MODELING P53 TRANSCRIPTIONAL REGULATION

by

TODD ROBERT RILEY

A Dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Computational Biology and Molecular Biophysics

Written under the direction of

Arnold J. Levine and Eduardo D. Sontag

and approved by

New Brunswick, New Jersey

October, 2008

ABSTRACT OF THE DISSERTATION

Modeling p53 Transcriptional Regulation

By Todd Robert Riley

Dissertation Directors:
Arnold J. Levine and Eduardo D. Sontag

The p53 protein has been called the “gatekeeper” of the cell. After DNA damage, p53 transcriptionally activates downstream pathways to prevent cancer from occurring. Target genes are activated to cause cell cycle arrest, DNA repair, cell senescence, and apoptosis. However, the exact transcriptional program that determines the specific outcome of a certain cell stress is not completely understood. We present an analysis to help shed light on the mechanisms of transcriptional regulation by the p53 protein.

First, we present a detailed analysis of the known modes of p53-regulation, and a dataset of 160 functional human p53-binding sites curated from the literature. Second, we present a new method (called p53HMM) to model p53-binding sites using Profile Hidden Markov Models (PHMMs). This new method is the most accurate predictor of functional p53 single-sites and cluster-sites to date. Third, we show that functional sites with low estimated relative affinity scores are highly correlated with distances from the TSS. Fourth, we show that the capability to fold into a non-linear cruciform DNA structure is an important predictor in estimating the overall binding affinity of a functional p53-binding site. We use UNAFold to calculate free energies and probabilities of p53-binding sites folding into different non-linear cruciform structures. Fifth, we present a new motif-finding algorithm (called PURE) that uses relative entropy to find over- and under-represented motifs near functional p53-binding sites. The goal is to find possible motifs (like co-factor motifs) that can help designate functional p53-binding sites, thereby reducing the false positive rate that currently plagues motif-finding algorithms.

Acknowledgements

I wish to thank my advisers Arnold Levine and Eduardo Sontag for their guidance and support. I was very fortunate to have such great and knowledgeable mentors to help guide my research.

I am grateful for my collaborations with Jiri Vanicek and Xin Yu, and for many thoughtful discussions with Gareth Bond, Michael Krasnitz, Raúl Rabadán, Gurinder Atwal, and Alexei Vazquez at the Simons Center for Systems Biology at the Institute for Advanced Study.

I am also grateful for the many thoughtful discussions and the camaraderie with Erhan Bilal, Amar Drawid, Kerri-Ann Norton, and Ariella Sasson in the BioMaPS Institute at Rutgers University.

I owe a debt of gratitude to Suzanne Christen, Paul Ehrlich, and Janice Pawlo. I have greatly appreciated their ever-friendly assistance over the years. I am also indebted to Diane DePiano, Alan Cheng, and Susan Higgins.

I also wish to thank the members of RUGCF that helped me maintain perspective in times of uncertainty.

I thank my mom and dad for their support and guidance through the years. They were and are wonderful parents and an inspiration for me to follow.

I thank my wife, Cari, for her love, support, patience, and wonderful mastery of English grammar. What a beautiful combination!

Dedication

To my father, my mentor, and my hero, a great man full of compassion, wisdom, humor, and love. I miss you every day.

To my mother, a wonderful woman full of laughter, joy, understanding, and love. Thank you for all you've done for me over the years.

To my wife, my rock, and my soul-mate, I thank God for bringing us together. Thank you for everything.

Table of Contents

Chapter 1

Introduction

The goal of this work is to increase our understanding of the mechanisms of transcriptional regulation by the p53 protein. The research is presented in five logical segments. Each segment (chapter) focuses on different characteristics of p53 transcriptional regulation and different computational approaches to help in our understanding.

In chapter 2, we present a detailed analysis of the known modes of p53 trans-activation and repression. We also present a detailed dataset of the 160 functional human p53-binding sites that we curated from the literature.

In chapter 3, we present a new method (called p53HMM) to model p53-binding sites using Profile Hidden Markov Models (PHMMs). This new method proposes a novel training method that leverages the redundant information in the repeated, palindromic p53-motif to increase predictive accuracy. We also present a new p53 cluster-site accuracy that correctly models experimental measurements. This new method is the most accurate predictor of functional p53 single-sites and cluster-sites to date.

In chapter 4, we show that functional sites with low estimated relative affinity scores are highly correlated with distances from the TSS. By using a Dynamic Acceptance Threshold during site searches, we are able to reduce the false positive rate three-fold.

In chapter 5, we show that a functional p53-binding site's capability to fold into a non-linear cruciform structure is an important predictor variable in estimating the overall binding affinity of the binding-site. We leverage UNAFold to calculate free energies and probabilities of p53-binding sites folding into different non-linear cruciform structures. We then construct linear models that use both sequence similarity and folding capacity to predict measured binding affinities. Differences between the models provide insight into the differences between the *in vitro* and *in vivo* conditions of the experiments.

In chapter 6, we present a new motif-finding algorithm (called PURE) that uses relative entropy to find over- and under-represented motifs near functional human p53-binding sites. The goal of *de novo* motif discovery is to find possible motifs (like co-factor motifs) that can help designate functional p53-binding sites, in the hope of reducing the false positive rate that currently plagues motif-finding algorithms. We also analyze the new algorithm by attempting to “re-discover” known transcriptional and post-transcriptional motifs in human genes.

Chapter 2

Review and Analysis of p53-binding Sites

The p53 pathway responds to a wide variety of cellular stress signals *(the input)* by activating p53 as a transcription factor (increasing its concentration and protein modifications) and transcribing a programme of genes *(the output)* to accomplish a variety of functions. Together, these functions prevent errors in the duplication process of a cell that is under stress, and as such the p53 pathway increases the fidelity of cell division and prevents cancers from arising. The goals of this review are: first, to bring together in one source a list of p53-regulated genes and the criteria that permits this classification; second, to analyze the p53 response elements (REs) in DNA that bind the p53 protein and promote transcriptional control; third, to organize and explore the functions of the p53-regulated genes; and finally, to review useful algorithms that can detect p53-regulated genes by their associated REs in DNA from various sources.

The value of this exercise is to bring together a large body of literature that has mostly been assembled one gene and one publication at a time. This has not permitted an appreciation of the cooperative and broad nature of the functions of many p53-regulated genes in altering the cell and the extracellular matrix (ECM), and the role of the p53 response in communicating with various organ systems of the body. There is good evidence that the nature of the stress signal *(the input)* and the cell type can both modulate the transcriptional pattern of p53-responsive genes that respond with a transcriptional programme *(the* *output)* [293, 71]. Because we have imperfect information about cell and tissue types and the nature of the stress signal for every gene discussed here, we can provide only a broad overview of the transcriptional programme regulated by the p53 protein. Where there is detailed information about cell and tissue type, and stress response, it will be discussed.

2.1 Criteria used to identify p53-responsive genes

Four sets of experimental criteria have been employed to identify a p53-regulated gene. The first is the presence of a p53 RE in the DNA near or in the gene. The second is a demonstration that the gene is indeed either up- or down-regulated at the RNA and protein levels by the activated wild type p53 protein (and not the mutant protein). The third line of evidence is to clone the p53 RE from that gene, place it near a test gene, such as luciferase, and demonstrate that the p53 protein can regulate the test gene. The fourth approach is to employ chromatin immunoprecipitation with a p53-specific antibody to demonstrate the presence of the p53 protein on the RE site in the DNA. In some cases a gel shift assay is also employed to demonstrate that the p53 protein binds *in vitro* to the p53 RE sequence from that gene.

These criteria may be modified by the cell or tissue specificity of some p53-regulated genes or by the nature of the stress signal that is responded to by the p53 pathway. In this review we have included a list of p53-responsive genes that have met a minimum of three out of four of these criteria. Based upon these criteria, Tables 2.3 on page 30 and 2.4 on page 38 contain 129 genes and 160 p53 REs from both the human and viral genomes (several of the genes contain more than one p53 RE). Table 2.3 on page 30 provides the gene name, the full description of this name, its accession number, a description of the p53 RE, and its spacer, if it employs one. Table 2.4 on page 38 provides the name of the gene, the location of the p53 RE, whether this RE functions as a transcriptional activator or a repressor, the distance from the transcriptional start site (TSS) of the p53 RE, the proposed functions of the gene product, and a reference to the publication that describes these properties of the p53-regulated gene. Table 2.2 lists the 15 p53 cluster sites that are present in tables 2.3 and 2.4, and the number of half-sites found in each (p53 REs with more than two half-sites are referred to as cluster sites). We should note that the p53-target list found in tables 2.3 and 2.4 is most probably not exhaustive, and will very likely grow as additional experimental evidence is acquired.

2.2 The p53 Consensus Motif

Two different groups first identified a p53 consensus sequence in the DNA to which the p53 protein bound with high affinity and specificity [79, 57]. The sequence was degenerate and was composed of $5'$-RRRCWWGYYY-$3'$ where R is a purine, Y a pyrimidine, W is either A or T (adenine or thymine) and G is guanine and C is cytosine (see Table 2.1 on the following page) [79, 57]. The p53-binding site in the genomes of many organisms is composed of a half-site RRRCWWGYYY followed by a spacer, usually composed of 0-21 base pairs, which is then followed by a second half-site RRRCWWCYYY sequence. By labeling each quarter-site RRRCW as $\longrightarrow$ and WGYYY as $\longleftarrow$, the first discovered p53 consensus sequence can be graphically represented by $\longrightarrow\longleftarrow_{spacer}\longrightarrow\longleftarrow$. This configuration of the four quarter-sites is often referred to at the head-to-head (HH) orientation. The two other possible orientations of the quarter-sites are tail-to-tail (TT, $\longleftarrow\longrightarrow_{spacer}\longleftarrow\longrightarrow$), and head-to-tail (HT, $\longrightarrow\longrightarrow_{spacer}\longrightarrow\longrightarrow$). (TH is not used since the complementary strand would contain an HT-oriented site).

Table 2.1: Original Data Used To Define The p53 Consensus Binding Site

		1^{st} Half-site		2^{nd} Half-site	
Clone	5′ Region	R R R C W W G Y Y Y	*Spacer*	R R R C W W G Y Y Y	3′ Region
s57	CGACCTGTCA caccg	G G G C C T G T C A		C A G C A T GaC C T	acctgtcacaccggg
N22	atttt CACCATGCTT	C T G C A T G T C T		A G G C A A G T C A	ccttctc CACTGGCC
11A2	ccccatcctccatcc	A A A C AaT G C C C		A G A C T T G T C T	ct CCGCCTGAAT ga
W211	tttgtcctaccatcc	A G G C A T G C C T		- - - - T T G C C T	CACTCGTTA tttcct
W7B2	tatct GTGCAGCTG t	G G G C A T G T T T	t	A G G C A A G C T T	cct GTGCTAGTTC cc
3H	AACTAGATC cttttc	A G A C A T G T T A		T A A C A A G T C A	GTACAAGTTT atttt
8A	gctggt GCACAAGAG	T G A C A T G T C C		C G A C G T G T T T	tgtc
532	CATCATGCCA cctgc	A G G C A T G T T C	tggat	G G G C - T G T C T	t GTGCTTTGTTG ttt
64A2	c AAACCAGGGT gtct	T G A C T T G C C T	atcctgggaggt	T G A C A T G T T C	ctccccttccccctc
W7A1	gccaaacataaccac	C A G C - T G C C A		A G G C A T G C A G	tacc ACGCTCAGCCC
s61	c	C A A C T T G T C T	attctgtgttgat	G G A C A T G T T C	ccgtttttggctatt
11B3	actgttgatgatgaa	A G A C A A G C C T	a	G G G C A G G T C C	tggggggtgggg
N42	gcagtgtggtggagg	A A A C A A G C C C	a	G G A T G T G C C C	a GGGCAGGCTG ggac
s201	tgttc ATACCTGTCC	A C A C T T G T C T		A T A C C T G C C T	ACACCTGTCT tgttt
s1583	ctttaattcagttgt	A A A C A T GaC T T	gttcattata	T G A C A T G T T C	aattacaattcgatt
s592I	ctcagttctcagctg	G G A C T T G C C C		T G G C C A G C C C	tgg GGTCACTGCTG c
s592II	tgcctcagcacctcc	A G G T TcT G C C -		G G G C T T G T T C	ctttcctttcagcat
2NB	gcctttgttgtgccc	T G A C T T G C C C		A G A C A T G T T T	gggaa TGTCTTGTGC
9H	gtattctcttttcct	A A G C A T G C C T		T G A C T T G T T C	tttcatctcctctga
CBE10d	tgaaagcaggtagat	T G C C T T G C C T		G G A C T T G C C T	GGCCTTGCCT tttct

This table presents the original DNA fragments from El-Deiry et al., that were collected from a genome-wide, p53-antibody immunoprecipitation and were used to define the head-to-head (HH) p53 Consensus Binding Site [57]. The yellow columns corresponding to the 1^{st} and 2^{nd} half-sites were used to define the consensus p53 motif. The p53-binding site is highly degenerative. Within the yellow columns, notice that seven of the 20 DNA target sites (35%) had at least one nucleotide insertion (green), deletion (red), or both (magenta) relative to the discovered 10bp-spacer-10bp consensus. Since insertions and deletions throw off the reading frame of a weight matrix, any PSSM approach will inherently mis-score at least 35% of these 20 sites. Alignments of the 160 experimentally validated p53-binding sites also reveal that any PSSM approach would inherently mis-score at least 30% of them as well. Another observation is that additional p53 half-sites are immediately adjacent (in yellow) to the ones used to define the consensus in 15 of the 20 target sites (75%). Since the genome-wide immunoprecipitation study was designed to pull down the highest affinity sites, the fact that 75% of the target sites are actually p53 cluster-sites is the first indication that cluster-sites of 3 or more half-sites confer higher binding affinity [20].

In almost all natural p53-binding sites, the two half-sites share the same orientations of their quarter-sites. Experiments have shown that the tetramer p53 protein can bind all three (HH, TT, HT) orientations of the quarter-sites with equally high affinity [57]. However, only a few of the experimentally validated p53-binding sites in this analysis do not have the head-to-head (HH) orientation. Due to allowed insertions and deletions (*indels*) relative to the consensus, half-sites can vary in size between 8 to 12 base pairs, although most have 10. Also, some p53 REs have more than two half-sites, and as such are referred to as cluster sites. A variety of experiments have shown that the level of binding affinity and subsequent trans-activation increases linearly with the number of adjacent half-sites [122, 20, 246]. Finally, some genes contain multiple p53-binding sites in different locations within the gene and promoter region, where each p53 RE can contribute to the p53 response. For example a $\longrightarrow\longrightarrow\longrightarrow\longleftarrow\ \longrightarrow$ cluster site is present in the promoter of the CDKN1A (p21) gene ≈ 900 base pairs $3'$ to a canonical $\longrightarrow\longleftarrow_{spacer}\longrightarrow\longleftarrow$ site, and both of these sites contribute to the induction of CDKN1A mRNA after a p53 stress response [58, 64, 215].

2.3 The p53 pathway and the functions of p53-regulated genes

The mechanisms of activation of the p53 pathway and the cellular outcomes produced by p53-activated genes are presented in Figure 2.1. Many proteins are involved in the p53 pathway in order to respond to stress signals and then to produce the proper response.

Stress signals determine the transcriptional programme. The p53 pathway responds to a wide variety of stress signals. These include several types of DNA damage: telomere shortening, hypoxia, mitotic spindle damage, heat or cold shock, unfolded proteins, improper ribosomal biogenesis, nutritional deprivation in a transformed cell, or even the activation of some oncogenes by mutation (see Figure 2.1) [267, 137]. These stress signals are detected by a variety of protein activities that mediate the information about cellular damage (via protein modifications) to the p53 protein or to its negative regulator, MDM2 – a ubiquitin ligase that both blocks p53 transcriptional activity directly (sterically) and mediates the degradation of the p53 protein [136].

The half-life of the p53 protein in many cells varies between 6 to 20 minutes. After a stress signal the MDM2 protein poly-ubiqutinates itself, and this results in the degradation of MDM2 and an increase in the half-life of the p53 protein from minutes to hours. Other mediators of the stress response act through protein modifications of p53. These rapid mechanisms of p53 protein modification and the greatly increased half-life of the p53 protein do not depend upon the slower mechanisms of transcription (of a damaged DNA template) or RNA transport. Thus the response to stress is rapid and it has been proposed (but not proven) that the nature of the stress signal determines the form of the protein modification and therefore the transcriptional programme of the p53 protein.

This is one way to integrate the nature of cellular stress signals at a single protein in the cell, whereby the activated p53 protein then binds to the p53 REs in the DNA and promotes a transcriptional programme that responds to that particular stress. There have been a number of experiments that suggest that, in addition to a transcriptional response to cellular damage, the p53 protein can act directly to trigger a response such as apoptosis [152]. While this is an active area of research, detailed mechanisms describing how p53 acts on or in the mitochondria to promote apoptosis are still lacking.

Outcomes of transcriptional activation. There are three major outcomes after the activation of p53: apoptosis, senescence, or cell cycle arrest. The first two are terminal for the cell, while cell cycle arrest can permit repair processes to act and damage to be reversed, so that the cell survives. The choice between these three outcomes in a stressed cell depends upon a number of other variables, which indicates that the p53 pathway is sensing the activities of other signal transduction pathways. For example, glucose starvation of normal cells results in the phosphorylation by AMP kinase of the p53 protein on serine 15 but no further activation of p53-mediated transcription. By contrast, glucose starvation of a transformed cell results in p53-mediated apoptosis [71]. p53 activation in some cell types that typically results in apoptosis can be reversed or reduced by the treatment of those cells with interleukin [206, 284]. The introduction of an activated RAS oncogene into a normal cell results in a p53-mediated senescence [281]. As part of this senescent state, p53-mediated transcripts produce cytokines that attract inflammatory cells, which in turn eliminate the RAS-transformed cell from an organ [281]. Thus, it is clear that elements of the p53 pathway are regulated by inputs of other signal transduction pathways, resulting in different programmes of transcription by the p53 protein.

While these three functional responses (apoptosis, senescence, and cell cycle arrest) are well appreciated, there are a number of other cellular processes that are altered by gene products regulated by the p53 protein. These include both positive and negative feedback loops in the p53 pathway [93], regulation of other signal transduction pathways and autophagy [71, 72], alterations in the extracellular matrix of cells, alterations in the cytoskeleton of cells, activation of the endosome compartment of cells with increased exosomal and endosomal activity [286], and the regulation of protein translation [37, 59, 196, 283], heat shock proteins [292, 2], and DNA repair processes [246, 172, 247, 232].

