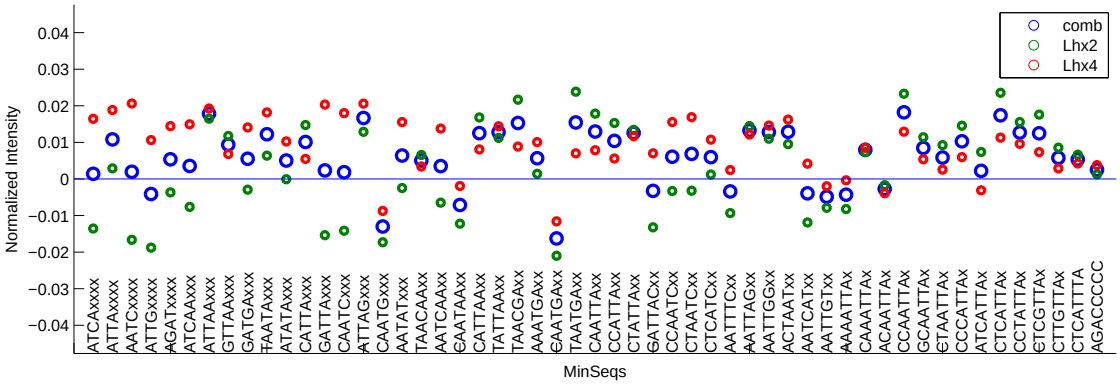
(b) Average pearson correlation between estimated and measured binding for COSME (first bar) and other computational methods

Figure 2.4: COSME to estimate DNA binding intensity

Binding differences between proteins of same family

Diff-COSME (algorithm 2.3) discovers binding differences in two different proteins of the homeodomain family- Lhx2 & Lhx4 [8]. Figure 2.5a shows the common PWM, whereas Figure 2.5b displays common and differential MinSeqs. When two of selected MinSeqs were plotted on a scatter plot with 8-mer binding score, differences became clearly visible and are novel differences missed by previous methods (Figure 2.6a). Lhx2 and Lhx4 differs only in non-DNA contacting residues, yet it still exhibits subtle difference in DNA binding. This observation indicates that even non-DNA contacting residues can play important role in determining DNA binding specificity. Our proposed method not only captures novel differences, but also extracted common binding sequences which PWM failed to capture.

(a) Common PWM



(b) MinSeqs (Common & Differential)

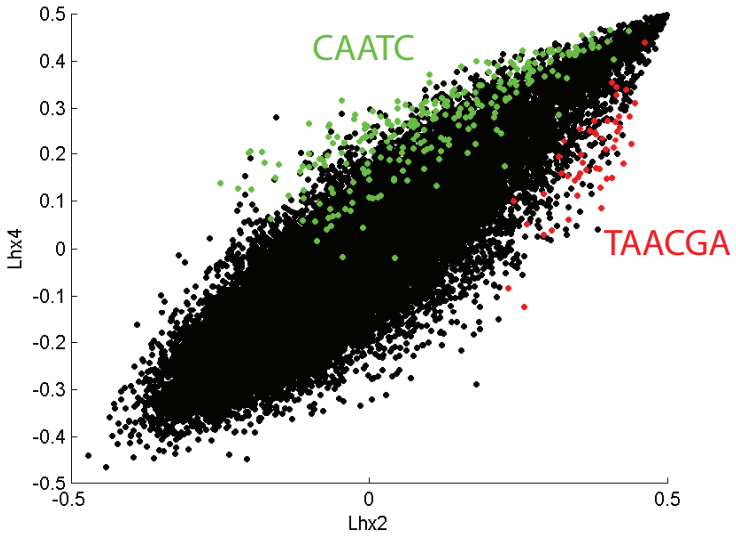
Figure 2.5: DNA binding of Lhx2 & Lhx4 captured as PWM and MinSeqs for Lhx2 & Lhx4

Exploring binding differences in variants of same protein

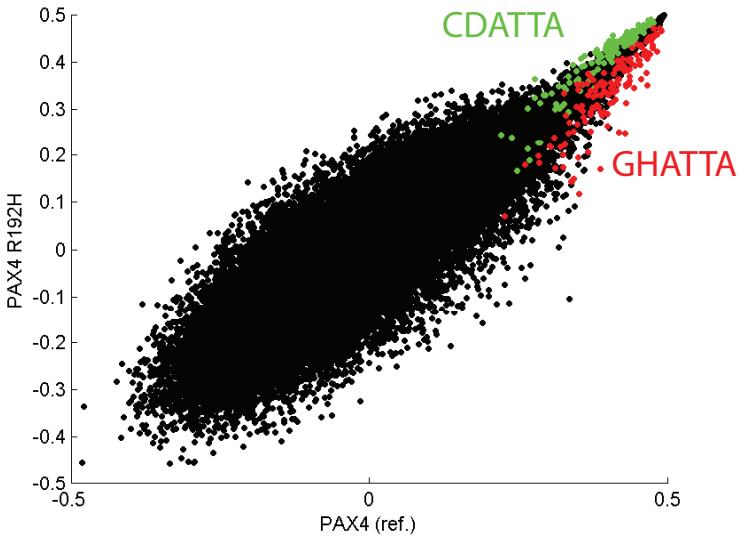
In a recent report on the DNA binding of different variants of proteins was studied [6]. We chose PAX4 protein and it's variant protein PAX R192H. PAX4 protein is a member of the paired box (PAX) family of transcription factors critical for fetal development and cancer growth. Diff-COSME found novel binding differences, and couple of results from analysis are plotted as scatter plot of 8mer binding data in Figure 2.6b. PAX4 protein prefers GHATTA whereas CDATTA preferred by PAX R192H (where D =A, G or T & H = A, C or T).

2.7 Conclusion

In this chapter we introduce a novel method to represent protein–DNA binding, called MinSeqs, that is a set of sequences with A, C, G, T or N and have corresponding binding intensity assigned to them. We also develop compressed sensing based motif extraction



(a) Binding differences for Lhx2 & Lhx4 protein from homedomain family: TAACGA preferred by Lhx4, whereas AATTA by Lhx2



(b) Binding differences for PAX4 (ref.) and PAX R192H variant: CDATTA preferred by PAX R192H whereas GHATTA by PAX4 (ref.) (where D =A, G or T & H = A, C or T)

Figure 2.6: Differences in binding of protein of same family and different variants of same protein (8-mer E Score plotted)

(COSME) as a tool to extract MinSeqs from protein–DNA binding data obtained from microarray. To find differences in binding of proteins of the same family, we also develop Differential-COSME or Diff-COSME. These algorithm captures protein–DNA binding and differences in the form of MinSeqs, when applied to published array binding data we find novel insights. The novel binding patterns and differences are the characteristics of corresponding proteins, which can be reflected in the form of various cellular functions.

Chapter 3

MinSeqFind Discovers Complex Protein–DNA Binding Motifs

DNA binding of proteins can be very complex. Many nuclear receptors (NRs) bind DNA with multiple preferences of monomer, homo-dimer and hetero-dimer with different gaps and orientations (Figure 1.3a) [19,21]. Microarray based experimental methods like PBM and CSI-array fall short in terms of number of different DNA molecules for which binding can be tested and thus cannot test binding corresponding to longer DNA sequences. We use HT-SELEX (high-throughput systematic evolution of ligands by exponential enrichment) [32,33] experimental method to capture complex binding of such proteins. There are many platform specific issues associated with HT-SELEX data as presented in next section. We introduce MinSeqFind, a novel method to deal with complex binding data and such platform specific issues. MinSeqFind discovers DNA binding from HT-SELEX data in form of MinSeqs (Figure 3.1). It explores complex binding patterns and captures multiple partners of NRs previously unreported, and also finds binding change to be linked to hundreds of disease associated genomic variants.

MinSeqFind uses steps of modeling and data normalization followed by a multivalued-reduction through orthogonal matching pursuit (OMP) extensively used for compressed sensing [22,48]. MinSeqFind (Figure 3.1) can be divided into multiple steps: 1) Reads for protein binding are first counted for number of occurrence (counts) of 10mer to 16mer with possible continuous stretch of gaps in middle represented by N. 2) A Markov model for mock control is constructed called Position Associated Gapped LOcation-specific (PAGLO) model. 3) PAGLO model is used to estimate occurrence of sequences with counts for protein binding above a given threshold. 4) Fold enrichment is then calculated using counts and estimated occurrence, which is number of times a sequence occurred in protein binding data in comparison to mock control. 5) A multivalued-reduction through orthogonal matching pursuit (OMP) is performed followed by a thresholding to get a reduced set of sequences with enrichment. Such a reduced set is

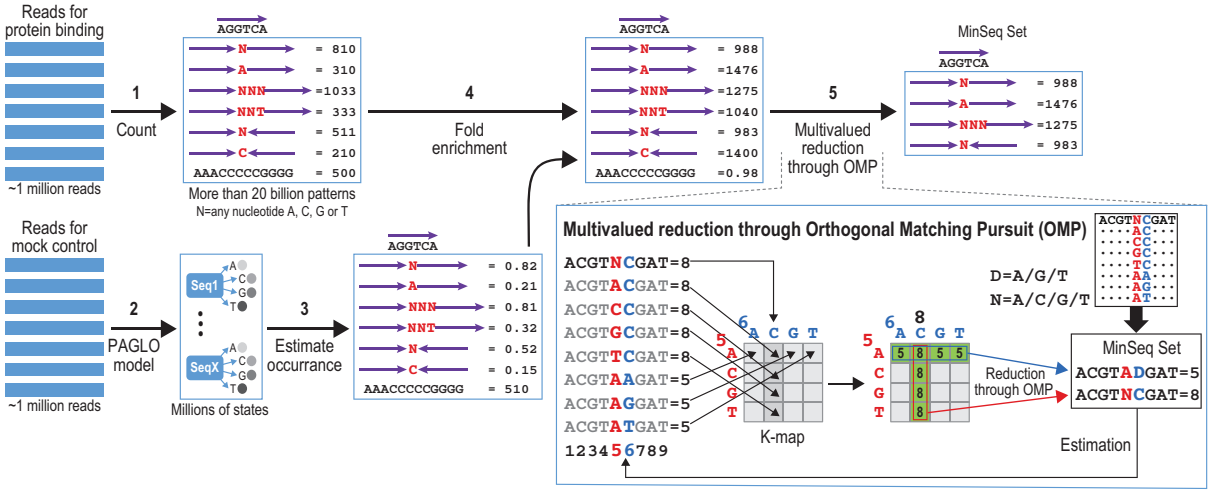


Figure 3.1: MinSeqFind algorithm- Extraction of complex DNA binding patterns using MinSeqs from HT-SELEX binding data: Sequences bound by protein from a 20 long random DNA library are obtained in the form of sequence reads (number of occurrence of each sequence) using HT-SELEX protocol. Shown by arrow are AGGTCA repeats and reverse complement TGACCT as reverse arrow.

termed as selected MinSeq set. In the following sections first problems associated with HT-SELEX experimental method are described and in further sections different steps of MinSeqFind are explained.

3.1 Problems associated with the HT-SELEX binding data

DNA binding measurement via HT-SELEX is preferred over microarray technology for following two reason mainly – 1) hundreds of DNA binding experiments can sequenced together by labeling them using DNA barcodes, and 2) to elucidate longer and more complex DNA binding, which arrays fails to capture. These experiments can be easily programmed to be performed by liquid handling robots makes them easier and faster [33]. HT-SELEX has advantages over microarray methods, but also have many issues associated with it.

HT-SELEX uses multiple rounds of DNA library selection using DNA binding protein followed by polymerase chain reaction (PCR) for amplification (Figure 1.1c and Figure C.1). After the final round, the recovered DNA is barcoded and sequenced. Out of more than billions of DNA molecules obtained after PCR step, only a small fraction of DNA is sequenced (<2 million DNA sequences) for each sample or experiment. Following steps involved in HT-SELEX make replication of the results is difficult-

1. **Probability of binding to DNA governs selection-** Since, SELEX includes process of selection of DNA using binding proteins in each round, this selection is based

on probability which is governed by protein's affinity for binding to various DNA sequences, protein concentration and relative concentration of different DNA sequences. Thus every repeat of the experiment would result in different set of DNA sequences selected on the basis of this probability.

2. **PCR contribution to variability-** PCR exponentially enriches the DNA. The DNA sequences that were isolated in the earlier PCR cycles will be continued to get enriched at exponentially higher rate than other sequences.

3. **Randomly picked small fraction gets sequenced-** Only a small fraction is sequenced in the end, which again is randomly picked by the sequencing machine. Thus even DNA from same experiment barcoded twice results in different output sequences.

Extending the approach of COSME to prune MinSeqs from previous chapter to SELEX based system is a challenge because of the following reasons-

1. **Sequencing reads instead of intensity-** Microarray intensities give a direct representation of protein's binding intensity, as however long is the probe sequence, one can compare the 2 sequences on the basis of array intensities. However in case of SELEX there are reads of k-mer sequence instead of intensities and most of the time any k-mer sequence appears maximum once or twice in the sequencing reads. If the sequence is less specific for binding it may not even show up in sequencing results. As explained above, only a small fraction gets randomly picked and sequenced. Sequencing reads give only a small sample of binding experiment and without grouping similar sequences together one can't assume a direct correlation between sequencing reads and binding affinity of sequences, especially when the reads corresponding to k-mers is very low. It has a Poisson distribution model for random picking of a k-mer, thus even if a k-mer appeared 1000 times there is a probability distribution corresponding to the protein binding affinity towards the k-mer and an exact affinity can't be assumed from that.

2. **Not enough representation-** If the total number of DNA reads are much larger in number compared to all the possible DNA sequences, then the normalized reads (normalized for library bias) can be used as intensities (with a probability distribution, as explained above), for example 8mer DNA library has 32,896 different DNA sequences and if sequencing outputs 100 million reads with at least 50 reads corresponding to each 8mer, then there is distribution of all 8mers in reads. But if reads are smaller in number then that can be misleading. For example a k-mer appeared once in reads, can be due to random picking of a non-sequence specific bound k-mer (low actual presence in the DNA pool) or due to real sequence spe-

cific binding of the k-mer (higher presence in the DNA pool, but got picked only once in sequencing).

3. **Binding to flanking region-** To sequence the selected DNA, the starting DNA library has to be flanked with constant DNA sequence primers on both sides. This region affects protein binding and should be taken into account.

4. **Gapped binding-** Many proteins bind DNA with gaps (any nucleotide in middle represented as N) eg. if a protein likes to bind a sequence ACACGTNNNNNNNNAC-GATC, that means it binds to ACACGT and ACGATC with gap of 7 nucleotides. Such binding preferences would show on k-mer reads with fixed ACACGT and ACGATC, but varying nucleotides in the middle.

In addition to the above mentioned issues, there are three more factors that affect output sequencing reads 1) the starting DNA library is not fully random and there is a preference for certain sequences over the other, 2) PCR step has it's own bias and amplifies a certain set of sequences more than other, and 3) in some experiments the protein of interest is bound to another kind of protein (for example Halo protein) that is needed to capture the protein–DNA complexes, and it has its own bias. These biases have to be filtered out (normalization step) to show true binding.

In MinSeqFind, we start with sub-sequences of gapped 6mer to gapped 16mers, such sub-sequences are called MinSeqs. We consider only those MinSeqs that appeared at least 50 times in sequencing reads. We then filter MinSeqs for biases introduced by initial DNA pool and PCR by normalizing MinSeq counts to expected counts from mock-control experiment. The effect of binding to flanking region is also considered for such normalization. After normalization and an additional weighting step, we use an iterative step to compress MinSeqs that is inspired by orthogonal matching pursuit (OMP). Compressed MinSeqs thus obtained are used for scoring to estimate protein binding to a new set of query sequences. We find multiple PWMs in a similar iterative manner to get a visual representation of binding. In the following sections we describe each step of MinSeqFind for HT-SELEX.

3.2 MinSeqs for HT-SELEX data

Definition: A **(k, g, l)-MinSeq** is a k-mer comprised of A, C, G, and T followed by a sequence of g >= 0 Ns (N=any nucleotide A, C, G, or T), followed an l-mer comprised A, C, G, and T. For example, the sequence AACGNNNGCTTA is a (4, 3, 5)-MinSeq because a 4-mer AACG is followed by NNN which is in turn followed by a 5-mer GCTTA. MinSeqs are used to capture protein–DNA binding into sequence intensity format, where sequences are of different length and exhibit gaps as well.

Definition: A k1-mer is said to be **left-contained** in k2-mer if k1<=k2 and the sequence corresponding to k1-mer is followed to the right by a sequence of A, C, G, and T to obtain the k2-mer. Likewise, a k1-mer is said to be **right-contained** in k2-mer if k1 <= k2 and the sequence corresponding to k1-mer is preceded to the left by a sequence of A, C, G, and T to obtain the k2-mer. For example, ACG is said to be left-contained in ACGATT but right-contained in ATTACG.

Definition: A (k1, g1, l1)-MinSeq is said to be a **direct subsequence** of a (k2, g2, l2)-MinSeq if exactly one of the following five conditions hold.

1. The two MinSeqs are identical OR
2. k1+1=k2, g1=g2, l1=l2, k1-mer is right-contained in k2-mer, and l1-mer is identical to l2-mer OR
3. k1=k2, g1=g2, l1+1=l2, l1-mer is left-contained in l2-mer, and k1-mer is identical to k2-mer OR
4. k1+1=k2, g1-1=g2, l1=l2, k1-mer is left-contained in k2-mer, and l1-mer is identical to l2-mer OR
5. k1=k2, g1-1=g2, l1+1=l2, l1-mer is right-contained in l2-mer and k1-mer is identical to k2-mer.

Definition: A (k1, g1, l1)-MinSeq is said to be a **subsequence** of (k2, g2, l2)-MinSeq if there is a sequence of MinSeqs m1, m2, …, mp, such that (k1, g1, l1)-MinSeq is a direct subsequence of m1, m1 is a direct subsequence of m2, …, mp-1 is a direct subsequence of mp, and mp is a direct subsequence of (k2, g2, l2)-MinSeq. We use the notation $\subseteq$ to denote this relationship. Example CGTNNA is a subsequence of sequence ACGTNNAAA, as CGTNNA is a direct subsequence of CGTNNAA, which is direct subsequence of CGTNNAAA, which is direct subsequence of ACGTNNAAA.

Note that MinSeqs (k, g, l) with g=0 can fall into multiple category of MinSeqs. Like MinSeq (k, 0, l) can also be written as MinSeq (k+1, 0, l-1) example - MinSeq ACGTAAA can regarded as MinSeq (4, 0, 3) - 4mer ACGT followed by no gap and a 3mer AAA, and can also be assigned as MinSeq (3, 0, 4) - 3mer ACG followed by no gap and a 4mer TAAAA. Thus MinSeq obtained in our analysis with no gap (g=0) are treated differently, if we get a MinSeq with g=0, we convert all MinSeqs into a format such that l=0, thus for the case ACGTAAA we use it as MinSeq (7,0,0) with 7mer ACGTAAA followed by no gap and a 0 length sequence. Further improvement in performance of MinSeqs can be obtained by use of not only A, C, G, T and N, but also K (G/T), M (A/C), R (A/G), Y (C/T), S (C/G), W (A/T), B (C/G/T), V (A/C/G), H (A/C/T), and D (A/G/T).

3.3 MinSeqFind: MinSeq extraction from HT-SELEX binding data

In this section we describe using MinSeqFind, how MinSeqs are obtained from HT-SELEX sequencing reads and is normalized against DNA library and other bias. Let us suppose the starting DNA library is of length M. Since DNA from different binding experiments were attached to a barcode first and then mixed together for sequencing, thus the first step is to de-multiplex the reads by matching DNA barcode corresponding to each experiment and then the reads are truncated to obtain the M-mer derived from the random DNA region (Figure C.1). Sequencing reads are obtained for a) just the M-mer DNA library, b) enriched library with mock control, and c) the enriched library with the TF.

First we study enrichment due to PCR and Halo-bead (mock control) by normalizing against starting random library, in the following two steps.

1. Model the relative abundance of all possible sequences in the library <=M of form MinSeq (k, g, l).

2. From the reads of mock control, consider all possible MinSeqs of (k, g, l) format >= count 50 and normalize against DNA library to analyze the binding of Halo-bead.

Second, we analyze the binding preferences of the TF. This analysis must account for the biases introduced by the "not-perfectly-random" library, the binding preferences of the Halo bead and bias introduced by PCR and other factors. After normalization of counts for MinSeqs we call the value as enrichment for protein–DNA binding. Here we characterize enrichment of TF compared to mock-control via the following two steps.

1. Model the relative abundance of all possible sequences in the mock-control <=M of form MinSeq (k, g, l).

2. Consider all possible MinSeqs of (k, g, l) format >= 50 and normalize against mock-control to analyze the binding of TF.

Thus, the above analysis normalizes randomness in the starting library as well as bias introduced by Halo-bead and PCR. The following sections explain how the sub-sequences are counted for protein–DNA binding and how modeling and normalization was done for a TF, against a mock-control (step 1 to step 4).

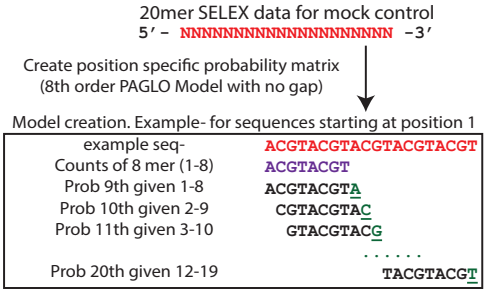
Step 1: Counting sub-sequences in protein–DNA binding data

From M-mer reads for the TF sample or protein–DNA binding data, MinSeqFind counts the occurrence of every sequence of type (k, g, l)- MinSeq with 5 <= k <= 8, 5 <= l <=

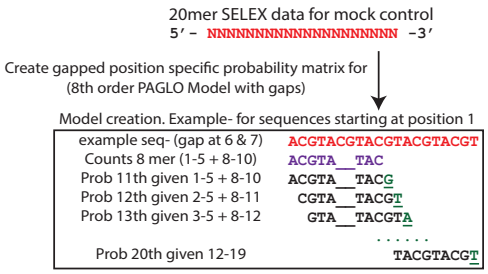
8, and k+g+l <= M (Figure 3.2). Counts for sequence and its reverse complement are merged together. If the count for a particular (k, g, l) sequence is below a minimum threshold of 50, then that sequence is discarded. The counts for others are retained for step 3.

Step 2: Position Associated Gapped LOcation-specific (PAGLO) model to characterize mock-control

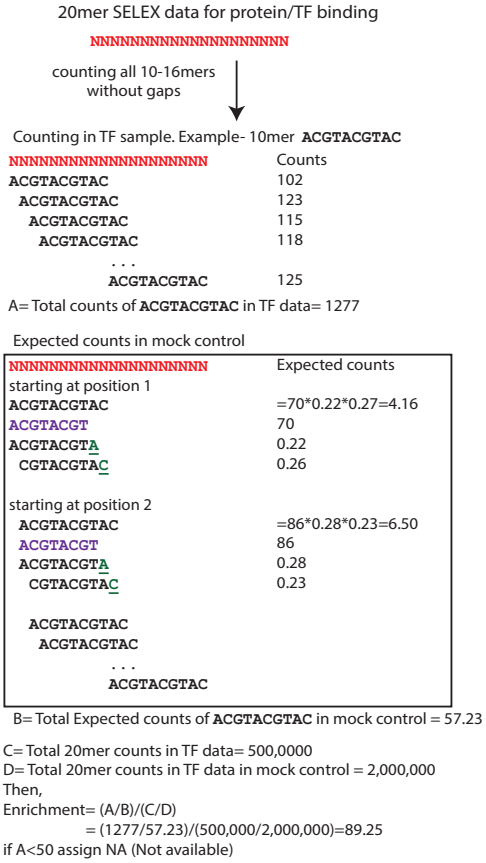
For a 20bp library there can be 4^{20} = 1 trillion different sequences in different concentrations, to get relative abundance of all 20mers we need to get >100 trillion reads, which is not feasible with current technology, and we get only a small portion or sample i.e. 1-2 million 20mer reads when 20bp library/sample is barcoded and sequenced. We create a model for relative abundance of all sequences of length <=M from limited reads obtained (for M-mer library). Even though a random M-mer library is used, the library isn't perfectly random. The probability of any nucleotide at a particular position depends on the previous nucleotides. Thus to capture biases in the library, binding of DNA sequences by the Halo-bead, and bias introduced by PCR, we construct a Position Associated Gapped LOcation-specific (PAGLO) model. From the M-mer reads of mock-control, we count the occurrence of every sequence of the form MinSeq (kx, gx, lx)- with 1 <= kx <= 8, 0 <= lx <= 8 such that, kx+gx+lx <= M and kx+lx <= 8. Counts for sequence and its reverse complement are merged together. If the count for a particular (kx, gx, lx) sequence was below a minimum threshold of 50, then that sequence was discarded. The counts for others are retained. In this study to characterize binding of a TF or mock-control, we use sequence of form MinSeq (k, g, l) with 5 <= k <= 8, 5 <= l <= 8, and k+g+l <= M. Since all possible such sequences will not have enough representation, we need an estimated abundance or counts for sequence (k, g, l). The estimate is calculated using PAGLO model as shown in Figure 3.2. For a 8th order PAGLO model, first Markov model of multiple smaller sequence (kx, gx, lx) with kx+lx =9 is used to estimate counts for longer sequence (k, g, l). In case if counts for sequence (kx, gx, lx) is less than 50, those counts are estimated using further shorter sequences with kx+lx <9 for which count>=50 i.e. estimating counts for a higher order Markov model (8th order) from a lower order Markov model [56]. Since the library is constructed 3' to 5' end of DNA, thus the model is created in 3' to 5' direction. The resulting model contains gaps, these gaps are location specific, means where the gap in the sequence is located is also taken into account while creating the model. Thus estimation for any sequence with a stretch of gap located anywhere with the sequence can be made. Also the model is built such that there is a separate model associated to different positions of a given sequence within an M-mer, means there will be different model for sequence starting at position 1 of M-mer and same sequence starting from



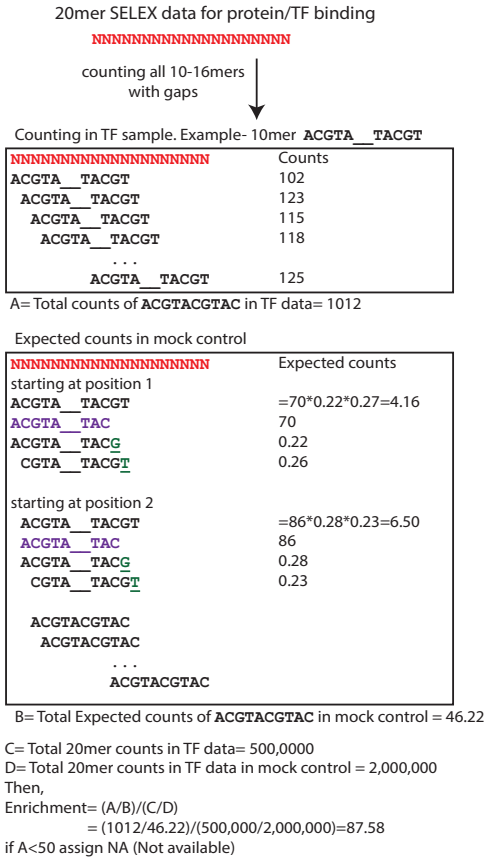
(a) PAGLO model for mock-control without gaps



(b) PAGLO model for mock-control with gaps



(c) Enrichment for 10-16mer without gaps



(d) Enrichment for 10-16mer with gaps

Figure 3.2: Enrichment calculation for MinSeq for DNA binding by HT-SELEX. a,b) Position specific PAGLO model is shown in the Figure (a- without gaps, b-with gaps). Means in the model, probability of A at position 9 given nucleotides at position 1-8 will be different from probability of A at 10 given same nucleotides at 2-9. In example PAGLO model starting from position 1 is shown. In purple is the 8mer for which counts and probabilities are calculated. c,d) First sub-sequences of length 10-16 with no gaps (d-with gaps) are counted for TF+Halo+Bead (TF sample) and then expected count for the same sequence are calculated for Halo+Bead (mock control) using PAGLO model and enrichment is calculated by dividing the two numbers.

position 2 of M-mer (Figure 3.2). This is done to normalize any possible binding to flanking primer region. Together the model is position associated, it can tolerate stretch of gaps in the middle and the gaps are location-specific.

Step 3: Calculating expected counts in mock-control

To normalize the counts of retained sequences from step 1 against the mock-control, first the expected counts of those sequences are calculated in mock-control using the PAGLO model as shown in Figure 3.2. Expected count is calculated for each retained sequence starting at different positions in the M-mer and then added to get total expected count for that sequence in mock-control.

Step 4: Enrichment calculation by normalizing against mock-control

The counts of retained sequences from protein–DNA binding are divided by total number of M-mer reads in protein–DNA binding data to obtain number of sequence occurrence per read in sample. Similarly, the total expected counts of those sequences from mock-control are divided by total number of M-mer reads in mock-control data to obtain estimated occurrence per read in mock-control. These numbers are then divided to get final enrichment values, which is number of times a sequence is enriched relative to the mock-control (Figure 3.2). Note that, TF or mock-control can bind to region flanking the random N-mer region partly or even completely. Since we consider a position associated model, such binding is normalized by the analysis pipeline described above (Figure 3.3).

3.4 Scoring with MinSeqFind model

Consider an l-mer, l large, and a set of sequence of type (k, g, l) MinSeq (maximum length M, equal to the length of random region), each with a weight. A moving window of length M is used to score a sequence of length l, resulting in l-M+1 sub-sequences. In algorithm 3 we describe procedure for scoring the l-mer.

Algorithm 3 Scoring with MinSeqFind model

```
For i = 1 to l-M+1 do
  For j = 0 to M do
    For k = 0 to M do
      Find sequences matching sub-seq i+j-th to i+k-th
    Endfor
  Endfor
  For M-mer starting at i-th position, assign max score from matching MinSeqs.
Endfor
```

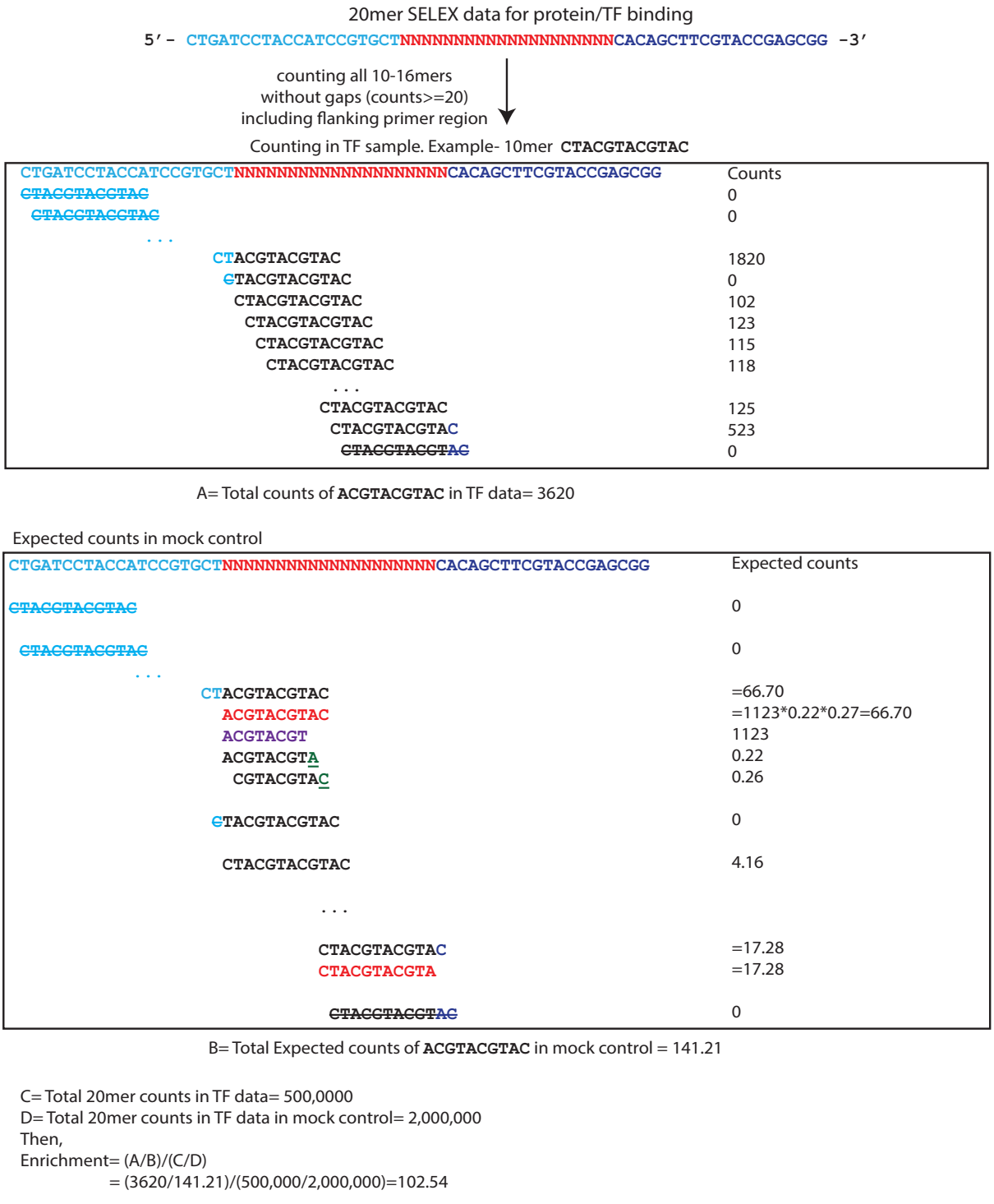


Figure 3.3: Normalization of possible binding to primer region flanking the random 20mer region. In example shown is a 10mer sequence CTACGTACGTAC without gap. It is aligned to match different regions of 60mer input DNA library and counted. Nucleotides are crossed if sequence can never appear at that position because of mismatched nucleotides. If a sequence like CTACGTACGTAC is bound by protein then it will have higher counts closer to primer A region because it matches partially to the primer region (CT).