The above processes all occur within or around a cell at the molecular and cellular levels, but there are also physiological or systemic consequences of a p53 response to a stress. Exosomes produced by p53 activation of the endosomal compartment in an apoptotic cell, after a p53 response to stress, combine with dendritic cells in the body and can enhance the immunization process for antigens in the stressed cell [286]. A variety of p53-regulated genes that express and act in the central nervous system can alter communication between neurons or result in neurodegeneration in some situations [70]. The p53 regulation of the LIF (leukemia inhibitory factor) gene in the uterus can directly regulate the efficiency of implantation of embryos in mice [106]. Thus, the p53-mediated transcriptional process can have systemic consequences in a host and communicate a stress signal throughout the body. These types of functions are reviewed in Table 2.4 on page 38.

2.4 Modes of p53 regulation

The p53 protein can either activate or repress the transcription of a gene. The major mode of transcriptional activation is through direct, sequence-specific DNA binding. A number of the genes listed in Table 2.4 on page 38 are transcriptionally repressed by p53. P53 employs both direct and indirect methods to repress gene transcription.

Activation by p53 through direct binding and co-factor recruitment. Almost all p53-activated genes have at least one putative DNA-binding site that moderately matches the consensus p53 response element. Through protein-protein interactions, p53 can bind to and then recruit general transcription proteins (TAFs) to the promoter-enhancer region of p53-regulated genes to induce transcription [67, 253]. Recent experiments have shown that p53 can also recruit the histone acetyltransferases (HATs) CBP, p300, and PCAF to the promoter-enhancer region of genes (via high-affinity protein-protein binding) [86, 85]. These

HATs acetylate lysine residues of histones within chromatin, which increases transcriptional activity.

Repression by p53 through direct and indirect means. In some genes, p53 binds to its response element (RE) resulting in direct repression of that gene. It is not clear as to what distinguishes an RE sequence from being a transcriptional activator site versus a repressor site. There are currently three generally accepted methods of direct p53-mediated repression: first, binding-site overlap (steric interference); second, p53 squelching of transcriptional activators; and third, p53-mediated recruitment of histone deacetylases (HDACs).

The p53-mediated repression by steric interference involves sequence specific DNA binding by p53 that overlaps the binding site of another (more powerful) transactivating protein. Examples of genes repressed by the method of p53 steric interference include: AFP, BCL2, HBV (Hepatitis B virus). In these examples, the corresponding activators that are occluded by DNA-bound p53 are FOXA1, POU4F1, and both RFX1/ABL1, respectively [134, 24, 188]. An entire family of cell cycle regulatory genes now appear to share the same squelching mechanism, whereby p53 binds to and suppresses bound and unbound activators of the CCAAT box, namely heterotrimeric NF-Y and CEBP. Examples of genes that share this mechanism are: cyclin A2, cdc25c, cdc2, hsp70, chk2, cdk1, FN1, BRCA1, and PTGS2 (COX2) [3, 11, 29, 30, 109, 288, 116, 159, 2, 112, 240].

The p53 squelching (inactivation) of other DNA-bound and DNA-unbound activators occurs through p53-mediated protein-protein interactions. Examples of p53-squelching of other transactivating genes are: Cyclin B1, TERT, IGF1R, ALB, and MMP1. The corresponding DNA-bound proteins that are inactivated by direct p53-binding are Sp1, Sp1, Sp1, CEBPB, and AP1, respectively [111, 117, 184, 129, 241]. Due to the observation that p53 binds the transcription machinery proteins TBP, TAF6 ($TAF_{II}70$), TAF9 ($TAF_{II}31$), and others *in vitro*, it was initially believed that p53 repression was achieved via p53 binding and suppression of these TATA-bound basal factors *in vivo* [220, 258, 67, 253]. Experimental evidence suggests that the preferred *in vivo* method of p53-mediated squelching is achieved by binding and inhibiting the transactivators of the CCAAT box [68, 2, 288]. However, it remains unclear whether or not these squelching mechanisms of repression are employed *in vivo* under normal physiological conditions.

The p53-mediated recruitment of histone deacetylases (HDACs) occurs through p53 binding to the repressor protein SIN3A, which in turn binds the histone deacetylase HDAC1 [176]. After p53-mediated recruitment to the promoter-enhancer region of a gene, HDAC1 deacetylates lysine residues of histones within chromatin, which represses gene transcription [176, 92]. Examples of genes repressed through this p53-mediated mechanism include MAP4, STMN1, and HSP90AB1 [176, 292].

There are two generally accepted modes of indirect p53-mediated repression. The first method of indirect repression by p53 comes about by p53-mediated activation of CDKN1A (p21), which in turn inhibits the cyclin D–CDK4 complex through direct binding. The consequence of this inhibition of cyclin D–CDK4 is the absence of hyperphosphorylation of the retinoblastoma (RB) protein in the $G1$ stage of the cell cycle [138]. Unphosphorylated RB represses the function of the E2F family of transcription factors through direct binding (forming an E2F–DP1–RB complex), thereby inhibiting the many downstream targets of E2F (including cyclin E, cyclin A, DNA polymerase, and thymidine kinase), and halting the cell cycle in $G1$ phase. It appears that many genes are fully or partially repressed through p53-mediated induction of CDKN1A and ensuing repression of E2F via RB unphosphorylation [138]. Table 2.4 on page 38 shows only those genes that are directly repressed by the p53 protein (and thus have an experimentally-validated p53 RE). In the second method of indirect p53-mediated repression, p53 binds to another transcription factor and, together, they repress a gene without a p53-specific RE.

2.5 Less established modes of p53 regulation

Investigators have also put forth other, and sometimes controversial, models for additional mechanisms of p53 repression and activation. One model proposes that the switch between p53 activation and repression is determined by the length of the spacer [100]. The hypothesis is that p53 proteins bound to a 3-bp spacer binding site are ineffective in recruiting the necessary additional activation proteins, while simultaneously occluding them from adjacent or overlapping REs. Investigators were able to convert direct p53-repression of the BIRC5 (survivin) gene into direct p53-activation by deleting the 3-bp spacer present in the p53 RE [100]. In this analysis, we show that experimentally validated repressor-sites do indeed have longer spacers (see Figure 2.2). However, many activator-sites have spacers of three or more base pairs as well.

Another model proposes that the existence of an adjacent response element (designated "EP" which binds RFX1 and ABL1 proteins) is sufficient to transform an activating p53 RE to a repressing RE [188]. Interestingly, Ori *et al.* succeeded in transforming the direct p53 repression found in the enhancer of HBV into direct p53 activation by mutating the adjacent EP response element. They also succeeded in transforming the direct p53 activation of mdm2 into direct p53 repression by inserting an EP response element adjacent to the p53 RE.

Yet another model proposes that the orientation of the quarter-sites within the p53 binding element determines activation versus repression. Johnson *et al.* propose that head-to-head (HH) p53-binding sites produce p53 activation, while head-to-tail (HT) sites produce p53 repression [115]. Interestingly, they succeeded in converting the p53-repressed ABCB1 gene into p53-activated by replacing the HT p53 RE in the promoter with an HH p53 RE. No experiments were performed with tail-to-tail (TT) p53-binding sites. However, it should be noted that all other experimentally validated repressing p53 REs in this analysis have a head-to-head (HH) configuration, and that the HT cluster site in the $5'$ UTR of the TP53i3 (PIG3) gene confers p53 transactivation, rather than repression. In the case of the heat shock gene HSP90AB1, investigators discovered a biphasic p53 regulatory system, where the co-factor p300 mediated p53 activation, and the co-factors SIN3A and HDAC1 mediated p53 repression [292]. Another important co-factor for p53 regulation in some genes, including CAV1, is E2F [17]. Combining these observations draws the following conclusions: first, properties of the p53 RE and adjacent co-factor REs confer the *potential* for direct p53 activation, respression, or both; and second, the induction of the right combination of p53 and co-factor proteins is required to regulate any potentially functional target site, when either activating or repressing.

2.6 Factors that affect p53 regulation

Experiments have shown that many factors can affect the mode and degree that p53 regulates different target genes. These factors include co-factors, spacer-lengths, quarter-site orientation, nucleosomes, and post-translational modifications of the p53 protein.

The Role of p53 post-translational modifications. An area of controversy is the role of post-translational modifications of p53 in determining the mode and efficacy of p53 transcriptional regulation. Experiments have shown that post-translational modifications of the p53 protein, such as phosphorylation, methylation, and acetylation, alter the stability and DNA-binding affinities of p53 [268, 148, 87, 34]. Investigators have shown that p53 needs post-translational modifications in the C-terminal domain to bind to naked DNA *in vitro*, but requires no modifications in the presence of chromatin to bind to p53 REs [64, 9]. This also coincides with experiments that showed that the deacetylated C-terminal domain inhibited binding to p53-binding sites in linear DNA and promoted binding to sites in non-linear, circularized DNA [163]. The p53-binding sites in circularized DNA segments mimic *in vivo* conditions where DNA is wrapped around histones. These experiments suggest that the C-terminal domain of the p53 protein confers DNA structure specificity (while the DNA-binding domain confers sequence specificity). In direct contradiction to these results, other investigators have shown that some of the experimentally validated p53-binding sites do not require any phosphorylation or acetylation of the p53 protein in order to confer high-affinity binding *in vitro* in the absence of chromatin [275].

Nevertheless, there are cases in which these post-translational modifications appear to play a major role. Investigators found that the induction of p53AIP1 is dependent upon the phosphorylation of the Ser46 residue of p53 [183]. Investigators also found that phosphorylation of the Ser15 and Ser392 residues conferred p53-activation of the APC gene, while un-phosphorylated p53 served as a repressor of APC [113].

The strongest evidence that supports that post-translational modifications of p53 are relevant to the p53 regulatory mechanism is the fact that HDAC inhibitors have been shown to simultaneously increase levels of acetylated p53 and induce apoptosis and senescence in cancerous and normal cells [92, 168]. HDAC inhibitors are currently in clinical trials as cancer chemotherapeutics and initial results look promising [168]. Although post-translational modifications of p53 are certainly important, the ability to properly quantify which ones are relevant, under which conditions, has been elusive. Further experimentation is needed to shed light on this complex mechanism of regulation in the p53 pathway.

The flexible CATG affect. It has been shown experimentally that in the head-to-head orientation, p53 greatly prefers the repeated RRRCATGYYY motif [89, 189]. Based upon X-ray crystallography studies of the p53 DNA-binding core domain bound to a p53-RE DNA sequence, the most critical bases for interactions with the p53 protein are the central RCWWGY, which come in close contact with the amino acids from the p53 core domain [33]. In conjunction with this, the most conserved positions after aligning all experimentally validated, functional p53 binding sites are exactly the central CWWG nucleotides within each half-site, especially the C and G (see Figure 3.3). Therefore, changes in the nucleotides in these central positions should affect binding affinity the most. Indeed, binding affinity measurements of 20 p53-binding sites revealed that 50% of the high-affinity sites contained the CATG at the center of both half-sites [275]. Investigators also found that replacing the central CATG with CTAG in both half-sites reduced transactivation 20-fold [110].

It is known that the CATG sequence element is unusually flexible and exhibits extreme bending and kinking in many DNA–protein complexes [7, 187]. Therefore, it is widely assumed that the flexibility of the p53 response element also affects binding affinities. p53-DNA binding affinity experiments have shown that p53 exhibits higher binding affinity for sites in cell cycle control target-genes than for sites in apoptotic target-genes, and that these differences coincide with the prevalence of the highly flexible CATG in both groups [275].

p53 RE sites that are not functional. Investigators have repeatedly found that p53 regulation of minimal promoters can be profoundly different from their respective full-length promoters. Examples include when experiments showed that p53 would no longer bind in the natural promoter [278], and even though experiments confirmed the presence of bound p53, the p53 RE was no longer functional in the natural promoter [41, 252]. These results indicate that the presence of co-factor sites and the p53-RE occlusion by nucleosomes or other proteins play a major role in p53 regulation. Examples of genes that contain p53-binding sites that have been shown *not* to be functional *in vivo* include: the intron 5 cluster site in AIFM2, the -328 site in TP53i3, and the promoter cluster site in human BAX [278, 41, 252]. In addition, experiments have shown that an adjacent SP1 RE is necessary to confer p53-mediated activation of the BBC3 (PUMA) and BAX genes [126, 251]. Clearly binding to DNA is not sufficient for transcription.

The affects of distance and DNA looping. It is well known that the distance between a cis-element binding site and the Transcription Start Site (TSS) can greatly affect the degree of regulation of a gene. In the case of p53, investigators showed that inserting an additional 200bp segment between a p53 RE and the TATA box eliminated a 45-fold p53-mediated induction [42]. It is also known that eukaryotic cells contain TF-binding proteins that bind together (“sticky” TF-proteins), and thereby mediate DNA looping. This process can bring distal TF-bound binding sites near the TATA box and thereby confer regulation. In the case of p53, investigators using electron microscopy techniques showed that p53 tetramers stack in register (on top of each other) when bound to a p53 RE, and thereby link distant p53-binding sites via DNA looping [237]. They also showed that distant p53-binding sites alone induced transcription poorly, but in the presence of a site proximal to the TSS, induction by the distal site is increased 25-fold [237]. p53-tetramer stacking translocates distally bound p53 protein to the promoter and increases the concentration of local p53 near the TSS.

In the absence of a proximal p53 RE, other "sticky proteins" may serve as a surrogate, provided that their response elements are present near the TSS and the distal p53 RE. An example of a proven "sticky protein" that mediates DNA looping is the known p53 co-factor SP1. An example may be found in the MDM2 gene, where a functional SNP (SNP309 T/G) within a cluster of SP1 binding sites affects the level of regulation of nearby estrogen and p53 REs, and has been associated with an early onset of breast cancer in pre-menopausal women [19]. In contradiction to this model, other investigators have hypothesized that distal sites may reduce transcription by attracting p53 proteins away from the start site of transcription [25]. For example, in the PLK2 gene the distant site is a repressor while nearer sites are activators [25]. Further investigation will be necessary to determine exactly how and when distant p53 REs regulate gene expression.

The effects of the spacers. Experiments have shown that the spacers separating the half-sites can greatly affect the binding affinity for the p53 protein. For example, Tan *et al.* showed that by mutating the spacer of a p53 RE from a GG to a T increased binding affinity 6.6-fold [246]. Two series of experiments using minimal promoter assays found a bimodal induction distribution, where the two induction peaks occurred with spacer-lengths of 0 and 10 bp [271, 42]. The authors theorized that optimum binding occurred with the half-sites aligned along the same face of the double-helix (stereospecific alignment), either with the half-sites adjacent or separated by a helical turn (10 bp). Other investigators showed that under certain experimental conditions, specific spacers with spacer lengths of size 4, 13, and 14 considerably decreased the RE's binding affinity for p53 as compared to having no spacer at all; however, a spacer length of 10 was not tested [254].

Unfortunately, only one spacer, as opposed to all possible spacers, of a certain length was tested for binding affinity in these experiments. Interestingly, our database of 160 functional p53-binding sites does not show a bimodal distribution of spacer lengths. It is possible that spacer lengths may affect binding affinity and regulatory function differently, in that high-binding affinity does not necessarily confer regulatory function. Although it is obvious that different spacers affect the function of p53 REs differently, the ability to quantify these effects has been elusive.

Rescue by p63 and p73. Yet another proposed p53-mediated activation mechanism is the rescue of weak p53-binding sites by the p53 homolog p63 and p73. Investigators have found that in mouse fibroblasts both p63 and p73 are required for p53-dependent transactivation of the NOXA and BAX genes [75].

To our knowledge, no experiments have been performed that may elucidate how these seemingly disparate determinants of p53 regulation (spacer-lengths, quarter-site orientation, co-factors, nucleosomes, and post-translational modifications of p53) may relate to each other in determining functional p53 repression and/or activation. It is obvious that our understanding of the mechanism(s) that determine p53 repression versus activation is not complete, and requires further study.

2.7 Experimental approaches and considerations

There are special considerations that need to be taken into account when attempting to experimentally validate putative p53-binding sites. Wei and his colleagues have employed chromatin immunoprecipitation with p53-specific antibodies to collect all of the tight binding sites for the p53 protein in the genome of a cancer cell line [274]. They then sequenced the DNA fragments selected for by p53 protein binding, and identified the genes in association with the p53 protein. They went on to validate, by other criteria, that a subset of these genes did indeed have a p53 RE and were regulated by p53. This has been a useful approach for identifying candidates, but it is clear from the outset that binding to an RE is not necessarily equivalent to regulating a gene. In addition, this approach requires tight binding and longer residence times of p53 at a site which could just store p53 proteins on the DNA for rapid use (not diffusion limited) at a regulated gene. Also, this method could identify p53-like sites on retroviruses and LINE elements (repetitive elements in the genome), both of which are observed by p53 RE algorithms [101].

In addition to this approach, others have employed RNA microarrays to explore the increases and decreases in the steady-state levels of RNAs in cells after the induction of p53 or exposure to a stress signal [293]. This too has been useful in identifying new p53-regulated genes, but they need to be shown to be directly regulated by p53, and not the consequence of a secondary event (such as the induction of a transcription factor by p53 that then acts upon other genes). In addition, any stress signal employed to induce p53 (such as UV exposure) may well induce the transcription of a gene by a pathway not involving p53. For example, the GADD45A gene is induced by p53 (following exposure to UV, as verified by CHIP) but is also induced by UV in a p53-null or mutant cell (by another mechanism). For these reasons, an inducible p53 gene or a temperature-sensitive p53 gene in a cell are often better employed to increase p53 levels and activity than a DNA-damaging agent. However, an inducible p53 gene may not contain the same protein modifications of the natural p53 protein that are observed in a stress response. Those modifications (acetylation, phosphorylation, methylation, ubiquitination, sumolation, etc.) could well lead to the choice of a transcriptional programme resulting from that particular stress signal (UV versus IR for example) [293]. Finally, the choice of cell lines to follow p53-regulated genes, used in these experiments for convenience, ignores the fact that there are cell-type and tissue-type specificities in the p53 response.

2.8 Conclusion

This analysis of the p53 REs, the genes they regulate and the properties they confer can add new information to our understanding of the p53 pathway and p53-mediated transcriptional control. First of all, the location of the p53 RE in a gene (Figure 2.3) is most commonly in the $5'$ promoter-enhancer region of the gene (50%) or in intron 1 (25%). More rarely it is located in introns 2 or 3 of a gene. Surprisingly, some functional p53 REs are in exon 1 or even exon 2. When this occurs, however, the p53 RE is predominately in the $5'$-UTR or the intron-exon boundary. Also, since $\approx$ 50% of the experimentally validated p53 sites are downstream of the TSS, intronic $5'$-UTR regions are equally important to promoter-enhancer regions in conferring p53 regulation. The p53 RE is commonly located near additional transcription factor RE sites.