We display such intensity data for l-N+1 sub-sequences of length M as colored bar plots and call them Genomescapes for genomic sequences, i.e. plot from the predicted binding to each sub-sequence in the genome. The maximum score over all the l-M+1 sub-sequences is used as binding score for the full l-mer.

3.5 Weighted Enrichment

In a MinSeq (k, g, l), k and l part of the sequence consists of nucleotides A, C, G or T, whereas g (gap) part consists of Ns. Number of different MinSeqs that can exhibit same (k, g, l) pattern, i.e. a k-length sequence followed by g-length stretch of Ns, followed by l-length sequence, thus is $4^{(k1+l1)}$, as there can be one out of 4 nucleotides at each position on k and l part. In a random situation, the probability of occurrence of MinSeq (k1, g1, l1) in a sequence of length L>=(k1+g1+l1) is thus $(4^{(k1+l1)})$*(L-k1-g1-l1+1). Unlike k-mers, MinSeqs are thus not equally likely to present, probability ratio of MinSeq (k1, g1, l1) and MinSeq (k2, g2, l2) is thus, $(4^{(k2+l2)-(k1+l1)})$*(L-k1-g1-l1+1)/(L-k2-g2-l2+1).

Why MinSeq weighting?

Ordering MinSeqs on the basis of their descending value of enrichment points out the best enrichment values and best binding sequence/MinSeqs possible, but if the k+l value (part of sequence containing A/C/G/T) for the best MinSeq is higher, then it is less likely to occur in a random case as well as in the genome and will carry lesser information. Thus there has to be a trade-off between desired higher-enrichment value and desired lower k+l value for a MinSeq.

This trade-off is the most important part in defining the value of a particular MinSeq in capturing the binding. We use minimum mean square error (MMSE) as the criteria, i.e. prefer/rank MinSeq (k1, g1, l1) over MinSeq (k2, g2, l2) if former gives a lesser MSE in predicting back the data (from which MinSeqs are derived) in comparison to the later.

Comparison of MinSeq1 (k1, 0, l1) & MinSeq2 (k1, 0, l2) with k1+l1=k2+l2

Lemma 1: In predicting back the data, if k1=k2, MinSeq1 (k1, 0, 0) gives lower MSE when compared to MinSeq2 (k2, 0, 0) if $E_{(k1, 0, 0)} > E_{(k2, 0, 0)}$, where $E_{(k, g, l)}$ is enrichment of MinSeq (k, g, l).

Proof: Considering enrichment for sequences of length k. MinSeq1 with enrichment a and MinSeq2 with b. Since there will not be any other k-mer sequence matching these two MinSeqs. MSE in estimation can be minimized by minimizing MSE just for

the two sequences matching MinSeq, MSE = $((a-a')^2 + (b-b')^2)/2$, where a' is estimated enrichment for a. Thus choosing MinSeq1 will give MSE $b^2/2$ whereas choosing MinSeq2 will give error $a^2/2$. Choose MinSeq1 over MinSeq2, if MSE for MinSeq1 is less i.e. $b^2/2<a^2/2$, which implies a>b (enrichment values are non-negative) or $E_{(k1, 0, 0)} > E_{(k2, 0, 0)}$, hence proved the Lemma 1.

The proof can be extended to a case where k1+l1=k2+l2. MMSE is achieved by ranking according to the higher enrichment (similar to k-mer enrichment).

Comparison of MinSeq1 (k1, g1, l1) & MinSeq2 (k1, g1, l1)

Lemma 2: In predicting back the data, MinSeq (k1, g1, l1) gives lower MSE when compared to MinSeq (k2, g2, l2) if $E_{(k1, g1, l1)}*2^{-(k1+l1)} > E_{(k2, g2, l2)}*2^{-(k2+l2)}$, where $E_{(k, g, l)}$ is enrichment of MinSeq (k, g, l).

Proof: Consider two MinSeqs 1- (k1, g1, l1) and 2- (k2, g2, l2), such that (k1, g1, l1) is direct subsequence of (k2, g2, l2), such that k1+1=k2, g1=g2, l1=l2.

Let us consider g1=g2=l1=l2=0, i.e. the two sequence doesn't contain any gap and aim is to minimize MSE for a k-mer binding data k=k2 in a context of a much longer genomic sequence L»k. Example MinSeq1=ACTA and MinSeq2=ACTAT, thus k1=4, k2=5 and trying to minimize MSE for a 5-mer binding data. MinSeq1 can be left contained sub-sequence of 4 different 5-mer sequences ACTAA, ACTAC, ACTAG and ACTAT, whereas MinSeq2 is a sub-sequence of only ACTAT. Note, although MinSeq1 is right contained sub-sequence of 4 other 5-mers- AACTA, CACTA, GACTA and TACTA, but we don't use those for MSE estimation as in a longer context (L»k), sequences ACTAA, ACTAC, ACTAG and ACTAT covers all possibilities in AACTA, CACTA, GACTA and TACTA, example in AAA<u>ACTAA</u>AAA - underlined sequence ACTAA is a 5-mer which has left contained sub-sequence ACTA, this also captures AACTA, which has ACTA as right contained sub-sequence.

Now, let us assume enrichment for ACTA, ACTAA, ACTAC, ACTAG and ACTAT is m, a, c, g and t respectively. Further assumption made for simplification, a=c=g and t>a. Given these details, we need to choose 1 MinSeq ACTA (m) or ACTAT (t) which minimizes MSE in estimating a, c, g and t. MSE in estimation = $((a-a')^2 + (c-c')^2 + (g-g')^2 + (t-t')^2)/4$, where (a' is estimated enrichment for a). Since enrichment of ACTA is average of enrichment of ACTAA, ACTAC, ACTAG and ACTAT -> 4m=a+c+g+t, this can be written as 4m=3a+t.

If m was picked as preferred MinSeq - a'=m, c'=m, g'=m, t'=m. Thus MSE in estimation = $((a-m)^2 + (c-m)^2 + (g-m)^2 + (t-m)^2)/4 = (3(a-m)^2 + (t-m)^2)/4$. If t was picked- a'=0, c'=0, g'=0, t'=t. Thus MSE in estimation = $((a-0)^2 + (c-0)^2 + (g-0)^2 + (t-t)^2)/4 = 3a^2/4$.

Choose t over m if-

=> $3a^2/4 < (3(a-m)^2+(t-m)^2)/4$, put t=4m-3a

=> $4m^2 - 8am + 3a^2 > 0$

=> 2m<a OR 2m > 3a

=> 2m+t<0 (not possible as enrichment has to be positive only) OR 2m<t

=> t>2m

Thus choose ACTAT (t) over ACTA (m) if t>2m to minimize MSE. Similarly can prove if t<2m choose m over t for MMSE.

A general case thus, select MinSeq (k1, g1, l1) over MinSeq (k2, g2, l2), where (k1, g1, l1) is direct subsequence of (k2, g2, l2), such that k1+1=k2, g1=g2=l1=l2=0, if $E_{(k1, g1, l1)}*2^{-k1} > E_{(k2, g2, l2)}*2^{-(k1+1)}$. And select MinSeq (k2, g2, l2) over MinSeq (k3, g3, l3) if $E_{(k2, g2, l2)}*2^{-(k1+1)} > E_{(k3, g3, l3)}*2^{-(k1+2)}$. This implies select MinSeq (k1, g1, l1) over MinSeq (k3, g3, l3) if $E_{(k1, g1, l1)}*2^{-k1} > E_{(k3, g3, l3)}*2^{-(k1+2)}$. This can be extended thus select MinSeq (k1, g1, l1) over MinSeq (k2, g2, l2), where (k1, g1, l1) is a subsequence of (k2, g2, l2), such that g1=g2=l1=l2=0 if $E_{(k2, g2, l2)}*2^{-k1} > E_{(k2, g2, l2)}*2^{-k2}$. With similar analogy constraint g1=g2=l1=l2=0 can be relaxed.

We used (k1, g1, l1) is a subsequence of (k2, g2, l2), this constraint is removed given Lemma 1. Thus, MinSeq (k1, g1, l1) gives lower MSE when compared to MinSeq (k2, g2, l2) if $E_{(k1, g1, l1)}*2^{-(k1+l1)} > E_{(k2, g2, l2)}*2^{-(k2+l2)}$. Till now we considered case where k1<=k2, but same equation holds for k1>=k2 as well. Hence proved the Lemma 2.

Thus the weighted enrichment of MinSeq (k, g, l) = $F_{(k, g, l)} = E_{(k, g, l)}*2^{-(k+l)}$ gives a better perspective when comparing MinSeqs of different length (k+l). Weighted enrichment ranks MinSeqs on the basis of their importance/predictive-capability in minimizing MSE.

3.6 Step 5: Multivalued-reduction through orthogonal matching pursuit (OMP)

Since MinSeq can be all k mer - g gap - l mer sequences with varying k, g and l, there is a lot of redundancy, for example if a protein prefers to bind only one 4mer sequence ACGA, then highest ranked MinSeq will be ACGA according to weighted enrichment, but there will be MinSeqs like ACGAA, ACGAC, …. as well in the list, all those which crossed the threshold of minimum counts. Thus, there is a a need to remove all 5mer and longer sequences in this example case. Given the binding is to ACGA, other sequences doesn't carry any additional information. Thus a multi-valued reduction is performed on enrichment of (k, g, l) sequences or MinSeqs obtained from step 4.

In an ideal case with limited binding patterns and no experimental variation a reduction is feasible with complete retrieval. Figure 3.4 is an example of such a reduction where there are eight different sequences with corresponding enrichment values of 5 or 8 as shown on the left. Except positions 5 and 6 all these sequences match exactly,

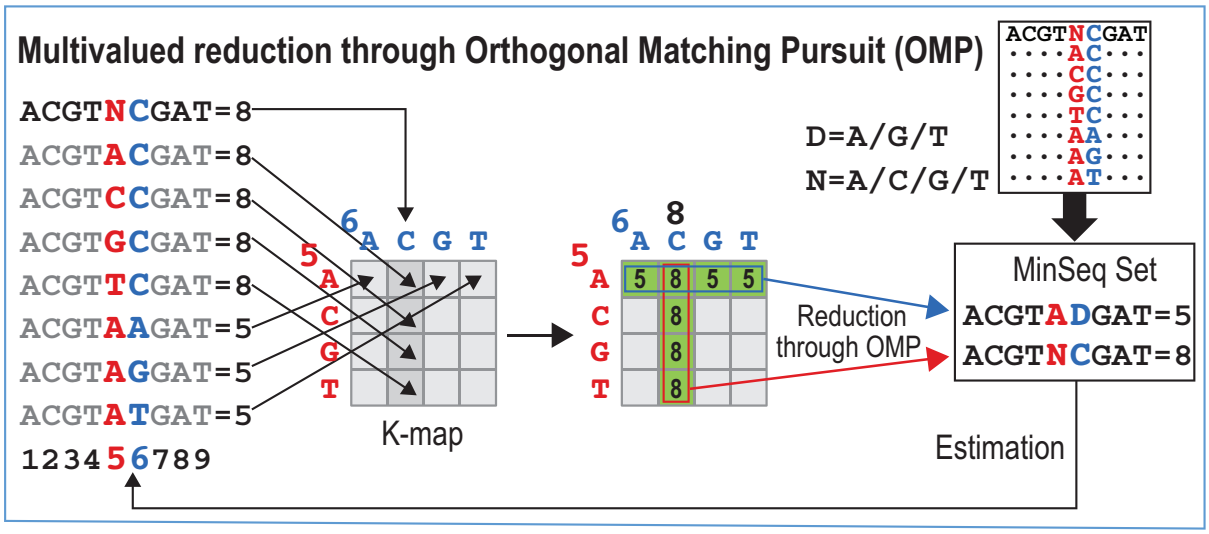


Figure 3.4: Multivalued-reduction through orthogonal matching pursuit (OMP)

nucleotides at these positions are used to create a two-dimensional K-map. A reduction step is performed through orthogonal matching pursuit (OMP) to obtain a selected MinSeq set. Given just the sequences of starting set, their corresponding enrichment values can be exactly estimated from this selected MinSeq set, thus it completely captures the starting set of sequences with enrichment.

In a realistic case there is no limit on number of binding patterns and there can be some experimental variation or noise and thus a complete retrieval isn't always needed or isn't always possible. Thus, instead of full reduction through OMP, compression is done through OMP by selecting one MinSeq at a time and putting a threshold on number of total MinSeqs. Given the information about MinSeq ranked 1, the information left in all other sequences changes. Thus to capture binding information in a smaller number of MinSeqs we need to remove such redundancies. We use compressed sensing (CS) based methods from signal processing for such pruning. We use a modified approach of orthogonal matching pursuit (OMP) [22,37,48] to retrieve K best sequences or MinSeqs, such a set in termed as selected MinSeq set. From a binding data- given are sequence of type (k, g, l) with their corresponding weighted enrichment $F_{(k,\,g,\,l)}$, and p = maximum number of final selected MinSeqs to be used, then algorithm 4 defines the extraction/compression using approach inspired by orthogonal matching pursuit (OMP) (Figure 3.5).

3.7 Results

We display binding as extracted by MinSeqFind using two different formats: gapped SELs (see appendix) and position weight matrix (PWM). PWM gives a neat visualization of binding and is used extensively, thus we use weighted MinSeqs and extract PWMs

Algorithm 4 MinSeqs compression for HT-SELEX binding data inspired by orthogonal matching pursuit (OMP) [48]

Initialize: Sequences (k, g, l), corresponding weighted enrichment F (as calculated in previous section), i=0, residual weighted enrichment R=F, set S={}, p = maximum MinSeqs
while i<p **do**
i=i+1
Choose MinSeq (ki, gi, li) with maximum residual weighted enrichment R.
Add it to set S i.e. S={S, MinSeq (ki, gi, li)}
Use enrichment for set S to score all given MinSeqs= N
Subtract to get residual R = F-N
end while
Set S defines the final MinSeqs.

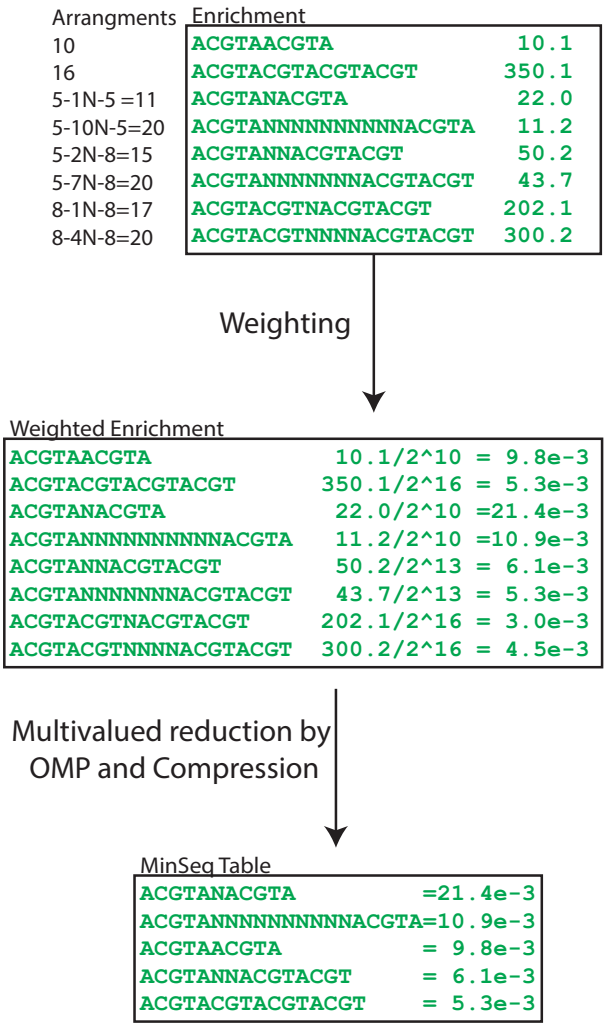


Figure 3.5: Weighted enrichment, compression and ordering. First enrichment of all MinSeqs (10-16mer with or without gap) is combined together in a list. Then enrichment is weighted by multiplying by $2^{-\text{number of nucleotides excluding N}}$. Only some sequences are selected and carried forward to get MinSeq table by compression and then ordered according to their weighted enrichment.

for binding. PWM is constructed using the best weighted MinSeq and residual signal is obtained as residual MinSeqs by subtracting out PWM's predictions (similar to OMP). By iterating this, multiple PWMs are obtained defining binding patterns (Figure 3.6). In following subsections we present results when MinSeqFind is applied on newly generated nuclear receptors DNA binding data and published protein–DNA binding data.

PAGLO model analysis

PAGLO model is a key component of MinSeqFind analysis. It consists of multiple Markov models. Figure 3.7a shows the maximum enrichment predicted by the PAGLO model when PAGLO model is trained on a mock control data and then normalized against mock control, different order Markov model were used for that in PAGLO model. Plotted is the enrichment, maximum out of all 6mer to 16mer sequences with any possible number of gaps in the mock control data, lower is better, ideal maximum enrichment should be 1. Plot thus captures the sensitivity of model in predicting the outcome of mock-control. The plot saturates at 4th and 5th order to maximum enrichment of 1.33. Thus, in worst case there can only be maximum of 33% possible error in prediction of mock control using a 5th order PAGLO model. Increasing order beyond that didn't give any improvement. Figure 3.7b and 3.7c shows the two PWM motifs obtained using zero order Markov model in PAGLO, in ideal condition we should not see any binding motif. It can be seen the PWM motifs are up to length 6 (with high information positions), which can be normalized using a 5th order Markov chain model, that is the reason of higher performance of 4th and 5th order Markov model in PAGLO model.

High throughput DNA binding for nuclear receptors

Nuclear receptors (NRs) are a family of transcription factors (TFs) that regulate many key cellular functions including steroid sensing, maintaining hormonal balance and embryonic development and thus they are target of over 13% of FDA approved drugs [46]. Drugs targeting NRs are used to effectively combat several diseases like receptor-induced cancers, asthma, hormonal imbalances, obesity and many other diseases [25]. A better understanding of the DNA binding preferences of NRs is crucial for devising new targeted drug therapies. Furtherer, in the presence of specific ligands (drugs) and/or other TF partner especially RXRA (Retinoid X receptor alpha), NRs get activated to up-regulate or down-regulate different genes [25].

We measure high-throughput (HT) DNA binding of all the full length human NRs. We use cognate site identification (CSI) by high-throughput systematic evolution of ligands by exponential enrichment (HT-SELEX) [32, 64] to examine DNA binding preference of NRs in cell-extract. A DNA library spanning the entire sequence space of

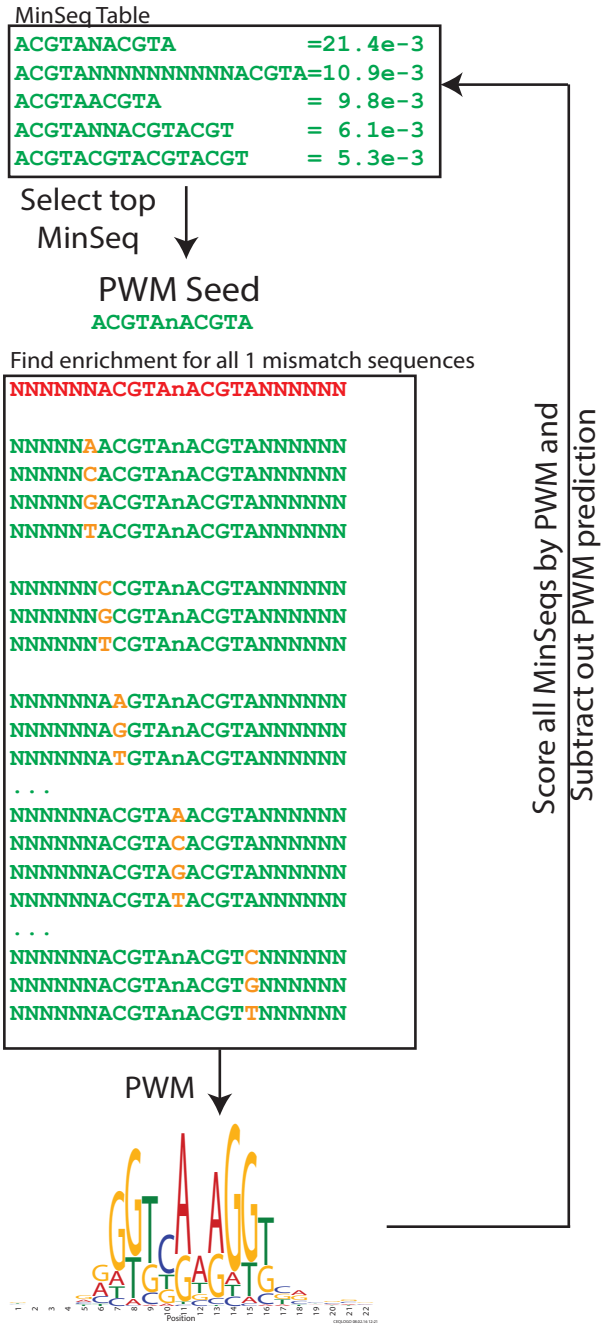
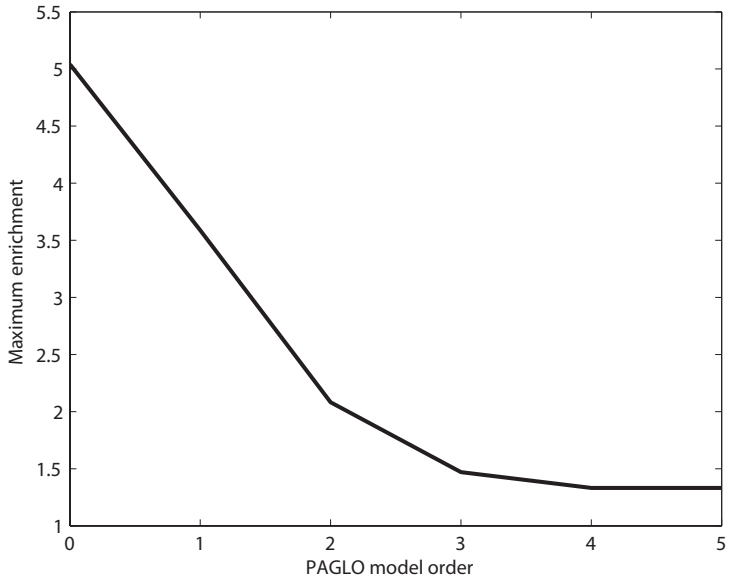


Figure 3.6: Iterative process of PWM extraction from MinSeqs: Top MinSeq is selected as the seed to get PWM. Seed is then extended by 3-6bp on each side by adding 'N' to consider flanking binding. Enrichment corresponding to seed and sequences exhibiting 1 mismatch to the seed are used to construct PWM using standard methods. Then from the calculated enrichment, the prediction made by PWM for all MinSeqs is subtracted out and the next top MinSeq from the weighted residual enrichment is chosen for next PWM and so on.



(a) Maximum enrichment obtained using different order PAGLO model

(b) Mock control first PWM using 0 order PAGLO model

(c) Mock control second PWM using 0 order PAGLO model

Figure 3.7: PAGLO model performance with different order of Markov model. Results are shown when PAGLO model is trained on a mock control and then normalized against itself.

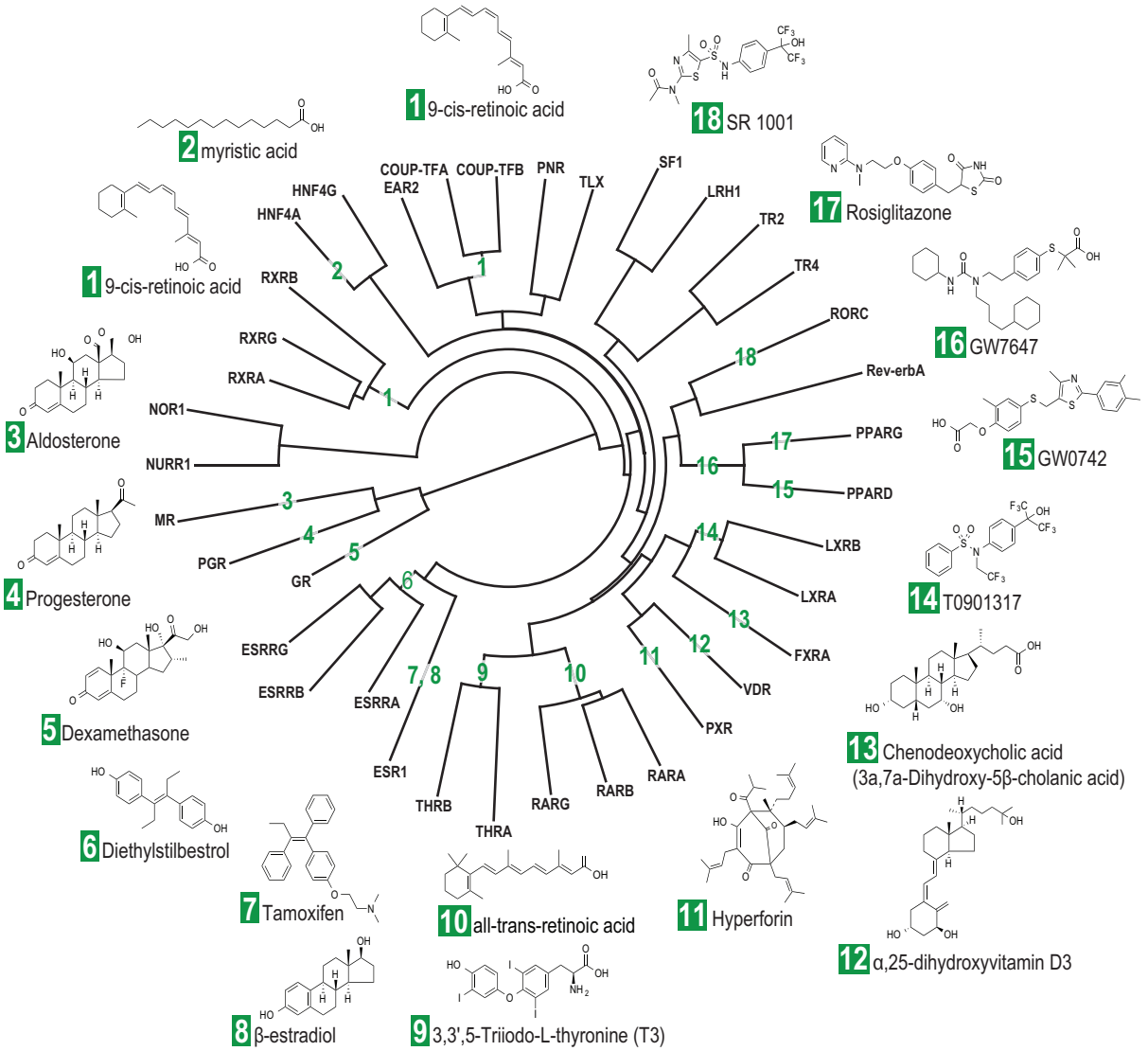


Figure 3.8: Circular phylogeny tree of nuclear receptors for which binding motif was observed using MinSeqFind algorithm. Ligands corresponding to each nuclear receptor for which binding motif was observed are also shown with their structure.

a 20-mer (10^{12} different sequence permutations) was independently incubated with all 48 members of human NR family. In combination we also use different ligands corresponding to each protein and RXRA as partner to study change in DNA binding preferences due to ligands and protein partners.

From sequence reads of CSI by HT-SELEX MinSeqFind discovers binding motifs of various lengths, dimers with various orientation and gaps. Out of all human (48) NRs tested in presence or absence of partner and/or ligand, using MinSeqFind we observe a binding motif in- 40(83%), 2(4%) (DAX1 & SHP1) doesn't have DNA binding domain and 6(13%) did not give any motif. A circular phylogeny tree of nuclear receptors for which binding motif was observed using MinSeqFind algorithm is shown in Figure 3.8, with corresponding ligands labeled around the tree.

Novel DNA targets for Nuclear Receptors

Figure 3.9 gives a comprehensive analysis of the preference of binding for NRs for different repeats and orientations of monomers, including NRs with protein partners and/or ligands. The proteins are arranged phylogenetically, with proteins with similar amino acid sequence shown together. As shown in the Figure 3.9 , a single insertion or deletion can cause a change in the gap between the monomers causing a different NR to bind, resulting in alteration in transcription causing disease. For example - a DNA sequence binding to Vitamin D receptor or VDR (DR3- direct repeats of monomer with 3 gap) can convert to preferred binding of THRB (DR4) if there is a single insertion between monomers, whereas to RARA:RXRA (RARA (retinoic acid receptor-alpha) binding in presence of RXRA partner protein) binding (DR5) by another insertion, popularly known as DR-3,4,5 rule [25].

In Figure 3.10 & 3.11, we display DNA binding in the form of newly developed gapped sequence energy landscapes (gapped-SELs). To plot them, first a seed sequence is chosen as combination of two monomers from top ranked MinSeqs or PWMs. The monomers are combined with multiple orientations (DR-direct repeat of monomers, IR-inverted repeat, ER-everted repeat as shown by arrows in Figure 3.9), and different gaps are placed between monomers. All sequences corresponding to gap=g are arranged along X-axis with Y coordinate=g. The same order of sequences is followed for all gaps along X-axis, example "AAGGTCANNAGGTCAT" and "AAGGTCANNNAGGTCAT" are sequences with direct repeats of AGGTCA with gap 2 (DR2) and 3 (DR3) respectively will have same x-coordinate, but have y=2 and y=3 Y-coordinate respectively. After deciding X and Y coordinate, the enrichment or binding intensity is plotted at that coordinate with height and color coded peak. Binding intensity of all the sequences for gapped-SELs is obtained using MinSeqs.

MinSeqFind discovers many novel binding patterns and binding differences, which are previously unreported [19]. Major findings are listed below.

1. MinSeqFind for HNF4G (Hepatocyte Nuclear Factor 4- gamma) protein, which is critical for liver development, identified two very similar looking binding motifs but had subtle difference as shown in Figure 3.12a, first motif is GGTCAAAGGTCA which is a known motif of HNF4G and new motif GGTCAAAGTCCA. Novel motifs for HNF4G leads to finding new genomic targets and genes regulated by it in connection with liver development.