Second, the distance of the p53 RE from the transcription start site (TSS) helps to determine the threshold for accepting any putative p53 RE based upon the normalized affinity score of the p53 protein for the known p53 REs (see Figure 2.4). Functional, low-affinity p53 RE sites in the DNA only exist around the TSS. Therefore, computational methods can employ a dynamic affinity-threshold to reduce false-positives during p53-site searches.

Third, $\approx$ 50% of the p53 RE sites have no spacer between the half sites, and the distribution of spacer lengths is relatively uniform for spacer lengths from 4 to 15 base pairs. This distribution contradicts *in vitro* experiments that would predict functional p53 RE sites based upon the half-sites being located on the same face of the DNA helix (see Figure 2.2). Interestingly, the distribution of spacer lengths in the p53 RE is different for genes that are transcriptionally activated by p53 and those that are repressed by p53 protein (see Figure 2.2). The spacer lengths in the p53 REs of repressed genes do not show a great preference for zero length or small spacers. This difference between spacer length and gene activation/repression is especially clear for those genes not involved in apoptosis of the cell (see Figure 2.5). Non-apoptotic p53-regulated genes that are repressed by p53 have no preference for zero length spacers.

The list of known p53-regulated genes collected in one place gives us a new feeling for the breadth of functions regulated in response to stress signals. After a p53-mediated response to stress, there are changes in the intracellular compartments, cytoskeleton, endosomal and exosomal functions, heat shock induction, and cellular repair processes. There are also changes in the extracellular matrix, increased secretion of exosomes and proteins that impact upon angiogenesis, growth factor functions, and the immune response. The p53 response sets up a series of positive and negative feedback loops that regulate p53-mediated functions, as well as other signal transduction pathways. In addition to these local effects of a p53 response, systemic signals are p53-regulated. Both exosomes and cytokines engage the immune response of the body. P53-mediated responses in the brain alter signal transmission in the central nervous system. Additionally, angiogenic signals and interactions with growth factors and their receptors can all have wider systemic impact. Clearly, the list of genes involved in a stress response mediated by p53 has a broad impact upon the host as well as the host cell.

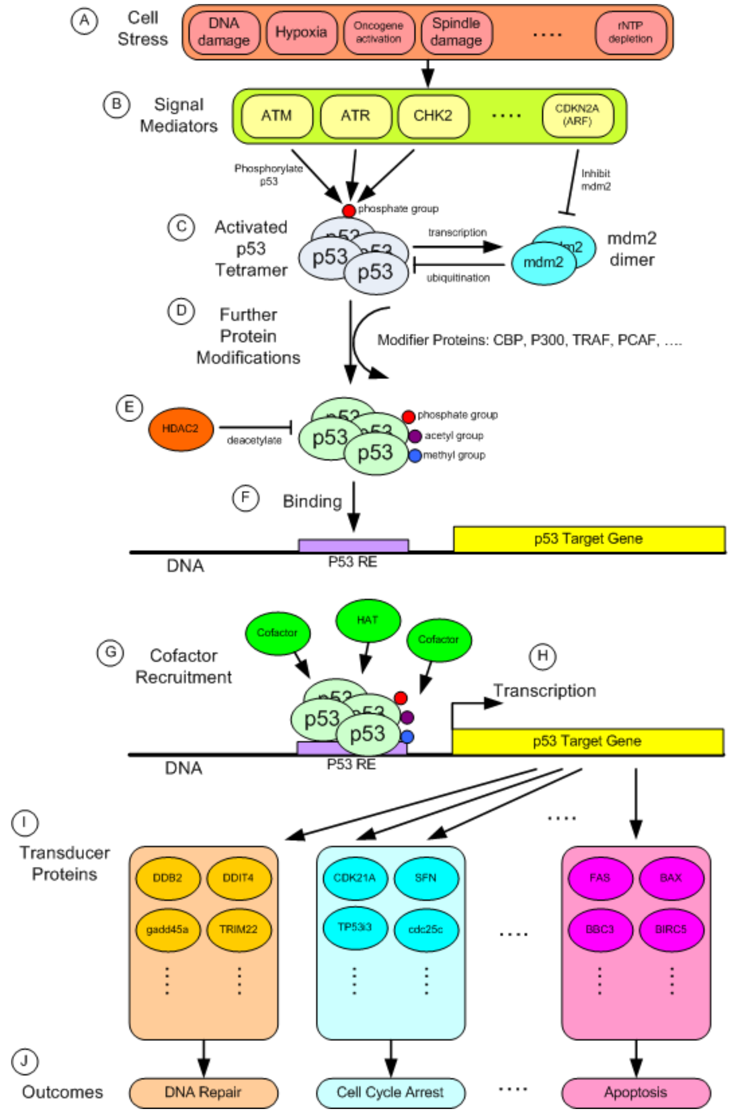


Figure 2.1: **Methods of p53 activation and regulation of downstream targets** (A) The cell undergoes stress that could lead to cancer. (B) Signal mediator proteins activate p53 by phosphorylating certain residues or inhibiting ubiquitination by MDM2. (C) Both processes increase the half-life of p53 protein (by inhibiting ubiquitination). The increased half-life, from minutes to hours, quickly produces higher p53 concentrations. (D) Further p53 protein modifications by acetyltransferases (CBP, p300, PCAF) and methyltransferases (SET9) can further stabilize the p53 protein and increase site-specific DNA binding. (E) The deacetylase HDAC2 can inhibit p53 binding to DNA by deacetylating the protein. (F) The p53 tetramer binds to a p53 response element to regulate transcription of a nearby gene. (G) p53 also recruits cofactors such as histone acetyltransferases (HATs) and TATA-binding protein associated factors (TAFs). (H) For this example, p53 mediates transactivation of its target gene. However, p53 can also mediate repression of transcription. (I) The p53 protein transactivates many genes whose protein products are involved in a variety of pathways. (J) The most important pathways involved in tumor suppression that are activated by p53 are DNA repair, cell cycle arrest, senescence, and apoptosis.

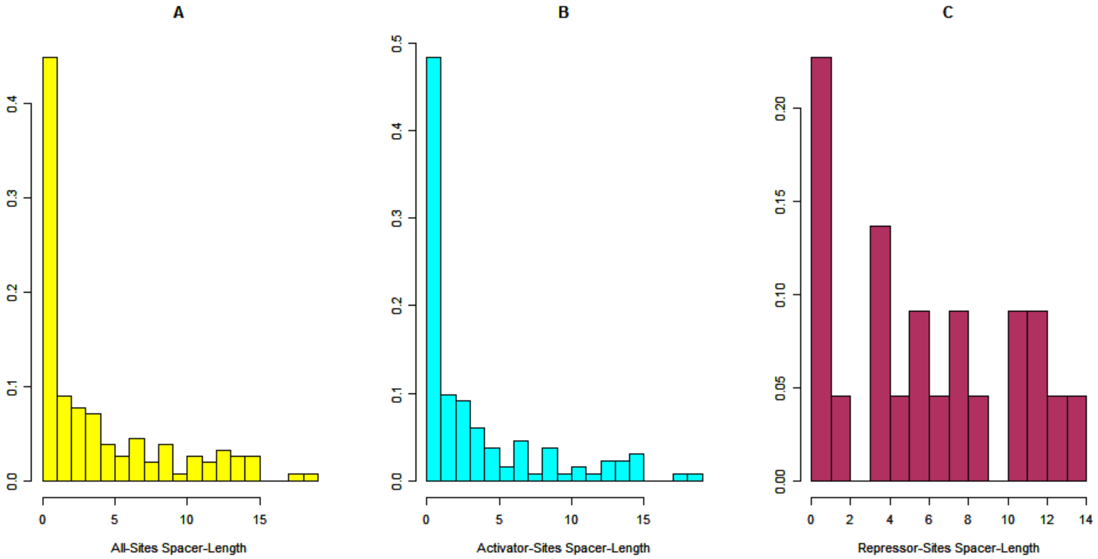


Figure 2.2: **Histograms of spacer-lengths by regulation types.** (A) The histogram of all 160 spacer-lengths of known, functional p53-binding sites reveals the following: *(1)* approximately 50% of the p53-binding sites have no spacer sequence (spacer-length = 0bp), and *(2)* the distribution is relatively uniform for spacer-lengths from 4 to 15 base pairs. This distribution does not match experimental results which would suggest a bimodal distribution with peaks at 0 and 10 base pairs, which would place the two half-sites on the same face of the DNA double-helix [271, 42]. (B,C) Repression-sites have a different distribution of spacer-lengths compared to activation-sites. Most importantly, repression-sites do not show a great preference for 0bp spacers.

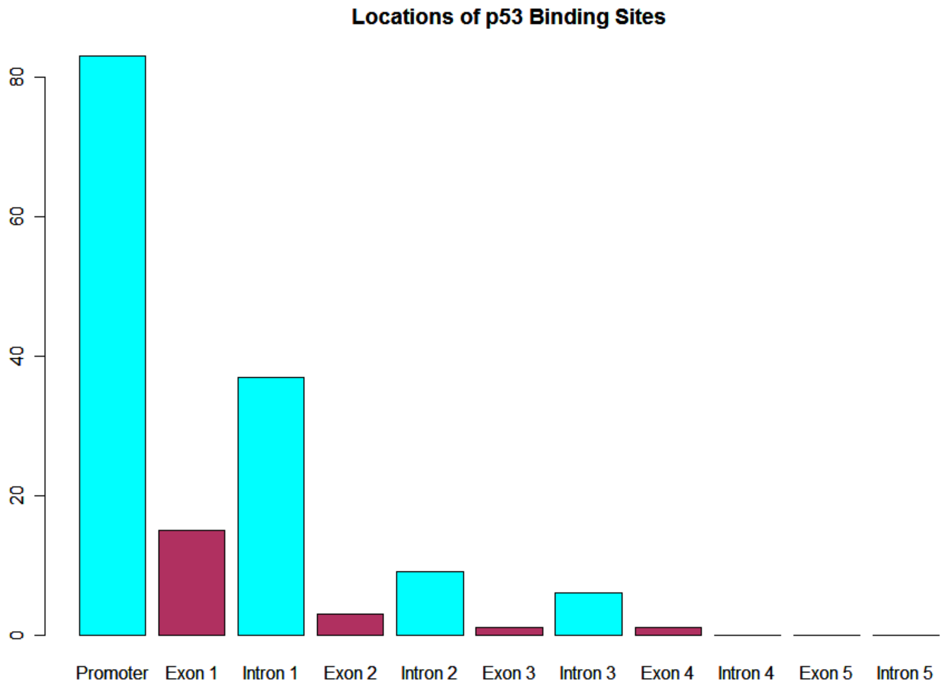


Figure 2.3: **Histogram of p53-binding sites by gene region.** The histogram of 160 functional p53-binding sites (by gene region) reveals the following: *(1)* there are slightly more p53 REs upstream of the TSS than downstream (83 of the 160 sites are completely in the promoter region, 3 straddle the TSS, and 74 are completely downstream of the TSS), *(2)* there are significantly more p53 REs in non-coding regions (cyan) than in coding regions (purple), and *(3)* there is an exponential decay of p53 REs as the distance from the TSS increases. 13 of the 15 Exon 1 REs (87%) are in the 5′-UTR region. (Note: some p53 REs straddle both coding and non-coding regions and are counted twice.)

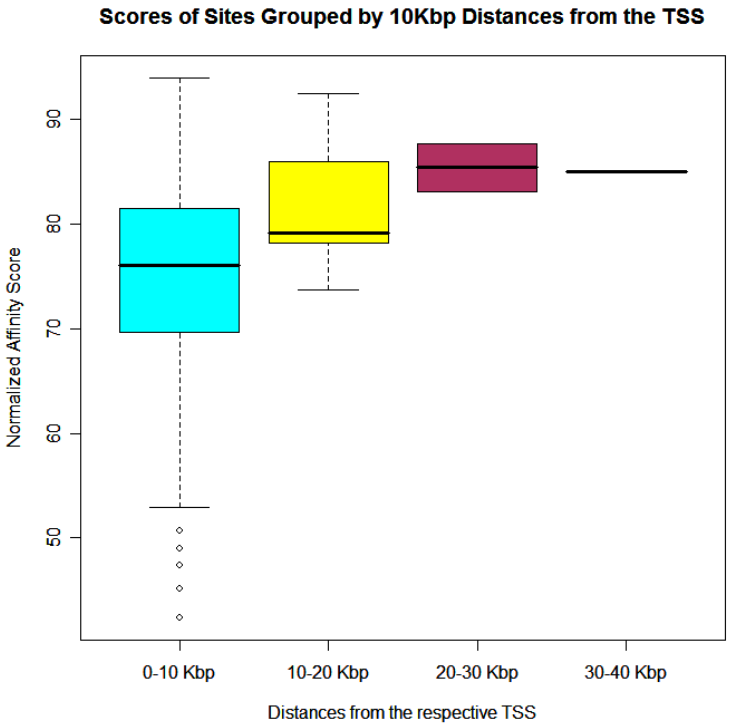
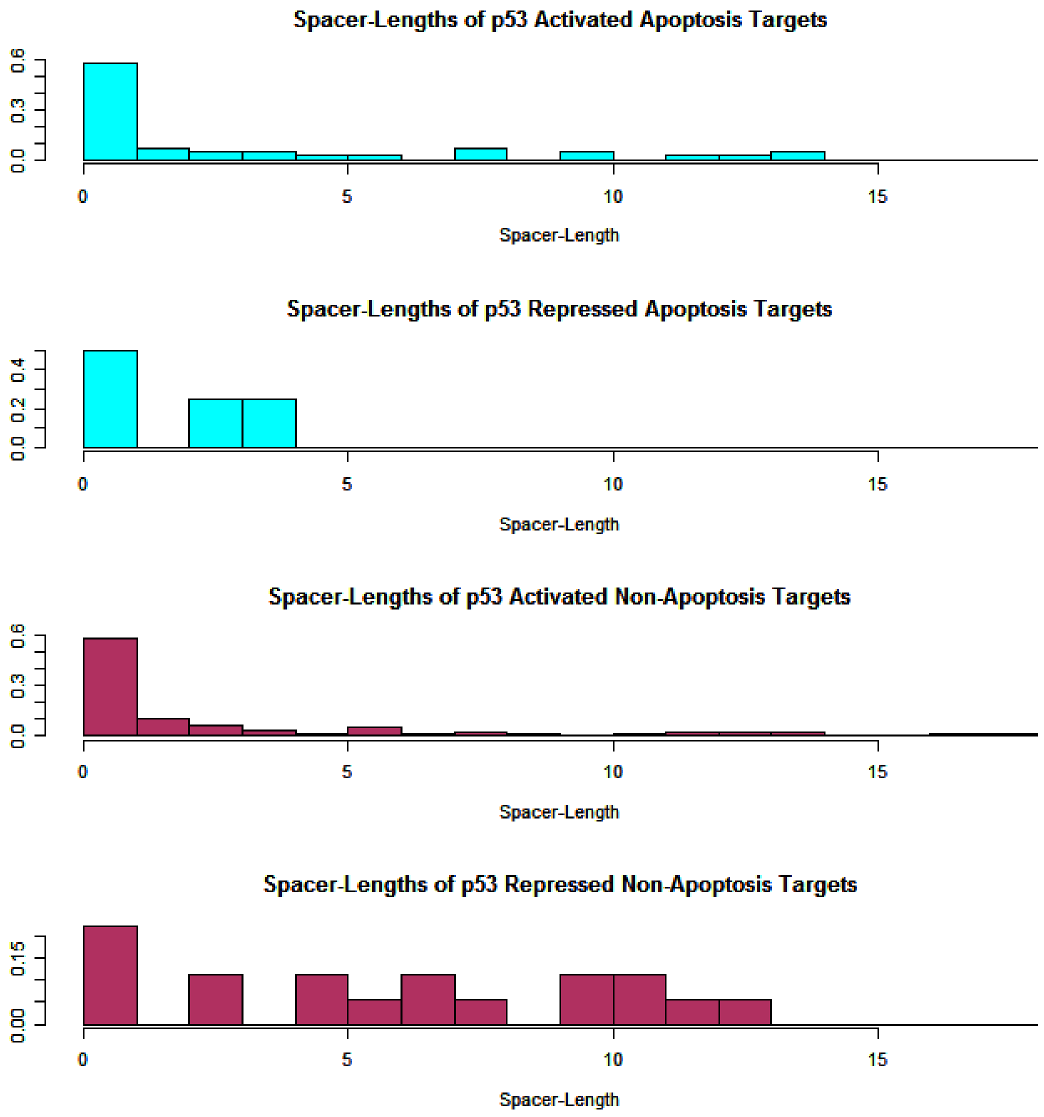


Figure 2.4: **Box plots of normalized affinity scores by 10Kb block distances from the TSS.** All low affinity sites are within the 1^{st} 10Kb block from the TSS. Median scores of the 10Kb blocks rise as a function of distance.



Spacer-Lengths of p53 Activated Apoptosis Targets

Spacer-Lengths of p53 Repressed Apoptosis Targets

Spacer-Lengths of p53 Activated Non-Apoptosis Targets

Spacer-Lengths of p53 Repressed Non-Apoptosis Targets

Figure 2.5: **Histogram of spacer-lengths by regulation type and gene-target function.** Non-apoptotic-target sites (in purple) have a higher frequency of repressor-sites compared to apoptotic-target sites (16.5% versus 8.5%). In addition, non-apoptotic-target sites have no preference for 0bp-length spacers (bottom histogram). Thus, p53-repressor binding sites have significantly longer spacers of average.

Gene Name(s)	Short Description	Half-site #
BTG2, TIS21	BTG family, member 2	4
CDKN1A, p21	cyclin-dependent kinase inhibitor 1A (p21, Cip1)	2.5
DDB2	damage-specific DNA binding protein 2, 48kDa	4
GML	GPI anchored molecule like protein	3
HRAS, c-Ha-Ras	Harvey rat sarcoma viral oncogene homolog	8
IGFBP3	insulin-like growth factor binding protein 3	11
mdm2	Mdm2, transformed 3T3 cell double minute 2	4
PCNA	proliferating cell nuclear antigen	5
SH2D1A, SAP	SH2 domain protein 1A, Duncans disease	4
TP53i3, Pig3	tumor protein p53 inducible protein 3	7.5
TP73, p73	tumor protein p73	3
TRPM2	transient receptor potential cation channel, M2	3
TYRP1, TRP-1	tyrosinase-related protein 1	6
VDR	vitamin D (1,25- dihydroxyvitamin D3) receptor	3
HBV	hepatitis B virus	3

Table 2.2: **cluster sites regulated by p53.** This table lists genes that contain cluster-site REs that have been shown experimentally to confer transcriptional regulation by p53. A cluster-site RE is defined as any RE which contains three or more half-sites, each separated by no more than 15 base-pairs.

Table 2.3: **Description of Genes Regulated by p53 I.** This table provides the genes names, the accession numbers, and the DNA response-elements (RE's) of experimentally validated p53-regulated genes. The RE's typically consist of two half-sites separated by a variable length spacer. Exceptional cases consist of only one apparent half-site. The RE's consisting of many (4+) half-sites are annotated as "Large cluster sites", and as such, are too large to include in the space provided.

#	Gene Name(s)	Short Description	Accession #	1st Half-site	Spacer	2nd Half-site	Reference
1	ABCB1, MDR1	ATP-binding cassette sub-family B member 1	AF016535	GGGCAGGAACA	gcgccggggcgt	GGGCTGAGCA	[115]
2	ACTA2	smooth muscle alpha-actin	NM_001613	AACCATGCCT		GCATCTGCCC	[269]
3	AIFM2, AMID	apoptosis-inducing factor, mitochondrion-assoc.	NM_032797	AGGCATGAGC	caccgtgcct	GGCCATGCCC	[278]
3	AIFM2, AMID	apoptosis-inducing factor, mitochondrion-assoc.	NM_032797	AGGTCTCGCTA	tgttgccc	AGGCTGGTCT	[278]
4	ANLN	anillin, actin binding protein	NM_018685	GAACTGGCTT	ttctga	GGGCCAGGCC	[169]
5	APAF1	apoptotic peptidase activating factor 1	NM_001160	AGACATGTCT	ggagaccctagga	CGACAAGCCC	[205]
6	APC	adenomatosis polyposis coli	NM_000038	GGGCATACCC	ccgaggggtacg	GGGCTAGGGCt	[113]
7	ARID3A, E2FBP1	AT rich interactive domain 3A (BRIGHT-like)	NM_005224	GGACACGCTG		GGACATGCCT	[150]
8	ATF3	activating transcription factor 3	NM_001674	AGTCATGCCG	ctggcttgggcaccatt	GGTCATGCCT	[290]
9	BAI1	brain-specific angiogenesis inhibitor 1	NM_001702	tGGCTGCCT		GGACATGTTC	[233]
10	BAX	BCL2-associated X protein	NM_004324	GGGCAGGCCC		GGGCTTGTCG	[252]
11	BBC3, PUMA	BCL2 binding component 3	NM_014417	CTGCAAGTCC		TGACTTGTCC	[179]
12	BCL2L14, BCL-G	BCL2-like 14 (apoptosis facilitator)	NM_030766	AGCCAAGGCT		GGTCTTGAAC	[167]
13	BCL6	B-cell CLL/lymphoma 6 (zinc finger protein 51)	NM_001706	AGACAGTGCTT	ggggggtgattc	GGGCTAGTCT	[153]
14	BDKRB2, BK2	bradykinin receptor B2	NM_000623	GGAagTGCCC		AGGaggcTga	[210]
15	BID	BH3 interacting domain death agonist	NM_197966	GGGCATGATG		GTGCATGCCT	[216]
16	BIRC5, survivin	baculoviral IAP repeat-containing 5 (survivin)	NM_001168	GGGCGTGCGC	tcc	CGACATGCCC	[100]
17	BNIP3L	BCL2/adenovirus E1B interacting protein 3-like	NM_004331	AAGCTAGTCT	cagtg	GcGCATGCCT	[69]

Continued on next page

#	Gene Name(s)	Short Description	Accession #	1st Half-site	Spacer	2nd Half-site	Reference
18	BTG2, TIS21	BTG family, member 2	NM_006763	AGTCCGGGCA	g	AGCCCGAGCA	[54]
19	C12orf5	chromosome 12 open reading frame 5	NM_020375	AGACATGTCC	ac	AGACTTGTCT	[114]
20	C13orf15, RGC32	chromosome 13 open reading frame 15	NM_014059	AGGCgAGTTT	aag	cAGCTTGTCC	[211]
21	CASP1	caspase 1, apoptosis-related cysteine peptidase	NM_033292	AGACATGCAT		ATGCATGCAca	[88]
22	CASP10	caspase 10, apoptosis-related cys-peptidase	NM_032977	AAACTTGCTg	gttta	AAtCTTGgCT	[201]
23	CASP6	caspase 6, apoptosis-related cysteine peptidase	NM_001226	AGGCAAGGAG	tttg	AGACAAGTCT	[151]
24	CAV1	caveolin 1, caveolae protein, 22kDa	NM_001753	GCCCAAGCAC	cccagcgcg	GGAGAaACGTTC	[17]
25	CCNG1	cyclin G1	NM_004060	GcACAAGCCC		AGGCTAGTCC	[62]
26	CCNK	cyclin K	NM_003858	AAACTAGCTT	gc	AGACATGCTg	[242]
27	CD82, KAI1	CD82 molecule	NM_002231	AGGCAAGCT	ggggca	GctCAAGCCT	[158]
28	CDC25C	cell division cycle 25 homolog C (S. pombe)	NM_001790	GGGCAAGTCT	taccatttcca	GAGCAAGCaC	[36]
29	CDKN1A, p21	cyclin-dependent kinase inhibitor 1A (p21, Cip1)	NM_000389	AGACTGGGCA		TGTCTGGGCA	[58]
29	CDKN1A, p21	cyclin-dependent kinase inhibitor 1A (p21, Cip1)	NM_000389	GAAgAAGaCT		GGGCATGTCT	[215]
29	CDKN1A, p21	cyclin-dependent kinase inhibitor 1A (p21, Cip1)	NM_000389	GAACATGTCC		cAACATGTTg	[215]
30	Chmp4C	chromatin modifying protein 4C	NM_152284	AAACAAGCCC	agtagcagcagctgctcc	GAGCTTGCCC	[287]
31	COL18A1	collagen, type XVIII, alpha 1	NM_030582	TGACATGTGT		GAGCATGTAT	[167]
31	COL18A1	collagen, type XVIII, alpha 1	NM_030582	TGACATGTGT		GAGCATGTAT	[167]
32	CRYZ	crystallin, zeta (quinone reductase)	NM_001889	ctGCAAGTCC	att	AAACcTGTTT	[169]
33	CTSD, IRDD	cathepsin D	NM_001909	AAcCTTGgTT		tgcAAgAgGCTT	[277]
33	CTSD, IRDD	cathepsin D	NM_001909	AAGCTgGgCC		GGGCTgaCCC	[277]
34	CX3CL1, fractalkine	chemokine (C-X3-C motif) ligand 1	NM_002996	GGGCATGTTC	c	CAGCTTGTGG	[225]
35	DDB2	damage-specific DNA binding protein 2, 48kDa	NM_000107	GAACAAGCCC	t	GGGCATGTTT	[246]

Continued on next page

#	Gene Name(s)	Short Description	Accession #	1st Half-site	Spacer	2nd Half-site	Reference
36	DDIT4, REDD1	DNA-damage-inducible transcript 4	NM_019058	AAACAAGTCT		TTCCTTGATC	[61]
37	DDR1	discoidin domain receptor family, member 1	NM_013994	GAGCTGGTCC		AGGCTTATCT	[212]
38	DKK1	dickkopf homolog 1 (Xenopus laevis)	NM_012242	AGCCAAGCTT	ttaatg	AACCAAGTTC	[226]
39	DNMT1	DNA (cytosine-5-)-methyltransferase 1	NM_001379	GCGCATGCGT	gttccct	GGGCATGGCC	[193]
40	DUSP1, MKP1	dual specificity phosphatase 1	NM_004417	GGTCCTGCCC	a	GGCAAATGGG	[139]
41	DUSP5	dual specificity phosphatase 5	NM_004419	CAACAAGCCC	t	TGTCTAGTGC	[261]
42	EDN2	endothelin 2	NM_001956	CTGCAAGCCC		GGGCATGCCC	[94]
43	EEF1A1	eukaryotic translation elongation factor 1 alpha 1	NM_001402	GGGCAGACCC	ga	GAGCATGCCC	[119]
43	EEF1A1	eukaryotic translation elongation factor 1 alpha 1	NM_001402	GGACACGTAG	attc	GGGCAAGTCC	[119]
43	EEF1A1	eukaryotic translation elongation factor 1 alpha 1	NM_001402	AAACATGATT	ac	AGGGACATCT	[119]
44	EGFR	epidermal growth factor receptor	NM_005228	GAGCTAGACG	tcc	GGGCAGCCCC	[147]
45	EphA2	EPH receptor A2	NM_004431	CACCATGTTG	gcc	AGGCATGTCT	[50]
46	FANCC, FAC	Fanconi anemia, complementation group C	NM_000136	GGACATGTTT	aaatacttga	GAGCTATTTT	[140]
47	FAS, CD95	Fas (TNF receptor superfamily, member 6)	NM_000043	GGACAAGCCC		TGACAAGCCA	[170]
48	FDXR	ferredoxin reductase	NM_024417	GGGCAgGagC		GGGCTTGCCC	[142]
49	GADD45A	growth arrest and DNA-damage-inducible, alpha	NM_001924	GAACATGTCT		AAGCATGCTG	[232]
50	GDF15, MIC-1	growth differentiation factor 15	NM_004864	AGCCATGCCC		GGGCAAGAAC	[245]
50	GDF15, MIC-1	growth differentiation factor 15	NM_004864	CATCTTGCCC		AGACTTGTCT	[118]
51	GML	GPI anchored molecule like protein	NM_002066	AtGCTTGCCC		AGGCATGTCC	[125]
52	GPX1	glutathione peroxidase 1	NM_000581	GGGCCAGACC		AGACATGCCT	[114]
53	HBV	hepatitis B virus	get_this	TTGCATGTAT	acaagct	AAACAGGCTT	[188]
54	HD, Huntington	huntingtin (Huntington disease)	NM_002111	ATGCTTGTTC	tacagaa	GAGCATGTTA	[70]

Continued on next page

#	Gene Name(s)	Short Description	Accession #	1st Half-site	Spacer	2nd Half-site	Reference
54	HD, Huntington	huntingtin (Huntington disease)	NM_002111	CGCCATGTTG	gcc	AGGCTGGTCT	[70]
54	HD, Huntington	huntingtin (Huntington disease)	NM_002111	GGGCCTGCTT	ccagtt	AAGCTTGCTT	[70]
55	HGF, SF	hepatocyte growth factor	NM_000601	ACACATGTAT		TTTCCTGTTT	[164]
56	HIC1	hypermethylated in cancer 1	NM_006497	GGGCGCTGCCC		TGGCACAGCTC	[22]
57	HRAS, c-Ha-Ras	Harvey rat sarcoma viral oncogene homolog	NM_176795	large	cluster	site	[46]
58	HSP90AB1, hsp90beta	heat shock protein 90kDa alpha B 1	NM_007355	GGGACTGTCT	gggtatcgga	AAGCAAGCCT	[292]
59	HSPA8	heat shock 70kDa protein 8	NM_006597	GcACTAGTTC	tggacctc	GcGCgTGCTT	[169]
60	IBRDC2, p53RFP	IBR domain containing 2	NM_182757	AGACAGGTCC		TGACAAGCAG	[180]
61	IER3, IEX-1	immediate early response 3	NM_003897	GCCACATGCCT		CGACATGTGCC	[108]
62	IGFBP3	insulin-like growth factor binding protein 3	NM_000598	large	cluster	site	[23]
62	IGFBP3	insulin-like growth factor binding protein 3	NM_000598	GGGCAAGACC		TGCCAAGCCT	[23]
62	IGFBP3	insulin-like growth factor binding protein 3	NM_000598	AAACAAGCCA	c	CAACATGCTT	[23]
63	IRF5	interferon regulatory factor 5	NM_032643	AGGCATGCCa	ca	AGGCATGgTC	[171]
64	KRT8, CK8	keratin 8	NM_002273	ccGCcTGCCT	cc	ActCcTGCCT	[175]
65	LGALS3, galectin-3	lectin, galactoside-binding, soluble, 3	NM_002306	GGGCTTGCAA	gctg	GAGCCTTGTTT	[197]
66	LIF	leukemia inhibitory factor	NM_002309	GGACATGTCG		GGACAGCTC	[106]
67	LRDD, PIDD	leucine-rich repeats and death domain containing	NM_018494	AGGCcTGCCT	gcgtgctg	GGACATGTCT	[141]
68	MAD1L1, MAD1	MAD1 mitotic arrest deficient-like 1 (yeast)	NM_003550	GATTCAAGCTG		ATACTGAGT	[35]
69	mdm2	Mdm2, transformed 3T3 cell double minute 2	NM_002392	AGTTAAGTCC		TGACTTGTCT	[289]
69	mdm2	Mdm2, transformed 3T3 cell double minute 2	NM_002392	GGTCAAGTTC		AGACACGTTC	[289]
70	MET	met proto-oncogene	NM_000245	ggacggacag	cacgcgaggcagac	AGACAcGTgC	[219]
71	MLH1	mutL homolog 1, colon cancer	NM_000249	AGGCATGTAC	a	GCGCATGCCC	[31]

Continued on next page

#	Gene Name(s)	Short Description	Accession #	1st Half-site	Spacer	2nd Half-site	Reference
72	MMP2	matrix metallopeptidase 2	NM_004530	AGACAAGCCT		GAACTTGTCT	[15]
73	MSH2	mutS homolog 2	NM_000251	GAcCTAGgCg	c	AGGCATGCgC	[272]
73	MSH2	mutS homolog 2	NM_000251	AGGCTAGTTT	tttttttgttttc	AAGTTTCCTT	[217]
74	NDRG1	N-myc downstream regulated gene 1	NM_006096	CCACATGCAC	acgcacgagcgc	GCACATGAAC	[236]
75	NLRC4, Ipaf	NLR family, CARD domain containing 4	NM_021209	AGACATGTTC		CTGGTAGTTT	[209]
76	NOS3	nitric oxide synthase 3 (endothelial cell)	NM_000603	GAGCcTcCCa	gcc	GGGCTTGTTC	[173]
77	ODC1	ornithine decarboxylase 1	NM_002539	GGACcAGTTC	caggc	GGGCgAGaCC	[169]
77	ODC1	ornithine decarboxylase 2	NM_002539	GGGCTcGCCT	tggtacagac	GAGCggGCCC	[169]
78	P2RXL1	purinergic receptor P2X-like 1, orphan receptor	NM_005446	GAACAAGggC	at	GAGCTTGTCT	[262]
79	P53AIP1	p53-regulated apoptosis-inducing protein 1	NM_022112	TCTCTTGCCC		GGGCTTGTCG	[183]
80	PCBP4, MCG10	poly(rC) binding protein 4	NM_020418	GgtCTTGgCCC		AGACTTAGCaC	[295]
80	PCBP4, MCG10	poly(rC) binding protein 4	NM_020418	GAACTT	aagaccgaggctct	GGACAAGTT	[295]
81	PCNA	proliferating cell nuclear antigen	NM_002592	GAACAAGTCC		GGGCATaTgT	[172]
82	PERP	PERP, TP53 apoptosis effector	NM_022121	AGGCAAGCTC		CAGCTTGTTC	[199]
83	PLAGL1, ZAC	pleiomorphic adenoma gene-like 1	BC074814	CAACTAGACT		AGACTAGCTT	[208]
84	PLK2, SNK	polo-like kinase 2 (Drosophila)	NM_006622	AGACATGgTg	tgt	AAACTAGCTT	[25]
84	PLK2, SNK	polo-like kinase 2 (Drosophila)	NM_006622	GGtCATGaTT	cct	tAACTTGCCT	[25]
84	PLK2, SNK	polo-like kinase 2 (Drosophila)	NM_006622	AAACATGCCT		GGACTTGCCC	[25]
85	PLK3	polo-like kinase 3 (Drosophila)	NM_004073	TAACATGCCC	gggcaa	AAGCGAGCGC	[114]
86	PML	promyelocytic leukemia	NM_002675	GcGCTgGCCT	ggagccag	GGGCATGTCC	[45]
87	PMS2	PMS2 postmeiotic segregation increased 2	NM_000535	ATACTTGATT	tg	TTTCTTGTAA	[31]
88	PPM1J, MGC19531	protein phosphatase 1J (PP2C domain containing)	NM_005167	GAACATGCCT		GAGCAAGCCC	[94]

Continued on next page

#	Gene Name(s)	Short Description	Accession #	1st Half-site	Spacer	2nd Half-site	Reference
89	PRDM1, BLIMP1	PR domain containing 1, with ZNF domain	NM_182907	GTGCAAGTCT		GGACATGTTT	[280]
90	PRKAB1, AMPKbeta1	protein kinase, AMP-activated, beta 1	NM_006253	GTTCTTGCCG		CGGCTTGCCT	[114]
91	PTEN	phosphatase and tensin homolog	NM_000314	GAGCAAGCCC	caggcagctacact	GGGCATGCTC	[234]
92	PTK2, FAK	PTK2 protein tyrosine kinase 2	NM_153831	AAGCAAGCC		no 2nd site	[83]
93	PYCARD, ASC	PYD and CARD domain containing	NM_013258	GTGCAAGCCC	ag	AGACAAGCAG	[185]
94	RABGGTA	Rab geranylgeranyltransferase, alpha subunit	NM_004581	CCTCTTGTGG	aacgtgca	AAGCCTGTCC	[114]
95	RB1	retinoblastoma 1 (including osteosarcoma)	NM_000321	GGGCGTGCCC	cgac	GTGCgcGCgC	[223]
96	RFWD2, COP1	ring finger and WD repeat domain	NM_022457	AGACTTGCCT	gt	GAACAGTCAC	[51]
97	RPS27L	ribosomal protein S27-like	NM_015920	GGGCATGTAG		TGACTTGCCC	[94]
98	RRM2B, p53R2	ribonucleotide reductase M2 B	NM_015713	tGACATGCCC		AGGCATGTCT	[247]
99	S100A2	S100 calcium binding protein A2	NM_005978	GGGCATGTgT		GGGCAcGTTC	[244]
100	SCARA3, CSR1	scavenger receptor class A, member 3	NM_016240	GGGCAAGCCC		AGACAAGTTg	[90]
101	SCD	stearoyl-CoA desaturase (delta-9-desaturase)	NM_005063	GGGCcgGTCC	t	GGGCTAGgCT	[169]
102	SCN3B	sodium channel, voltage-gated, type III, beta	NM_018400	TGACTTGCTC		TGCCTTGCCT	[1]
102	SCN3B	sodium channel, voltage-gated, type III, beta	NM_018400	TGGCAAGGCT		GAGCTAGTTC	[1]
103	SERPINB5, maspin	serpin peptidase inhibitor, clade B, member 5	NM_002639	GAACATGTTg	g	AGGCcTtTTg	[296]
104	SERPINE1	serpin peptidase inhibitor, clade E, member 1	NM_000602	AcACATGCCT		cAGCAAGTCC	[130]
105	SESN1, PA26	sestrin 1	AF033120	GGACAAGTCT		CCACAAGTCa	[266]
106	SFN, 14-3-3sigma	stratifin	NM_006142	AGCATTAGCCC		AGACATGTCC	[96]
107	SH2D1A, SAP	SH2 domain protein 1A, Duncans disease	NM_002351	GGCTGGCTC	agctgt	CAGCTTGCTT	[177]
107	SH2D1A, SAP	SH2 domain protein 1A, Duncans disease	NM_002351	GGGCTGGCTC		GGCTGGCTC	[177]
107	SH2D1A, SAP	SH2 domain protein 1A, Duncans disease	NM_002351	CAACACTGCAC	tagt	GGGCTGGCTC	[177]