2. Liver X receptor (LXR) class of proteins is an important regulator of cholesterol, fatty acid, and glucose homeostasis. LXRA member of it gave binding motif as dimer of RGGTTAC & MGGTCA (where M=(C/A)) with a gap of 3 (Figure 3.12b), which is new finding as LXRA is supposed to bind as dimers of RGKTCA (where, R=(G/A) & K=(G/T)). Further MinSeqFind also discovers that LXRB, which is

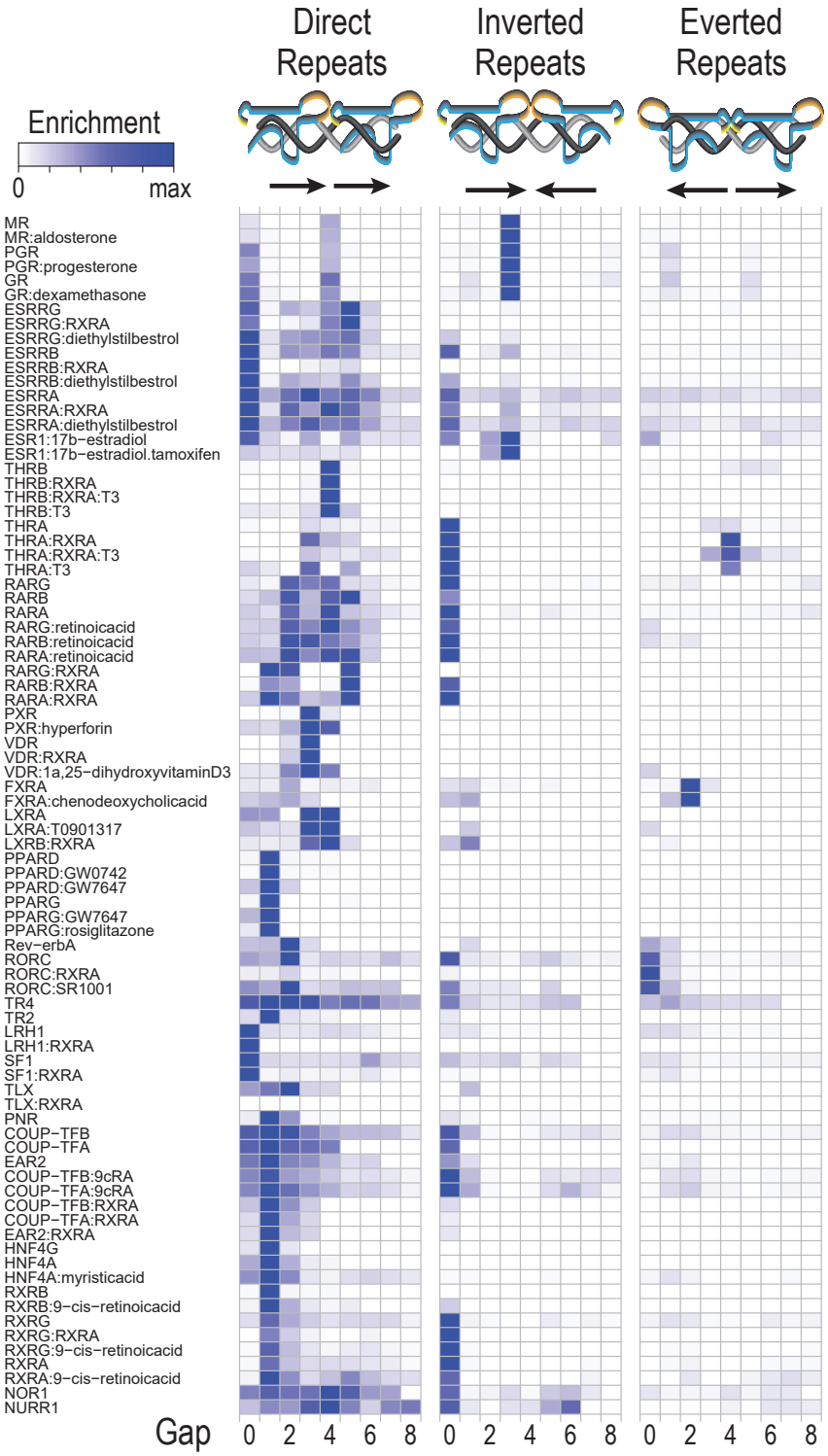
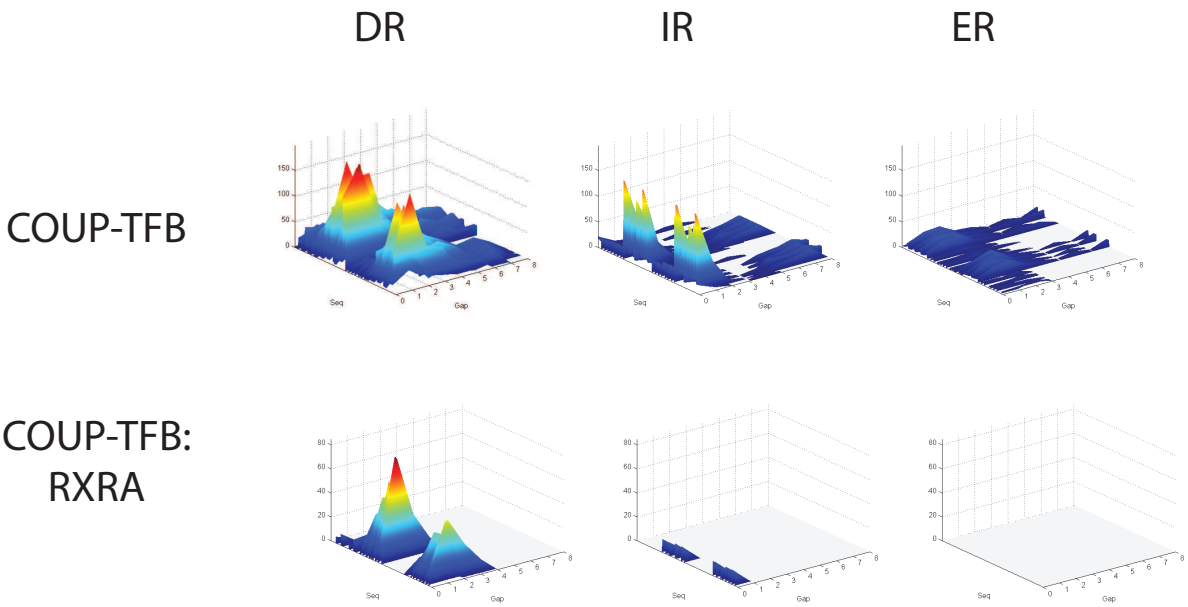
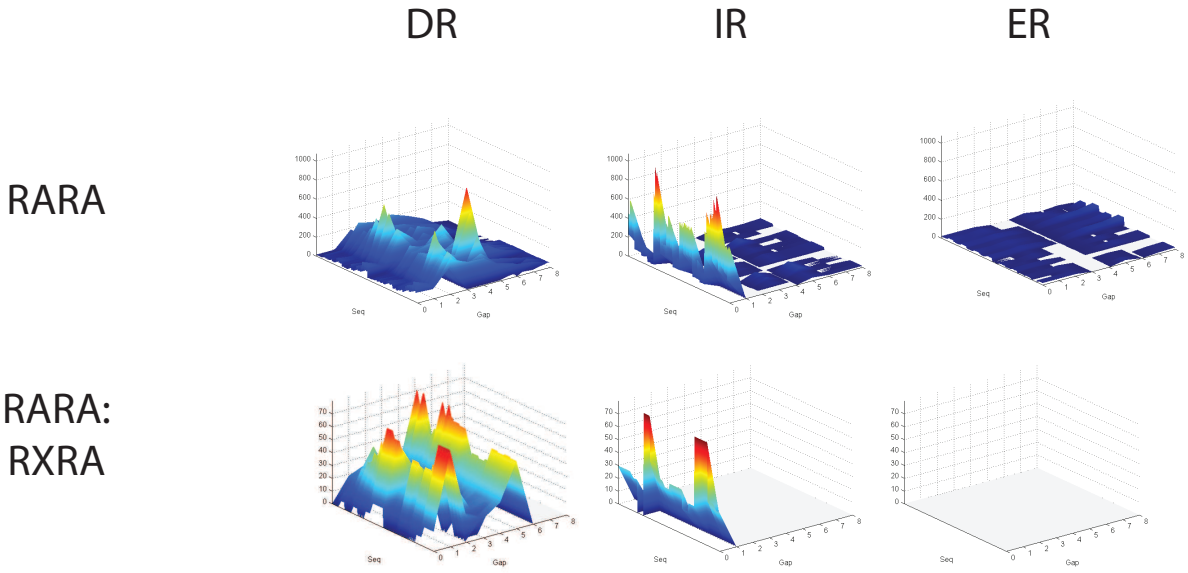


Figure 3.9: Different repeat preferences for NR as calculated by MinSeqs. With direct repeats, inverted repeats & everted repeats of gap 0 to 8, enrichment is displayed in form of color coded squares for different repeat preference, with maximum enrichment normalized to same value for each sample or experiment represented by single row. All the samples are ordered on the basis of phylogeny tree of NR family. If multiple NRs or ligands are used in binding experiment, then for naming each one is separated by colon (:).

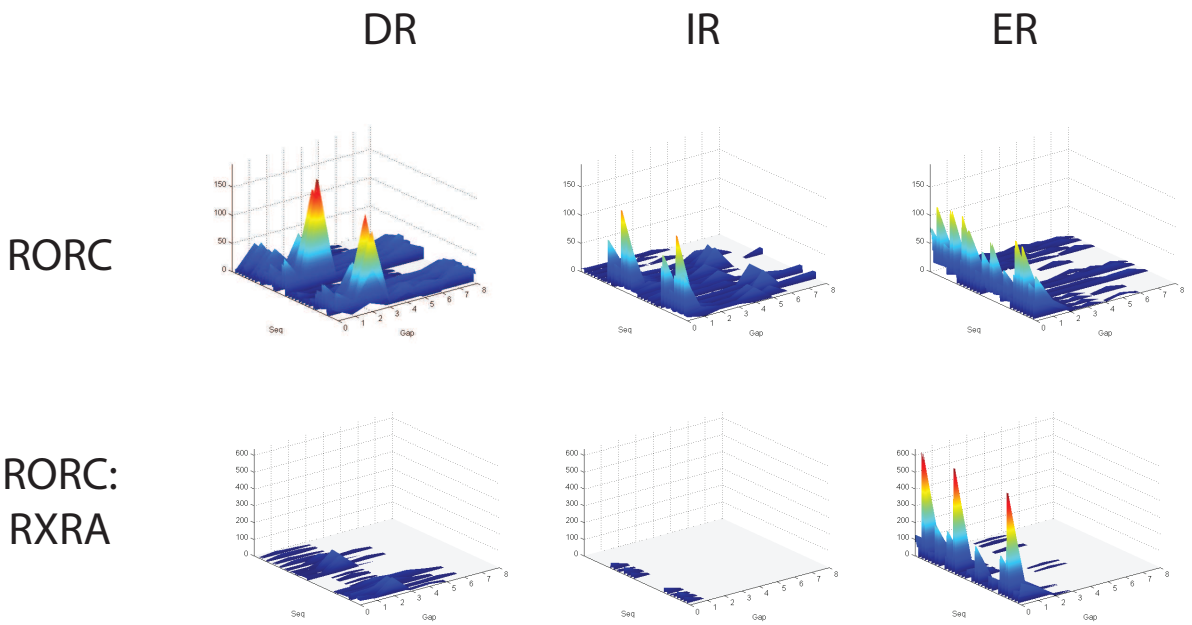


(a) Gapped SEL with RGKTCR as monomer, shows difference in binding of COUP-TFA due to presence of RXRA protein

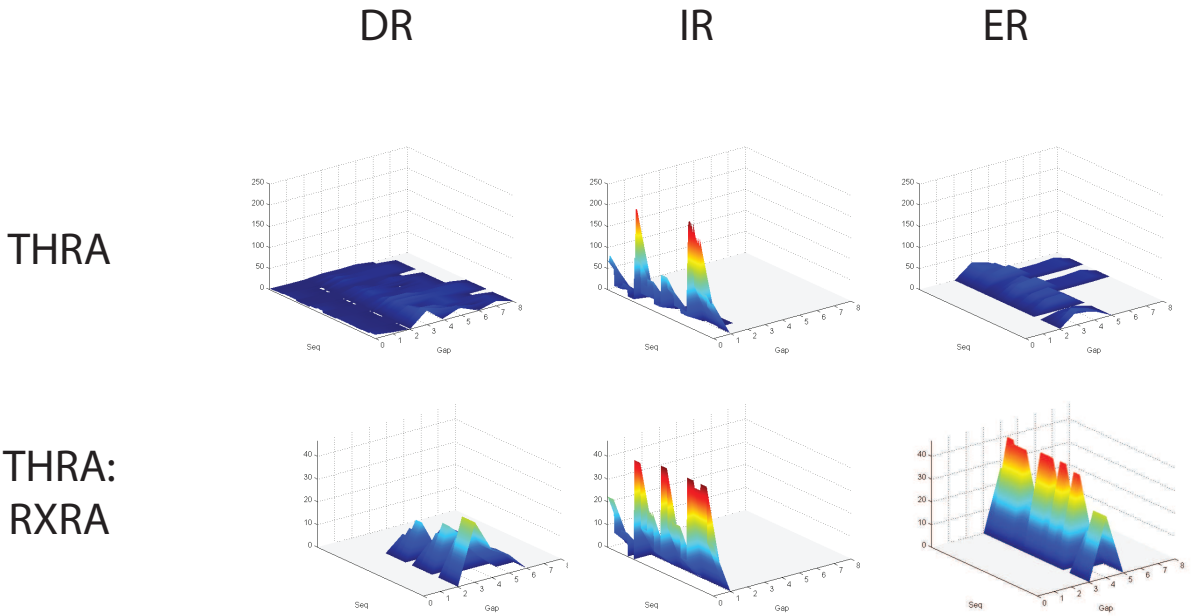


(b) Gapped SEL with RGGTCR as monomer, shows difference in binding of RARA due to presence of RXRA protein

Figure 3.10: Gapped SEL uncovered hetero-dimer formation of COUP-TFA and RARA protein with RXRA

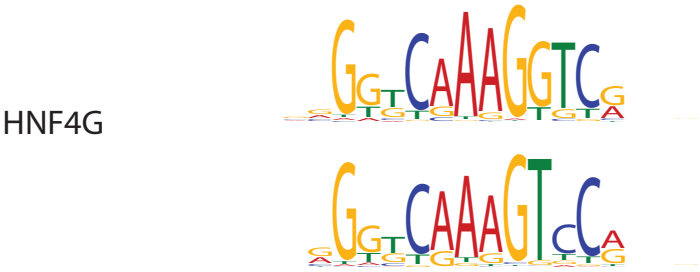


(a) Gapped SEL with RGKTCR as monomer, shows difference in binding of RORC due to presence of RXRA protein

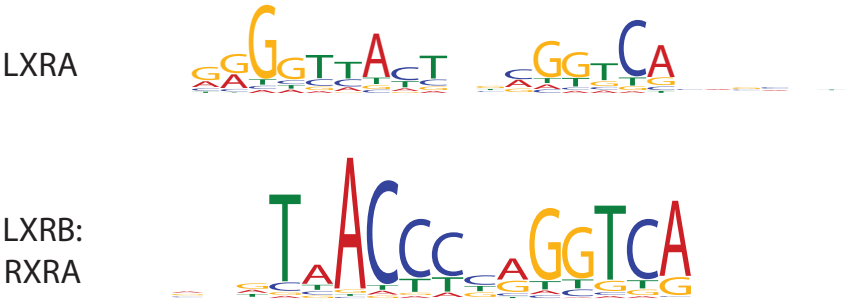


(b) Gapped SEL with RGGTCR as monomer, shows difference in binding of THR due to presence of RXRA proteinA

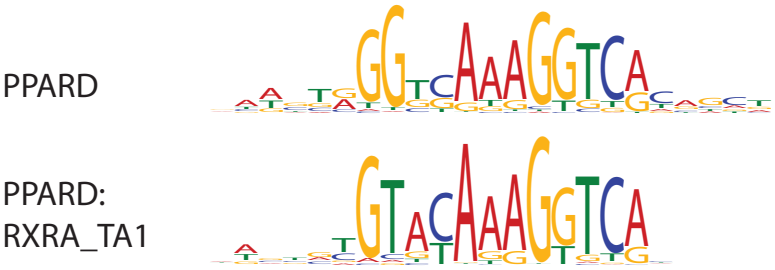
Figure 3.11: Gapped SEL uncovered hetero-dimer formation of RORC and THR protein with RXRA



(a) HNF4G multiple binding patterns



(b) LXRA and LXRB binds as dimers of RGGTTAC and AGGTCA



(c) Change in binding of PPARD due to combination of partner protein and ligand

(d) ESRRA novel binding motif

Figure 3.12: PWMs derived from MinSeqFind discovers novel binding patterns of NRs

another protein of same sub-family binds TAACCY-RGGTCA (where Y=(C,T)) in presence of RXRA, i.e. inverted repeats of dimers TGGTTA & RGGTCA with gap of 1 (IR1). Finding such motifs changes the way regulation of cholesterol, fatty acid, and glucose is interpreted.

3. Peroxisome proliferator-activated receptors (PPARs) play essential roles in the regulation of cellular differentiation, development, and metabolism. DNA binding of PPARD a memeber of PPAR family changes from direct repeated dimer of RGGTCA & RGGTCA to dimer of TGTACA & RGGTCA, both with gap of 1 in presence of RXRA partner and TA1 ligand (Figure 3.12c). This finding emphasizes role of ligands in DNA binding, thus TA1 can change specificity of PPARD in presence of RXRA, leading to activation/repression of different genes.

4. ESRRA and other members of ESRR sub-family showed a new motif that resembles the overlap of two dimers in presence of RXRA specially (Figure 3.12d). It is a dimer of AGGTCA & TGACCT i.e inverted repeat of AGGTCA with gap of -3 i.e. IR-3 to give motif like AGGT(C/G)ACCT.

5. Chicken ovalbumin upstream promoter-transcription factors (COUP-TFs) play critical roles in the development of organisms. From gapped-SELs of COUP-TFB & COUP-TFB (members of COUP-TF family) with RXRA (Figure 3.10a), it is evident that COUP-TFB hetero-dimerizes with RXRA to bind DNA as DR1 of RGKTCR and prefers that over it's homo-dimer binding of IR0 and DR2. Thus, in presence of RXRA COUP-TFs target different genomic regions and thus RXRA is affecting their gene regulation. This was true for all the members of that sub-family.

6. Similar to COUP-TFs, members of retinoic acid receptor (RAR) sub-family like RARA prefers DR1, DR5 and IR0 of RGGTCR as hetero-dimer (Figure 3.10b).

7. RORC (RAR-related orphan receptor-gamma) protein which plays critical role in lymph node development and immune response, with RXRA it likes to bind ER0 of RGKTCR (Figure 3.11a).

8. Thyroid hormone receptors (THRs) regulate metabolism and heart rate as well as play critical roles in the development of organisms. A member of THRs sub-family, THRA likes to bind ER4 of RGGTCR, by hetero-dimerizing with RXRA (Figure 3.11b).

9. Although THRA (thyroid hormone receptor- alpha) and THRB (thyroid hormone receptor- beta) have exactly the same DNA binding helix and known to bind similar DNA sequences [19], still THRA preferred inverted repeat of RGGTCR with no gap (IR0), whereas THRB binds direct repeat of RGGTCR with gap of 4 nucleotides

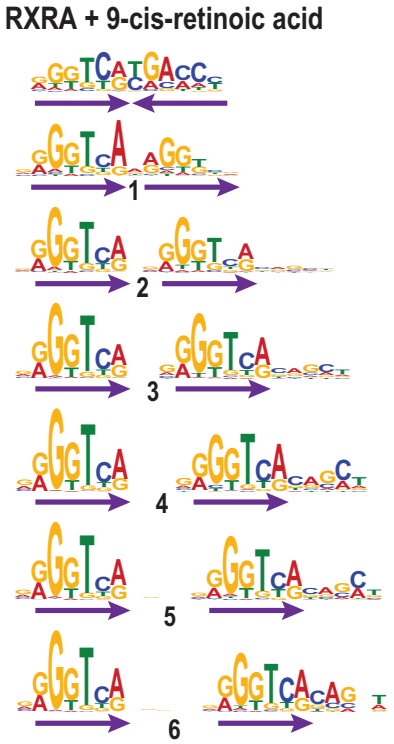
(DR4), which can be due to allosteric effects of non-DNA contacting residues and/or different preferences of dimerization as homo-dimer and hetero-dimer.

RXRA heterodimerizes with other NRs giving rise to multiple binding motifs

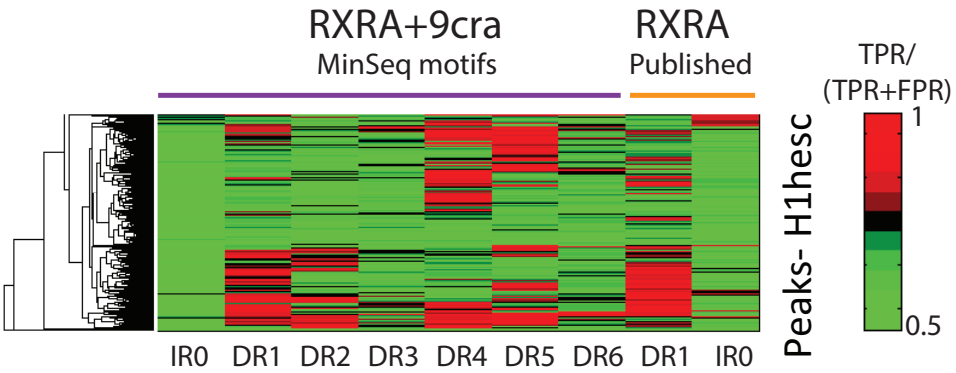
RXRA has a unique property that it hetero-dimerizes with many NRs and binding motif varies according to the partner protein. In the experimental design of the CSI experiments employed, each nuclear receptor was overexpressed in a human cell line, and the extract is used to perform CSI. Thus, it is possible that the nuclear receptor of interest can dimerize with protein partners found in the extract. To determine whether this occurs, we search for different binding motifs adjacent to the RXRA motif. In Figure 3.13a various PWM motifs are shown for RXRA. To validate MinSeqFind discoveries, we score published in vivo binding peaks using results from MinSeqFind. Bound regions are used as positive and two random permutations of each peak are used as negative and are scored using PWMs to plot receiver operating characteristics (ROC) curve. In Figure 3.13b top 500 ChIP-Seq peaks in H1hesc cell line for RXRA protein [10] are scored using PWMs obtained from PWMs from MinSeqFind for HT-SELEX binding data for RXRA with 9cra ligand. Each peak was assigned a score S equal to maximum of TPR/(TPR+FPR) where it was first detected as positive peak in ROC (TPR-true positive rate & FPR- false positive rate). A heatmap was then generated on the basis of these results, clustered according to their score. Figure 3.13b shows multiple motifs IR0, DR1-6 make prediction of different in vivo peaks which were not found in published results of RXRA binding [33].

SNPs associated with diseases related to NR binding

Most single nucleotide polymorphism (SNPs) associated with disease occur in non-coding regions, making their function unclear. To determine if any of these SNPs would alter the binding of NRs, we studied a compendium of 5076 carefully curated SNPs sites associated with diseases and quantitative traits [39] and predicted change (gain/loss) in binding of NRs at these SNPs. MinSeqFind scores 20 bp region around reference allele (hg19) and alternate allele of these SNPs using MinSeqFind and calculated fold change in predicted binding of NRs. As shown in Figure 3.14 via genomescapes, MinSeqFind detect that at MODY1 (maturity onset diabetes of the young-1) SNP rs1893217 which is associated with Type I diabetes, change from nucleotide adenine (reference allele hg19) to guanine (alternate allele) removes the best binding site for HNF4A in a region of 5000 bp around that SNP. Thus, a strong contributor to the Type I diabetes caused by that SNP and is consistent with what was discovered by earlier studies [39]. In all there are a total 353 SNPs (out of 5076 [39]) predicted to cause two fold or more change in binding of at least one NR and are displayed in Figure 3.15 in a heatmap as columns,



(a) MinSeqFind extracted binding motifs of RXRA with 9cra ligand



(b) In vivo binding as predicted by different motifs of RXRA using MinSeqFind Vs published (rows are top 500 ChIP peaks and columns are different motifs, heat-map shows different peaks predicted by different motifs with red as predicted, green as not predicted).

Figure 3.13: Multiple binding preferences of RXRA observed by MinSeqFind are found in vivo

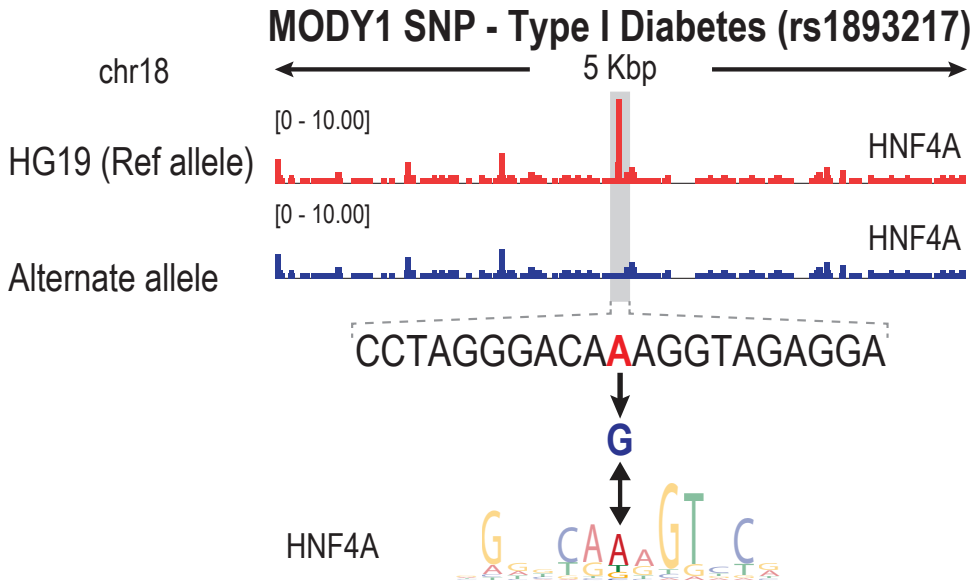


Figure 3.14: Genomescape plot of DNA binding of HNF4A for MODY1 SNP associated to Type I Diabetes (rs1893217)

with rows corresponding to different NRs. Out of these more than 300 were previously not annotated to NR binding (RegulomeDB [13]).

Published HT-SELEX binding data

We use available HT-SELEX data used by the authors in paper Jolma et al. [33]. Since authors of Jolma et al. didn't use any mock control thus we use here previous round data or round 0 data to normalize against the provided round data of the TF to get enriched MinSeqs and PWMs using MinSeqFind. Due to this reason MinSeqFind will capture not only DNA binding due to transcription factors, but also unintended bias introduced by factors like PCR and magnetic beads, which could have been resolved had authors used mock control instead.

Figure 3.16a compares PWM motifs extracted by MinSeqFind from HT-SELEX binding data to those from DeepBind [2], which is the best published computational method. DeepBind captured DNA binding for nuclear receptor as direct repeat of RGGTCA (R=A/G) with gap of 1 (DR1) and failed to capture inverted repeat of same with gap 0 (IR0) which is a known RXRG binding motif. MinSeqFind discovered both the binding preferences for RXRG – DR1 and IR0 of RGGTCA from the same data. Figure 3.16b compares MinSeqFind to DeepBind using area under receiver operating characteristic (AUROC) metric for nuclear receptor proteins. In vivo bound ChIP peaks are used as true positive and two random permutations as true negatives and peaks are scored using corresponding protein's in vitro DNA binding as measured by HT-SELEX and modeled by MinSeqFind and DeepBind algorithms. Figure 3.16b clearly shows the

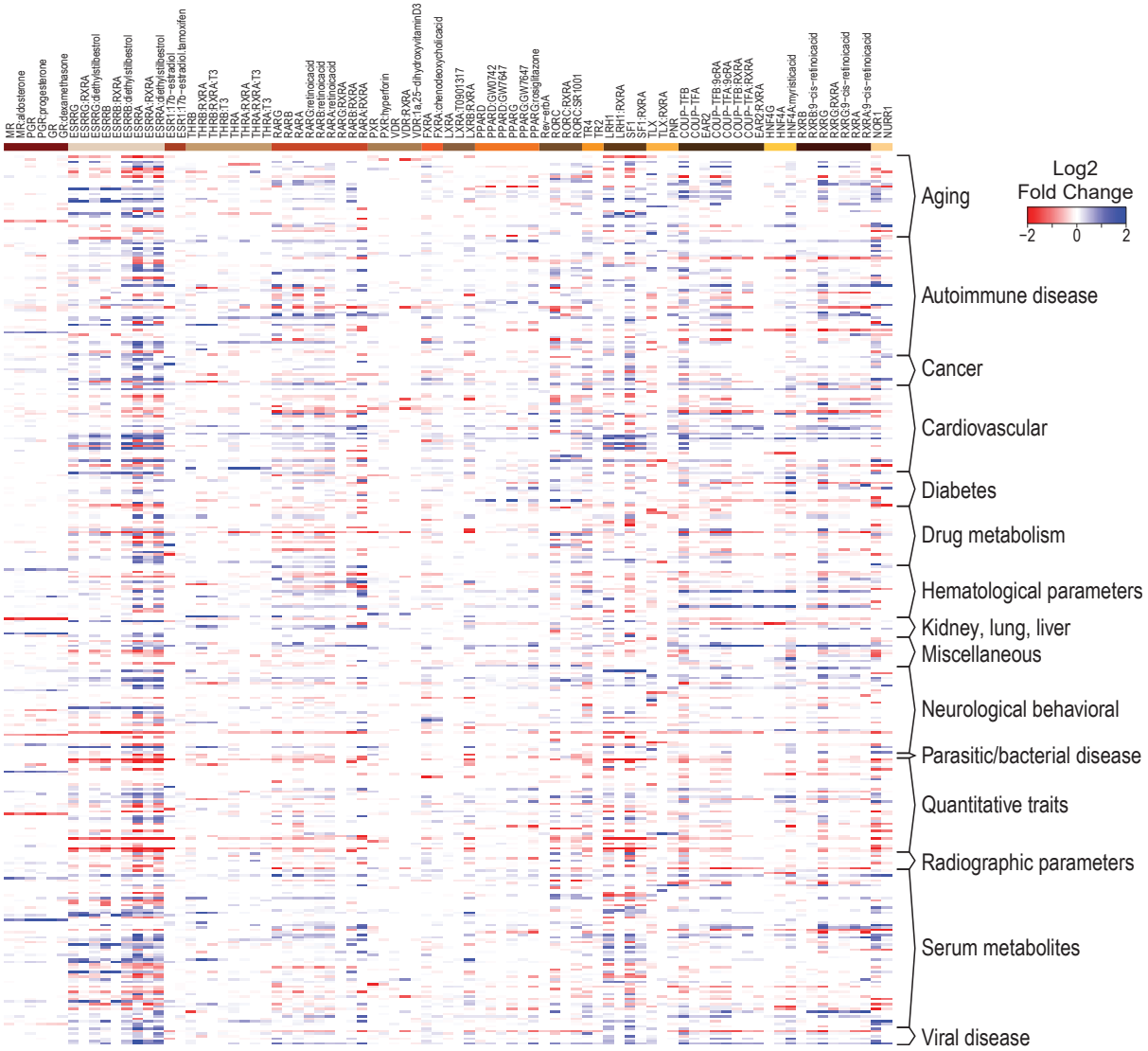
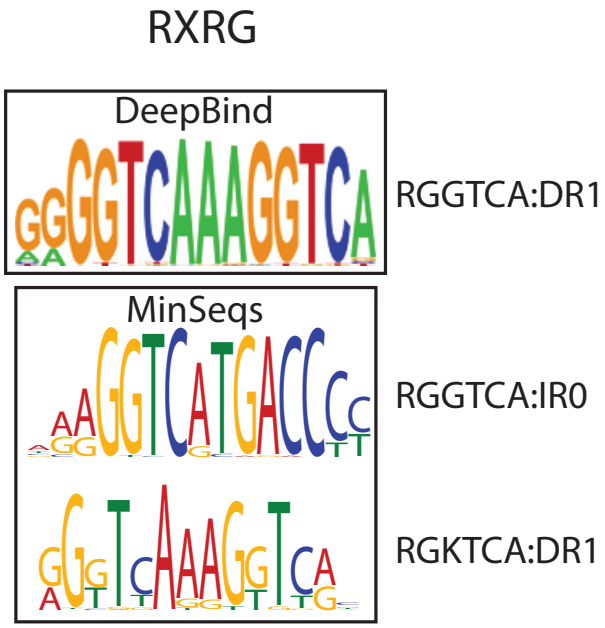
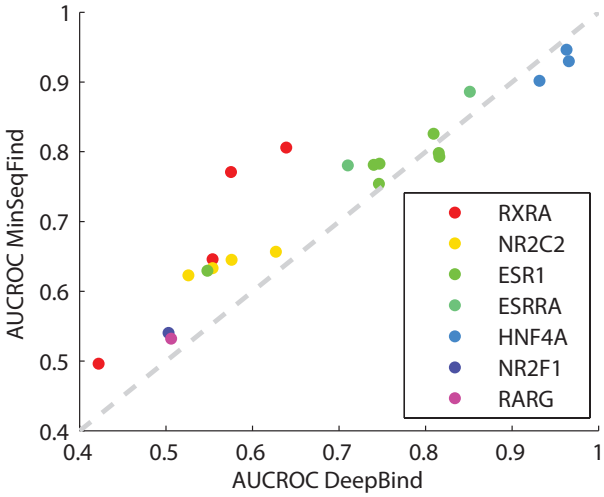


Figure 3.15: Predicted change in DNA binding of NRs due to SNPs linked to disease and quantitative traits. Columns are different NRs with partner and/or ligands and rows are SNPS that confer two-fold change in binding by at least one NR. SNPs are arranged by corresponding associated disease class or trait class. Heatmap plots log2 fold change in DNA binding of NRs as predicted by MinSeqFind, red displays loss in DNA binding of NRs whereas blue gain in binding at these SNPs.



(a) DeepBind fails to capture known binding motifs for published data



(b) Area under receiver operating characteristic (AUROC) as obtained by MinSeqFind and DeepBind model of HT-SELEX binding data on ChIP peaks

Figure 3.16: Comparison of MinSeqFind to state of the art method DeepBind [2] in modeling published in vitro protein–DNA binding of nuclear receptors

greater performance of MinSeqFind in comparison to DeepBind in modeling DNA binding with gaps as exhibited by most of the nuclear receptor proteins.

3.8 Conclusion

MinSeq is a novel way of representing multiple complex binding patterns, from monomer to dimer with gaps and multiple orientations. MinSeqFind presented here not only captures such complex motifs in form of MinSeqs, but also filters out many platform

related biases. Our analysis of NR DNA binding using MinSeqFind is a novel study providing HT in vitro DNA binding for NRs. MinSeqFind captures many similarities and differences among NRs of same sub-family, which cannot be predicted just on the basis of their amino-acid sequence homology [69]. Many repeat preferences of NRs has been reported in previous studies [19], but not the relative binding of these repeats. MinSeqFind discovers novel binding patterns and those which don't fall in the category of monomer or perfect dimers reported earlier. RXRA is well known to dimerize with other NR to cause change in their activation/repression function. MinSeqFind discovers that by dimerizing with RXRA NRs change DNA binding and thus RXRA exhibits multiple binding patterns which are also found in vivo. Evaluation of 5076 disease linked SNPs [39] using MinSeqFind resulted in mapping of 353 SNPs responsible for change in DNA binding of NRs. This repertoire of NR DNA binding data using MinSeqFind provides a useful resource for several genomic targets of NR and diseases caused by them, which will prove to be of immense help in development of drugs targeting NR DNA binding.

Chapter 4

Summary & Future Directions

4.1 Summary

DNA binding proteins are responsible for multiple cellular functions. Most DNA-binding proteins bind to DNA and control gene expression. Such proteins are known as transcription factors (TFs). DNA binding of TFs can be very complex as shown for the case of nuclear receptor proteins, for example RXRA binds to DNA as a dimer of itself or other nuclear receptors in many different patterns. Such multiple complex patterns are required for a better study of protein–DNA binding and gene-regulatory networks. Existing methods do not fully comprehend the complexity of protein–DNA binding. We propose MinSeqs as novel way to capture such multiple complex binding, each MinSeq is a stretch of nucleotide sequence containing DNA binding information in the form of weighted MinSeq score.

In Chapter 2 we use compressed sensing algorithms like compressive sampling matching pursuit (CoSaMP) for sparse solution and develop Compressed sensing based motif extraction (COSME) as a tool to extract MinSeqs from protein–DNA binding data obtained from microarray. To find differences in binding of proteins of same family, we also develop Differential-COSME or Diff-COSME. Diff-COSME not only captures complex DNA binding, but also facilitates as tool to differentiate binding of multiple TFs having similar binding profile. State of the art of methods fail to capture such subtle differences especially among the proteins of same family and different alleles of the same protein. Such differences becomes key when proteins with similar binding profile compete for same genomic location inside a cell and is deciding factor for which protein will bind to activate/repress gene expression leading to any genotypic or phenotypic change.

Proteins like nuclear receptors exhibit complex DNA binding as explained earlier, such binding measurement cannot be accommodated in a microarray. Thus, to study complex binding patterns we use HT-SELEX binding data and develop MinSeqFind

to extract such binding patterns. MinSeqFind solves many platform-related issues of HT-SELEX data, like pseudo-random library, protein bias for binding to primer region and PCR. MinSeqFind uses a novel way to weight binding for sequences of different lengths and use orthogonal matching pursuit (OMP) based approach to get compressed MinSeqs providing normalized enrichment values of binding. MinSeqFind when applied to members of nuclear receptor protein family yields amazing new findings. MinSeqFind capture multiple complex binding patterns and finds differences due to RXRA hetero-dimerization and ligand addition. These new patterns discovered made better prediction of protein binding in vivo (Figure 3.13b). MinSeqFind identifies more than 300 new diseases and quantitative traits related SNPs linked to NRs, which were previously unannotated to NR binding. Such novel and comprehensive study of DNA binding serves as a great resource by capturing many published and new motifs for NRs. In this study we focus mostly on NRs as it is one of the most drug targeted protein family and serves as a good example of complex DNA binding, but MinSeqs and MinSeqFind are versatile tools. We can extend these approaches to thousands of other proteins for which HT DNA binding data is available.

4.2 Future directions

DNA Binding in HT-SELEX using replicate data

Unlike microarray binding data, HT-SELEX has ability to capture longer and more complex DNA binding and it also provides an experimental protocol to multiplex hundreds of experiments together. But with these additional benefits there are many problems associated with HT-SELEX data as presented in previous chapter. In microarray binding experiments, different DNA probes can be used in the binding experiment independent to other probes and thus part of the binding data can be used to train model and part of it to reject over-fitting in form of cross-validation. Since the same random N-mer sequence was used for HT-SELEX binding thus dividing the data into different parts and modeling will not help to reject over-fitting, as they don't serve the purpose of independent experiment like probes in a microarray.

Due to above limitations any model for HT-SELEX can face the problem of over-fitting. The two proposed solutions are a) using experimental replicates of HT-SELEX data- replicates are used a lot in the field, but haven't been used to get a better DNA binding model for HT-SELEX data, and b) new experimental design- using more than one DNA library in same experiment and after sequencing, separate data from these libraries. In both cases we can use one part for training and other to reject over-fitting. We expect computational methods like MinSeqFind can take advantage of such strategy and can build a better binding model free from over-fitting.

Additional parameters for DNA binding

Protein–DNA binding can also be affected by other factors like shape of DNA sequence and DNA methylation. a) DNA Shape – Although most of the genomic DNA is double stranded helix, it doesn't follow the same shape pattern. The shape of even free DNA also depends on the nucleotide sequence and each DNA sequence exhibit specific three-dimensional shape. In recent studies the shape of DNA sequences has been evaluated using multiple DNA structures [55]. DNA binding of proteins get affected by such shape features, particularly like stretches of A or T nucleotides example AAAAA or TTTTT [17]. Modeling these shape based effects will give another dimension to MinSeqs other than A, C, G and T nucleotides and can capture protein–DNA binding better especially for the shape affected. b) DNA methylation – A methyl group attaches to different nucleotides of DNA, most commonly to cytosine, and in many cases alters DNA binding. Each methylated base has to be treated as a new nucleotide beyond the unmodified four bases [67].

In vivo protein–DNA binding

Till now we used MinSeqs to capture in vitro DNA binding motifs and use them for prediction of in vivo bound peaks for comparison. Computational approaches can be developed to extract information from in vivo binding in the form of MinSeqs. Multiple factors has to be accounted for, including a) genomic DNA is wrapped around histones and thus not all of it is accessible to proteins, thus a model of histone binding and accessible DNA region is required, b) cell-type and cell-state specific factors – there are many proteins in different concentrations at different time points in different cell types, making binding very specific to that particular cellular environment. Thus, there can be millions of factors that can contribute to unusual binding and errors. Also, reproducibility of in vivo data has always been an big issue. When modified for in vivo data, we expect to discover multiple binding partners and co-factors of proteins in form of MinSeqs similar to discovery of RXRA partners in this study. Identifying co-factors and partners in different cellular environment can help in distinguishing cell-specific behavior, and cell-specific gene-network and metabolic-network.

4.3 Conclusion

Here we address data extraction and compression techniques for protein–DNA binding affinity data. MinSeqs introduced here serves as a novel way to characterize binding. Computational tools like Diff-COSME and MinSeqFind developed to extract MinSeqs, gave novel insights to DNA binding of proteins exhibiting simple as well as those having complex binding patterns. Diff-COSME and MinSeqFind found binding differences

among closely related family members crucial for differentiating them. A detailed study of DNA binding of members of nuclear receptor protein family using MinSeqs has been provided here. MinSeqFind found RXRA dimerizes with different nuclear receptor proteins to bind DNA in multiple orientations and with different gaps in middle, which are also found in vivo. Many novel patterns and motifs extracted by MinSeqFind predicted hundreds of disease associated SNPs affected by NR DNA binding, which were never earlier mapped to NR. Use of MinSeqs can be further extended to in vivo binding as well to study DNA binding with multiple partner proteins and in effect of cellular environment.

Appendix A

Protein–DNA binding data and analysis

A.1 Types of binding data

There are following different types of datasets that is obtained from various experimental methods for studying protein–DNA binding-

Type 1: Complete k-mer intensity

Output data where all permutations of k-mer DNA sequences are there with corresponding binding intensities for array data, are considered in this category. CSI array (Figure 1.1b) yield data of such kind, 4^{10} sequences and corresponding binding intensities are obtained after experiment and normalization.

Type 2: Complete k-mer count data

In this category we consider sequencing based methods for which corresponding to each k-mer DNA sequence there is a 'count' that represents how many times that particular sequence has appeared in sequencing reads. This category is valid when there is enough representation of all k-mer sequences, so that computational tools can be applied here similar to previous category. HT-SELEX for 10mers [63,71] is a sequencing based method with 10 base pair random DNA sequence. The protein binding DNA strands were sequenced deeply that there were enough representation for each 10bp sequence.

Type 3: Limited k-mer intensity data

Array data with not all possible k-mers represented on the array, comes under this category eg. ChIP-chip and PBM (Figure 1.1a). In PBM array all 10bp DNA sequences are arranged on limited number of separate array features 35bp long using de-bruijn based method [8]. Note that this is not equivalent to first category where separate independent intensity for each 10mer is available. Thus in PBM arrays finally 8bp sequence intensity data is used after processing 35bp data.

Normalized 8-mer intensity data thus obtained can used as "Complete k-mer intensity" data for some methods.

Type 4: Limited k-mer count data

All other sequencing based methods fall in this category. In this case there is not enough representation of k-mer data. For example- ChIP-Seq [31,40], HT-SELEX [32,33] (Figure 1.1c), SELEX-Seq [59] and Bacterial 1 Hybrid system [43,44]. In such datasets not all k-mer have counts mainly because k is too large or sequenced with less reads.

Type 5: Best Binders or Peak format

Most of the ChIP-Seq data is converted to peaks in the genome and then best of those are used to get binding motif [10]. All the other data formats can be easily converted to best binding sequence by putting a threshold to binding intensity or counts and rejecting all peaks below that.

A.2 Computational Methods for Protein–DNA binding analysis

Table A.1: Computational methods for protein–DNA binding data analysis-

Method	Publication Year	Algorithm description and usage
MEME [4,5]	1995	Based on expectation maximization, it is a tool for discovering motifs in sets of protein or DNA sequences. It can search for multiple motifs of different lengths. Used generally for ChIP peaks (type-5).
Weeder [49] [50]	2001	Uses best binding sequence and considers fixed number of mutations to that and search among the mutated sequences to get the best sequence motif (type-5).
BioProspector [35]	2001	Uses Gibbs sampling strategy to search for motif. In addition it uses 0 to 3rd order Markov model for background. It accounts for multiple motif length and gaps between them. Used for co-expressed genes mainly (type-5).

Table A.1: Computational methods for protein–DNA binding data analysis-

Method	Publication Year	Algorithm description and usage
MDScan [36]	2002	Start with searching motif from highly enriched ChIP Peaks and then update and refine (type-5).
Seed-and-Wobble [9]	2006	Uses E-scores of best 8-mer and its single mismatch variants to construct PWM motif (type-3).
MatrixREDUCE [26]	2006	Performs a least square fit of data to affinity logo (biophysical model defined in [61]) (type-3).
RankMotif++ [18]	2007	Maximizes the likelihood of set of binding preference on the basis of the rank of k-mers (type-3).
CSI-Tree [34]	2008	Since normal PWM is a compressed version of all the data, CSI-Tree uses regression tree based approach to divide PWM into set of PWMs which are not compressed (type-1).
BEEML [71]	2009	Finds maximum likelihood estimate of parameters to a biophysical PWM or dinucleotide model, including the TF's chemical potential, nonspecific binding affinity (type-2).
SSL [17]	2010	Sequence specificity landscape is a visualization of all k-mer based binding data arranged in form of concentric circles around a seed motif (type-1 & 2).
BEEML-PBM [72]	2011	BEEML modified for PBM data which also includes probe position-specific effects (type-3).
Annala et al. [3]	2011	For PBM constructed an indexing matrix of contiguous k-mers (size 4-8) on each 35bp probe. Applied conjugate gradient method to fit intensity to k-mers. They include top few 7mers and 8mers to reduce number of variables (type-3).

Table A.1: Computational methods for protein–DNA binding data analysis-

Method	Publication Year	Algorithm description and usage
AutoSeed [33]	2013	Semi-automatic method for HT-SELEX/SELEX-Seq data, it starts with a seed sequence and builds a PWM from all sequences with 1 mismatch (type-4).
FeatureREDUCE [53]	2015	Combines a biophysical free energy model (PWM or dinucleotide) with a contiguous k-mer background model (type-3).
DeepBind [2]	2015	Uses deep learning based approach to get upto 16 PWMs motif with different weights.

Appendix B

Compressed Sensing Algorithms

Following are two algorithms to solve a linear under-determined system represented by equation B.1 (similar to equation 2.12), given the knowledge of sparsity of the system.

$$y = Hx + \eta \tag{B.1}$$

B.1 Orthogonal Matching Pursuit (OMP)

Taken from [48] [22]

Algorithm 5 Orthogonal Matching Pursuit

Input: CS matrix H, measurement vector y
Output: Sparse representation $\hat{x}$
  Initialize: $\hat{x}_0 = 0, r = y, \Omega = \emptyset, i = 0$
  while halting criterion false **do**
    $i \leftarrow i + 1$
    $b \leftarrow H^T r$ {form residual signal estimate}
    $\Omega \leftarrow \Omega \cup \text{supp}(\tau(b, 1))$ {update support with residual}
    $\hat{x}_i|_\Omega \leftarrow H_\Omega^\dagger y, \hat{x}_i|_{\Omega^c} \leftarrow 0$ {update signal estimate}
    $r \leftarrow y - H\hat{x}_i$ {update measurement residual}
  end while
  return $\hat{x} \leftarrow \hat{x}_i$

B.2 Compressive Sampling Matched Pursuit (CoSaMP)

Taken from [41] [22]

Algorithm 6 Compressive Sampling Matched Pursuit

Input: CS matrix H, measurement vector y, sparsity K
Output: K-sparse approximation $\hat{x}$ to true signal x

Initialize: $\hat{x}_0 = 0, r = y, \Omega = \emptyset, i = 0$
while halting criterion false **do**
  $i \leftarrow i + 1$
  $e \leftarrow H^T r$ {form residual signal estimate}
  $\Omega \leftarrow supp(\tau(e, 2K))$ {prune residual}
  $T \leftarrow \Omega \cup supp(\hat{x}_{i-1})$ {merge supports}
  $b|_T \leftarrow H_T^\dagger y, b|_{T^C} \leftarrow 0$ {form signal estimate}
  $\hat{x}_i \leftarrow \tau(b, K)$ {prune signal using model}
  $r \leftarrow y - H\hat{x}_i$ {update measurement residual}
end while
return $\hat{x} \leftarrow \hat{x}_i$

Appendix C

Experiments Performed

Here different experimental procedure for Nuclear Receptor Cloning and CSI by HT-SELEX is explained -

C.1 Cloning and Expression (performed by Jacqui Mendez)

Plasmids containing N-terminus HaloTag fusions of human nuclear receptors were obtained from Kazusa DNA Research Institute (Kisarazu, Chiba, Japan). HEK293T cells were grown in DMEM media supplemented with 10% FBS at 37 Celsius in an atmosphere of 5% CO2. Cells were transiently transfected using FuGENE HD Transfection Reagent (Promega, Madison, WI, USA) following the manufacturer's protocol. After 24-48 hr at 37 Celsius and 5% CO2, cells were washed with ice cold PBS, scraped and collected in a conical centrifuge tube. Cells were lysed in Mammalian Lysis Buffer (50 mM Tris-HCl pH 7.5, 150 mM NaCl, 1% Triton X-100, 0.1% sodium deoxycholate) supplemented with protease inhibitors. Cell lysates were centrifuged, the clear supernatant was transfer to a clean microcentrifuge tube, flash frozen in $N_2(l)$, and stored at -80 Celsius. Expression of the HaloTag fusions was confirmed by SDS-PAGE.

C.2 Cognate Site Identification (CSI) by HT-SELEX (performed by Jose A. Rodriguez-Martinez)

Cognate binding sites for HaloTag-human nuclear receptor (HaloTag-hNR) transcription factors were determined by HT-SELEX. A DNA library with a 20 bp random region flanked by constant sequences to allow PCR amplification was used. In vitro selections were performed by incubating the DNA library (100 nM in 20 μL) with cell lysate overexpressing a HaloTag-hNR in binding buffer (25 mM HEPES (pH 7.4), 80 mM KCl,

0.2 mM EDTA, 1 mM MgCl2, 0.1 mM ZnSO4, 2.5 mM DTT, 50 ng/ul poly dI-dC, 0.1% BSA) for 1 hr at room temperature. HaloTag-hNR bound DNA was enriched using Magne HaloTag magnetic particles (Promega) following manufacturers specifications. After immobilization on the magnetic particles, three quick washes with 100µL of ice-cold binding buffer were performed to remove unbound DNA. The magnetic beads were resuspended in a PCR master mix (EconoTaq PLUS 2X Master Mix, Lucigen) and the DNA was amplified for 18 cycles. Amplified DNA was purified (QIAGEN), quantified by UV absorbance at 260 nm, and used for subsequent binding rounds. A total of 3 rounds of selection were performed. After selection, an additional PCR was done to incorporate a 6 bp 'barcode' and Illumina sequencing adapters. The starting library (Round 0) was also barcoded. Samples were combined and sequenced in an Illumina HiSeq 2000 instrument.

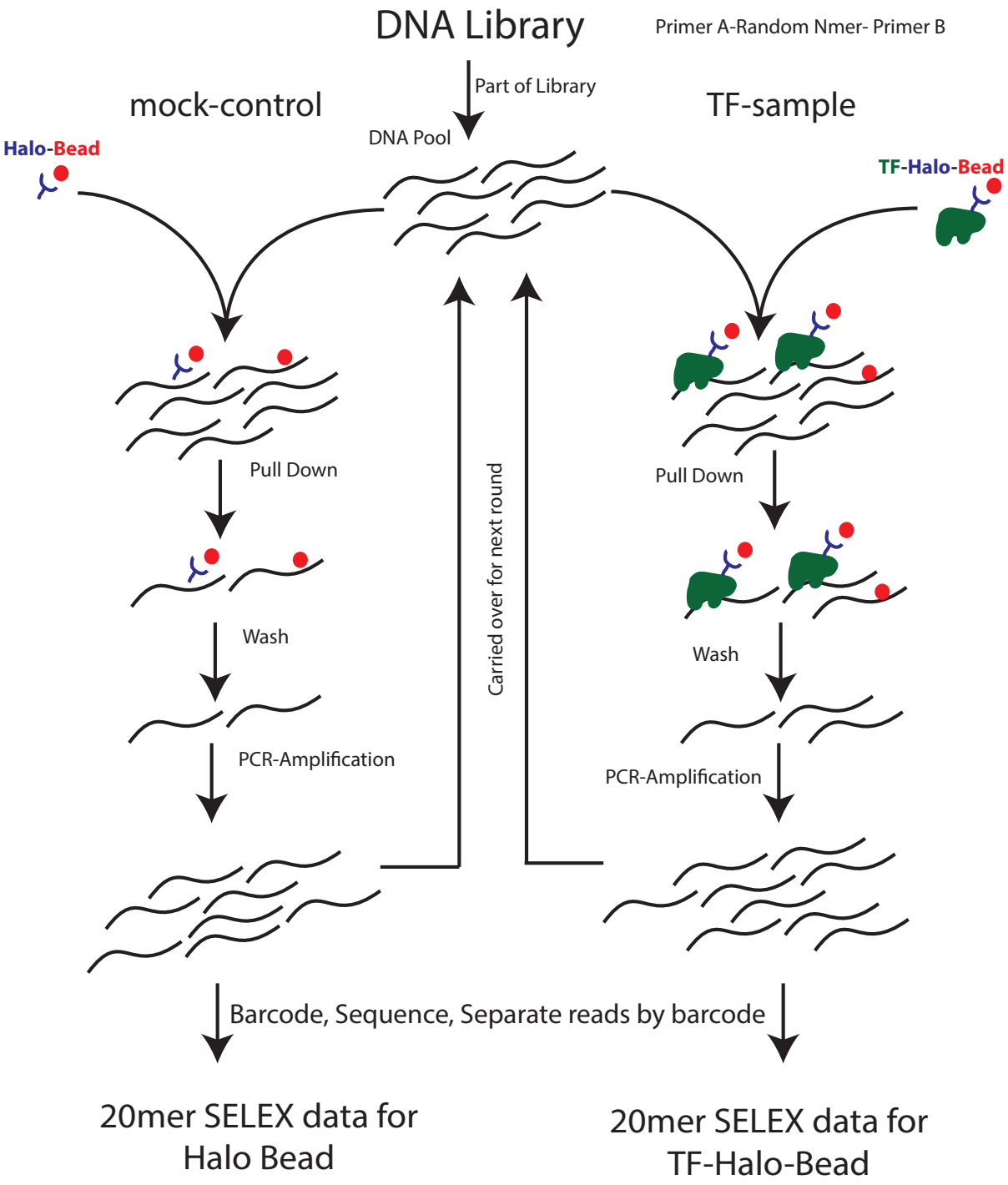


Figure C.1: HT-SELEX protocol [32,59,64]- Experiment performed by Jose A. Rodriguez-Martinez

Appendix D

Sequence Specificity and Energy Landscapes of Nuclear Receptor DNA Binding

Specificity and energy landscapes (SELs) provides three dimensional display of high-throughput protein–DNA (or protein-RNA) binding data through series of concentric rings [14, 17, 64]. The height of each color-coded peak corresponds to the binding intensity, which can be measured by different experimental platforms. SEL for binding of all k-mers is built around a seed sequence or motif as reference, relative to which sequences are arranged on SEL. The seed sequence is derived from top scored MinSeqs or PWM. Sequences are arranged in such a way that similar sequences appear together and no sequence is repeated.

Gapped Sequence Specificity and Energy Landscapes (gapped-SELs)

We develop gapped SELs to display DNA binding of proteins which binds to sequences with multiple gaps like nuclear receptors in this study-

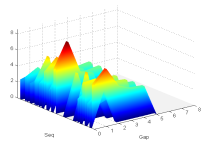
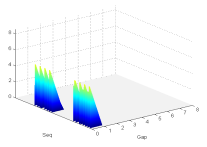
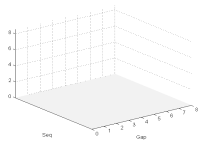
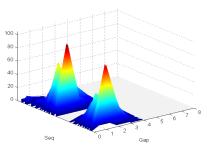
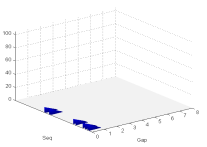
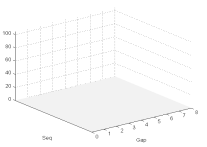
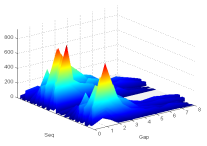
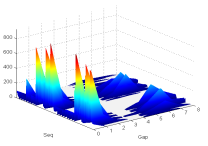
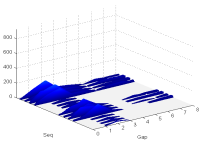
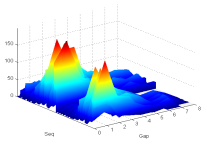
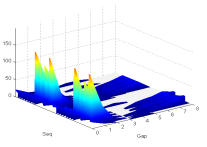
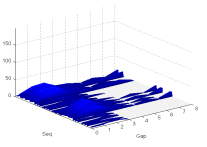
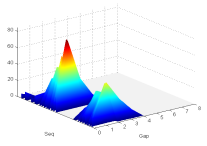
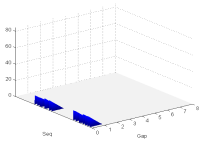
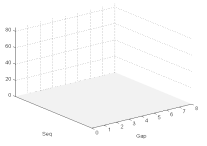
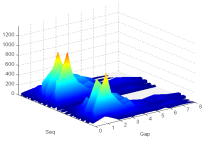
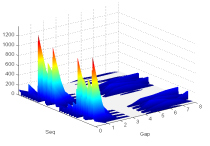
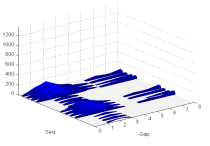
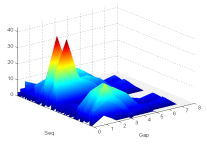
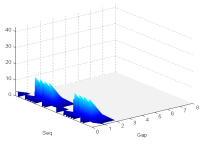
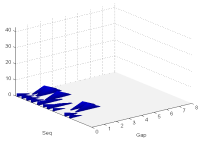
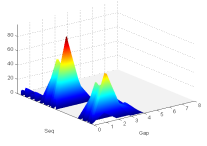
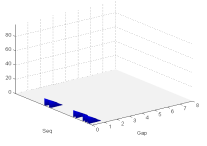
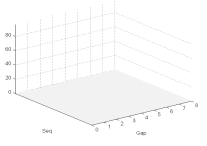
1. First a seed sequence is chosen as combination of two monomers. Monomers are chosen from top ranked MinSeqs or PWMs. The monomers are combined with multiple orientations and different gaps to construct multiple seeds. For example we take RGGTCR (R=A/G) as our starting monomer for RXRA. By adding multiple gaps between 2 such monomers with Ns surrounding those (a single N was added on opposite sides in this case) we get - NRGGTCRRGGTCRN, NRGGTCRxRGGTCRN, NRGGTCRxxRGGTCRN, …as seed for direct repeats of GGTCA with gap 0, 1 and 2. Similarly NRGGTCRYGACYN, NRGGTCRxYGACYN, NRGGTCRxxYGACYN, …as seeds for inverted repeats of GGTCA with gap 0, 1, and 2 (Y=C/T). Gaps in gapped-SELs is represented by x.

2. All the sequences matching the seeds are then obtained by replacing N and other nucleotides (excluding gap which is represented by x) with degenerate nucleotides A, C, G and T, for example N with A, C, G and T, R with A and G, Y with C and T. In the above example, each seed will give rise to total 2^4*4^2 = 256 different matching sequences.

3. A 3D plot for the matching sequences is made. All the sequences corresponding to gap=g are arranged along X-axis with Y coordinate=g. The sequences along X-axis are plotted in order with preference to degenerate nucleotides from the monomer, and then to the flanking bases (Ns) i.e. in NRGGTCRRGGTCRN, preference is given to what is replacing R (by A or)G over what is replacing N (by A, C, G or T). The sequences are ordered in the order of first A, then C, G and T. The same order of sequences is followed for all gaps along X-axis, example AAGGTCAAGGTCAT and AAGGTCAxAGGTCAT will have same x-coordinate, but have y=0 and y=1 Y-coordinate respectively. After deciding X and Y coordinate, the enrichment then is plotted at that coordinate with height and color representative of it. Enrichment of all the sequences for gapped-SELs is obtained using MinSeqFind for NRs. Gapped-SELs for the NRs are displayed in following sections, the corresponding monomer seed that is used as reference to build gapped-SELs is also shown at the top.

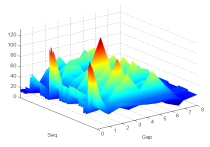
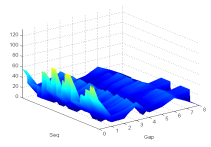
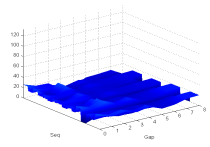
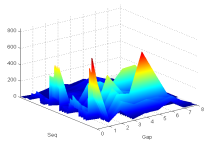
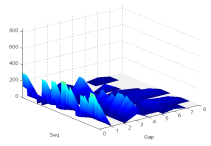
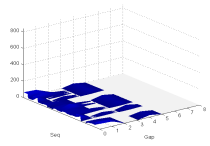
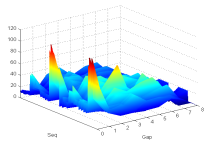
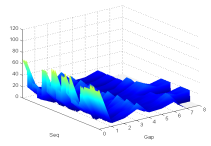
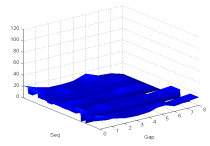
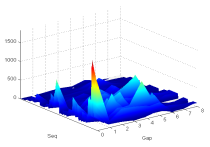
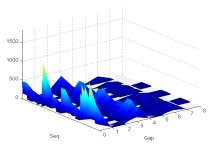
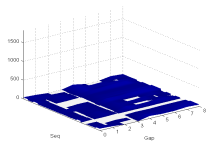
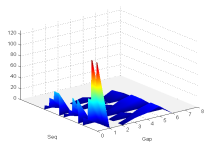
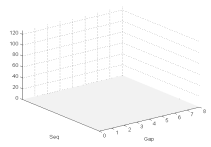
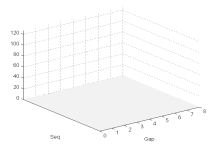
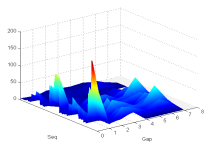
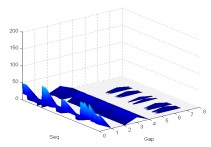
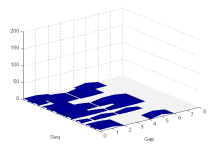
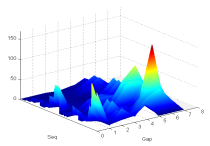
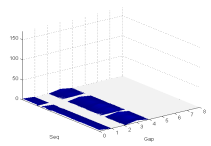
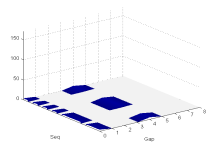
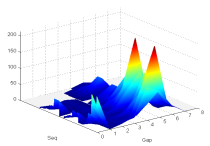
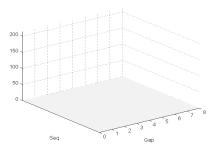
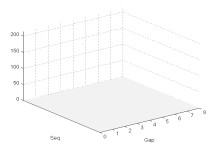
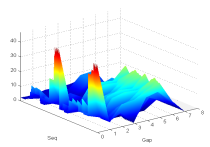
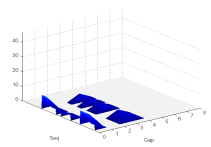
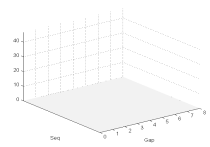
D.1 COUP/EAR (NR2Fs) family

Monomer seed: RGKTCR

Protein (with partner and/or ligand)	Enrichment of Direct repeats	Enrichment of Inverted repeats	Enrichment of Everted repeats
COUP-TFA			
COUP-TFA:RXRA			
COUP-TFA:9cRA			
COUP-TFB			
COUP-TFB:RXRA			
COUP-TFB:9cRA			
EAR2			
EAR2:RXRA			

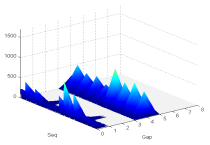
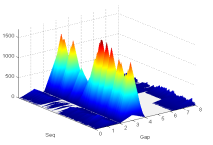
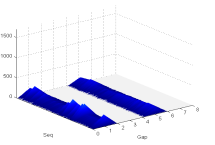
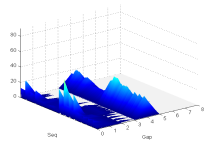
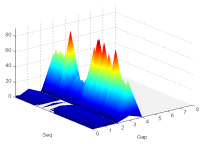
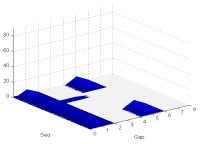
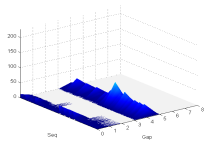
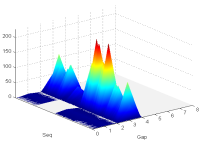
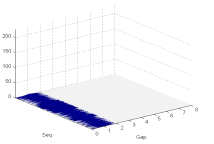
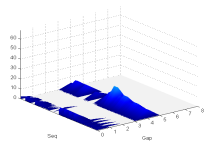
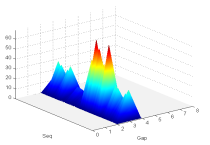
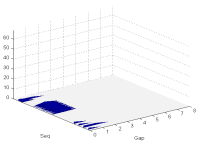
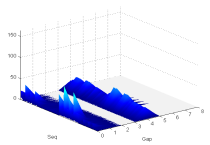
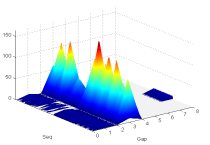
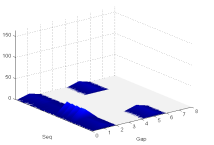
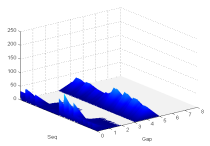
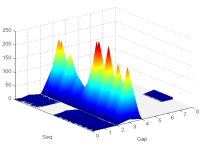
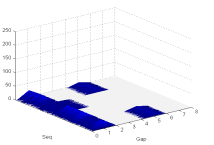
D.2 Estrogen related receptor family (ESRRs or NR3Bs)

Monomer seed: RGGTCR

Protein (with partner and/or ligand)	Enrichment of Direct repeats	Enrichment of Inverted repeats	Enrichment of Everted repeats
ESRRA			
ESRRA:RXRA			
ESRRA:diethylstilbestrol			
ESRRB			
ESRRB:RXRA			
ESRRB:diethylstilbestrol			
ESRRG			
ESRRG:RXRA			
ESRRG:diethylstilbestrol			

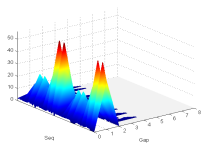
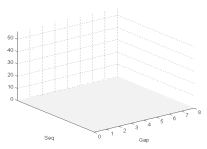
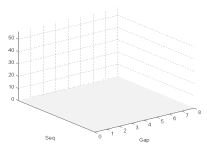
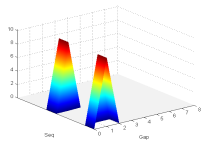
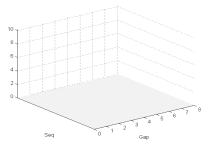
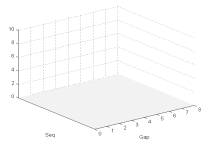
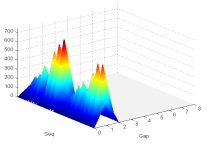
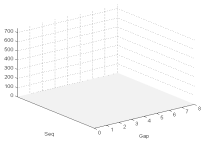
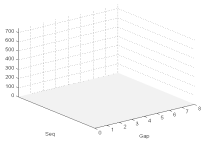
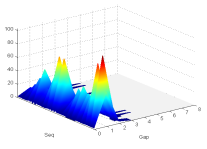
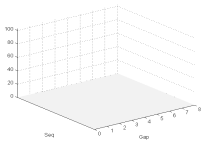
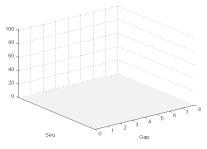
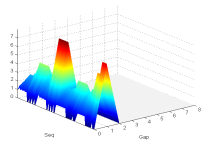
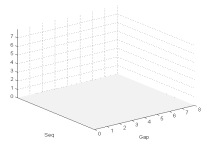
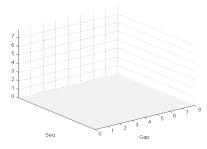
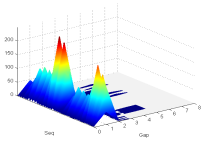
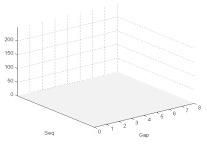
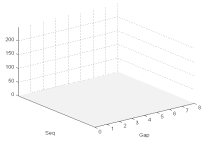
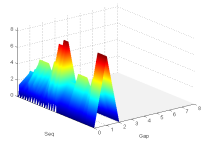
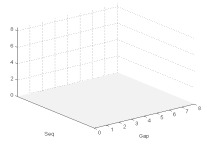
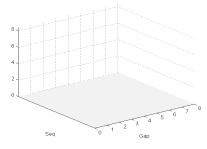
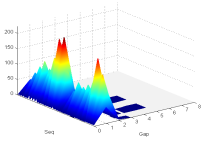
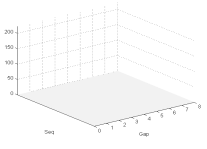
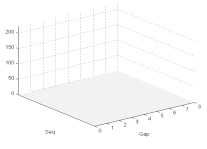
D.3 3-Ketosteroid receptors family (NR3Cs)

Monomer seed: GNACR

Protein (with partner and/or ligand)	Enrichment of Direct repeats	Enrichment of Inverted repeats	Enrichment of Everted repeats
GR			
GR:dexamethasone			
MR			
MR:aldosterone			
PGR			
PGR:progesterone			