Continued on next page

#	Gene Name(s)	Short Description	Accession #	1st Half-site	Spacer	2nd Half-site	Reference
108	SLC38A2	solute carrier family 38, member 2	NM_018976	AAcCATGCTg	ttacacgcacc	AGCTTGTCC	[169]
109	STEAP3, TSAP6	STEAP family member 3	NM_001008410	AGACAAGCAT	ag	GGACATGCTC	[191]
110	TAP1	transporter 1, ATP-binding cassette	NM_000593	GGGCTTGgCC	ctgccg	GGACTTGCCT	[98]
111	TGFA	transforming growth factor, alpha	NM_003236	GGGCAGGCCC		TGCCTAGTCT	[224]
112	TNFRSF10A, DR4	tumor necrosis factor receptor superfamily, 10a	NM_003844	GGGCATGTCC		GGGCAgGagg	[145]
113	TNFRSF10B, DR5	tumor necrosis factor receptor superfamily, 10b	NM_003842	GGGCATGTCC		GGGCAAGaCg	[243]
114	TNFRSF10C, DcR1	tumor necrosis factor receptor superfamily, 10c	NM_003841	GGGCATGTCC		GGGCAGGACG	[146]
115	TNFRSF10D, DcR2	tumor necrosis factor receptor superfamily, 10d	NM_003840	GGGCATGTCT		GGGCAGGACG	[146]
116	TP53, p53	tumor protein p53 (Li-Fraumeni syndrome)	NM_000546	TTACTTGCCC		TTACTTGTCA	[12]
117	TP53i3, Pig3	tumor protein p53 inducible protein 3	NM_004881	large	cluster	site	[41]
118	TP53INP1	tumor protein p53 inducible nuclear protein 1	NM_033285	GAACTTGggg		GAACATGTTT	[186]
119	TP63, TP73L	tumor protein p63, p73-like Delta N variant	AF075433	TAACTTGTTA	ttg	AAACATGCTC	[91]
120	TP73, p73	tumor protein p73	NM_005427	GtACTTGCCg	tccgggga	GAACTTGCag	[32]
120	TP73, p73	tumor protein p73	NM_005427	GAACTTGCag	agtaagctgga	GAGCTTGaaT	[32]
121	Tp73:Delta	tumor protein p73 Delta N variant	AY040827	GGGCAAGCT	gaggcctgcccc	GGACTTGGAT	[178]
122	TRIAP1, p53CSV	TP53 regulated inhibitor of apoptosis 1	NM_016399	CTTCATGTCC		GTGCATGCCT	[190]
123	TRIM22, Staf50	tripartite motif-containing 22	NM_006074	TGACATGTCT		AGGCATGTAG	[182]
124	TRPM2	transient receptor potential cation channel, M2	NM_003307	GGCCTTGCCT	tgctc	AGGCCTGCTT	[169]
124	TRPM2	transient receptor potential cation channel, M2	NM_003307	GAGCAGGTCT	gacctgcttccca	GGGCCTGCTT	[169]
124	TRPM2	transient receptor potential cation channel, M2	NM_003307	TGCCTTGCTC		AGGCCTGCTT	[169]
125	TSC2	tuberous sclerosis 2	NM_000548	TAACAAGCTC	g	GGGCTAGCCC	[71]
125	TSC2	tuberous sclerosis 2	NM_000548	AGGCTAGTCT	gaaactcctgggc	TGACGTGAC	[71]

Continued on next page

#	Gene Name(s)	Short Description	Accession #	1st Half-site	Spacer	2nd Half-site	Reference
125	TSC2	tuberous sclerosis 2	NM_000548	GGGCATGGTG		GCACATGCCT	[71]
126	TYRP1, TRP-1	tyrosinase-related protein 1	NM_000550	CGCCTAGTTT	gggt	GAGCAGATT	[181]
126	TYRP1, TRP-1	tyrosinase-related protein 1	NM_000550	GAGCAGATT	tgggattaattatc	AGGCAGCAA	[181]
126	TYRP1, TRP-1	tyrosinase-related protein 1	NM_000550	CCACATGCAC	t	TAACAGTTC	[181]
126	TYRP1, TRP-1	tyrosinase-related protein 1	NM_000550	AGACCAGCCC	cc	CGCCTAGTTT	[181]
126	TYRP1, TRP-1	tyrosinase-related protein 1	NM_000550	AGGCAGCAA	t	CCACATGCAC	[181]
127	UBD, FAT10	ubiquitin D	NM_006398	AGGCATGCTC		AGTGGCGTGG	[291]
128	VCAN, CSPG2	versican	NM_004385	AGACTTGCC	a	CAGACAAGTCC	[285]
129	VDR	vitamin D (1,25- dihydroxyvitamin D3) receptor	NM_000376	TAACTAGTTT		GAACAAGTTG	[157]
129	VDR	vitamin D (1,25- dihydroxyvitamin D3) receptor	NM_000376	AGGTTAGATG	tac	TAACTAGTTT	[157]

Table 2.4: **Description of Genes Regulated by p53 II.** This table provides additional information on the gene set found in Table 2.3. This table provides the relative locations in the gene, the distances to the transcription start site (TSS), the gene functions, and the references. Functions are as follows: A = apoptosis, C = cell cycle control, S = senescence, CytoS = cytoskeleton, E = endosome and exosome compartment, ECM = extra-cellular matrix, F+/- = positive/negative feedback loops for p53, R+/- = regulation by p53 upon other signal transduction pathways, T = transcription and translation, DNA-R = DNA repair, CytoR = cytokine and inflammatory regulator, CNS = central nervous system regulator, GR = growth factor regulator, HSP = heat shock protein.

#	Gene Name(s)	Location	Type	bp from TSS	Function	PHMM Score	Reference
1	ABCB1, MDR1	Promoter	repressor	-89 to -57bp		63.34	[115]
2	ACTA2	Promoter	activator	-330 to -311bp	CytoS	67.61	[269]
3	AIFM2, AMID	Promoter	activator	-596 to -567bp	A	81.34	[278]
3	AIFM2, AMID	Promoter	activator	-682 to -654bp	A	66.15	[278]
4	ANLN	Promoter	repressor	-866 to -841bp	CytoS	73.89	[169]
5	APAF1	Promoter	activator	-603 to -571bp	A	90.33	[205]
6	APC	Promoter	both	-230 to -198bp	R-	74.23	[113]
7	ARID3A, E2FBP1	Intron 2	activator	4240 to 4259bp	C	83.53	[150]
8	ATF3	Promoter	activator	-388 to -352bp	T	80.67	[290]
9	BAI1	Intron 9	activator	17444 to 17462bp	ECM	79.09	[233]
10	BAX	Intron 1	activator	354 to 373bp	A	87.61	[252]
11	BBC3, PUMA	Promoter	activator	-145 to -126bp	A	79.34	[179]
12	BCL2L14, BCL-G	5-UTR, Intron 1	activator	1612 to 1631bp	A	64.87	[167]
13	BCL6	5-UTR, Intron 1	activator	696 to 728bp	-F	78.49	[153]
14	BDKRB2, BK2	Promoter	activator	-86 to -67bp	CytoR	45.12	[210]
15	BID	Intron 1	activator	17277 to 17296bp	A	80.44	[216]
16	BIRC5, survivin	5-UTR, Exon 1	repressor	34 to 56bp	A	81.44	[100]

Continued on next page

#	Gene Name(s)	Location	Type	bp from TSS	Function	PHMM Score	Reference
17	BNIP3L	Downstream (4476)	activator	34574 to 34598bp	A	85	[69]
18	BTG2, TIS21	Promoter	activator	-25 to -5bp	C, DNA-R	42.33	[54]
19	C12orf5	Intron 1	activator	411 to 432bp	?	93.39	[114]
20	C13orf15, RGC32	Intron 2	activator	1116 to 1138bp	C	76.75	[211]
21	CASP1	Promoter	activator	-99 to -79bp	CytoR	79.94	[88]
22	CASP10	Promoter	activator	-1082 to -1058bp		70.01	[201]
23	CASP6	Intron 3	activator	5974 to 5997bp	A	76.81	[151]
24	CAV1	Promoter, 5-UTR, Exon 1	activator	-17 to 13bp	E,C	52.91	[17]
25	CCNG1	5-UTR, Intron 1	activator	356 to 375bp	C	86.92	[62]
26	CCNK	5-UTR, Intron 1	activator	2887 to 2908bp	C	81.53	[242]
27	CD82, KAI1	Promoter	activator	-886 to -862bp	ECM	77.17	[158]
28	CDC25C	Promoter	repressor	-155 to -125bp	C	87.3	[36]
29	CDKN1A, p21	Promoter	activator	-1373 to -1354bp	C, S	49	[58]
29	CDKN1A, p21	Promoter	activator	-1378 to -1359bp	C, S	70.59	[215]
29	CDKN1A, p21	Promoter	activator	-2260 to -2241bp	C, S	82.92	[215]
30	Chmp4C	Promoter	activator	-497 to -460bp	E	90.68	[287]
31	COL18A1	Promoter	activator	-2836 to -2817bp	ECM	77.62	[167]
31	COL18A1	Promoter	activator	-2360 to -2341bp	ECM	77.62	[167]
32	CRYZ	5-UTR, Intron 1	repressor	7721 to 7743bp	?	74.99	[169]
33	CTSD, IRDD	Promoter	activator	-373 to -352bp	A	53.25	[277]
33	CTSD, IRDD	Promoter	activator	-144 to -125bp	A	66.26	[277]
34	CX3CL1, fractalkine	Promoter	activator	-279 to -259bp	CytoR	78.38	[225]

Continued on next page

#	Gene Name(s)	Location	Type	bp from TSS	Function	PHMM Score	Reference
35	DDB2	5-UTR, Exon 1	activator	18 to 38bp	DNA-R	92.09	[246]
36	DDIT4, REDD1	Promoter	activator	-302 to -283bp	DNA-R	66.03	[61]
37	DDR1	Promoter	activator	-1494 to -1475bp	GR, R+	74.81	[212]
38	DKK1	Promoter	activator	-2136 to -2111bp	A	75.34	[226]
39	DNMT1	5-UTR, Exon 1	repressor	29 to 55bp	?	82.01	[193]
40	DUSP1, MKP1	Intron 2	activator	1235 to 1255bp	C, A	63.61	[139]
41	DUSP5	Promoter	activator	-1127 to -1107bp	C, CytoS	69.28	[261]
42	EDN2	Intron 3	activator	2197 to 2216bp	?	88.14	[94]
43	EEF1A1	Exon 4, CDS	activator	1869 to 1890bp	CytoS, A	81.43	[119]
43	EEF1A1	Exon 2, CDS	activator	1044 to 1067bp	CytoS, A	78.98	[119]
43	EEF1A1	Exon 3, CDS	activator	1670 to 1691bp	CytoS, A	61.94	[119]
44	EGFR	Promoter, 5-UTR, Exon 1	activator	-19 to 3bp	C, R+	72.16	[147]
45	EphA2	Promoter	activator	-1541 to -1519bp	A	78.25	[50]
46	FANCC, FAC	Promoter	activator	-1286 to -1257bp	A, DNA-R	71.11	[140]
47	FAS, CD95	Intron 1	activator	779 to 798bp	A	84.08	[170]
48	FDXR	Promoter	activator	-43 to -24bp		80.48	[142]
49	GADD45A	Intron 3	activator	1576 to 1595bp	DNA-R	86.2	[232]
50	GDF15, MIC-1	5-UTR, Exon 1, CDS	activator	12 to 31bp	A	79.68	[245]
50	GDF15, MIC-1	Promoter	activator	-866 to -847bp	A	80.39	[118]
51	GML	Promoter	activator	-18969 to -18950bp	C	90.16	[125]
52	GPX1	Promoter	activator	-182 to -163bp	DNA-R	83.67	[114]
53	HBV		repressor			69.91	[188]

Continued on next page

#	Gene Name(s)	Location	Type	bp from TSS	Function	PHMM Score	Reference
54	HD, Huntington	Intron 2	activator	15233 to 15259bp	CNS	78.76	[70]
54	HD, Huntington	Promoter	activator	-1855 to -1833bp	CNS	72.93	[70]
54	HD, Huntington	Intron 3	activator	25968 to 25993bp	CNS	83.11	[70]
55	HGF, SF	Promoter	activator	-324 to -305bp	C, R+	59.08	[164]
56	HIC1	5-UTR, Intron 1	activator	555 to 576bp	F	67.63	[22]
57	HRAS, c-Ha-Ras	5-UTR, Intron 1	activator	735 to 851bp	C	79.63	[46]
58	HSP90AB1, hsp90beta	5-UTR, Exon 1	repressor	16 to 45bp	HSP	80.95	[292]
59	HSPA8	5-UTR, Intron 1	repressor	648 to 675bp	HSP	70.91	[169]
60	IBRDC2, p53RFP	Promoter	activator	-168 to -149bp	C	74.51	[180]
61	IER3, IEX-1	Promoter	repressor	-247 to -226bp	A	77.27	[108]
62	IGFBP3		activator				[23]
62	IGFBP3	Intron 2	activator	4090 to 4109bp	R-	77.64	[23]
62	IGFBP3	Intron 1	activator	3170 to 3190bp	R-	78.69	[23]
63	IRF5	Exon 2, CDS	activator	4007 to 4028bp	CytoR	84.55	[171]
64	KRT8, CK8	5-UTR, Exon 1	activator	30 to 51bp	CytoS	63.36	[175]
65	LGALS3, galectin-3	Intron 2	repressor	8239 to 8263bp	A	74.97	[197]
66	LIF	Intron 1	activator	873 to 891bp	CytoR	80.9	[106]
67	LRDD, PIDD	5-UTR, Exon 2	activator	804 to 831bp	A	90.45	[141]
68	MAD1L1, MAD1	Promoter	repressor	-316 to -297bp	C	47.36	[35]
69	mdm2	5-UTR, Intron 1	activator	762 to 781bp	F-	70.03	[289]
69	mdm2	5-UTR, Intron 1	activator	724 to 743bp	F-	77.32	[289]
70	MET	Promoter	activator	-232 to -199bp	C, R+	67.42	[219]

Continued on next page

#	Gene Name(s)	Location	Type	bp from TSS	Function	PHMM Score	Reference
71	MLH1	Intron 1	activator	269 to 289bp	DNA-R	87.36	[31]
72	MMP2	Promoter	activator	-1645 to -1626bp	EMC	89.83	[15]
73	MSH2	Promoter	activator	-173 to -153bp	DNA-R	71.85	[272]
73	MSH2	Promoter	activator	-378 to -346bp	DNA-R	68.75	[217]
74	NDRG1	Promoter	activator	-373 to -342bp	A	65.88	[236]
75	NLRC4, Ipaf	Promoter	activator	-169 to -150bp	A	67.51	[209]
76	NOS3	5-UTR, Intron 1	repressor	2575 to 2597bp	CytoR	72.37	[173]
77	ODC1	Promoter	repressor	-334 to -310bp	C	73.5	[169]
77	ODC1	5-UTR, Intron 1	repressor	585 to 614bp	C	73.55	[169]
78	P2RXL1	Downstream (1631)	activator	15281 to 15302bp	CNS	78.99	[262]
79	P53AIP1	5-UTR, Intron 1	activator	2002 to 2021bp	A	73.6	[183]
80	PCBP4, MCG10	Promoter	activator	-891 to -870bp	A	70.04	[295]
80	PCBP4, MCG10	Promoter	activator	-1852 to -1824bp	A	63.97	[295]
81	PCNA	5-UTR, Intron 1	activator	6428 to 6447bp	C, DNA-R	77.17	[172]
82	PERP	Intron 1	activator	3361 to 3380bp	A	84	[199]
83	PLAGL1, ZAC	Promoter	activator	-861 to -842bp	C, F-	73.58	[208]
84	PLK2, SNK	Promoter	activator	-2258 to -2236bp	C	75.96	[25]
84	PLK2, SNK	Promoter	activator	-1303 to -1281bp	C	73.01	[25]
84	PLK2, SNK	Promoter	repressor	-2033 to -2014bp	C	93.13	[25]
85	PLK3	Promoter	activator	-439 to -414bp	C	73.15	[114]
86	PML	Intron 1	activator	643 to 670bp	T, S, A	85.42	[45]
87	PMS2	Intron 1	activator	2977 to 2998bp	DNA-R	50.66	[31]

Continued on next page

#	Gene Name(s)	Location	Type	bp from TSS	Function	PHMM Score	Reference
88	PPM1J, MGC19531	Downstream (6082)	activator	11355 to 11374bp	?	92.48	[94]
89	PRDM1, BLIMP1	Promoter	activator	-356 to -337bp	CytoR	87.21	[280]
90	PRKAB1, AMPKbeta1	5-UTR, Exon 1	activator	65 to 84bp	F-	74.37	[114]
91	PTEN	Promoter	activator	-117 to -84bp	A	93.98	[234]
92	PTK2, FAK	Promoter	repressor	-968 to -960bp	C, R	-18.5	[83]
93	PYCARD, ASC	Promoter	activator	-80 to -59bp	A	79.38	[185]
94	RABGGTA	5-UTR, Exon 1	activator	226 to 253bp	?	66.72	[114]
95	RB1	5-UTR, Exon 1	activator	59 to 82bp	C	73.47	[223]
96	RFWD2, COP1	Promoter	activator	-2198 to -2177bp	F-	75.83	[51]
97	RPS27L	Intron 1	activator	223 to 242bp	?	81.59	[94]
98	RRM2B, p53R2	Intron 1	activator	2259 to 2278bp	DNA-R	91.36	[247]
99	S100A2	Promoter	activator	-1850 to -1831bp	C	82.35	[244]
100	SCARA3, CSR1	Intron 2	unknown	17074 to 17093bp	DNA-R	87.81	[90]
101	SCD	Promoter	repressor	-199 to -179bp	ECM	78.12	[169]
102	SCN3B	Promoter	activator	-9137 to -9118bp	A	75.77	[1]
102	SCN3B	Intron 3	activator	13595 to 13614bp	A	77.54	[1]
103	SERPINB5, maspin	Promoter	activator	-224 to -204bp	ECM	64.63	[296]
104	SERPINE1	Promoter	activator	-226 to -207bp	ECM	82.89	[130]
105	SESN1, PA26	Intron 1	activator	511 to 530bp	C, S	77.67	[266]
106	SFN, 14-3-3sigma	Promoter	activator	-1812 to -1792bp	C	77.49	[96]
107	SH2D1A, SAP	Promoter	activator	-1884 to -1860bp	C, CytoR	70.91	[177]
107	SH2D1A, SAP	Promoter	activator	-1894 to -1876bp	C, CytoR	72.95	[177]

Continued on next page

#	Gene Name(s)	Location	Type	bp from TSS	Function	PHMM Score	Reference
107	SH2D1A, SAP	Promoter	activator	-1909 to -1885bp	C, CytoR	66.65	[177]
108	SLC38A2	Downstream (532)	repressor	15079 to 15108bp	?	75.78	[169]
109	STEAP3, TSAP6	5-UTR, Intron 1	activator	21225 to 21246bp	ECM, E	87.66	[191]
110	TAP1	Exon 1, CDS	activator	643 to 668bp	R	89.35	[98]
111	TGFA	Promoter	activator	-84 to -65bp	C, R	77.98	[224]
112	TNFRSF10A, DR4	Intron 1	activator	479 to 498bp	A	77.36	[145]
113	TNFRSF10B, DR5	Intron 1	activator	538 to 557bp	A	86.85	[243]
114	TNFRSF10C, DcR1	Intron 1	activator	369 to 388bp	A	83.07	[146]
115	TNFRSF10D, DcR2	Intron 1	activator	351 to 370bp	A	82.02	[146]
116	TP53, p53	Promoter, 5-UTR, Exon 1	activator	-12 to 7bp	A, C, S, DNA-R, F+	66.94	[12]
117	TP53i3, Pig3	5-UTR, Exon 1	activator	441 to 515bp	DNA-R	66.29	[41]
118	TP53INP1	Intron 3	activator	10562 to 10581bp	A	73.72	[186]
119	TP63, TP73L	Promoter	activator	-756 to -734bp	R-	76.02	[91]
120	TP73, p73	Promoter	activator	-2630 to -2603bp	R-	73.37	[32]
120	TP73, p73	Promoter	activator	-2612 to -2582bp	R-	68.78	[32]
121	Tp73:Delta	Promoter	activator	-75 to -45bp	R-	84.87	[178]
122	TRIAP1, p53CSV	Exon 1, CDS	activator	56 to 75bp	A	76.63	[190]
123	TRIM22, Staf50	5-UTR, Intron 1	activator	694 to 713bp	DNA-R	80.91	[182]
124	TRPM2	Promoter	repressor	-2251 to -2227bp	CNS, C	81.77	[169]
124	TRPM2	Promoter	repressor	-1878 to -1846bp	CNS, C	82.07	[169]
124	TRPM2	Promoter	repressor	-2246 to -2227bp	CNS, C	74.96	[169]
125	TSC2	Intron 11	activator	13579 to 13599bp	R-	84.19	[71]

Continued on next page

#	Gene Name(s)	Location	Type	bp from TSS	Function	PHMM Score	Reference
125	TSC2	Intron 2	activator	3921 to 3952bp	R-	71.77	[71]
125	TSC2	Intron 2	activator	2579 to 2598bp	R-	80.62	[71]
126	TYRP1, TRP-1	Promoter	activator	-122 to -100bp	protective	59.86	[181]
126	TYRP1, TRP-1	Promoter	activator	-108 to -77bp	protective	54.39	[181]
126	TYRP1, TRP-1	Promoter	activator	-75 to -56bp	protective	61.9	[181]
126	TYRP1, TRP-1	Promoter	activator	-134 to -113bp	protective	73.8	[181]
126	TYRP1, TRP-1	Promoter	activator	-85 to -66bp	protective	60.42	[181]
127	UBD, FAT10	Promoter	repressor	-239 to -220bp	A	59.67	[291]
128	VCAN, CSPG2	5-UTR, Intron 1	activator	684 to 704bp	C	87.31	[285]
129	VDR	5-UTR, Intron 1	activator	4720 to 4739bp	C, A	73.09	[157]
129	VDR	5-UTR, Intron 1	activator	4707 to 4729bp	C, A	55.95	[157]

Chapter 3

Modeling p53-binding Sites with PHMMs

3.1 Computational Methods to model DNA binding sites

Several algorithms have been devised to detect p53 REs in the DNA of all organisms and identify possible p53-responsive genes [101, 8, 49, 239, 256, 273, 282]. These computational algorithms have been used extensively to help the experimental process of finding functional p53-binding sites, transcriptional gene targets of p53, and functional SNPs in the p53 pathway. Although these algorithms have been extremely useful, they also have serious drawbacks. All the algorithms attempt to approximate the relative binding affinity of a putative p53-binding site by training a probabilistic model from a dataset of experimentally validated, functional p53 REs ('training set'). Therefore, the strength and predictive power of any such model is completely dependent on the sampling size and quality of the training set.

In addition, experiments have shown repeatedly that relative binding affinity is not the only thing when it comes to response elements (see above). Other important variables that affect the degree of function of a p53 RE include: adjacent co-factor binding sites, spacer-length, distance from the TSS, nucleosome positioning, and possibly others. For these reasons, all the algorithms that approximate relative binding affinity alone have very high false positive rates (for most TF-binding sites) [273]. In order to seriously boost predictive power, future algorithms will need to include at least some of the additional variables mentioned above.

The common Position Specific Score Matrix. By far the most common computational method for predicting p53 REs (and all other response elements) is the *Position Specific Score Matrix* (PSSM or weight matrix), which attempts to estimate the binding affinity of a putative site [239]. Besides the drawbacks mentioned above, PSSMs have other serious limitations in their attempts to approximate relative binding affinity. The PSSM model contains the probabilities of each nucleotide at each position in the motif (or the logarithms of the probabilities), and is therefore static in length. So, PSSMs cannot model allowed nucleotide insertions into, or allowed deletions from, the consensus motif, since any nucleotide insertion or deletion (*indels*) throws off the PSSM reading frame. This is clearly a problem because the p53 RE is very degenerate and $\approx 30\%$ of the 160 functional p53-binding sites in Table 2.4 on page 38 have at least one nucleotide insertion or deletion (*indel*) relative to the consensus. Any PSSM approach would thereby mis-score at least 30% of the binding sites in the dataset. Examples of genes that contain these degenerate sites are: BAI1, CAV1, EEF1A1, HSP90AB1, PCBP4, SH2D1A, TYRP1, and LIF.

3.2 PHMMs can model nucleotide insertions and deletions

Profile Hidden Markov Models provide a coherent theory for probabilistic modeling of degenerate binding sites where random nucleotide insertions into and deletions from the motif are tolerated at certain positions [127, 55]. Natural selection suggests that critical nucleotides are conserved over evolutionary time, while non-critical nucleotides (including tolerated insertions in the motif) are not conserved. The match state emissions of the PHMM serve to model the critical positions in the motif with their observed nucleotide frequencies. The additional hidden deletion and insertion states at each position enable the model to train for (relatively rare) observed deletions and insertions (*indels*) at different positions in the motif (see Figure 3.1 on page 53). Although the probability of any particular insertion or deletion of a nucleotide at a certain position in a functional motif may be rare, the accumulated probability over all the positions in the motif that an insertion or deletion event may occur can be significant. The training set of observed insertions and deletions (*indels*) serves to fine-tune the model to be properly sensitive to tolerated deviations from the most prevalent consensus motif. The main strength of the PHMM is this *trained flexibility* to properly model variable length motifs. The major drawback is that more data is required to train the extra parameters not found in weight matrices (PSSMs).

3.2.1 The Theory of Modeling TF-Binding Sites with Profile Hidden Markov Models

Given a set S of experimentally validated binding sites s for a TF-protein (and a few assumptions) it is possible to use the set S to estimate the relative binding free energy $-\Delta G(x)$ of any putative site x (without having to perform direct experimental measurements of binding constants). This bioinformatic approach using PHMMs (and PSSMs) is an attractive alternative, if a sufficient set S of experimentally validated binding sites is available.

The Assumptions:

1. The positions of a binding site contribute independently and additively to the binding free-energy.

2. Background DNA sequences are generally random samples from some k-mer distribution.

Neither of these assumptions are always true [238]. The first assumption can be relaxed by calculating di-nucleotide, tri-nucleotide,....,n^{th}-nucleotide frequencies from the training set S, but at some point an additivity assumption must be applied. Also, genomes are generally not random, but can be closely approximated by a 3^{rd} or 4^{th} Order Markov Model [248]. For simplicity in the examples here, we will assume that the background DNA can be modeled by a simple 0^{th} Order Markov Model (i.e. by mononucleotide content alone). This assumption greatly simplifies the calculation of the partition function [238].

From the additivity assumption we have that for any putative site x:

$$-\Delta G(x) = \sum_{i=1}^{length(x)} -\Delta G_j(b)$$

where we define...

$$-\Delta G_j(b) = \text{the independent contribution of base } b \text{ observed at position } j \text{ to the over-all binding free energy} \qquad (3.1)$$

The Profile Hidden Markov Model (PHMM) provides a completely probabilistic model for observing a sequence x within the modeled motif. The PHMM achieves this by incorporating the probabilities of different nucleotide insertions, deletions, and motif matches at each position in the motif [127]. In this application, the PHMM model is used to calculate the probability $P_{hmm}(x)$ of observing the putative site x in a real transcription factor binding site that is modeled by the PHMM. The probability $P_{hmm}(x)$ is used to find the *site log-odds score* of a putative site x. The *site log-odds score* $G^s(x)$ calculated by a PHMM trained by S is given by:

$$\begin{aligned} G^s(x) &= \ln\left(\frac{P_{hmm}(x)}{P_{background}(x)}\right) \qquad \text{(Site Log-odds Score)} \\ &= \sum_{j=1}^{\text{length}(x)} G_j^s(b) \end{aligned}$$

where we define:

$$\begin{aligned} G_j^s(b) &= \ln\left(\frac{P_{hmm}(j,b)}{P_{background}(j,b)}\right) \qquad \text{(Nucleotide Log-odds Score)} \\ j &= \text{position in the sequence } x,\ j \in \{1\ldots\ \text{length}(x)\} \\ b &= \text{observed nucleotide base, } b \in \{A, C, G, T\} \\ P_{hmm}(j,b) &= \text{probability of base } b \text{ at position } j \text{ in the PHMM model} \\ P_{background}(j,b) &= \text{probability of base } b \text{ at position } j \text{ in the null (background) model} \end{aligned} \tag{3.2}$$

With these definitions, and assuming independence of positions, we have:

$$\begin{aligned} P_{hmm}(x) &= \text{probability of candidate site } x \text{ in the PHMM model} \\ P_{background}(x) &= \text{probability of candidate site } x \text{ in the null (background) model} \end{aligned}$$

The *Site Log-odds Score* $G^s(x)$ can be considered proportional to the relative binding free energy $-\Delta G(x)$ when the Fermi-Dirac Equation for the equilibrium probability of a protein-bound binding site can be approximated by the Maxwell-Boltzmann Equation [49]. Another assumption is that the training set S consists of a proper sampling of functional binding sites that were collected under similar experimental conditions (like temperature T). However, this is likely not the case. A last assumption is that we are able to perfectly train the PHMM from our training set S, so that we can accurately predict the probability $P_{hmm}(x)$ for all possible putative sites x. However, properly training a PHMM from a limited training set S is a challenging problem. But with our idealizations and assumptions, the *Nucleotide Log-odds Score* $G^s_j(b)$ (calculated by our perfectly trained PHMM) is directly proportional to the binding free energy contribution of each observed base b at each position j in the sequence x.

Thus, under ideal conditions the log-odds scores that a trained Profile Hidden Markov Model calculates for any candidate site x is directly proportional to the free energy of binding to that candidate site. (Typically, proper scaling of G^s if not performed to make $G^s(x) \approx -\Delta G(x)$. Instead, G^s is only proportional to $-\Delta G(x)$.) [53] If the Profile Hidden Markov Model has no insertion or deletion states, then the PHMM is essentially a PSSM (weight matrix), and the probability $P_{hmm}(j, b)$ is equivalent to the $(b, j)^{th}$ entry in the (probability) weight matrix.

Three dynamic programming algorithms are used to calculate the probability $P_{hmm}(x)$ of observing the putative site x in the model. The *forward* and *backward* algorithms calculate $P_{hmm}(x)$ by summing up the probability of observing x for all possible paths π through the model:

$$forward(x) \quad = \quad backward(x) \quad = \quad P_{hmm}(x) \quad = \quad \sum_{\pi}^{all\ paths} P(x, \pi) \tag{3.3}$$

The *Viterbi* algorithm calculates both the optimal alignment of the putative site x which produces the path $\pi^*(x)$ with the highest log-odds score, and the probability $P^{\pi^*}_{hmm}(x)$ of observing that optimal path in the model. These two results of the Viterbi algorithm are commonly referred to as the *Viterbi path* and the *Viterbi score*, respectively:

$$\begin{aligned} Viterbi\ path(x) &= \pi^*(x) &= \underset{\pi}{\text{argmax}}\,[P(x,\pi)] \\ Viterbi\ score(x) &= P_{hmm}^{\pi^*}(x) &= P_{hmm}(x,\pi^*(x)) \end{aligned}$$

In the case of modeling transcription factor binding sites, it is commonly assumed that the log-odds score of the optimal path that best aligns the putative site x to the model is the only significant contributor to the over-all log-odds score. When this is indeed true, the *Viterbi score* can be used as a good approximation to $P_{hmm}(x)$:

$$Viterbi\ score(x) \;=\; P_{hmm}(x,\pi^*(x)) \;\approx\; \sum_{\pi}^{all\ paths} P(x,\pi) \;=\; P_{hmm}(x) \;=\; forward(x) \tag{3.4}$$

However, we see that this assumption is not true when modeling p53 cluster sites, where experiments suggest that the p53 protein can bind to overlapping combinations of adjacent half-sites. In this scenario, the true probability $P_{hmm}(x)$ provided by the *forward* and *backward* algorithms is needed to properly model experimental results.

All three dynamic programming algorithms are highly efficient, and when applied to PHMMs run in $O(NM)$ time and $O(NM)$ space for a PHMM with M states and a sequence of length N [55]. For further details about the *forward*, *backward*, and *Viterbi* algorithms please see [53].

Example: Log-Odds Scoring of the sequence ACCG in Figure 3.1. Let's assume that the PHMM in Figure 3.1 models a transcription factor binding site, and that the depicted path through the PHMM is the optimal path for the candidate sequence ACCG obtained by the Viterbi algorithm. Let's further assume that the log-odds score for this optimal alignment is a good approximation for the overall log-odds score $G^s(ACCG)$ (which would include the probabilities of observing ACCG for all paths through the model, not just the optimal path). Now, with the optimal alignment of the sequence ACCG, we can

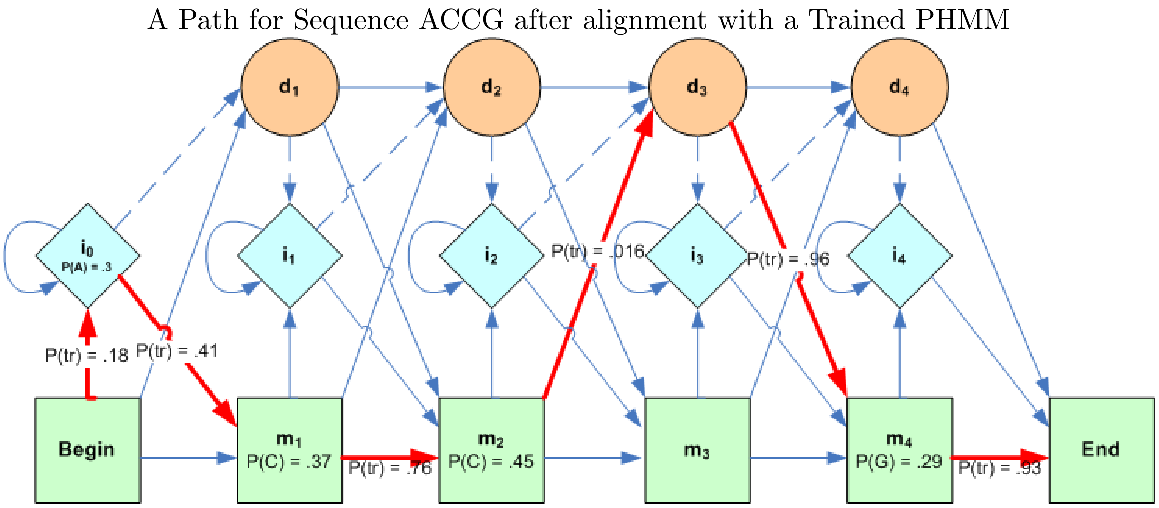


Figure 3.1: **A State Path through the topology of a PHMM.** A possible path for the sequence ACCG through a trained PHMM is highlighted in red. The match, insertion, and deletion states are green, blue, and orange respectively. The arrows represent transitional probabilities between the states. P(tr) and P(N) are the probabilities of a transition between states and the probabilities of emitting nucleotide N within a state, respectively. In this path, A is an inserted nucleotide, both C's are matches to the consensus, consensus position 3 is absent in the sequence, and G matches the final consensus position. A model with only the match states (green), and no insertion or deletion states, would be synonymous with a weight matrix (PSSM).

calculate the log-odds scores of the sequence and get an estimate of the relative binding affinity for this site. Here, we will assume a uniform background distribution at each position so that $P_{background}(j, b) = P_{background}(b) = .25$ for each base b.

$$\begin{aligned}
G_1^s(A) &= \ln\left(\frac{P_{hmm}(1,A)}{P_{background}(A)}\right) = \ln\left(\frac{.18 \cdot .3}{.25}\right) = \ln(.18)+\ln(.3)-\ln(.25) = -1.532 \\
G_2^s(C) &= \ln\left(\frac{P_{hmm}(2,C)}{P_{background}(C)}\right) = \ln\left(\frac{.41 \cdot .37}{.25}\right) = \ln(.41)+\ln(.37)-\ln(.25) = -.499 \\
G_3^s(C) &= \ln\left(\frac{P_{hmm}(3,C)}{P_{background}(C)}\right) = \ln\left(\frac{.76 \cdot .45}{.25}\right) = \ln(.76)+\ln(.45)-\ln(.25) = .313 \\
G_4^s(G) &= \ln\left(\frac{P_{hmm}(4,G)}{P_{background}(G)}\right) = \ln\left(\frac{.016 \cdot .96 \cdot .29 \cdot .93}{.25}\right) = -4.100 \\
G^s(ACCG) &= \sum_{j}^{length(ACCG)} G_j^s(b) = -1.532+-.4995+.313+-4.100 = -5.819 \qquad (3.5)
\end{aligned}$$

Therefore the Log-Odds Score of sequence ACCG is -5.819, which is not a good score. A negative score signifies a site that better fits the random (uniform) background distribution than it fits the trained Profile Hidden Markov Model (trained by the set S of known binding sites). This means that this site has a binding affinity that is on par with the binding affinity of a random sequence from the given background distribution. Therefore, ACCG is not a potential binding site in this example. Notice that the biggest contributors to this negative score are the inserted and deleted nucleotides relative to the motif. It is often the case that nucleotide insertions and deletions (*indels*) within a binding site have a heavy associated cost that greatly lessons the total site score.