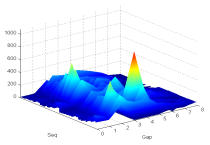
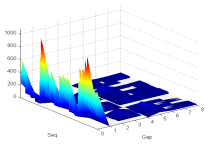
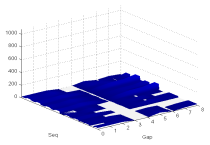
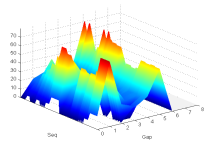
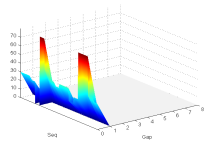
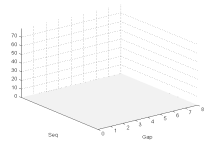
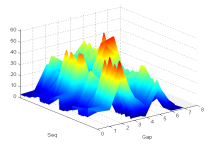
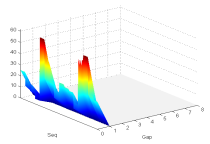
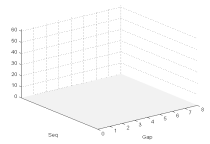
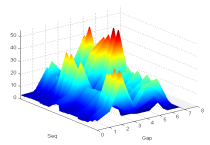
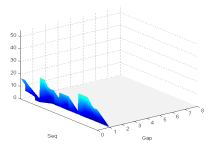
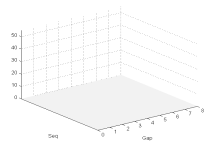
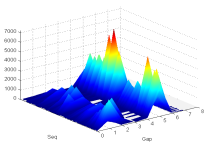
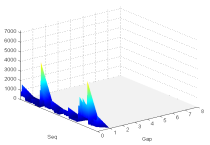
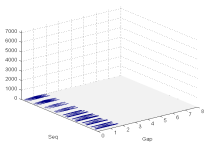
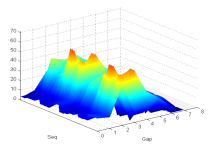
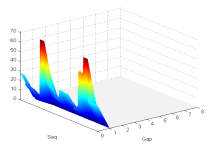
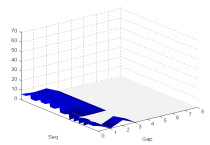
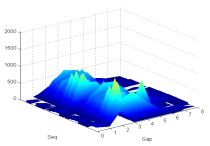
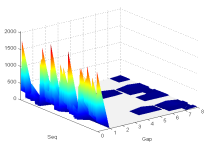
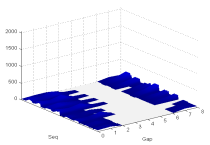
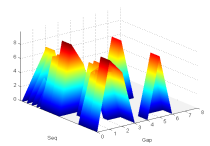
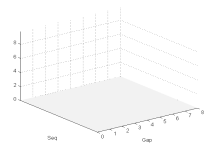
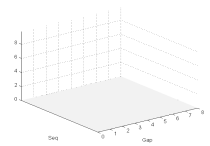
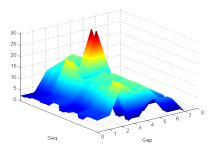
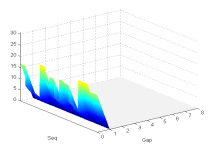
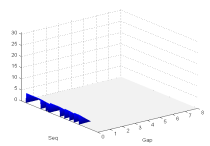
D.4 Peroxisome proliferator-activated receptor family (PPARs or NR1Cs)

Monomer seed: RGGTCR

Protein (with partner and/or ligand)	Enrichment of Direct repeats	Enrichment of Inverted repeats	Enrichment of Everted repeats
PPARD			
PPARD:RXRA:T1AM			
PPARD:RXRA:TA1			
PPARD:GW0742			
PPARD:GW7647			
PPARG			
PPARG:GW7647			
PPARG:rosiglitazone			

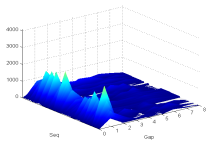
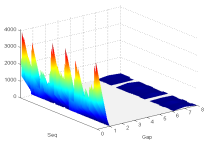
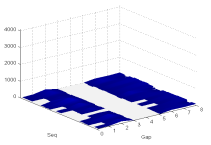
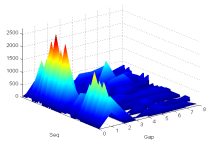
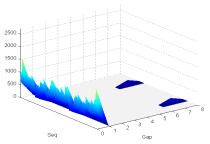
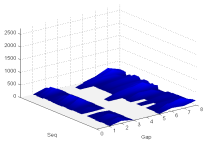
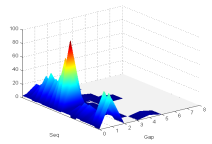
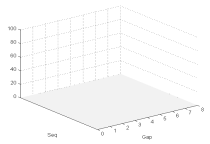
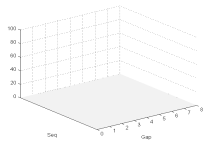
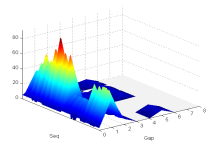
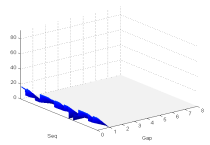
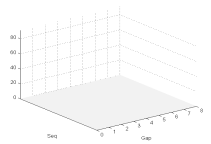
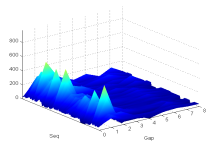
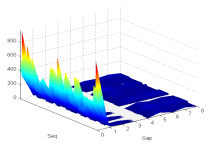
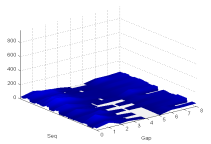
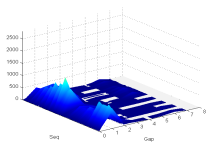
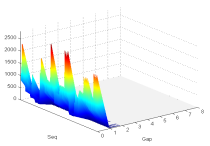
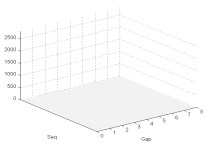
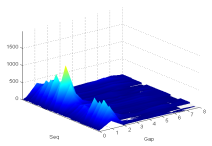
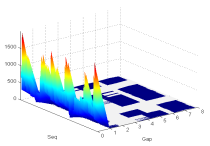
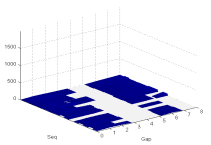
D.5 Retinoic acid receptor family (RARs or NR1Bs)

Monomer seed: RGGTCR

Protein (with partner and/or ligand)	Enrichment of Direct repeats	Enrichment of Inverted repeats	Enrichment of Everted repeats
RARA			
RARA:RXRA			
RARA:retinoicacid			
RARB			
RARB:RXRA			
RARB:retinoicacid			
RARG			
RARG:RXRA			
RARG:retinoicacid			

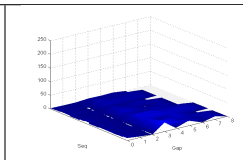
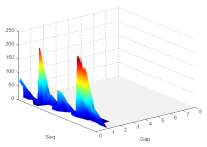
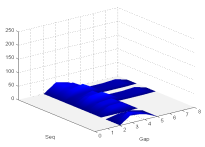
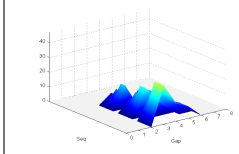
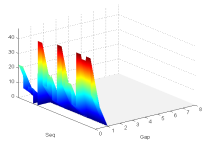
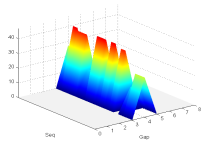
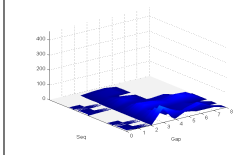
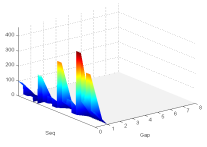
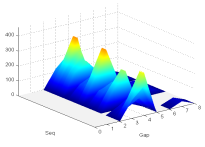
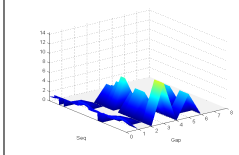
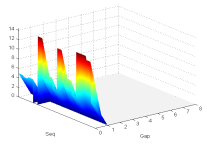
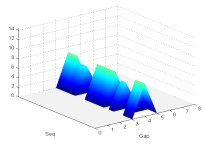
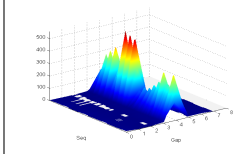
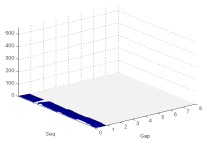
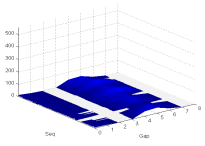
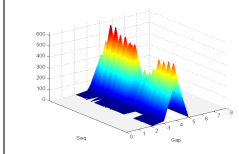
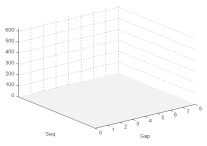
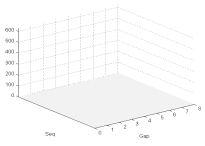
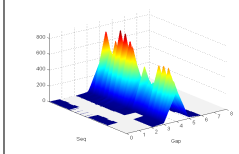
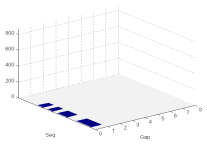
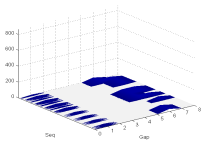
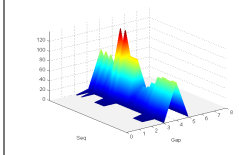
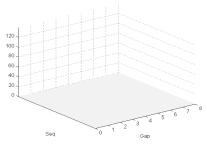
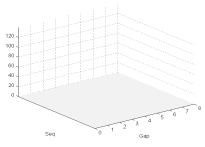
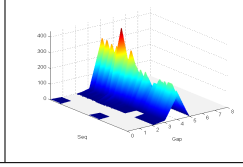
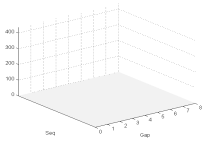
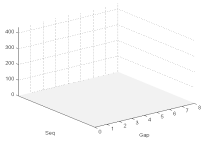
D.6 Retinoid X receptor family (RXRs or NR2Bs)

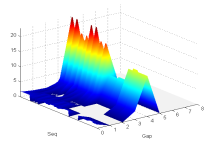
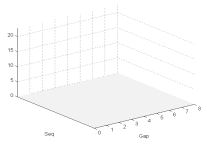
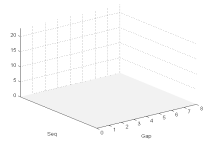
Monomer seed: RGGTCR

Protein (with partner and/or ligand)	Enrichment of Direct repeats	Enrichment of Inverted repeats	Enrichment of Everted repeats
RXRA			
RXRA:9-cis-retinoicacid			
RXRB			
RXRB:9-cis-retinoicacid			
RXRG			
RXRG:RXRA			
RXRG:9-cis-retinoicacid			

D.7 Thyroid hormone receptor family (THRs or NR1As)

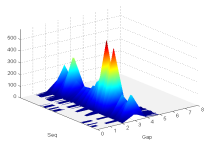
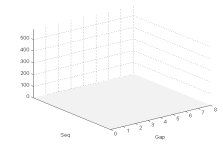
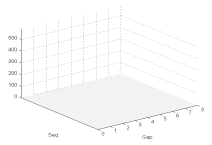
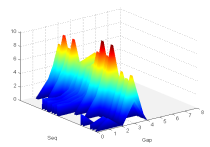
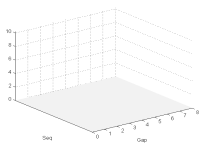
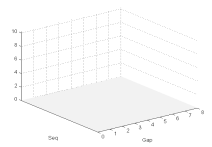
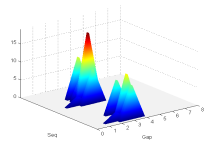
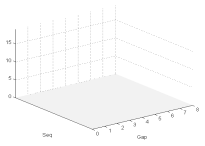
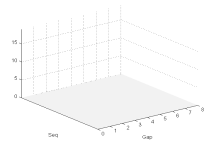
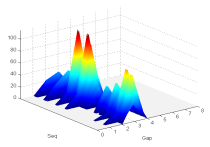
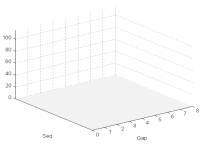
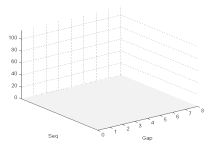
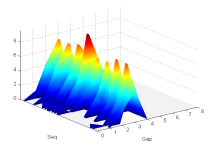
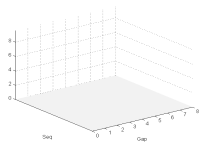
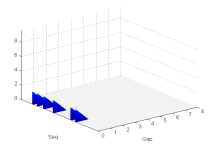
Monomer seed: RGGTCR

Protein (with partner and/or ligand)	Enrichment of Direct repeats	Enrichment of Inverted repeats	Enrichment of Everted repeats
THRA			
THRA:RXRA			
THRA:RXRA:T3			
THRA:T3			
THRB			
THRB:RXRA			
THRB:RXRA:GC1			
THRB:RXRA:T3			
THRB:RXRA:T4			

Protein (with partner and/or ligand)	Enrichment of Direct repeats	Enrichment of Inverted repeats	Enrichment of Everted repeats
THRB:T3			

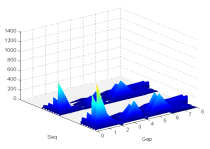
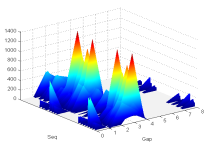
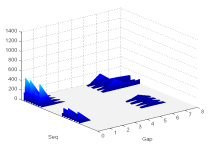
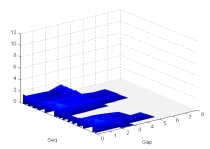
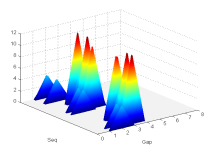
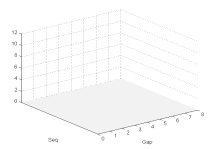
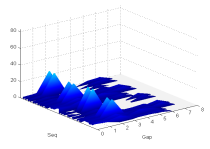
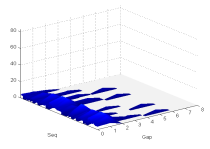
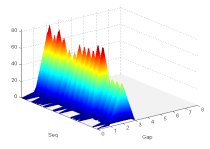
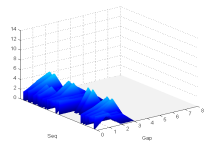
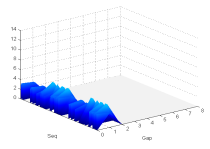
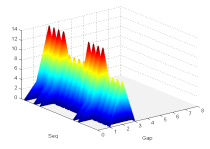
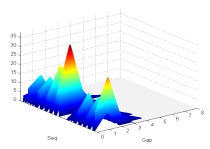
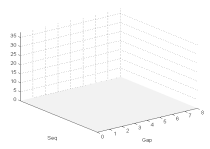
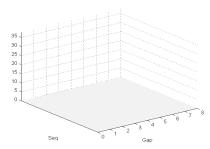
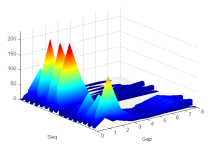
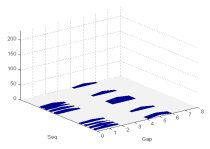
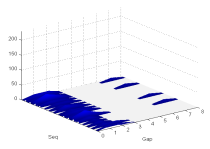
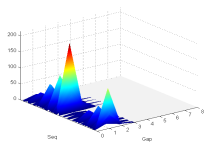
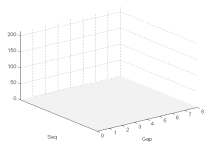
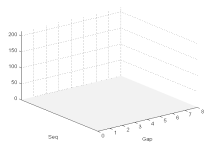
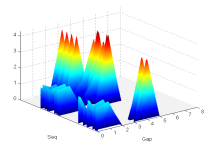
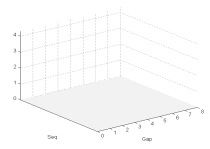
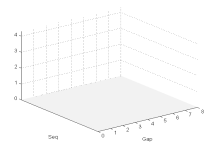
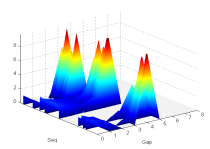
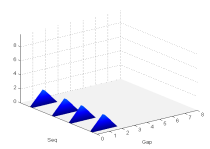
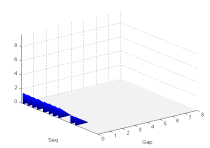
D.8 Vitamin D receptor-like family (NR1Is)

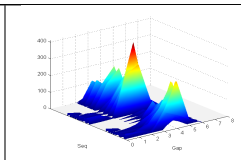
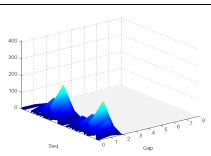
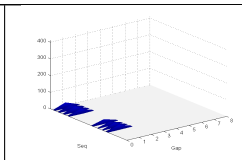
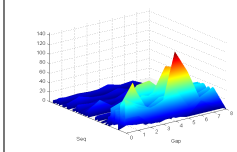
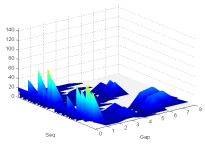
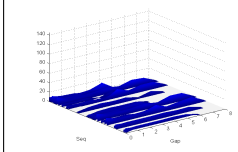
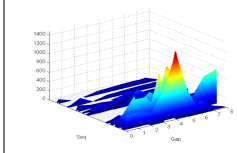
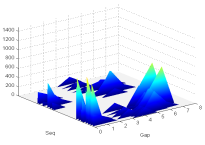
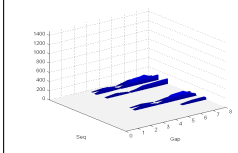
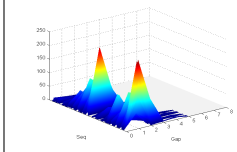
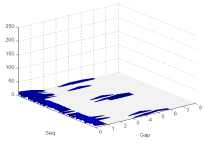
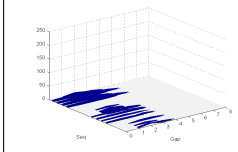
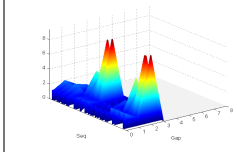
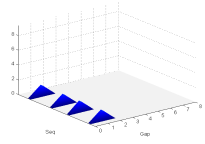
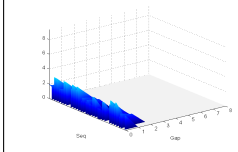
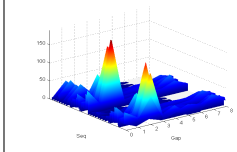
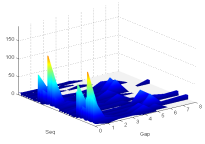
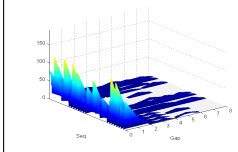
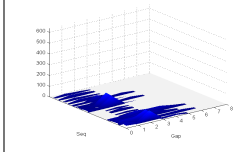
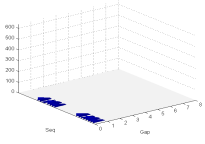
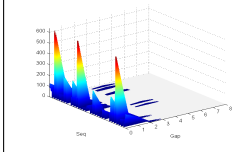
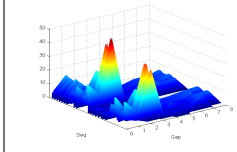
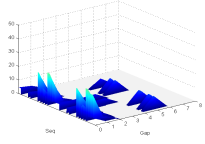
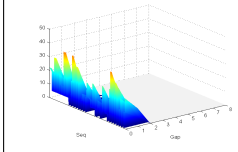
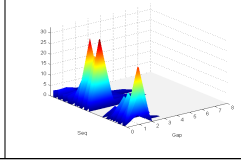
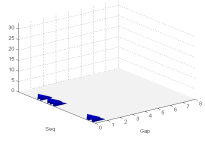
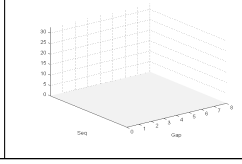
Monomer seed: RGKTCR

Protein (with partner and/or ligand)	Enrichment of Direct repeats	Enrichment of Inverted repeats	Enrichment of Everted repeats
PXR			
PXR:hyperforin			
VDR			
VDR:RXRA			
VDR:1a25-dihydroxyvitaminD3			

D.9 Others

Monomer seed: RGKTCR

Protein (with partner and/or ligand)	Enrichment of Direct repeats	Enrichment of Inverted repeats	Enrichment of Everted repeats
ESR1:17b-estradiol			
ESR1:17b-estradioltamoxifen			
FXRA			
FXRA:chenodeoxycholicac			
HNF4A			
HNF4A:myristicacid			
HNF4G			
LXRA			
LXRA:T0901317			

Protein (with partner and/or ligand)	Enrichment of Direct repeats	Enrichment of Inverted repeats	Enrichment of Everted repeats
LXRB:RXRA			
NOR1			
NURR1			
PNR			
Rev-erbA			
RORC			
RORC:RXRA			
RORC:SR1001			
TR2			

Appendix E

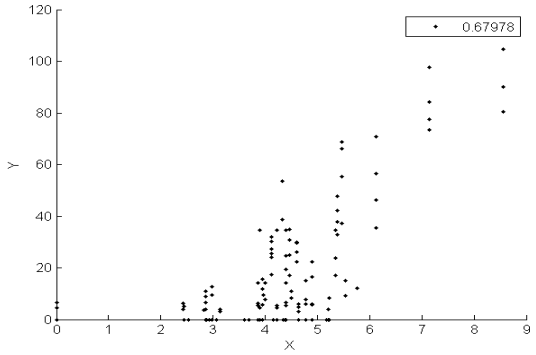
Differential Energy Landscapes of Nuclear Receptor DNA Binding

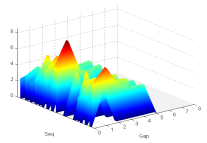
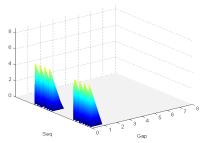
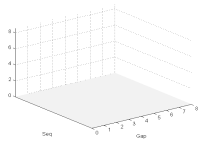
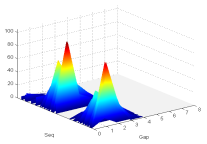
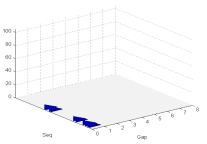
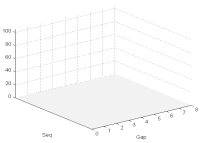
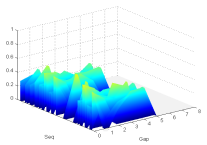
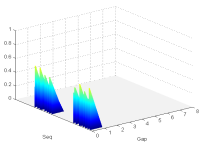
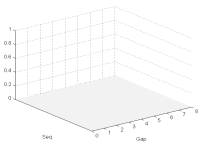
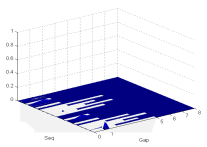
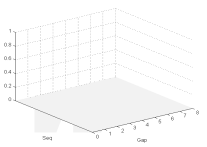
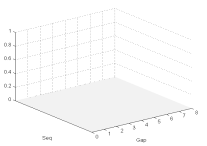
Similar to gapped-SELs we bring a new visualization to compare binding to two different proteins, gapped differential sequence specificity and energy Landscapes (gapped-DiSELs). For each NR sample we first normalize maximum enrichment to 1 by dividing the enrichment values for all sequences in gapped-SEL by highest enrichment and then subtract such normalized enrichment of one sample from the other and plot as gapped-SEL to call gapped-DiSEL. Following sections present few of the DiSEL comparisons between different samples. DiSELs are plotted to compare DNA binding of NR with and without ligand, with and without RXRA and also two different NRs of sample family.

First is the scatter plot between enrichment of the two samples. All the sequences that are plotted on Gapped-SEL are used for scatter plot, with pearson correlation between their enrichment values on top right corner. Then are the gapped-SELs for both the samples and then corresponding gapped-DiSEL X (first) over Y (second) and Y over X, representing sequences preferred by first sample over second sample and vice-versa.

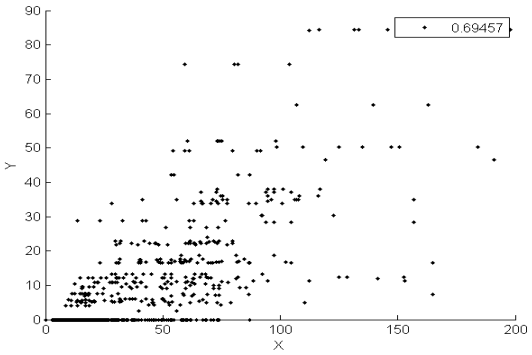
E.1 COUP/EAR (NR2Fs) family

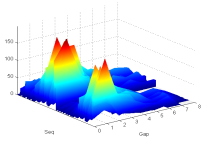
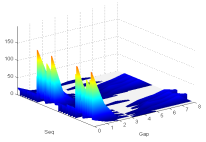
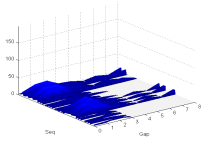
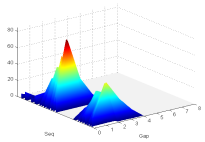
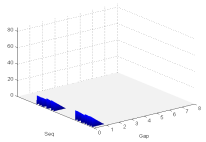
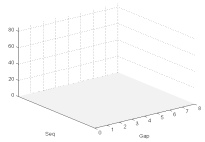
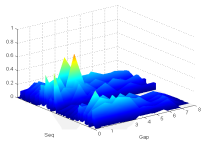
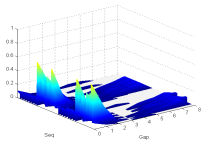
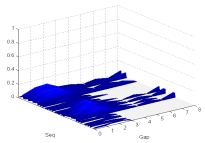
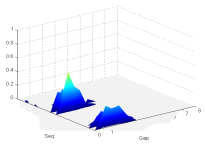
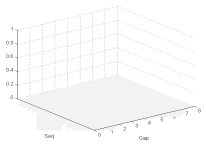
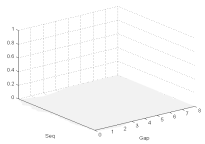
COUP-TFA Vs COUP-TFA:RXRA



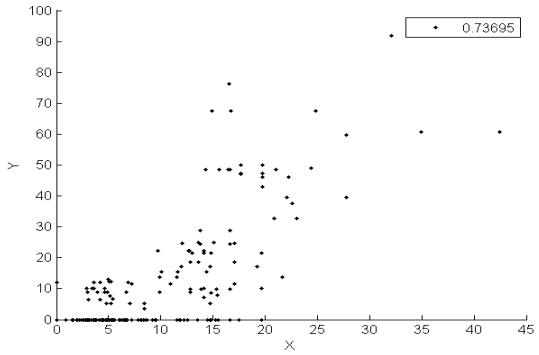
Gapped SELs and DiSELs	Enrichment of Direct repeats	Enrichment of Inverted repeats	Enrichment of Everted repeats
COUP-TFA (Gapped SEL: RGKTCR = X)			
COUP-TFA:RXRA (Gapped SEL: RGK-TCR = Y)			
Gapped DiSEL: RGK-TCR = X over Y			
Gapped DiSEL: RGK-TCR = Y over X			

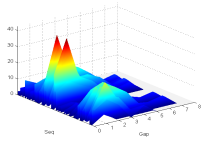
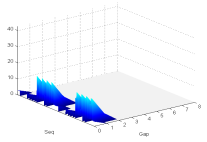
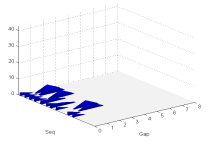
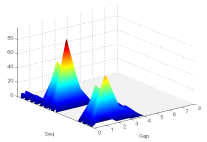
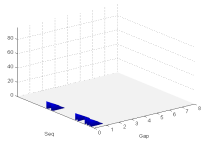
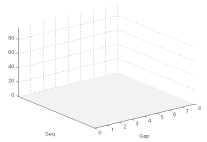
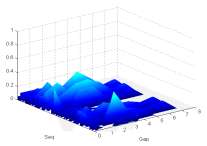
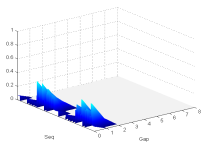
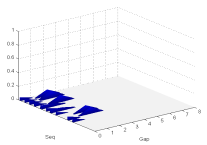
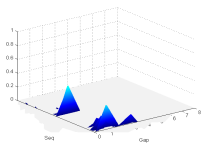
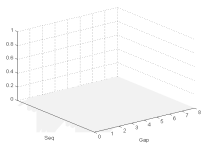
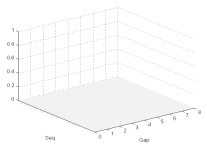
COUP-TFB Vs COUP-TFB:RXRA



Gapped SELs and DiSELs	Enrichment of Direct repeats	Enrichment of Inverted repeats	Enrichment of Everted repeats
COUP-TFB (Gapped SEL: RGKTCR = X)			
COUP-TFB:RXRA (Gapped SEL: RGK-TCR = Y)			
Gapped DiSEL: RGK-TCR = X over Y			
Gapped DiSEL: RGK-TCR = Y over X			

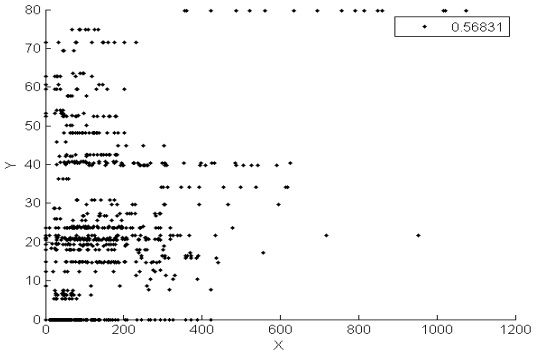
EAR2 Vs EAR2:RXRA

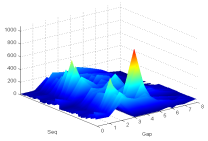
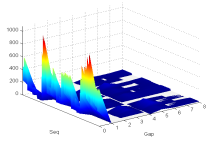
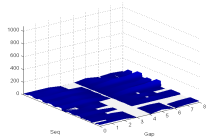
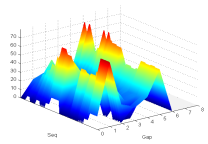
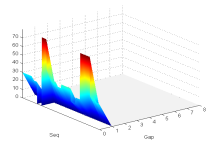
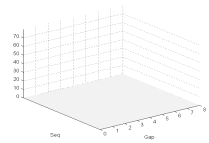
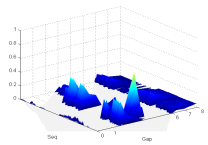
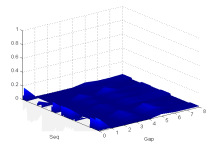
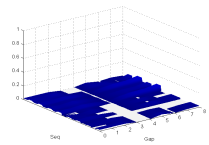
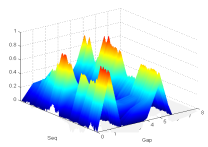
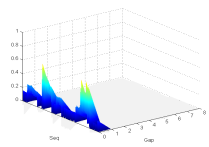
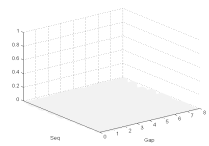


Gapped SELs and DiSELs	Enrichment of Direct repeats	Enrichment of Inverted repeats	Enrichment of Everted repeats
EAR2 (Gapped SEL: RGKTCR = X)			
EAR2:RXRA (Gapped SEL: RGKTCR = Y)			
Gapped DiSEL: RGK-TCR = X over Y			
Gapped DiSEL: RGK-TCR = Y over X			

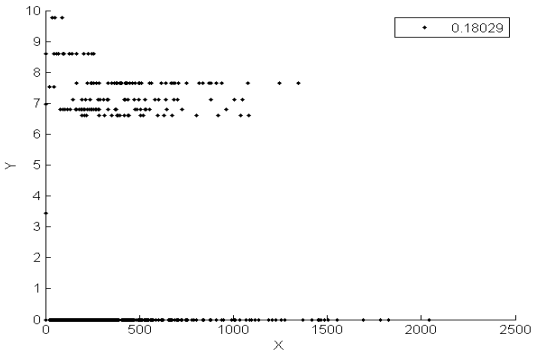
E.2 Retinoic acid receptor family (RARs or NR1Bs)

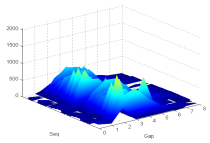
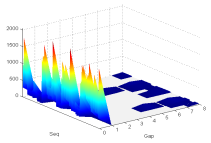
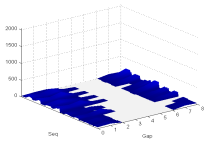
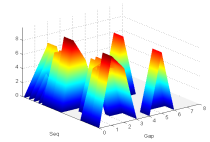
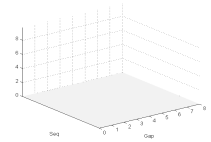
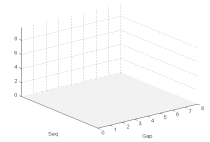
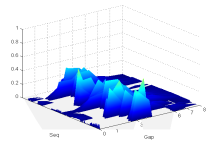
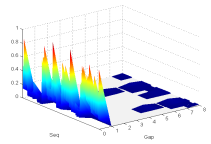
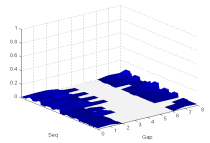
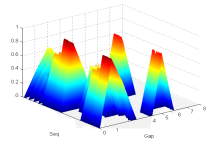
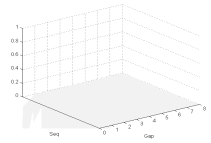
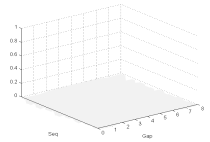
RARA Vs RARA:RXRA