3.2.2 The Proof that the Log-odds Score $G^s(x)$ is proportional to $-\Delta G(x)$

It has been shown experimentally that in general, transcription factor proteins have a weak affinity for background DNA (any non-consensus sequence) and a strong affinity for consensus sites. Within the nucleus (or general cell in prokaryotes) the DNA concentration is high enough that an activated TF-protein is bound somewhere on the DNA essentially all the time (to a 1^{st} approximation) [239]. Therefore, the binding specificity (the ability of the TF protein to distinguish a functional site from background DNA) must be adequately high for proper regulation to occur [239]. The goal is to quantify the free energy of binding to a candidate site x through statistical mechanics, thermodynamics and Information Theory.

We start with the mass action kinetics of a TF-protein binding to a site:

$$\begin{aligned}
p &= \text{transcription factor protein} \\
x &= \text{a candidate DNA binding site} \\
px &= \text{Bound Protein-Binding Site Complex} \\
k^{+} &= \text{forward equilibrium binding constant} \\
k^{-} &= \text{backward equilibrium binding constant} \\
p + x &\underset{k^{-}}{\overset{k^{+}}{\rightleftharpoons}} px \\
K_{eq}^{x} &= \frac{k^{+}}{k^{-}} \quad = \text{equilibrium association constant for site } x
\end{aligned} \tag{3.6}$$

We normalize K_{eq}^{x} in order to obtain the specific association constant K_{s}^{x} that quantifies specificity:

1. K_{eq}^{avg} = Average K_{eq} for all sites x
2. $K_{s}^{x} \quad = \quad \frac{K_{eq}^{x}}{K_{eq}^{avg}}, \quad (avg(K_{s}^{x}) = 1)$
3. Specificity of Valid Site: $K_{s}^{\text{valid site}} \approx 10^{6}$
4. Specificity of Background: $K_{s}^{\text{background}} < 1$

In experiments performed in E. Coli cells, with about 5×10^{6}bp of DNA, a single TF-protein and a single binding site with a specificity of 10^{6} will be bound together only about 20% of the time. During the other 80% of the time, the protein will be transiently bound to other random places along the genome. However, with 20 copies of the protein the binding site will be occupied about 99% of the time [73].

The specific association constant K_s^x is related to the binding free energy $-\Delta G(x)$ by the following:

$$-\Delta G(x) \quad = \quad -k_\beta \cdot T \cdot \ln(K_s^x)$$

and

$$-K_s^x \quad = \quad \frac{k^+}{k^- \cdot K_{eq}^{avg}} \quad = \quad e^{-\Delta G(x)/k_\beta T} \tag{3.7}$$

Now lets estimate the probability that a putative binding site x is bound by a TF-protein in a well-mixed solution at equilibrium. Let $P(x \text{ bound})$ be the probability that the binding site x is bound by a TF-protein. Then we have:

$$\begin{aligned} P(x \text{ bound}) &= \frac{\text{binding rate}}{\text{binding rate} + \text{unbinding rate}} \\ &= \frac{[p] \cdot k^+}{[p] \cdot k^+ + k^-} \\ &= \frac{[p] \cdot K_{eq}^{avg} \cdot e^{-\Delta G(x)/k_\beta T}}{[p] \cdot K_{eq}^{avg} \cdot e^{-\Delta G(x)/k_\beta T} + 1} \end{aligned} \tag{3.8}$$

which can be re-written into the form known as the Fermi-Dirac Equation, where $\mu = k_\beta T \ln(K_{eq}^{avg} \cdot [p])$ is the *chemical potential* dependent on the protein concentration $[p]$:

$$P(x \text{ bound}) = \frac{1}{e^{(\Delta G(x) - \mu)/k_\beta T} + 1} \qquad \text{(Fermi-Dirac)}$$

In the low concentration limit the Fermi-Dirac Equation for the probability $P(x \text{ bound})$

can be approximated by the Maxwell-Boltzmann Equation:

$$\begin{aligned} P(x \text{ bound}) \approx & \quad \frac{1}{e^{(\Delta G(x)-\mu)/k_\beta T}} \quad \text{when } \Delta G(x) \gg \mu \\ \approx & \quad e^{\mu/k_\beta T} \cdot e^{-\Delta G(x)/k_\beta T} \qquad \text{(Maxwell-Boltzmann)} \\ \approx & \quad z e^{-\Delta G(x)/k_\beta T} \qquad (z = e^{\mu/k_\beta T} = \text{fugacity}) \end{aligned} \tag{3.9}$$

Now we are ready to analyze a sampling set S of known transcription factor binding sites for a given TF-protein. A version of this proof exists for weight matrices (PSSMs) in [99, 49]. Here we provide a general proof that it is applicable for any fully probabilistic model that calculates $P_{background}(x)$ and $P_{setS}(x)$.

Assume that we attain the set S from a single experiment so that all the sites are collected under identical conditions. Assume that we have a very large number of DNA sequences of roughly similar length from a given genome mixed in solution with a certain concentration of TF-proteins. At equilibrium some of the DNA sequences with bound TF-protein are extracted (precipitated) and sequenced to create our sampling set S.

The probability of observing exactly the set S is given by:

$$\begin{aligned} P(\text{observing the set } S) &= \prod_{x \in S} (P_{exist}(x) \cdot P_{bound}(x) \cdot P_{extract}(x)) \cdot \prod_{x \notin S} (1 - P_{exist}(x) \cdot P_{bound}(x) \cdot P_{extract}(x)) \\ &\approx \prod_{x \in S} (P_{exist}(x) \cdot P_{bound}(x) \cdot P_{extract}(x)) \cdot e^{\sum_{x \notin S} (P_{exist}(x) \cdot P_{bound}(x) \cdot P_{extract}(x))} \end{aligned} \tag{3.10}$$

The likelihood function $\mathscr{L}$ for the P(observing the set S) can now be approximated:

$$
\begin{aligned}
\mathscr{L} &= \ln[P(\text{observing the set } S)] \\
&\approx \ln\left[\prod_{x\in S}(P_{exist}(x)\cdot P_{bound}(x)\cdot P_{extract}(x))\cdot e^{\sum_{x\notin S}(P_{exist}(x)\cdot P_{bound}(x)\cdot P_{extract}(x))}\right] \\
&\approx \sum_{x\in S}\ln(P_{exist}(x)\cdot P_{bound}(x)\cdot P_{extract}(x)) - \sum_{x\notin S}(P_{exist}(x)\cdot P_{bound}(x)\cdot P_{extract}(x)) \qquad (3.11)
\end{aligned}
$$

Now plug-in the Maxwell-Boltzmann approximation $ze^{-\Delta G(x)/k_\beta T}$ for $P(x \text{ bound})$, and for simplicity assume that $P_{extract}(x) = P_{extract}$ is identical for all x:

$$
\begin{aligned}
\mathscr{L} &\approx \sum_{x\in S}\ln\Big(P_{exist}(x)\cdot ze^{-\Delta G(x)/k_\beta T}\cdot P_{extract}\Big) - \sum_{x\notin S}\Big(P_{exist}(x)\cdot ze^{-\Delta G(x)/k_\beta T}\cdot P_{extract}\Big) \\
&\approx N_s\cdot\ln(z\cdot P_{extract}) + \sum_{x\in S}\left(\ln(P_{exist}(x))\cdot\frac{-\Delta G(x)}{k_\beta T}\right) - z\cdot P_{extract}\sum_{x\notin S}\Big(P_{exist}(x)\cdot e^{-\Delta G(x)/k_\beta T}\Big) \qquad (3.12)
\end{aligned}
$$

Where N_s is the size of the sampling set S. We are now ready to maximize the likelihood function $\mathscr{L}$ by taking the partial derivatives with respect to $zP_{extract}$ and $\Delta G_i(b)$ and setting them equal to 0. We have From the additivity assumption that for any putative site x:

$$
-\Delta G(x) = \sum_{i=1}^{length(x)} -\Delta G_i(b)
$$

where we define...

$$
\begin{aligned}
-\Delta G_i(b) &= \text{the independent contribution of base } b \text{ observed at position } i \\
-\Delta G_i(x,b) &= -\Delta G_i(b)\cdot x(i,b) \\
x(i,b) &= 1 \text{ if } x_i = b, \text{ and } 0 \text{ if } x_i \neq b \qquad (3.13)
\end{aligned}
$$

After taking the partial derivatives we have:

$$\begin{aligned}
\frac{\partial \mathscr{L}}{\partial (zP_{extract})} &= \frac{N_s}{z\cdot P_{extract}} - \sum_{x\notin S}\left(P_{exist}(x)\cdot e^{-\Delta G(x)/k_\beta T}\right) &= 0\\
\frac{\partial \mathscr{L}}{\partial (\Delta G_i(b))} &= \frac{\sum\limits_{x\in S} x(i,b)}{k_\beta T} - \left[\frac{z\cdot P_{extract}}{k_\beta T}\cdot P_{exists}(i,b)\cdot e^{-\Delta G_i(b)/k_\beta T}\cdot \prod_{j\neq i}\sum_{b'} P_{exists}(j,b')\cdot e^{-\Delta G_j(b')/k_\beta T}\right] &= 0
\end{aligned} \tag{3.14}$$

We can combine the results from the partial derivatives to obtain:

$$\frac{1}{N_s}\sum_{x\in S} x(i,b) = \frac{P_{exists}(i,b)\cdot e^{-\Delta G_i(b)/k_\beta T}\cdot \prod\limits_{j\neq i}\sum\limits_{b'} P_{exists}(j,b')\cdot e^{-\Delta G_j(b')/k_\beta T}}{\sum\limits_{x\notin S}\left(P_{exist}(x)\cdot e^{-\Delta G(x)/k_\beta T}\right)} \tag{3.15}$$

If we make the observation that:

$$\sum_{x\notin S}\left(P_{exist}(x)\cdot e^{-\Delta G(x)/k_\beta T}\right) = \sum_{b'} P_{exists}(i,b')\cdot e^{-\Delta G_i(b')/k_\beta T}\cdot \prod_{j\neq i}\sum_{b'} P_{exists}(j,b')\cdot e^{-\Delta G_j(b')/k_\beta T}$$

then we have that:

$$\begin{aligned}
\frac{1}{N_s}\sum_{x\in S} x(i,b) &= \frac{P_{exists}(i,b)\cdot e^{-\Delta G_i(b)/k_\beta T}\cdot \prod\limits_{j\neq i}\sum\limits_{b'} P_{exists}(j,b')\cdot e^{-\Delta G_j(b')/k_\beta T}}{\sum\limits_{b'} P_{exists}(i,b')\cdot e^{-\Delta G_i(b')/k_\beta T}\cdot \prod\limits_{j\neq i}\sum\limits_{b'} P_{exists}(j,b')\cdot e^{-\Delta G_j(b')/k_\beta T}}\\
&= \frac{P_{exists}(i,b)\cdot e^{-\Delta G_i(b)/k_\beta T}}{\sum\limits_{b'} P_{exists}(i,b')\cdot e^{-\Delta G_i(b')/k_\beta T}}\\
&= \frac{P_{exists}(i,b)\cdot e^{-\Delta G_i(b)/k_\beta T}}{C}\\
\frac{\frac{1}{N_s}\sum\limits_{x\in S} x(i,b)}{P_{exists}(i,b)}\cdot C &= e^{-\Delta G_i(b)/k_\beta T}\\
\ln\left[\frac{\frac{1}{N_s}\sum\limits_{x\in S} x(i,b)}{P_{exists}(i,b)}\right] + \ln C &= -\frac{\Delta G_i(b)}{k_\beta T}\\
\ln\left[\frac{\frac{1}{N_s}\sum\limits_{x\in S} x(i,b)}{P_{exists}(i,b)}\right] &\approx\propto -\Delta G_i(b)
\end{aligned} \tag{3.16}$$

Now we make the following observations:

$$\begin{aligned}
\frac{1}{N_s}\sum_{x\in S} x(i,b) &= \text{probability of observing base b at position i in our set } S \\
&= P_{setS}(x_i(b)) \\
P_{exists}(i,b) &= P_{background}(i,b) \qquad (3.17)
\end{aligned}$$

So now we have:

$$\begin{aligned}
G_i^s(b) &= \ln\left[\frac{P_{setS}(x_i(b))}{P_{background}(i,b)}\right] && \approx\propto && -\Delta G_i(b) \\
G^s(x) &= \ln\left[\frac{P_{setS}(x)}{P_{background}(x)}\right] && \approx\propto && -\Delta G(x) \qquad \text{(by the additivity assumption)}
\end{aligned}$$

□

Training a PHMM with validated binding sites. Before a PHMM can be used to estimate the relative binding affinity for any putative binding site, the PHMM must be trained to properly model a functional binding site of interest. When training a PHMM for a particular motif, the goal is to choose the parameters of the model in order to maximize the likelihood of the sequences in the training set, without over-fitting. Again, under ideal conditions the log-odds score (*log-likelihood ratio*) $G^s(x)$ to be maximized for the collection of binding sites in the training set is proportional to the estimated binding free energy $-\Delta G(x)$ of these binding sites. When the state paths for the training sequences are not known, no known closed form solution exists for the parameter estimations [53]. The *Baum-Welch* algorithm is the most commonly used iterative Expectation Maximization (EM) method to train the parameters of the model. The Baum-Welch algorithm always climbs the gradient (to increase the combined scores of the training set) and uses the optimized dynamic programming *forward* and *backward* algorithms [53].

3.3 Results and Discussion

3.3.1 p53HMM: using a training method that boosts predictive power.

To increase the predictive power of our p53-motif PHMMs, we attempt to exploit the *a priori* knowledge that when proteins bind as homodimers or homotetramers, their corresponding binding sites typically have a *palindromic*, *repeat*, and/or *reverse complement* structure (see Figure 3.2). This prior knowledge can be used to correspond (fully or partially tie) the parameters between positions in order to exploit the inherent redundancy in the information of the motif. Within a set of corresponding positions, the updating of emission and transition probabilities can borrow strength from each other by sharing information. In addition, the degree of sharing of information for any set of corresponding positions can be optimized during training. The process of corresponding parameters can greatly reduce the parameter search-space during the training of the model, and provide the ability to train for rare occurrence insertion and deletion events (See Figure 3.3).

This general technique has been effectively used when HMMs have been applied to speech and handwriting recognition problems, and has been referred to as *parameter tying* [133]. We introduce an extension to this method that allows for the setting or training for an optimal level of partial or full parameter tying. In the domain of protein-DNA binding sites, even if a palindromic, repeat, or reverse complement structure of a binding site is not known *a priori*, all the known structural motifs can be tested, and the structure can be *discovered* (inferred) from the ROC curve that maximizes predictive accuracy. For example, of the six structural models tested for the p53-binding motif, the palindromic motif that completely corresponds the two half-sites is the *discovered* motif, since it is the best classifier (see Figure 3.5).

3.4 The Corresponded Baum-Welch algorithm

In order to include the prior knowledge of the structural motif (or in an attempt to discover it), a novel "Corresponded Baum-Welch" algorithm is proposed to enforce or learn the optimal correspondence between expectations of parameters for corresponding positions after each iteration of the Baum-Welch algorithm (see Methods). For example, assume that we have prior knowledge that a transcription factor protein binds to the DNA in homodimer form, where each monomer interacts with 5 DNA base pairs. Then a corresponding palindromic motif for the nucleotide positions would be: *1 2 3 4 5 5 4 3 2 1*, while a reverse-complement palindromic motif would be: $\mathit{1}\ \mathit{2}\ \mathit{3}\ \mathit{4}\ \mathit{5}\ \tilde{\mathit{5}}\ \tilde{\mathit{4}}\ \tilde{\mathit{3}}\ \tilde{\mathit{2}}\ \tilde{\mathit{1}}$ (where $\tilde{a}$ has the complement nucleotide emission distribution of a). All the emission distributions for each of the five sets of synonymous positions would be made corresponding, as well as all the transition probabilities between synonymous positions. In this example, if all the parameters between synonymous positions were fully corresponding (tied), then the parameter search space would be roughly cut in half. The level of correspondence between the parameters for synonymous positions can be given *a priori*, or trained for if the training set is sufficiently large. One optimal level of correspondence, c, can be calculated for the whole motif (for all the corresponding positions), or a separate one can be found for each set of corresponding positions.

3.4.1 Details of the Corresponded Baum-Welch Algorithm

The standard Baum-Welch EM algorithm is used to estimate the expected transition and emission probabilities from the training set. The Baum-Welch algorithm is an optimized, iterative EM method that always climbs the gradient and uses the dynamic programming *forward* and *backward* algorithms [53].