Gapped SELs and DiSELs	Enrichment of Direct repeats	Enrichment of Inverted repeats	Enrichment of Everted repeats
RARA (Gapped SEL: RGGTCR = X)			
RARA:RXRA (Gapped SEL: RGGTCR = Y)			
Gapped DiSEL: RGGTCR = X over Y			
Gapped DiSEL: RGGTCR = Y over X			

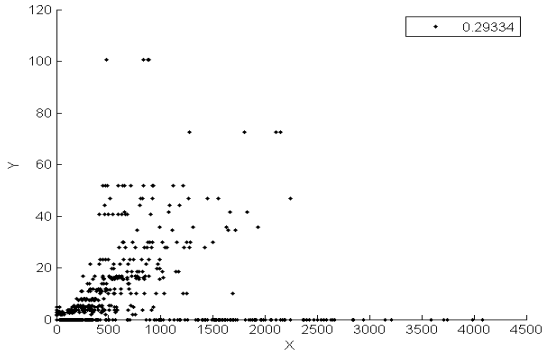
RARG Vs RARG:RXRA

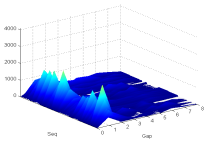
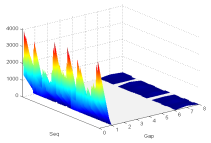
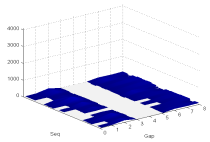
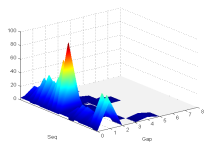
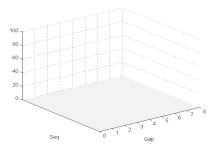
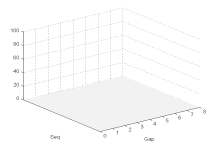
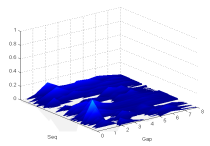
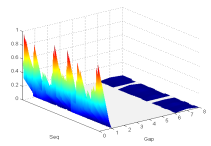
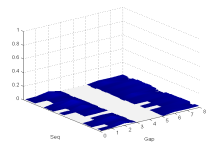
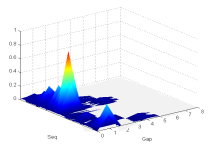
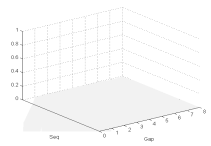
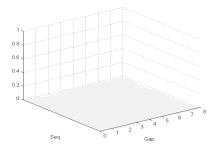


Gapped SELs and DiSELs	Enrichment of Direct repeats	Enrichment of Inverted repeats	Enrichment of Everted repeats
RARG (Gapped SEL: RGGTCR = X)			
RARG:RXRA (Gapped SEL: RGGTCR = Y)			
Gapped DiSEL: RGGTCR = X over Y			
Gapped DiSEL: RGGTCR = Y over X			

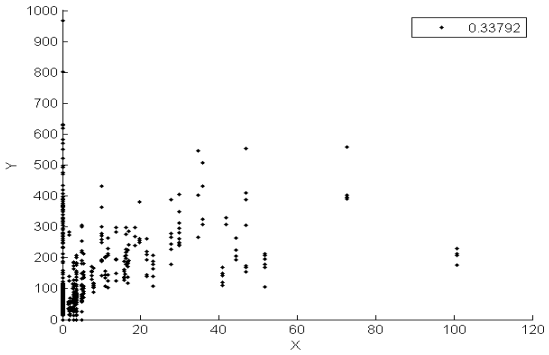
E.3 Retinoid X receptor family (RXRs or NR2Bs)

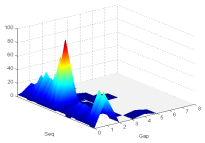
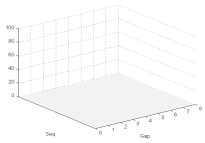
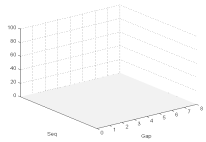
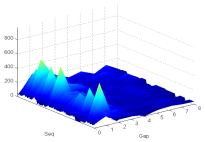
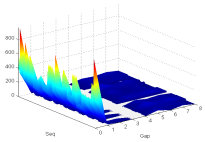
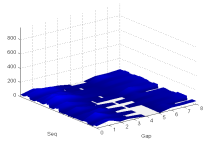
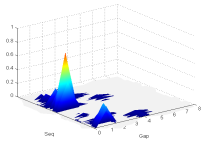
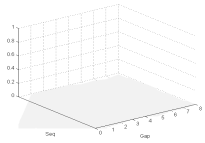
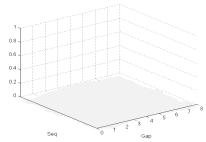
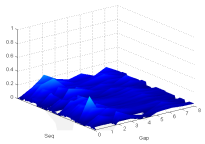
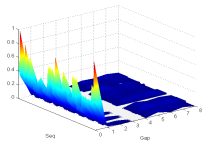
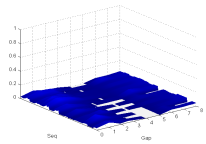
RXRA Vs RXRB



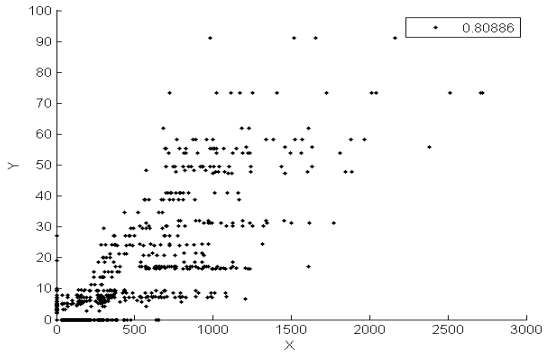
Gapped SELs and DiSELs	Enrichment of Direct repeats	Enrichment of Inverted repeats	Enrichment of Everted repeats
RXRA (Gapped SEL: RGGTCR = X)			
RXRB (Gapped SEL: RGGTCR = Y)			
Gapped DiSEL: RGGTCR = X over Y			
Gapped DiSEL: RGGTCR = Y over X			

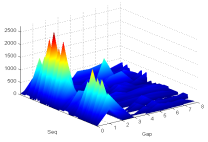
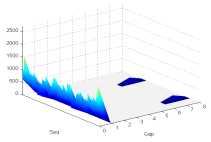
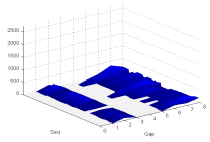
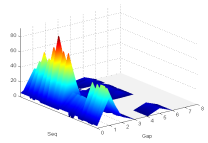
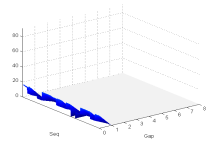
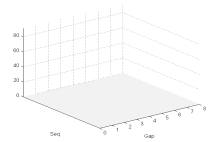
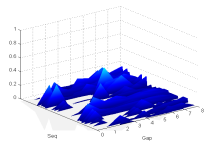
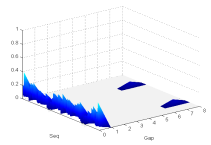
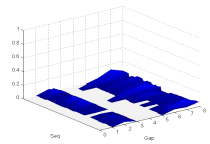
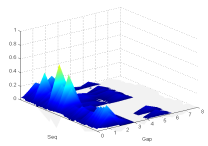
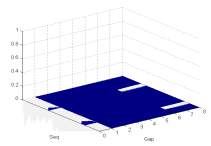
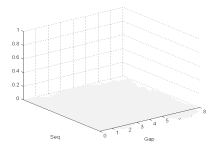
RXRB Vs RXRG



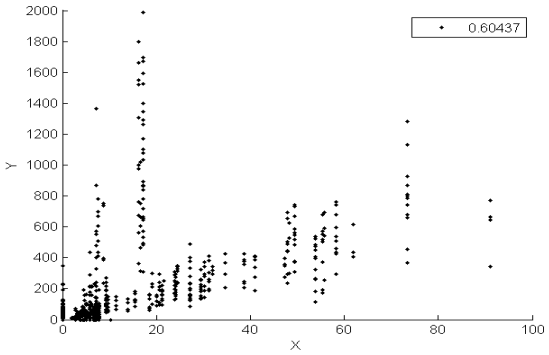
Gapped SELs and DiSELs	Enrichment of Direct repeats	Enrichment of Inverted repeats	Enrichment of Everted repeats
RXRB (Gapped SEL: RGGTCR = X)			
RXRG (Gapped SEL: RGGTCR = Y)			
Gapped DiSEL: RGGTCR = X over Y			
Gapped DiSEL: RGGTCR = Y over X			

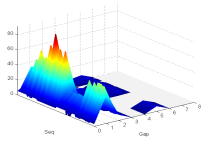
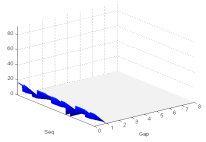
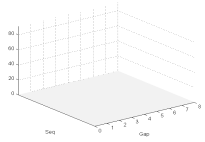
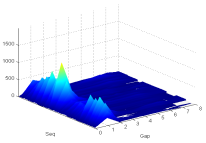
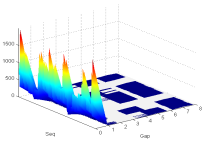
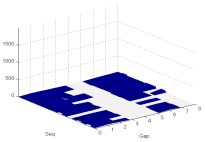
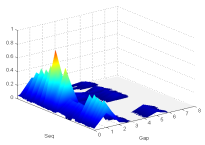
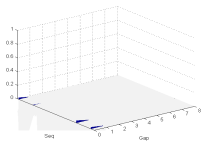
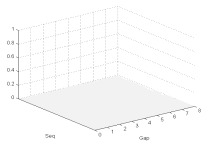
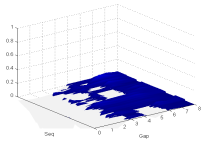
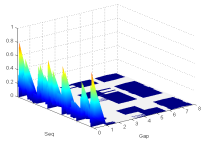
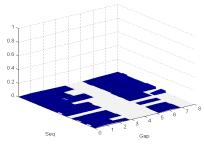
RXRA:9-cis-retinoicacid Vs RXRB:9-cis-retinoicacid



Gapped SELs and DiSELs	Enrichment of Direct repeats	Enrichment of Inverted repeats	Enrichment of Everted repeats
RXRA:9-cis-retinoicacid (Gapped SEL: RGGTCR = X)			
RXRB:9-cis-retinoicacid (Gapped SEL: RGGTCR = Y)			
Gapped DiSEL: RGGTCR = X over Y			
Gapped DiSEL: RGGTCR = Y over X			

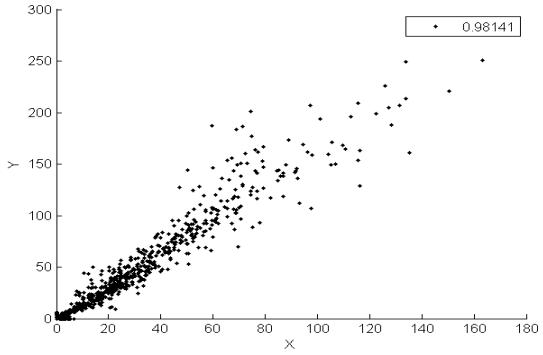
RXRB:9-cis-retinoicacid Vs RXRG:9-cis-retinoicacid

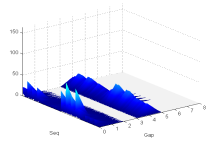
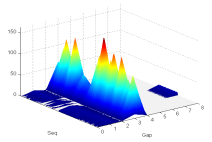
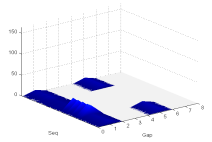
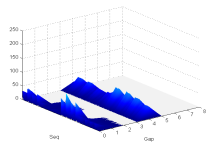
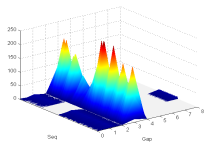
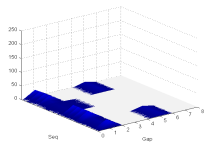
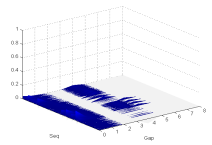
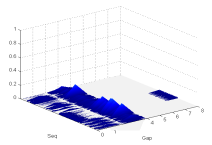
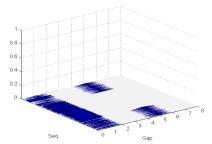
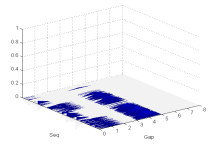
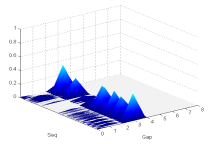
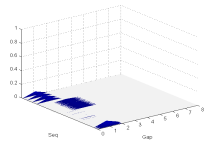


Gapped SELs and DiSELs	Enrichment of Direct repeats	Enrichment of Inverted repeats	Enrichment of Everted repeats
RXRB:9-cis-retinoicacid (Gapped SEL: RGGTCR = X)			
RXRG:9-cis-retinoicacid (Gapped SEL: RGGTCR = Y)			
Gapped DiSEL: RGGTCR = X over Y			
Gapped DiSEL: RGGTCR = Y over X			

E.4 3-Ketosteroid receptors family (NR3Cs)

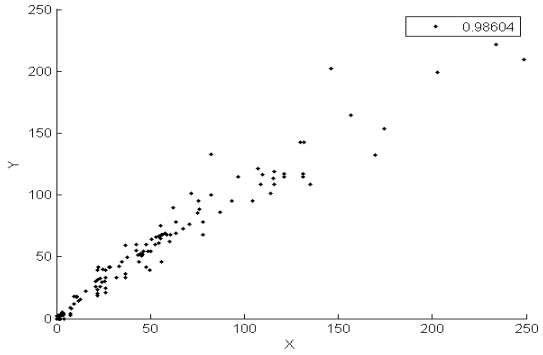
PGR Vs PGR:progesterone

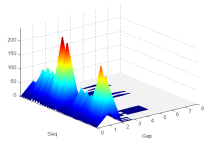
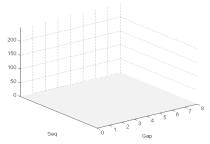
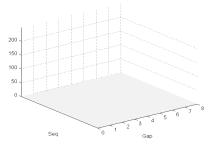
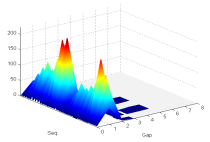
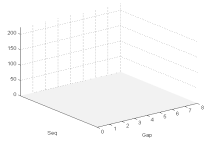
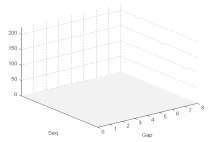
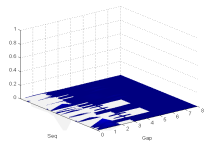
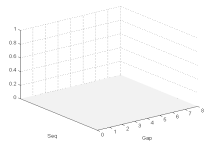
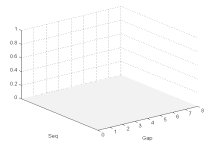
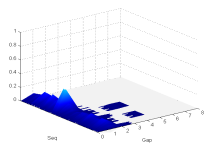
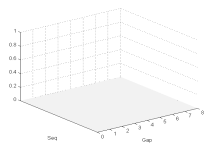
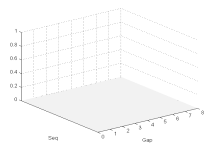


Gapped SELs and DiSELs	Enrichment of Direct repeats	Enrichment of Inverted repeats	Enrichment of Everted repeats
PGR (Gapped SEL: GNACR = X)			
PGR:progesterone (Gapped SEL: GNACR = Y)			
Gapped DiSEL: GNACR = X over Y			
Gapped DiSEL: GNACR = Y over X			

E.5 Peroxisome proliferator-activated receptor family (PPARs or NR1Cs)

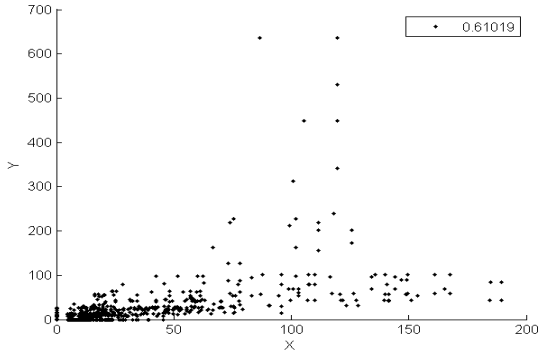
PPARG Vs PPARG:rosiglitazone

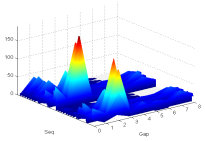
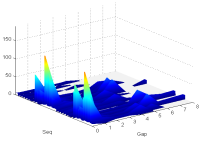
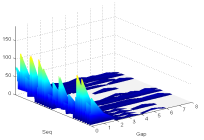
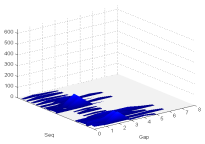
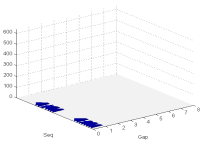
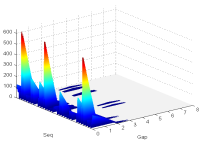
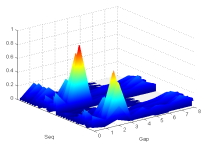
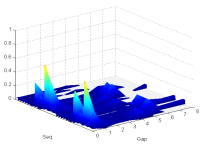
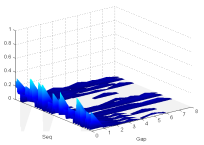
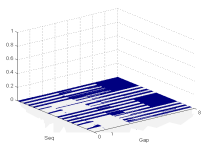
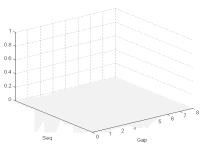
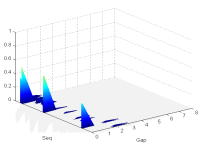


Gapped SELs and DiSELs	Enrichment of Direct repeats	Enrichment of Inverted repeats	Enrichment of Everted repeats
PPARG (Gapped SEL: RGGTCR = X)			
PPARG:rosiglitazone (Gapped SEL: RGGTCR = Y)			
Gapped DiSEL: RGGTCR = X over Y			
Gapped DiSEL: RGGTCR = Y over X			

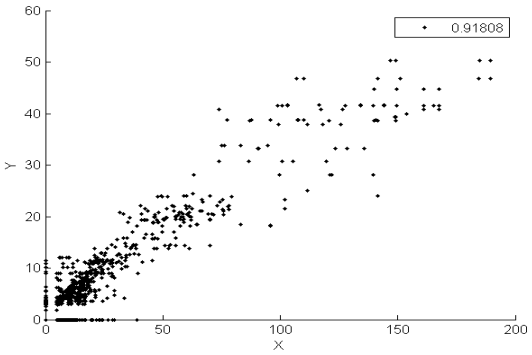
E.6 Others

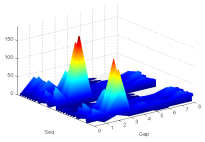
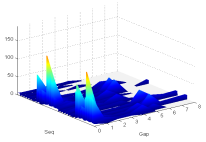
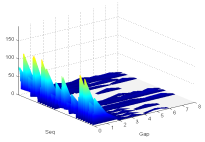
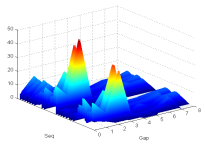
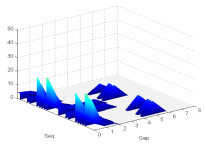
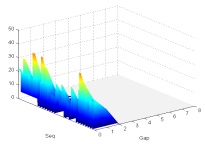
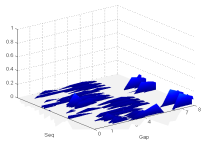
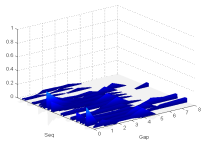
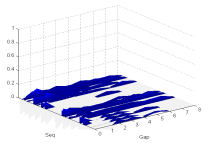
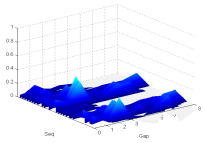
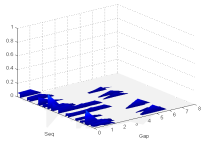
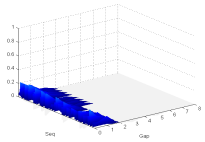
RORC Vs RORC:RXRA



Gapped SELs and DiSELs	Enrichment of Direct repeats	Enrichment of Inverted repeats	Enrichment of Everted repeats
RORC (Gapped SEL: RGKTCR = X)			
RORC:RXRA (Gapped SEL: RGK-TCR = Y)			
Gapped DiSEL: RGK-TCR = X over Y			
Gapped DiSEL: RGK-TCR = Y over X			

RORC Vs RORC:SR1001



Gapped SELs and DiSELs	Enrichment of Direct repeats	Enrichment of Inverted repeats	Enrichment of Everted repeats
RORC (Gapped SEL: RGKTCR = X)			
RORC:SR1001 (Gapped SEL: RGK-TCR = Y)			
Gapped DiSEL: RGK-TCR = X over Y			
Gapped DiSEL: RGK-TCR = Y over X			

Appendix F

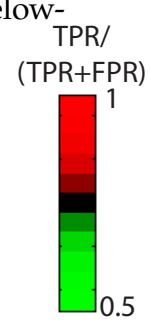
Nuclear Receptor In Vitro DNA Binding Compared to In Vivo Binding

Different labs have tested NR's binding in vivo using ChIP-chip and ChIP-Seq techniques. Here we use ChIP-Seq peak data from ENCODE consortium [10] and LoVo cell line [70]. The genomic sequence underlying ChIP-Seq peaks are then used to generate ROC curves. In this analysis, ChIP-Seq peaks are taken as positives and two random permutations (moving positions of DNA bases) of each peak are used as known negatives. Each peak (all positives and negatives) is then scored using MinSeqFind. An ROC curve between false positive rate (FPR) and true positive rate (TPR) is plotted by varying a moving threshold, positive peaks scored above that threshold (true positives) are used to get TPR (true positives over total positives) and negative peaks scored above threshold (false positives) are used to get FPR (false positives over total negatives). Area under ROC (AUROC) curve is used to analyze how well in vitro data correlates to a set of ChIP-Seq peaks. Where AUROC=1 means complete prediction and AUROC=0.5 means random prediction. AUROC is used to do a first level of comparison between two different sets of data in predicting set of ChIP-Seq peaks. For deeper peak-by-peak comparison, we assigned a score S to each peak. S for a peak is defined as maximum value of TPR/ (TPR + FPR) at which a true positive peak is detected as positive peak (similar to precision TP/ (TP+FP)). Score S represents predictability of each peak using a given DNA binding data as opposed to randomized region when considering all the ChIP-Seq peak. The scale varies from the 1 (highest predictability i.e. peaks detected as positive at FPR=0) to 0.5 (lowest predictability, peaks detected as positive at FPR=TPR). The S score for a given set of ChIP-Seq peaks is represented as a heatmap in MATLAB. In case of random prediction i.e. a diagonal ROC curve (AUC-ROC=0.5), there will not be a single peak that will be assigned as positive detected even at 0.6 S score (this is what is intended), if we choose FPR cutoff as our metric we will get 10% peaks detected positive at FPR cutoff 0.1 (which is considered as better prediction). Thus we used newly developed S

score instead of FPR cutoff.

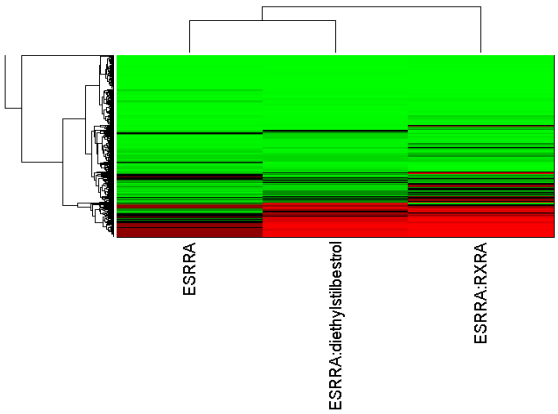
In the following section multiple cell line ChIP peaks (all peaks and top 500 peaks) are scored using MinSeqFind for NR binding. Clustergram of a protein is plotted between it's ChIP peaks in a particular cell line and condition, and different samples of NR binding with same protein. Green to red represent no prediction to best prediction of peaks using MinSeqFind. If there exist only one sample of NR against ChIP peaks are present in vivo, then ROC is plotted instead of clustergram with corresponding AUC-ROC displayed on top right corner.

Naming of section is as follows- example "A549 GR treatment:Dex 50nm" represents ChIP peaks for GR protein in A549 cell line with Dex 50nm treatment. In few cases an intersection of peaks is used, example "LoVo ESR1:RXRA" represents peaks present in both for ESR1 protein and RXRA protein in LoVo cell line. ChIP peak intersection is done use bedops tool [42]. A common coloring scheme is used for all the plots shown below-

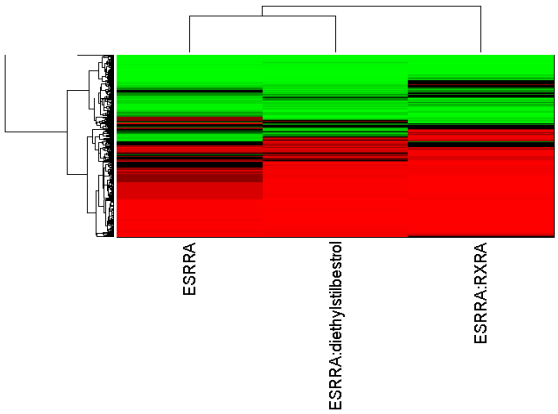


F.1 LoVo ESRRA

All peaks

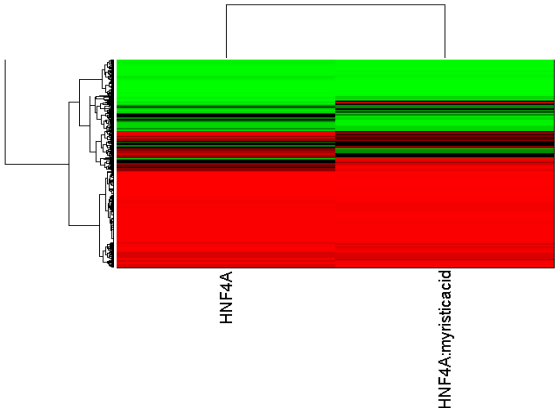


Top 500 peaks

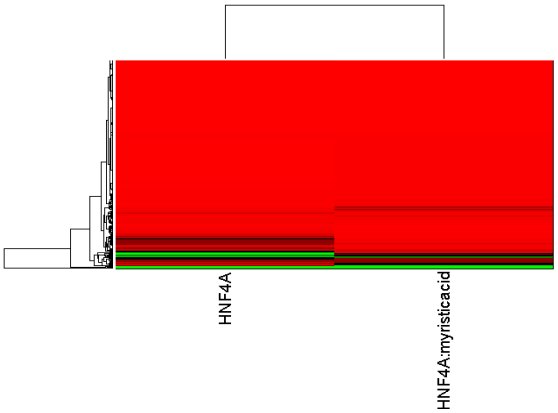


F.2 LoVo HNF4A

All peaks

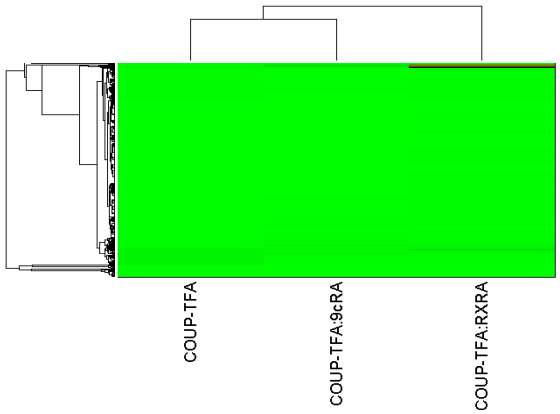


Top 500 peaks

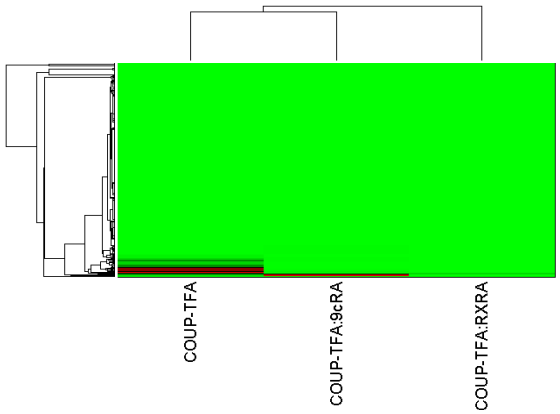


F.3 LoVo NR2F1

All peaks

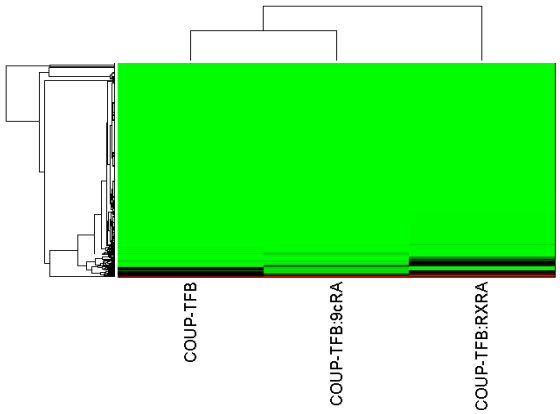


Top 500 peaks

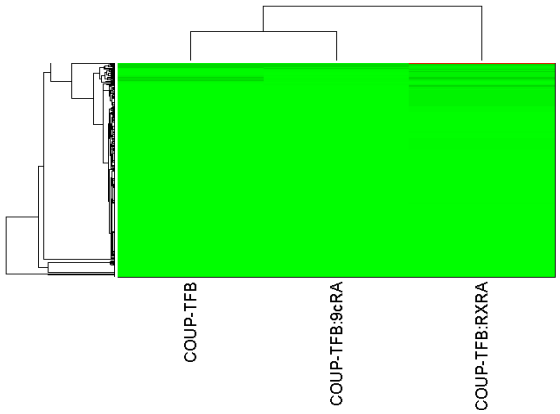


F.4 LoVo NR2F2

All peaks

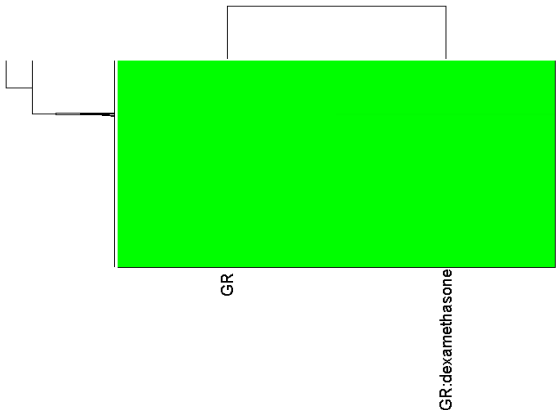


Top 500 peaks

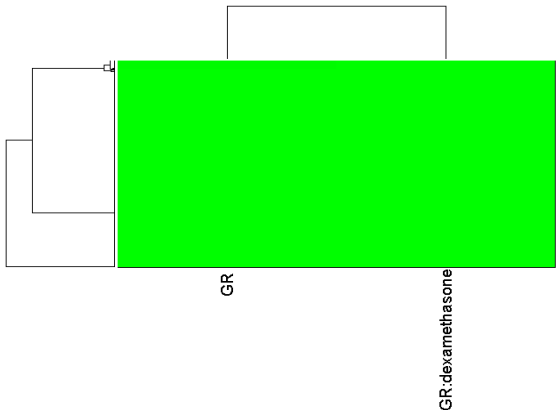


F.5 LoVo NR3C1

All peaks

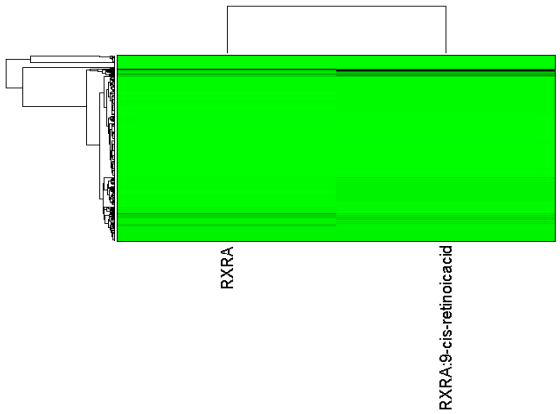


Top 500 peaks

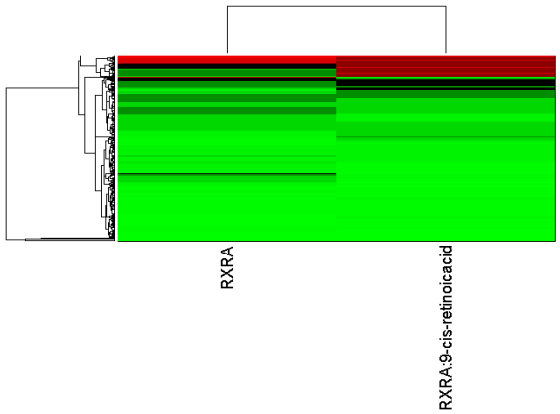


F.6 LoVo RXRA

All peaks

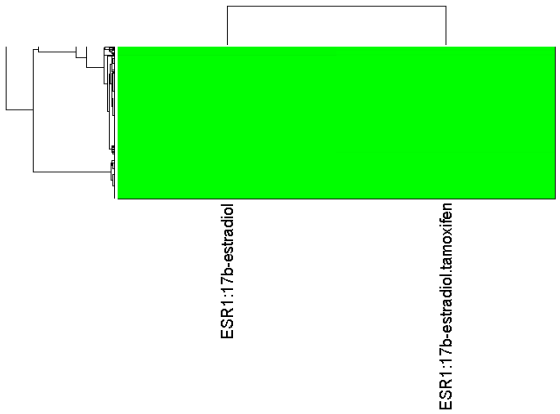


Top 500 peaks

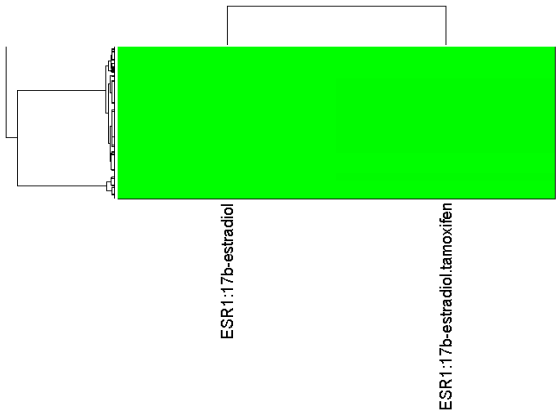


F.7 LoVo ESR1

All peaks

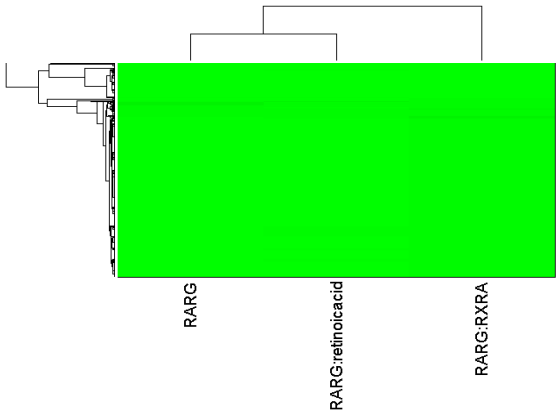


Top 500 peaks

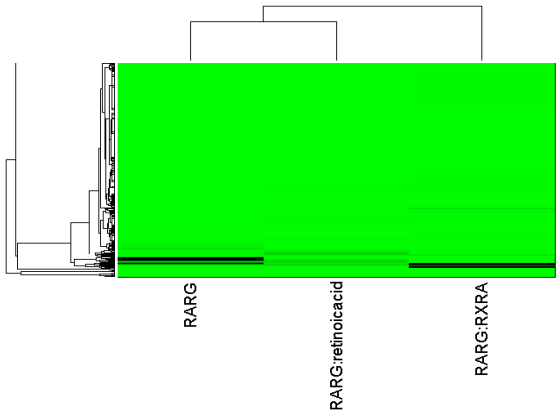


F.8 LoVo RARG

All peaks

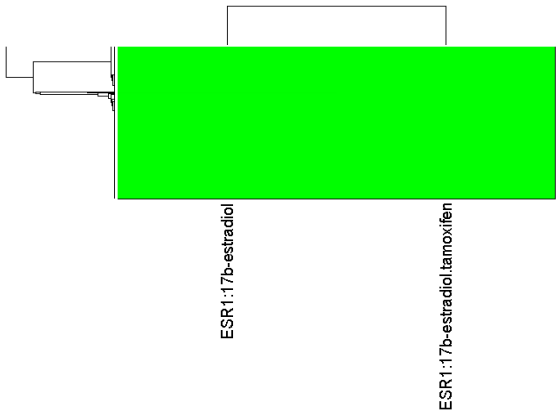


Top 500 peaks



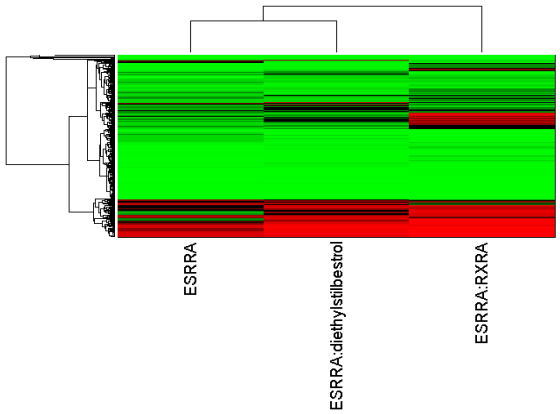
F.9 LoVo ESR1:RXRA

All peaks



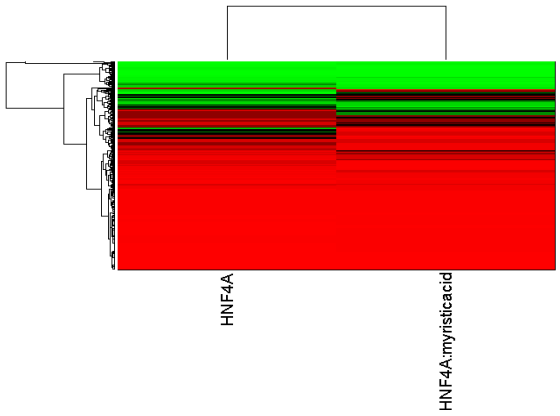
F.10 LoVo ESRRA:RXRA

All peaks



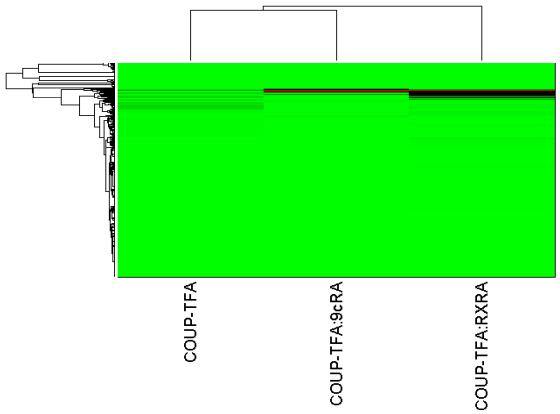
F.11 LoVo HNF4A:RXRA

All peaks



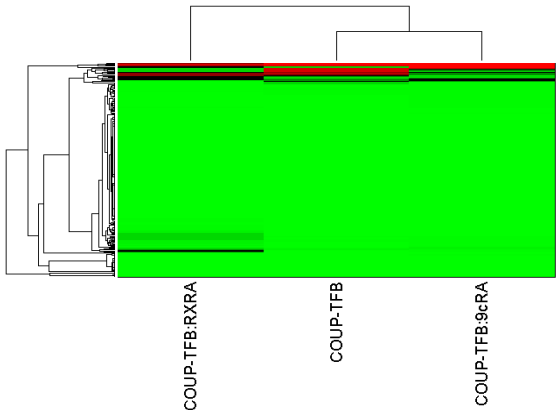
F.12 LoVo NR2F1:RXRA

All peaks



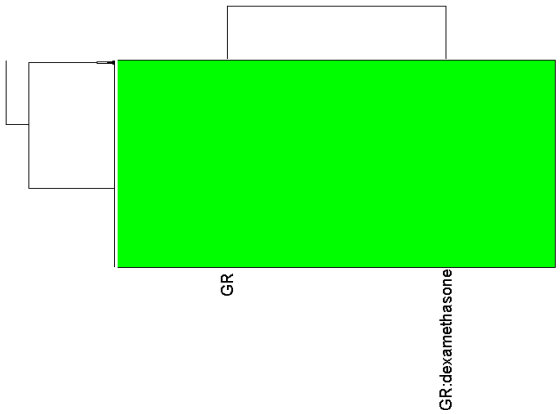
F.13 LoVo NR2F2:RXRA

All peaks



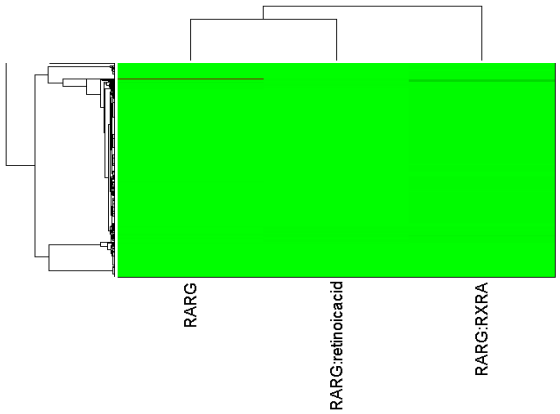
F.14 LoVo NR3C1:RXRA

All peaks



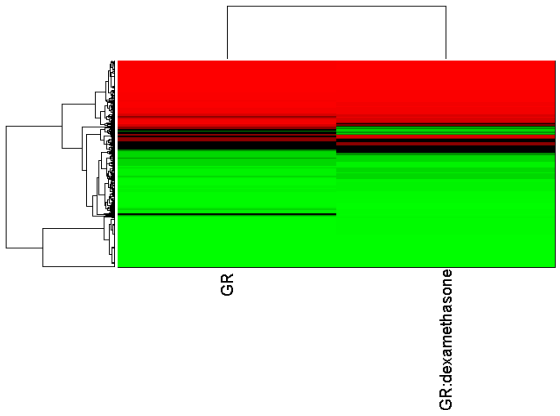
F.15 LoVo RARG:RXRA

All peaks

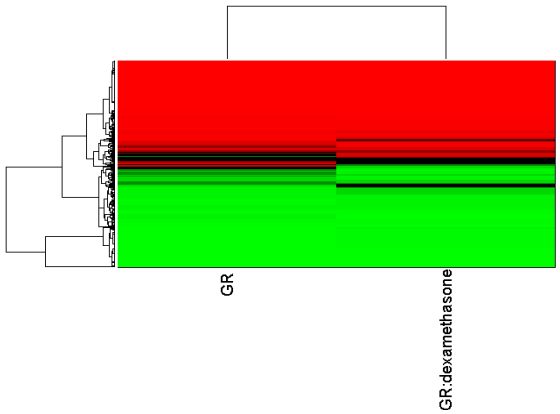


F.16 A549 GR treatment:Dex 500pm

All peaks

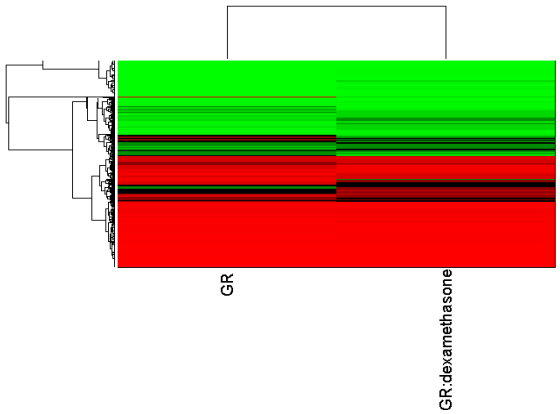


Top 500 peaks

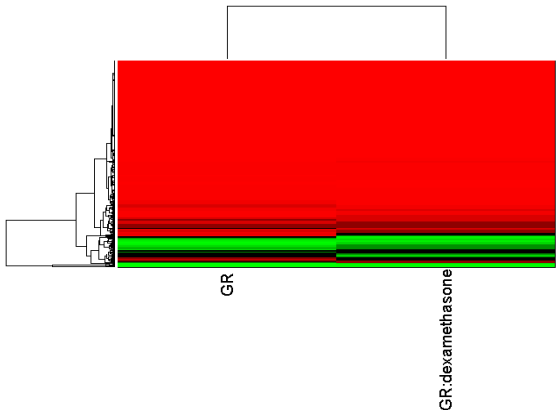


F.17 A549 GR treatment:Dex 50nm

All peaks

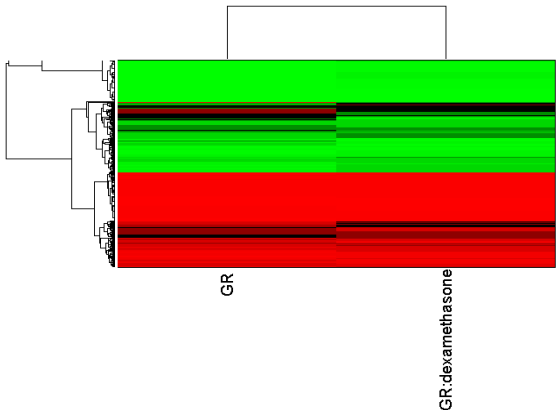


Top 500 peaks

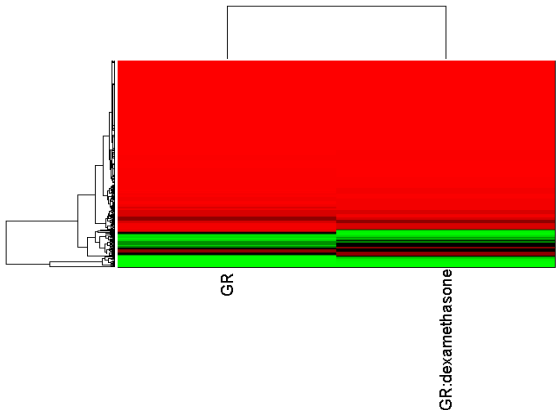


F.18 A549 GR treatment:Dex 5nm

All peaks

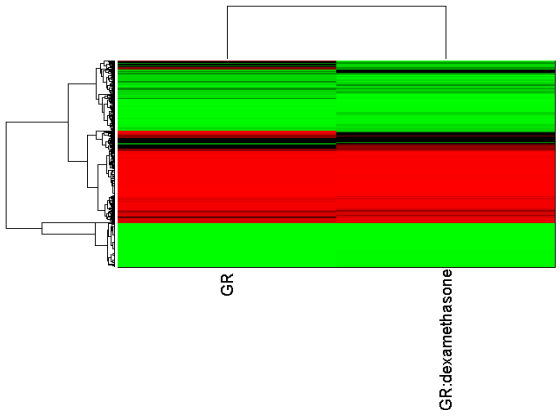


Top 500 peaks

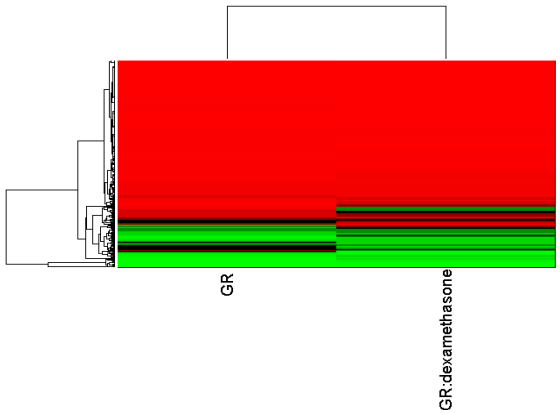


F.19 A549 GR treatment:Dex 100nm

All peaks

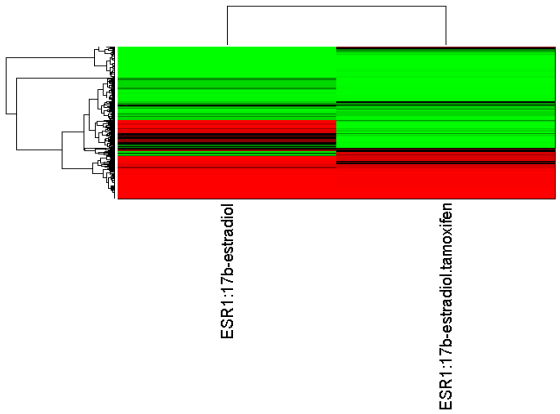


Top 500 peaks

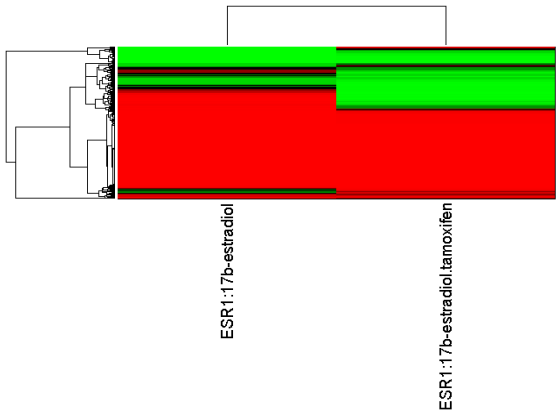


F.20 ECC-1 ERRA treatment=BPA 100nM

All peaks

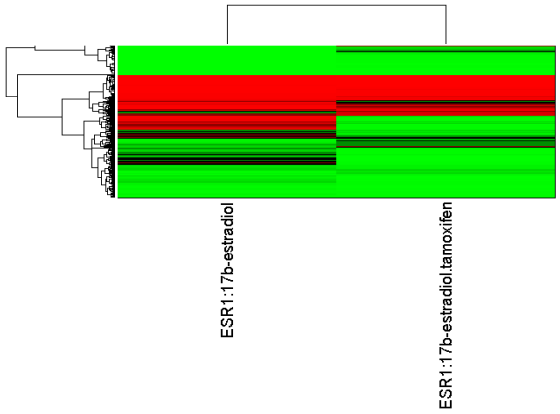


Top 500 peaks

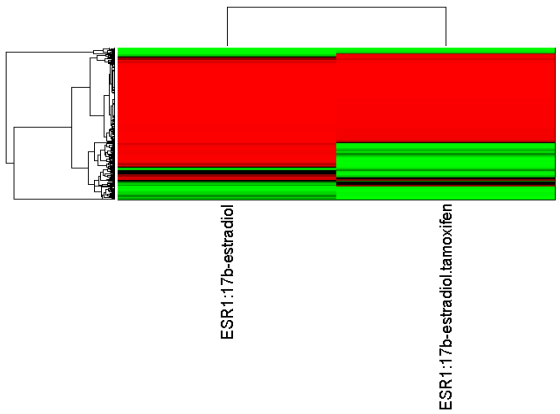


F.21 ECC-1 ERRA treatment=Estradiol 10nM

All peaks

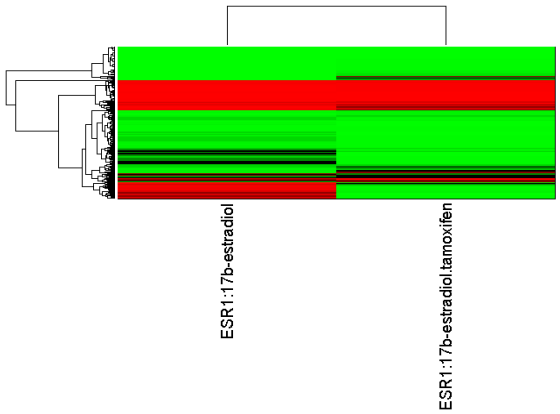


Top 500 peaks

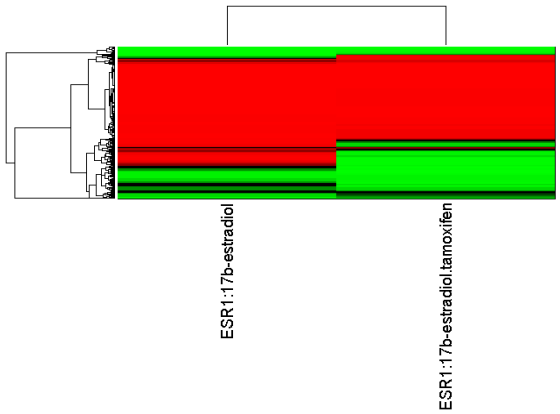


F.22 ECC-1 ERRA treatment=Genistein 100nM

All peaks

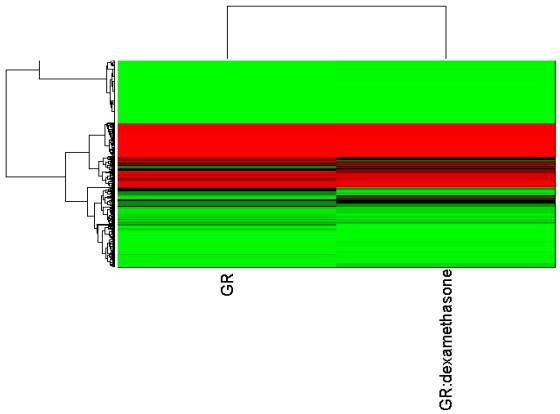


Top 500 peaks

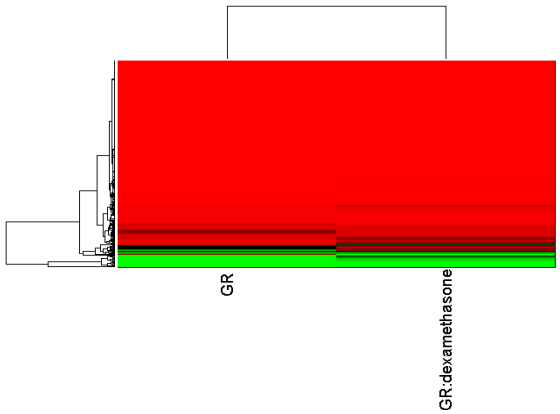


F.23 ECC-1 GR treatment=DEX 100nM

All peaks

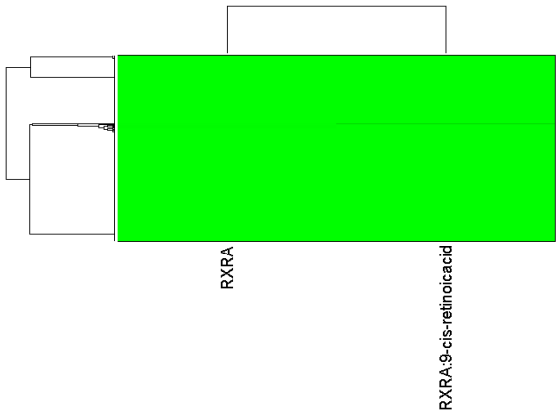


Top 500 peaks

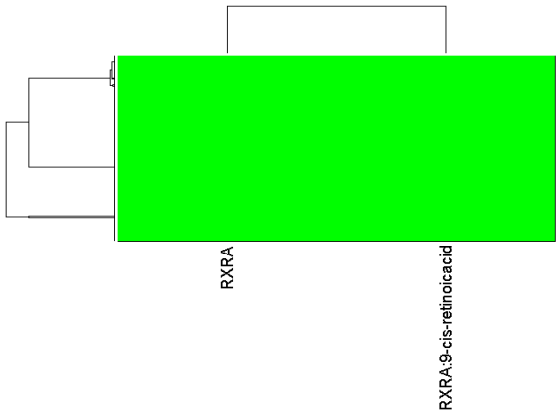


F.24 GM12878 RXRA

All peaks

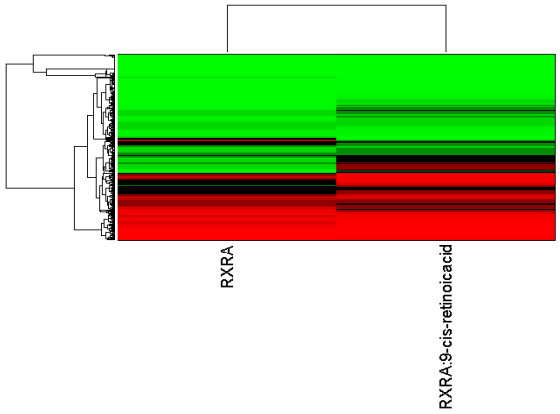


Top 500 peaks

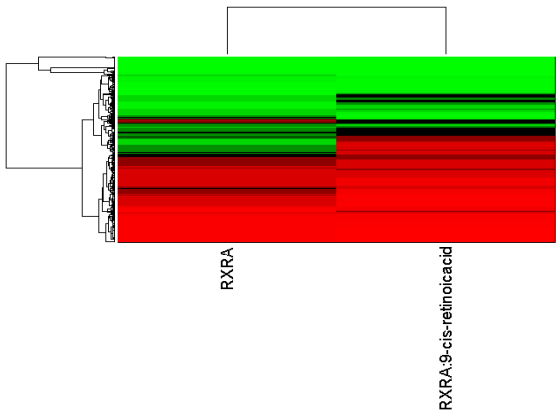


F.25 H1-hESC RXRA

All peaks

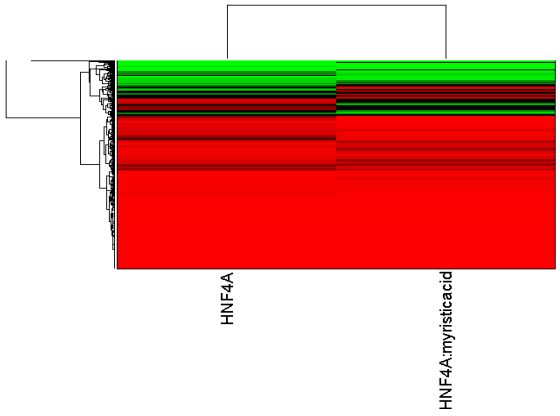


Top 500 peaks

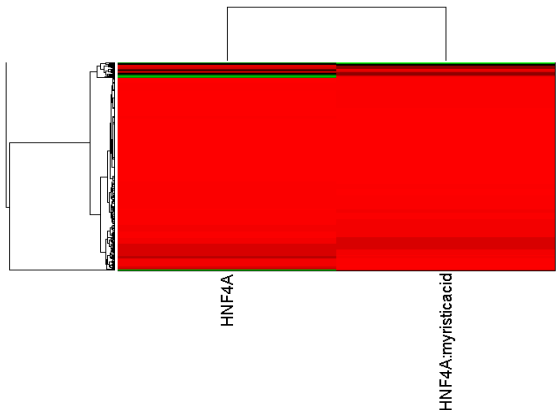


F.26 HepG2 HNF4A (SC-8987)

All peaks

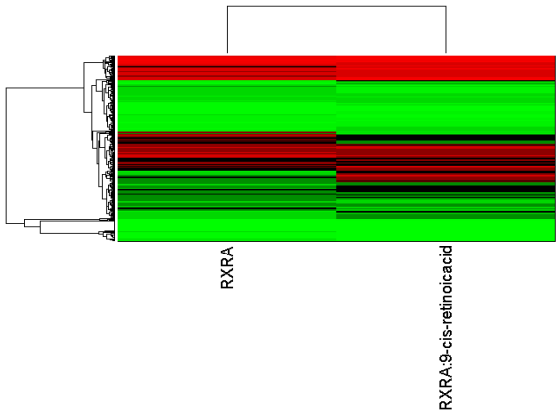


Top 500 peaks

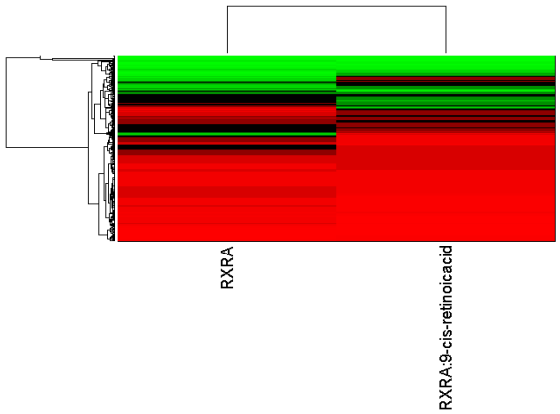


F.27 HepG2 RXRA

All peaks

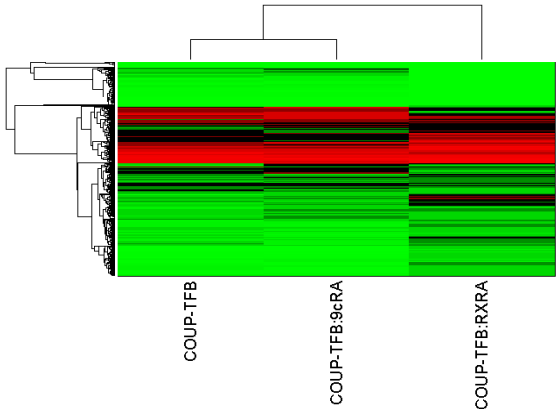


Top 500 peaks

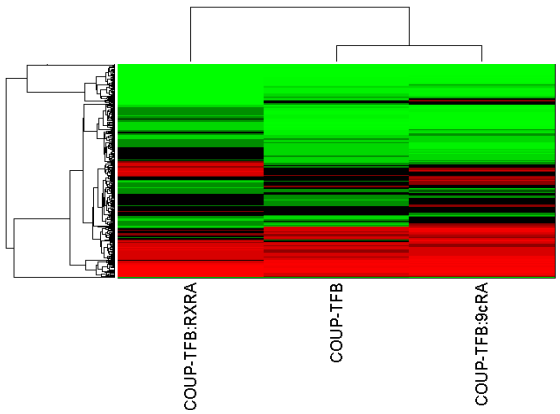


F.28 K562 NR2F2 (SC-271940)

All peaks

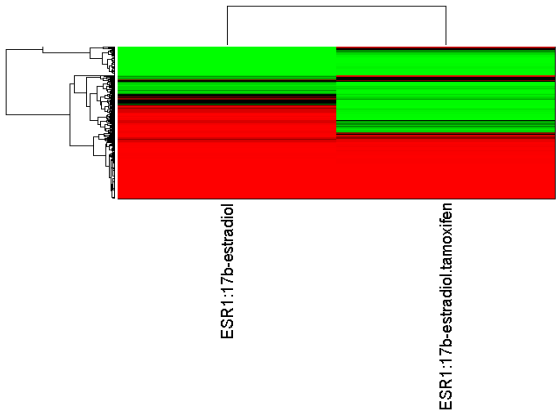


Top 500 peaks

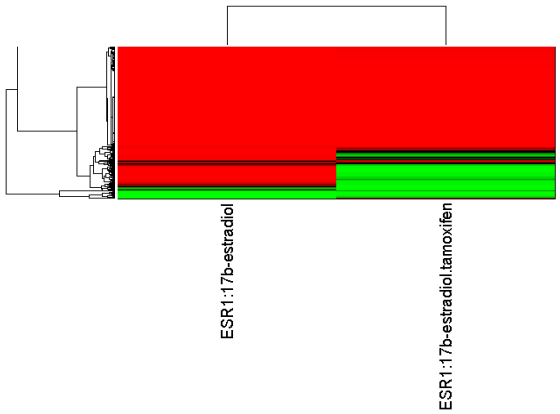


F.29 T-47D ERRA treatment=BPA 100nM

All peaks

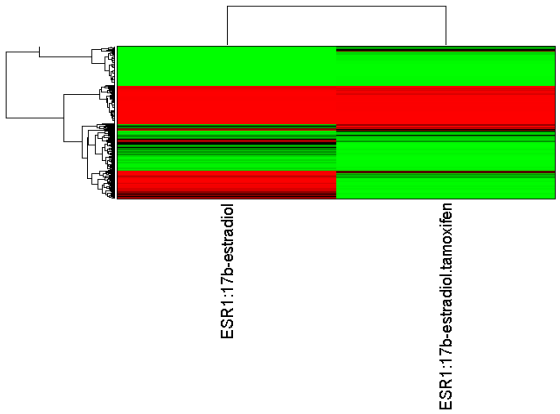


Top 500 peaks

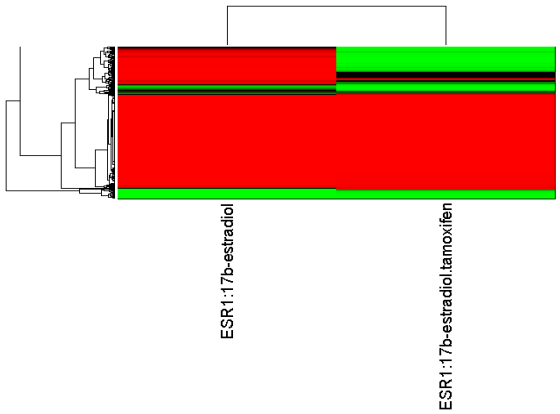


F.30 T-47D ERRA treatment=Genistein 100nM

All peaks

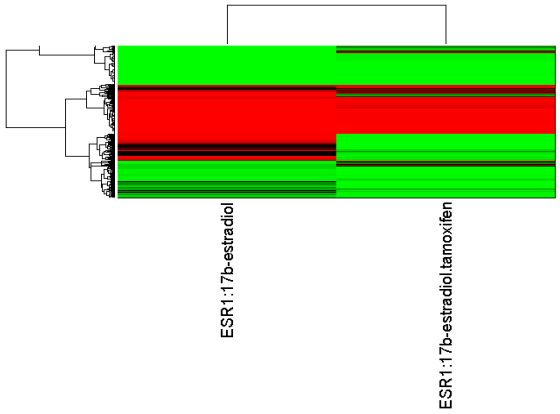


Top 500 peaks

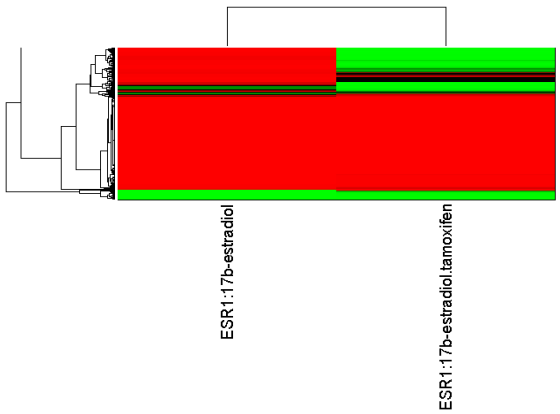


F.31 T-47D ERRA treatment=Estradiol 10nM

All peaks

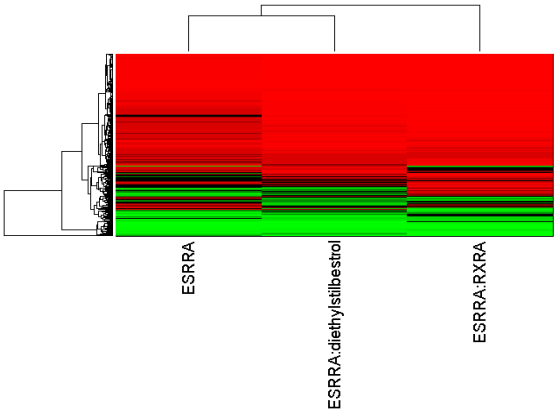


Top 500 peaks

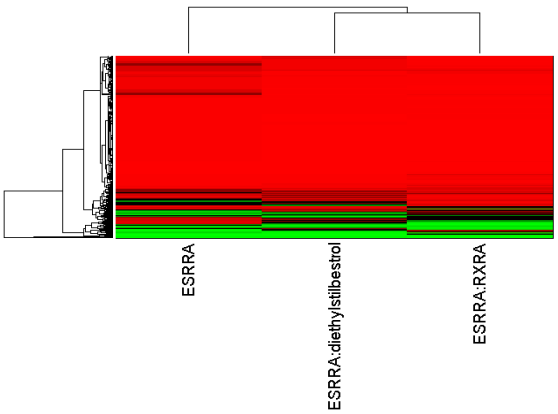


F.32 HepG2 ERRA treatment=forskolin

All peaks

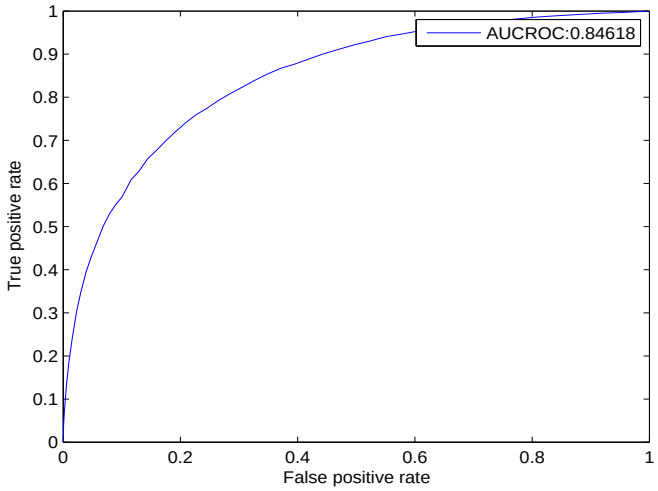


Top 500 peaks

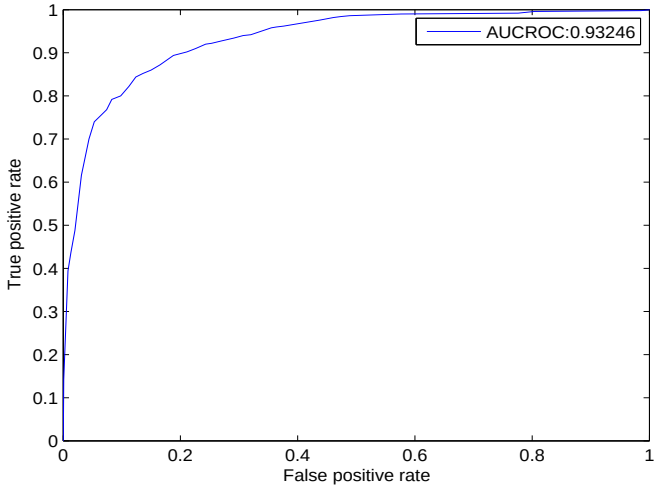


F.33 HepG2 HNF4G (SC-6558)

Predicted by in vitro DNA binding of HNF4G
All peaks

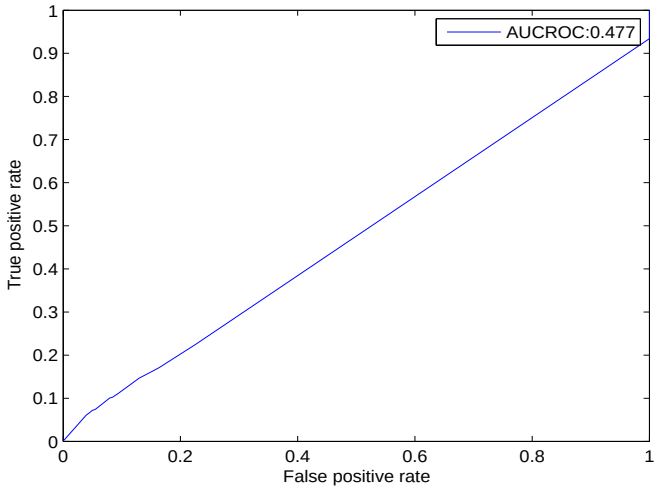


Top 500 peaks

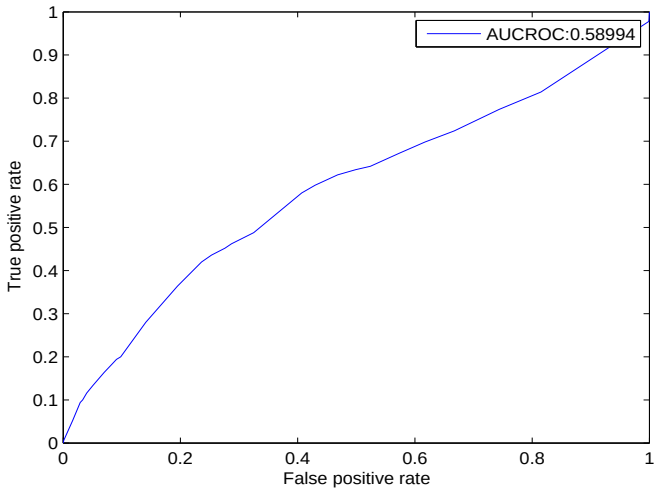


F.34 GM12878 TR4

Predicted by in vitro DNA binding of TR4
All peaks

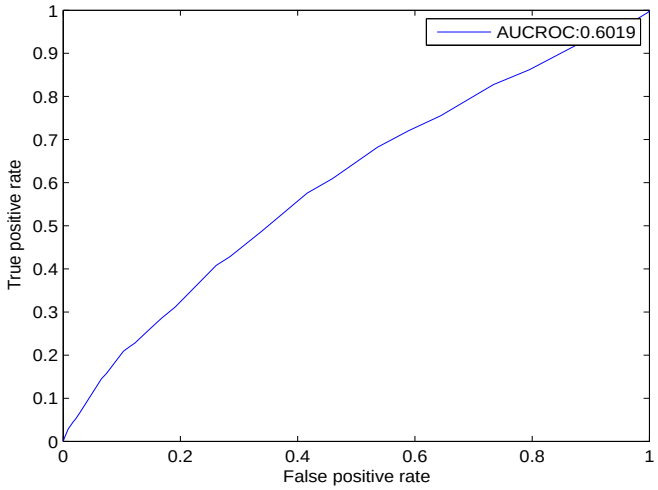


Top 500 peaks

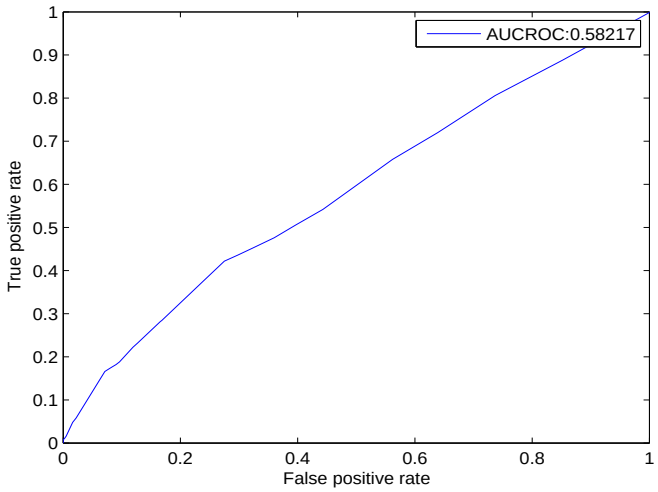


F.35 HeLa-S3 TR4

Predicted by in vitro DNA binding of TR4
All peaks

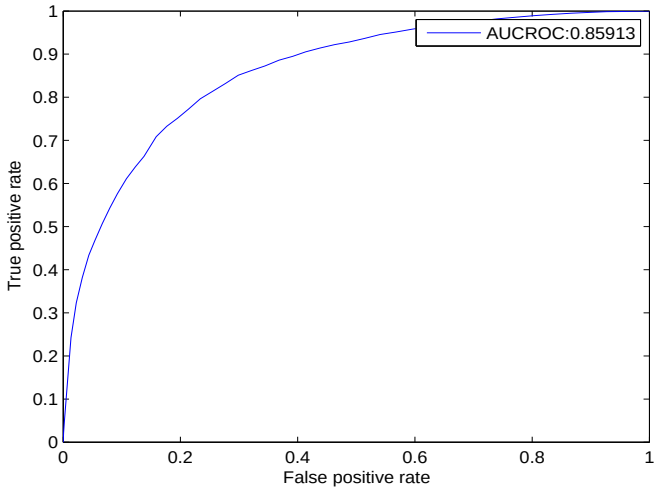


Top 500 peaks

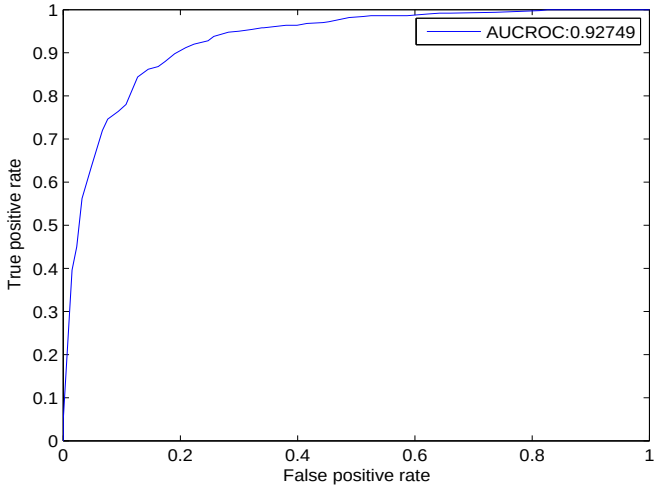


F.36 HepG2 HNF4A treatment=forskolin

Predicted by in vitro DNA binding of HNF4A
All peaks

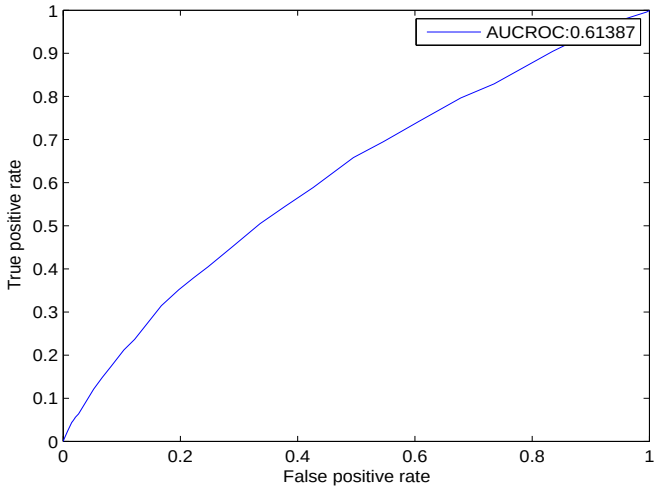


Top 500 peaks

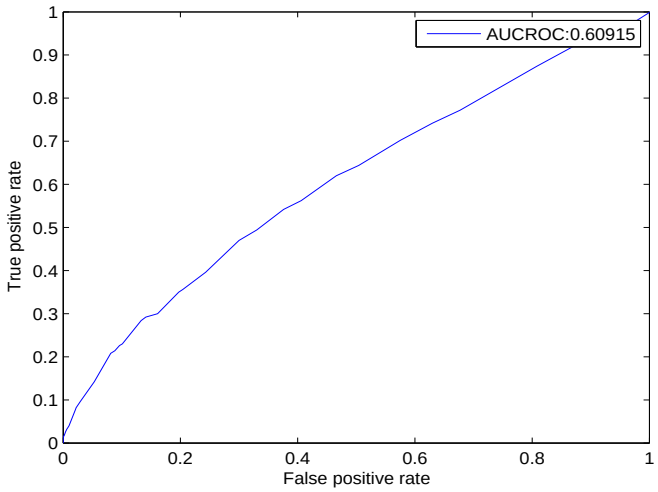


F.37 HepG2 TR4

Predicted by in vitro DNA binding of TR4
All peaks

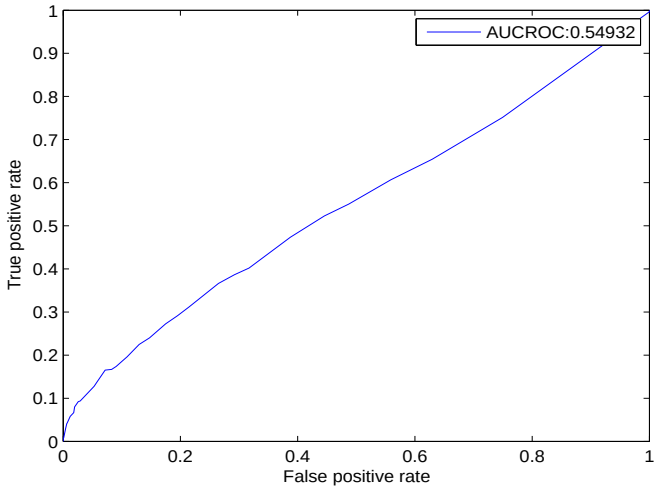


Top 500 peaks

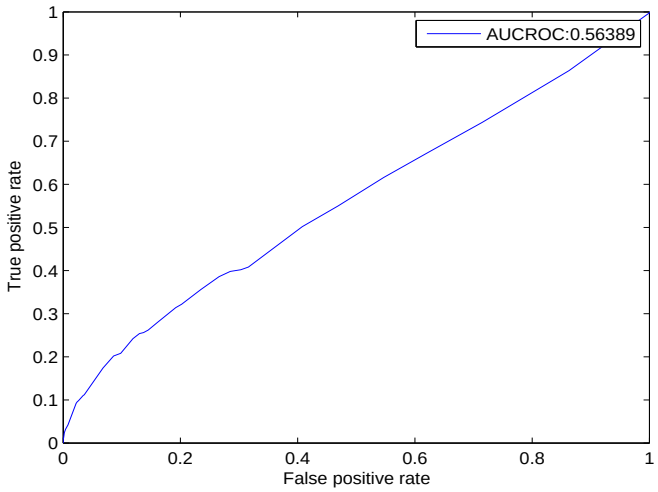


F.38 K562 TR4

Predicted by in vitro DNA binding of TR4
All peaks



Top 500 peaks



Bibliography

[1] M. M. Aagaard, R. Siersbæk, and S. Mandrup, "Molecular basis for gene-specific transactivation by nuclear receptors," *Biochimica et Biophysica Acta - Molecular Basis of Disease*, vol. 1812, no. 8, pp. 824–835, 2011.

[2] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey, "Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning," *Nature Biotechnology*, vol. 33, no. 8, pp. 831–838, 2015.

[3] M. Annala, K. Laurila, H. Lähdesmäki, and M. Nykter, "A linear model for transcription factor binding affinity prediction in protein binding microarrays," *PLoS ONE*, vol. 6, no. 5, pp. e20059, January 2011.

[4] T. L. Bailey and C. Elkan, "The value of prior knowledge in discovering motifs with MEME," *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*, vol. 3, pp. 21–29, 1995.

[5] T. L. Bailey, N. Williams, C. Misleh, and W. W. Li, "MEME: discovering and analyzing DNA and protein sequence motifs," *Nucleic Acids Research*, vol. 34, no. Web Server issue, pp. W369–73, July 2006.

[6] L. A. Barrera, A. Vedenko, J. V. Kurland, J. M. Rogers, S. S. Gisselbrecht, E. J. Rossin, J. Woodard, L. Mariani, K. H. Kock, S. Inukai, T. Siggers, L. Shokri, R. Gordân, N. Sahni, C. Cotsapas, T. Hao, S. Yi, M. Kellis, M. J. Daly, M. Vidal, D. E. Hill, and M. L. Bulyk, "Survey of variation in human transcription factors reveals prevalent DNA binding changes," *Science*, vol. 351, no. 6280, pp. 1450–1454, 2016.

[7] D. R. Bentley, "Whole-genome re-sequencing," *Current opinion in genetics & development*, vol. 16, pp. 545–552, 2006.

[8] M. F. Berger, G. Badis, A. R. Gehrke, S. Talukder, A. a. Philippakis, L. Peña-Castillo, T. M. Alleyne, S. Mnaimneh, O. B. Botvinnik, E. T. Chan, F. Khalid, W. Zhang, D. Newburger, S. a. Jaeger, Q. D. Morris, M. L. Bulyk, and T. R. Hughes, "Variation in Homeodomain DNA Binding Revealed by High-Resolution Analysis of Sequence Preferences," *Cell*, vol. 133, no. 7, pp. 1266–1276, June 2008.

[9] M. F. Berger, A. a. Philippakis, A. M. Qureshi, F. S. He, P. W. Estep, and M. L. Bulyk, "Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities," *Nature Biotechnology*, vol. 24, no. 11, pp. 1429–1435, November 2006.

[10] B. E. Bernstein, E. Birney, I. Dunham, E. D. Green, C. Gunter, and M. Snyder, "An integrated encyclopedia of DNA elements in the human genome," *Nature*, vol. 489, no. 7414, pp. 57–74, September 2012.

[11] D. Bhimsaria, J. A. Rodriguez-Martinez, E. N. Korkmaz, Q. Cui, P. Ramanathan, and A. Z. Ansari, "Specificity Landscapes unmask differential sequence preferences of homologous transcription factors," *In review*.

[12] Y. Blat and N. Kleckner, "Cohesins bind to preferential sites along yeast chromosome III, with differential regulation along arms versus the centric region," *Cell*, vol. 98, no. 2, pp. 249–259, 1999.

[13] A. P. Boyle, E. L. Hong, M. Hariharan, Y. Cheng, M. a. Schaub, M. Kasowski, K. J. Karczewski, J. Park, B. C. Hitz, S. Weng, J. M. Cherry, and M. Snyder, "Annotation of functional variation in personal genomes using RegulomeDB," *Genome research*, vol. 22, no. 9, pp. 1790–7, September 2012.

[14] Z. T. Campbell, D. Bhimsaria, C. T. Valley, J. A. Rodriguez-Martinez, E. Menichelli, J. R. Williamson, A. Z. Ansari, and M. Wickens, "Cooperativity in RNA-Protein Interactions: Global Analysis of RNA Binding Specificity," *Cell Reports*, vol. 1, no. 5, pp. 570–581, May 2012.

[15] E. Candes and T. Tao, "Decoding by Linear Programming," *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[16] E. J. Candès, "Compressive sampling," *Proceedings of the International Congress of Mathematicians*, pp. 1433–1452, 2006.

[17] C. D. Carlson, C. L. Warren, K. E. Hauschild, M. S. Ozers, N. Qadir, D. Bhimsaria, Y. Lee, F. Cerrina, and A. Z. Ansari, "Specificity landscapes of DNA binding molecules elucidate biological function," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 10, pp. 4544–9, March 2010.

[18] X. Chen, T. R. Hughes, and Q. Morris, "RankMotif++: A motif-search algorithm that accounts for relative ranks of K-mers in binding transcription factors," In *Bioinformatics*, volume 23, pp. i72–9, July 2007.

[19] D. Cotnoir-White, D. Laperrière, and S. Mader, "Evolution of the repertoire of nuclear receptor binding sites in genomes," *Molecular and cellular endocrinology*, vol. 334, no. 1-2, pp. 76–82, March 2011.

[20] J. Crocker, N. Abe, L. Rinaldi, A. P. McGregor, N. Frankel, S. Wang, A. Alsawadi, P. Valenti, S. Plaza, F. Payre, R. S. Mann, and D. L. Stern, "Low affinity binding site clusters confer HOX specificity and regulatory robustness," *Cell*, vol. 160, no. 1-2, pp. 191–203, December 2015.

[21] B. Deplancke, D. Alpern, and V. Gardeux, "The Genetics of Transcription Factor DNA Binding Variation," *Cell*, vol. 166, no. 3, pp. 538–554, 2016.

[22] M. F. Duarte and Y. C. Eldar, "Structured compressed censing: From theory to applications," *IEEE Transactions on Signal Processing*, vol. 59, no. 170, pp. 4053–4085, 2011.

[23] G. S. Erwin, D. Bhimsaria, A. Eguchi, and A. Z. Ansari, "Mapping Polyamide-DNA Interactions in Human Cells Reveals a New Design Strategy for Effective Targeting of Genomic Sites," *Angewandte Chemie (International ed. in English)*, vol. 53, no. 38, pp. 1–6, July 2014.

[24] G. S. Erwin, M. P. Grieshop, D. Bhimsaria, A. Eguchi, J. A. Rodríguez-Martínez, and A. Z. Ansari, "Genome-wide Mapping of Drug-DNA Interactions in Cells with COSMIC (Crosslinking of Small Molecules to Isolate Chromatin)," *Journal of visualized experiments : JoVE*, no. 107, pp. e53510, 2016.

[25] R. M. Evans and D. J. Mangelsdorf, "Nuclear Receptors, RXR, and the Big Bang," *Cell*, vol. 157, no. 1, pp. 255–66, March 2014.

[26] B. C. Foat, A. V. Morozov, and H. J. Bussemaker, "Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE," In *Bioinformatics*, volume 22, pp. e141–9, July 2006.

[27] P. M. Fordyce, D. Gerber, D. Tran, J. Zheng, H. Li, J. L. DeRisi, and S. R. Quake, "De novo identification and biophysical characterization of transcription-factor binding sites with microfluidic affinity analysis," *Nature Biotechnology*, vol. 28, no. 9, pp. 970–5, September 2010.

[28] M. Fried and D. M. Crothers, "Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis," *Nucleic Acids Research*, vol. 9, no. 23, pp. 6505–6525, 1981.

[29] M. M. Garner and A. Revzin, "A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system," *Nucleic Acids Research*, vol. 9, no. 13, pp. 3047–3060, 1981.

[30] K. E. Hauschild, J. S. Stover, D. L. Boger, and A. Z. Ansari, "CSI-FID: high throughput label-free detection of DNA binding molecules," *Bioorganic & medicinal chemistry letters*, vol. 19, no. 14, pp. 3779–82, July 2009.

[31] D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold, "Genome-wide mapping of in vivo protein-DNA interactions," *Science*, vol. 316, no. 5830, pp. 1497–1502, June 2007.

[32] A. Jolma, T. Kivioja, J. Toivonen, L. Cheng, G. Wei, M. Enge, M. Taipale, J. M. Vaquerizas, J. Yan, M. J. Sillanpää, M. Bonke, K. Palin, S. Talukder, T. R. Hughes, N. M. Luscombe, E. Ukkonen, and J. Taipale, "Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities," *Genome Research*, vol. 20, no. 6, pp. 861–873, April 2010.

[33] A. Jolma, J. Yan, T. Whitington, J. Toivonen, K. R. Nitta, P. Rastas, E. Morgunova, M. Enge, M. Taipale, G. Wei, K. Palin, J. M. Vaquerizas, R. Vincentelli, N. M. Luscombe, T. R. Hughes, P. Lemaire, E. Ukkonen, T. Kivioja, and J. Taipale, "DNA-binding specificities of human transcription factors," *Cell*, vol. 152, no. 1-2, pp. 327–339, January 2013.

[34] S. Keles, C. L. Warren, C. D. Carlson, and A. Z. Ansari, "CSI-Tree: A regression tree approach for modeling binding properties of DNA-binding molecules based on cognate site identification (CSI) data," *Nucleic Acids Research*, vol. 36, no. 10, pp. 3171–3184, June 2008.

[35] X. Liu, D. Brutlag, and J. Liu, "BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes," *Pacific Symposium on Biocomputing*, vol. 138, pp. 127–138, 2001.

[36] X. S. Liu, D. L. Brutlag, and J. S. Liu, "An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments," *Nature Biotechnology*, vol. 20, no. 8, pp. 835–9, August 2002.

[37] S. G. Mallat and Z. Zhang, "Matching Pursuits With Time-Frequency Dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, 1993.

[38] E. R. Mardis, "Next-generation DNA sequencing methods," *Annual review of genomics and human genetics*, vol. 9, pp. 387–402, 2008.

[39] M. T. Maurano, R. Humbert, E. Rynes, R. E. Thurman, E. Haugen, H. Wang, A. P. Reynolds, R. Sandstrom, H. Qu, J. Brody, A. Shafer, F. Neri, K. Lee, T. Kutyavin, S. Stehling-Sun, A. K. Johnson, T. K. Canfield, E. Giste, M. Diegel, D. Bates, R. S. Hansen, S. Neph, P. J. Sabo, S. Heimfeld, A. Raubitschek, S. Ziegler, C. Cotsapas,

N. Sotoodehnia, I. Glass, S. R. Sunyaev, R. Kaul, and J. A. Stamatoyannopoulos, "Systematic Localization of Common Disease-Associated Variation in Regulatory DNA," *Science*, vol. 337, no. 6099, pp. 1190–1195, 2012.

[40] T. S. Mikkelsen, M. Ku, D. B. Jaffe, B. Issac, E. Lieberman, G. Giannoukos, P. Alvarez, W. Brockman, T.-K. Kim, R. P. Koche, W. Lee, E. Mendenhall, A. O'Donovan, A. Presser, C. Russ, X. Xie, A. Meissner, M. Wernig, R. Jaenisch, C. Nusbaum, E. S. Lander, and B. E. Bernstein, "Genome-wide maps of chromatin state in pluripotent and lineage-committed cells," *Nature*, vol. 448, no. 7153, pp. 553–560, 2007.

[41] D. Needell and J. A. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Applied and Computational Harmonic Analysis*, pp. 1–25, 2008.

[42] S. Neph, J. Vierstra, A. B. Stergachis, A. P. Reynolds, E. Haugen, B. Vernot, R. E. Thurman, S. John, R. Sandstrom, A. K. Johnson, M. T. Maurano, R. Humbert, E. Rynes, H. Wang, S. Vong, K. Lee, D. Bates, M. Diegel, V. Roach, D. Dunn, J. Neri, A. Schafer, R. S. Hansen, T. Kutyavin, E. Giste, M. Weaver, T. Canfield, P. Sabo, M. Zhang, G. Balasundaram, R. Byron, M. J. MacCoss, J. M. Akey, M. a. Bender, M. Groudine, R. Kaul, and J. a. Stamatoyannopoulos, "An expansive human regulatory lexicon encoded in transcription factor footprints," *Nature*, vol. 489, no. 7414, pp. 83–90, September 2012.

[43] M. B. Noyes, R. G. Christensen, A. Wakabayashi, G. D. Stormo, M. H. Brodsky, and S. a. Wolfe, "Analysis of Homeodomain Specificities Allows the Family-wide Prediction of Preferred Recognition Sites," *Cell*, vol. 133, no. 7, pp. 1277–1289, June 2008.

[44] M. B. Noyes, X. Meng, A. Wakabayashi, S. Sinha, M. H. Brodsky, and S. a. Wolfe, "A systematic characterization of factors that regulate Drosophila segmentation via a bacterial one-hybrid system," *Nucleic Acids Research*, vol. 36, no. 8, pp. 2547–60, May 2008.

[45] R. Nutiu, R. C. Friedman, S. Luo, I. Khrebtukova, D. Silva, R. Li, L. Zhang, G. P. Schroth, and C. B. Burge, "Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument," *Nature Biotechnology*, vol. 29, no. 7, pp. 659–664, June 2011.

[46] J. P. Overington, B. Al-Lazikani, and A. L. Hopkins, "How many drug targets are there?," *Nature reviews. Drug discovery*, vol. 5, no. 12, pp. 993–6, 2006.

[47] P. J. Park, "ChIP-seq: advantages and challenges of a maturing technology," *Nature reviews. Genetics*, vol. 10, no. 10, pp. 669–80, October 2009.

[48] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," *Proceedings of 27th Asilomar Conference on Signals, Systems and Computers*, pp. 1–5, 1993.

[49] G. Pavesi, G. Mauri, and G. Pesole, "An algorithm for finding signals of unknown length in DNA sequences," *Bioinformatics*, vol. 17 Suppl 1, pp. S207–S214, 2001.

[50] G. Pavesi, P. Mereghetti, G. Mauri, and G. Pesole, "Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes," *Nucleic Acids Research*, vol. 32, pp. W199–W203, 2004.

[51] F. D. Porter, J. Drago, Y. Xu, S. S. Cheema, C. Wassif, S. P. Huang, E. Lee, a. Grinberg, J. S. Massalas, D. Bodine, F. Alt, and H. Westphal, "Lhx2, a LIM homeobox gene, is required for eye, forebrain, and definitive erythrocyte development," *Development (Cambridge, England)*, vol. 124, no. 15, pp. 2935–44, August 1997.

[52] B. Ren, F. Robert, J. J. Wyrick, 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*, vol. 290, no. 5500, pp. 2306–2309, 2000.

[53] T. R. Riley, A. Lazarovici, R. S. Mann, and H. J. Bussemaker, "Building accurate sequence-to-affinity models from high-throughput in vitro protein-DNA binding data using FeatureREDUCE," *eLife*, vol. 4, pp. e06397, December 2015.

[54] J. A. Rodriguez-Martinez, Reinke Aaron W., Bhimsaria Devesh, A. E. Keating, and A. Z. Ansari, "Combinatorial bZIP dimers define complex DNA-binding specificity landscapes," *In review*.

[55] R. Rohs, S. M. West, A. Sosinsky, P. Liu, R. S. Mann, and B. Honig, "The role of DNA shape in protein-DNA recognition," *Nature*, vol. 461, no. 7268, pp. 1248–53, October 2009.

[56] S. L. Salzberg, a. L. Delcher, S. Kasif, and O. White, "Microbial gene identification using interpolated Markov models," *Nucleic Acids Research*, vol. 26, no. 2, pp. 544–8, January 1998.

[57] T. D. Schneider and R. M. Stephens, "Sequence logos: a new way to display consensus sequences," *Nucleic Acids Research*, vol. 18, no. 20, pp. 6097–6100, 1990.

[58] J. Shendure and H. Ji, "Next-generation DNA sequencing," *Nature Biotechnology*, vol. 26, pp. 1135–1145, 2008.

[59] M. Slattery, T. Riley, P. Liu, N. Abe, P. Gomez-Alcala, I. Dror, T. Zhou, R. Rohs, B. Honig, H. J. Bussemaker, and R. S. Mann, "Cofactor binding evokes latent differences in DNA binding specificity between hox proteins," *Cell*, vol. 147, no. 6, pp. 1270–1282, December 2011.

[60] M. J. Solomon, P. L. Larsen, and A. Varshavsky, "Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene," *Cell*, vol. 53, no. 6, pp. 937–947, 1988.

[61] G. D. Stormo, T. D. Schneider, and L. Gold, "Quantitative analysis of the relationship between nucleotide sequence and functional activity," *Nucleic Acids Research*, vol. 14, no. 16, pp. 6661–79, 1986.

[62] G. D. Stormo, "Maximally Efficient Modeling of DNA Sequence Motifs at All Levels of Complexity," *Genetics*, vol. 1224, pp. 1219–1224, February 2011.

[63] G. D. Stormo and Y. Zhao, "Determining the specificity of protein-DNA interactions," *Nature reviews. Genetics*, vol. 11, no. 11, pp. 751–60, September 2010.

[64] J. R. Tietjen, L. J. Donato, D. Bhimsaria, and A. Z. Ansari, "Sequence-specificity and energy landscapes of DNA-binding molecules," *Methods in Enzymology*, vol. 497, pp. 3–30, January 2011.

[65] P. H. von Hippel and O. G. Berg, "On the specificity of DNA-protein interactions," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 83, no. 6, pp. 1608–12, March 1986.

[66] C. L. Warren, N. C. S. Kratochvil, K. E. Hauschild, S. Foister, M. L. Brezinski, P. B. Dervan, G. N. Phillips, and A. Z. Ansari, "Defining the sequence-recognition profile of DNA-binding molecules," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 4, pp. 867–872, January 2006.

[67] C. L. Warren, J. Zhao, K. Glass, V. Rishi, A. Z. Ansari, and C. Vinson, "Fabrication of duplex DNA microarrays incorporating methyl-5-cytosine," *Lab on a chip*, vol. 12, no. 2, pp. 376–80, January 2012.

[68] M. T. Weirauch, A. Cote, R. Norel, M. Annala, Y. Zhao, T. R. Riley, J. Saez-Rodriguez, T. Cokelaer, A. Vedenko, S. Talukder, H. J. Bussemaker, Q. D. Morris, M. L. Bulyk, G. Stolovitzky, and T. R. Hughes, "Evaluation of methods for modeling transcription factor sequence specificity," *Nature Biotechnology*, vol. 31, no. 2, pp. 126–34, January 2013.

[69] M. T. Weirauch, A. Yang, M. Albu, A. G. Cote, A. Montenegro-Montero, P. Drewe, H. S. Najafabadi, S. A. Lambert, I. Mann, K. Cook, H. Zheng, A. Goity, H. van

Bakel, J.-C. Lozano, M. Galli, M. G. Lewsey, E. Huang, T. Mukherjee, X. Chen, J. S. Reece-Hoyes, S. Govindarajan, G. Shaulsky, A. J. Walhout, F.-Y. Bouget, G. Ratsch, L. F. Larrondo, J. R. Ecker, and T. R. Hughes, "Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity," *Cell*, vol. 158, no. 6, pp. 1431–1443, September 2014.

[70] J. Yan, M. Enge, T. Whitington, K. Dave, J. Liu, I. Sur, B. Schmierer, A. Jolma, T. Kivioja, M. Taipale, and J. Taipale, "Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites," *Cell*, vol. 154, no. 4, pp. 801–813, August 2013.

[71] Y. Zhao, D. Granas, and G. D. Stormo, "Inferring binding energies from selected binding sites," *PLoS Computational Biology*, vol. 5, no. 12, pp. e1000590, December 2009.

[72] Y. Zhao and G. D. Stormo, "Quantitative analysis demonstrates most transcription factors require only simple models of specificity," *Nature Biotechnology*, vol. 29, no. 6, pp. 480–483, 2011.