Let:

$s = \textit{binding site}$	The nucleotide sequence of a binding site
$s_i = \textit{nucleotide}$	The i^{th} nucleotide in the binding site s
$S = \textit{training set}$	The training set of binding sites s_j
$\pi = \textit{path}$	The state sequence of a binding site s
$\pi_i = \textit{state}$	The i^{th} state in the path π
$ps_{kl} = \textit{pseudocount}$	Prior bias of probability of transition from k to l
$ps_k(b) = \textit{pseudocount}$	Prior bias of probability of emitting symbol b in state k
$\psi = \{ps_{kl}, ps_k(b)\}, \forall k, l, b$	The set of all pseudocounts in the model
$a_{kl} = P(\pi_i = l \vert \pi_{i-1} = k)$	The probability of transition from state k to state l
$e_k(b) = P(s_i = b \vert pi_i = k)$	The probability of emitting symbol b in state k
$\theta = \{a_{kl}, e_k(b)\}, \forall k, l, b$	The set of all parameters in the model
$a_{kl}^{background} = P_{background}(\pi_i = l \vert \pi_{i-1} = k)$	The probability of transition from state k to state l in the null (background) model
$e_k^{background}(b) = P_{background}(s_i = b \vert pi_i = k)$	The probability of emitting symbol b in state k in the null (background) model
$A_{kl} = \textit{expected } a_{kl} \textit{ counts}$	Number of transitions from k to l in the training set
$E_k(b) = \textit{expected } e_k(b) \textit{ counts}$	Number of emissions of b from state k in the training set
$f_k(i) = P(s_1 \ldots s_i, \pi_i = k)$	The probability of the sequence up to and including s_i, requiring that $\pi_i = k$
$f_k(i+1) = e_k(s_{i+1}) \cdot \sum_j^{states}(f_j(i) \cdot a_{jk})$	Recursive formula for $f_k(i+1)$ going forward
$b_k(i) = P(s_i \ldots s_L, \pi_i = k)$	The probability of the sequence from s_i to the end, requiring that $\pi_i = k$, $L =$ length of the sequence s
$b_k(i-1) = e_k(s_{i-1}) \cdot \sum_j^{states}(b_j(i) \cdot a_{jk})$	Recursive formula for $b_k(i-1)$ going backward

The goal is to choose the parameters θ of the model in order to maximize the log-likelihood of the sequences s in the training set S, without over-fitting. To avoid over-fitting, the goal is to find the Posterior Mean Estimator (PME), a Bayesian approach that uses the pseudo-counts ψ as a prior from a Dirichlet family of distributions and all the paths π for all sequences s in the training set S [53]:

$$\theta^{PME} = \underset{\theta}{\text{argmax}} \left[\sum_{s \in S} \log P(s \vert \theta, \psi) \right] = \underset{\theta}{\text{argmax}} \left[\sum_{s \in S} \sum_{\pi} \log P(s, \pi \vert \theta, \psi) \right]$$

The Baum-Welch algorithm climbs the gradient during each iteration and is guaranteed to converge within some epsilon to a local maximum, which may or may not be the PME [53]. Theoretically, the Corresponded Baum-Welch algorithm has the advantage of using prior motif knowledge to greatly reduce the parameter space and to potentially "flatten" the space. Both of these improvements can increase the probability of the algorithm converging to the PME.

In each iteration, the Baum-Welch algorithm calculates the expected number of times each transition and emission is used by the training set sequences (calculates A_{kl} and $E_k(b)$), given the current model parameters (a_{kl} and $e_k(b)$). Then the model parameters are updated to the new posterior mean estimators a'_{kl} and $e'_k(b)$, calculated from the new expectation counts (A_{kl} and $E_k(b)$).

Notice that the probability that a_{kl} is used at position i of binding site sequence s with current model parameters θ is given by:

$$P(\pi_i = k, \pi_{i+1} = l|s, \theta) = \frac{f_k(i) \cdot a_{kl} \cdot e_l(s_{i+1}) \cdot b_l(i+1)}{P(s)}$$

By summing over all training sequences and positions, we can derive A_{kl} and $E_k(b)$, the expected number of times that a_{kl} and $e_k(b)$ are used by the training set, given the current model parameters θ:

$$
\begin{aligned}
N =\ & \text{number of training sequences} \\
L =\ & \text{length of the sequence } s^j \\
W(s^j) =\ & \text{sequence weight of } s^j \\
A_{kl} =\ & \sum_{s^j \in S}^{N} \frac{W(s^j)}{P(s^j)} \sum_{i=1}^{L} f_k^j(i) \cdot a_{kl} \cdot e_l(s_{i+1}^j) \cdot b_l^j(i+1) \\
E_k(b) =\ & \sum_{s^j \in S}^{N} \frac{W(s^j)}{P(s^j)} \sum_{i | s_i^j = b}^{L} f_k^j(i) \cdot b_k^j(i)
\end{aligned} \tag{3.18}
$$

The sequence weight $W(s^j)$ is used to vary the importance of different sequences in the training set S and to vary their influence in training the model. A weight $W(s^j) > 1$ increases the expected counts in sequence s^j, and a weight $W(s^j) < 1$ decreases them. Sequence weights are used when we do not fully trust that the training set S provides a proper distribution of valid binding sites, and we attempt to remedy that deficiency by weighting the known sequences. Most sequence weighting methods attempt to penalize the expected counts of similar sequences and to enhance the expected counts of distant sequences [53].

Additionally, the process by which the training set S was ascertained may be biased toward a certain subset of sites independent of their sequences (*ascertainment bias*). In the derivation for our approximation for $-\Delta G(x)$ in the next section, we relied on the assumption that the probability $P_{extract}(x)$ of extracting a TF-bound binding site was independent of the sequence in or around x. This may not always be the case. For example, if we know that a certain antibody preferentially binds to adjacent binding sites compared to ones with no neighbors, then after precipitation our training set S would be biased toward adjacent binding sites that appear in tight clusters in the DNA. We could attempt to compensate for this inherent bias by penalizing those sequences found adjacent to each other in the genome and promoting the ones with no neighbors. Different sequence weighting schemes can be found in [250, 81, 5, 227, 56, 95, 128].

From these new expected counts, we can now calculate new maximum likelihood estimators for each position:

$$\begin{aligned} a'_{kl} &= \frac{A_{kl}}{\sum_{m}^{states} A_{km}} \\ e'_{k}(b) &= \frac{E_k(b)}{\sum_{n}^{\{A,C,G,T\}} E_k(n)} \end{aligned} \tag{3.19}$$

However, if we believe the training set S to be incomplete and intend to avoid overfitting the data, we add pseudocounts as priors to our expected counts. Here, pseudocounts are distributed in proportion to the null (background) model. The pseudocount weight w represents how many counts from the null (background) model we want to include in the expected counts of our model. From the expected counts, we calculate the new posterior mean estimators using pseudocounts for each position:

$$\begin{aligned}
w &= \text{pseudocount weight} \\
ps_{kl} &= w \cdot a_{kl}^{background} \\
ps_k(b) &= w \cdot e_k^{background}(b) \\
a'_{kl} &= \frac{ps_{kl} + A_{kl}}{w + \sum_{m}^{states} A_{km}} \\
e'_k(b) &= \frac{ps_k(b) + E_k(b)}{w + \sum_{n}^{\{A,C,G,T\}} E_k(n)}
\end{aligned} \tag{3.20}$$

Now we use the prior knowledge (or make a guess) of the repeat and/or palindromic motif and correspond (partially or fully tie) the new posterior mean estimators based upon corresponding positions. This prior knowledge can be used to reduce the parameter space and increase the statistical accuracy of the model. The degree of sharing of information between corresponding positions is controlled by a correspondence factor c, which can be fixed or trained to an optimum value. One can estimate a correspondence factor based on the initial conditions by the following:

$$\begin{aligned}
dist &= \text{a probability distribution in the set of corresponding distributions} \\
var &= \text{a variable in the probability distributions} \\
N &= \text{number of corresponding distributions} \\
\overline{P(var)} &= \text{average probability of a variable over all corresponding distributions} \\
c_0 &= \text{initial correspondence factor} \\
&= 1 \quad - \quad \frac{1}{N-1} \sum_{dist} \sum_{var} \left| \overline{P(var)} - P(var) \right| \qquad (3.21)
\end{aligned}$$

We calculate the corresponding posterior mean estimator (PME) after calculating the average emission and transition probabilities for all the corresponding positions:

$$\begin{aligned}
c &= \text{correspondence factor} \\
\overline{a'} &= Avg(a'_{kl}) \quad \text{(over all transitions from } k \text{ to } l \text{ in the set of corresponding positions)} \\
\overline{e'(b)} &= Avg(e'_k(b)) \quad \text{(over all emissions in the set of corresponding positions)} \\
a''_{kl} &= a'_{kl} + c \left[\overline{a'} - a'_{kl} \right] \\
e''_k(b) &= e'_k(b) + c \left[\overline{e'(b)} - e'_k(b) \right] \qquad (3.22)
\end{aligned}$$

If we wish to train for the optimum correspondence factor, then we calculate a new c' for each emission and transition probability at each position in the set of corresponding positions:

$$\begin{aligned} c'_{kl} &= \frac{c \cdot \overline{a'}}{a'_{kl} + c\left[\overline{a'} - a'_{kl}\right]} &= \frac{c \cdot \overline{a'}}{a''_{kl}} \\ c'_k(b) &= \frac{c \cdot \overline{e'(b)}}{e'_k(b) + c\left[\overline{e'(b)} - e'_k(b)\right]} &= \frac{c \cdot \overline{e'(b)}}{e''_k(b)} \end{aligned} \tag{3.23}$$

Now, we can calculate a new correspondence factor c' by averaging over sets of the c'_{kl} and $c'_k(b)$ values. The one optimum correspondence factor for the whole motif or separate correspondence factors for sets of corresponding positions are obtained by averaging over different sets:

$$c' = \overline{c'_k(b)} \quad \text{(over all bases } b \text{ and all emissions and transitions } k\text{)}$$

$$\text{or}$$

$$\text{(over all bases } b \text{ and corresponding emissions and transitions } k\text{)} \tag{3.24}$$

We can now update the parameters of the model to the new posterior mean estimators that have been made corresponding (fully or partially tied) by our prior knowledge (or guess) of the motif:

$$\begin{aligned} a_{kl} &\Longrightarrow a''_{kl} \\ e_k(b) &\Longrightarrow e''_k(b) \\ c &\Longrightarrow c' \end{aligned} \tag{3.25}$$

This process is then iterated to obtain new A_{kl} and $E_k(b)$ values from the new model parameters. At each iteration the log likelihood of the training set increases to a local maximum. Since convergence is in a continuous-valued space, the maximum is never actually reached. Typically, the iterations are stopped when the change in the total log likelihood is sufficiently small or after some fixed number of iterations, whichever comes first [53].

Derivation of finding optimum correspondence. The method of finding the locally optimum degree of correspondence (sharing of information) between corresponding positions starts by introducing the new parameter c for each set of corresponding positions. If we interpret the correspondence factor c as the probability $P(\text{identical})$ that the positions are completely synonymous, then we can interpret that every emission and transition probability $P(x)$ for each corresponding position in the model can now be replaced by a new probability $P'(x)$:

$$\begin{aligned} P'(x) &= P(\text{identical}) \cdot \overline{P(x)} \quad + \quad (1 - P(\text{identical})) \cdot P(x) \\ &= P(x) + c\left[\overline{P(x)} - P(x)\right] \end{aligned} \tag{3.26}$$

where $\overline{P(x)}$ is the average of the corresponding emission and transition probabilities. Now we can calculate new correspondence factors c' for each corresponding emission and transition probability in the set of corresponding positions:

$$\begin{aligned} c' &= \frac{P(\text{identical}) \cdot \overline{P(x)}}{P(\text{identical}) \cdot \overline{P(x)} \quad + \quad (1 - P(\text{identical})) \cdot P(x)} \\ &= \frac{c \cdot \overline{P(x)}}{P(x) + c\left[\overline{P(x)} - P(x)\right]} \\ &= \frac{c \cdot \overline{P(x)}}{P'(x)} \end{aligned} \tag{3.27}$$

Now we can calculate a new correspondence factor c'' for the set of corresponding parameters by averaging over the new c' for all the corresponding emission and transition probabilities:

$$c'' \;=\; \overline{c'} \quad \text{(over all } c' \text{ in the set of corresponding positions)} \tag{3.28}$$

Example. Assume that we have prior knowledge (or we guess that) the binding motif of a 10-bp binding site is singly palindromic: *1 2 3 4 5 5 4 3 2 1*. Then the positions that have been made corresponding are: *1* and *10*, *2* and *9*, *3* and *8*, *4* and *7*, *5* and *6*. (There are five sets of corresponding positions in this example.) First, each of the 10 distributions of the posterior mean emission probabilities for each of the 10 positions in the motif are now corresponding and sharing data with its partner position. Then the posterior mean transition distributions between positions are similarly made corresponding (for example *1-2* and *2-1*). Separate correspondence calculations are performed for each of the sets of corresponding positions. A correspondence factor of $c = 1$ would fully correspond (tie) the parameters between synonymous positions to the average over all corresponding parameters. (In this case, the parameter space would roughly be cut in half, and the training data per parameter would roughly double.) A correspondence factor of $c = 0$ would not change the initial distributions of emission and transition probabilities at a position at all, thus creating no correspondence between the positions. The correspondence factor c can be regarded as our *known* prior belief in the level of correspondence between synonymous positions in a palindromic, repeat, and/or reverse-complement binding-site motif. Alternatively, the correspondence factor c can be regarded as the *unknown* probability of correspondence between synonymous positions that needs to be determined. In the latter case, the Corresponded Baum-Welch algorithm will converge on the (locally) optimum c that maximizes the total log likelihood of the training set.

Comparing the different p53 corresponding (structural) motifs. Since the 20bp-tetrameric p53-binding site has a repeated and nested palindromic structure, different correspondence motifs can be constructed to train the PHMM models, and cross validation can be used to compare their predictive properties. The motifs that are compared are: the repeat or T-coupled motif ($\mathit{1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 10\ 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 10}$), the (reverse-complement) palindromic or H-coupled motif ($\mathit{1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 10\ \widetilde{10}\ \tilde{9}\ \tilde{8}\ \tilde{7}\ \tilde{6}\ \tilde{5}\ \tilde{4}\ \tilde{3}\ \tilde{2}\ \tilde{1}}$), the independently (reverse-complement) palindromic or Q-coupled motif ($\mathit{1\ 2\ 3\ 4\ 5\ \tilde{5}\ \tilde{4}\ \tilde{3}\ \tilde{2}\ \tilde{1}\ 6\ 7\ 8\ 9\ 10\ \widetilde{10}\ \tilde{9}\ \tilde{8}\ \tilde{7}\ \tilde{6}}$, and the repeated, fully-palindromic or combined motif ($\mathit{1\ 2\ 3\ 4\ 5\ \tilde{5}\ \tilde{4}\ \tilde{3}\ \tilde{2}\ \tilde{1}\ 1\ 2\ 3\ 4\ 5\ \tilde{5}\ \tilde{4}\ \tilde{3}\ \tilde{2}\ \tilde{1}}$) (see Figure 3.2) [149]. We perform 1000 iterations of ten-fold random-split cross validation on each model to gain statistics on their predictive accuracy. The positive set contains 160 experimentally validated p53 binding sites from [202], and the negative set contains 40bp random samples from the mononucleotide content of the training set. Then we utilize Receiver Operating Characteristic (ROC) curves in order to compare the predictive power of the classifiers in an unbiased, threshold-independent (non-parametric) manner. This is achieved by calculating the true positive and false positive rates for all possible threshold values for each model. The summary statistic for comparing the ROC curves is the AUC (Area Under the Curve). AUC values lie somewhere between 1.0 and 0.5 (where an AUC of 1.0 would correspond to a perfect classifier, and an AUC of 0.5 would correspond to a classifier that is no better than random coin flipping.)

Training Insert State Emissions. A major consideration when training Profile Hidden Markov Models (PHMMs) is which parameters to train for at each position, and which parameters to fix at each position to the over-all average. The more non-fixed parameters that must be trained for at each position in the motif, the more data that is needed to properly train the model. Ideally, a sufficiently large training set is available to be able to train for all the parameters in the PHMM at each position. Unfortunately, in the case of transcription factor binding sites, this is rarely the case. Typically, when using PHMMs to model DNA binding sites, both the insert probabilities and insert state nucleotide emissions probabilities are set to the binding site averages, since there are rarely enough examples of these rare occurrence events at a particular position to train those parameters for that position alone [154]. By corresponding (fully or partially tying) positions and in effect increasing the training data for each position, it may be possible to train the insertion state emissions distributions for these corresponding positions. This could possibly boost predictive power of the models, if the p53 protein is selective as to which nucleotides can be inserted into the motif at certain positions without compromising the binding affinity of the site. A common example of such selective sequence insertions can be found in functional protein families, whereby hydrophobic or hydrophilic amino acid insertions may be tolerated at certain positions, provided that the insertions are present either in the core or at the surface of the protein, respectively, after folding. Notice that fixing the insertion state emission distributions at every position to the amino-acid average for the whole sequence would be very inappropriate in this example.

The final results. The (reverse-complement) palindromic or H-coupled motif ($\mathit{1}\ \mathit{2}\ \mathit{3}\ \mathit{4}\ \mathit{5}\ \mathit{6}\ \mathit{7}\ \mathit{8}\ \mathit{9}\ \mathit{10}\ \widetilde{\mathit{10}}\ \tilde{\mathit{9}}\ \tilde{\mathit{8}}\ \tilde{\mathit{7}}\ \tilde{\mathit{6}}\ \tilde{\mathit{5}}\ \tilde{\mathit{4}}\ \tilde{\mathit{3}}\ \tilde{\mathit{2}}\ \tilde{\mathit{1}}$) outperforms all other structural motifs (see Figures 3.4 and 3.5). However, the repeat and independently palindromic motifs perform nearly as well. Finally, the combined motif ($\mathit{1}\ \mathit{2}\ \mathit{3}\ \mathit{4}\ \mathit{5}\ \tilde{\mathit{5}}\ \tilde{\mathit{4}}\ \tilde{\mathit{3}}\ \tilde{\mathit{2}}\ \tilde{\mathit{1}}\ \mathit{1}\ \mathit{2}\ \mathit{3}\ \mathit{4}\ \mathit{5}\ \tilde{\mathit{5}}\ \tilde{\mathit{4}}\ \tilde{\mathit{3}}\ \tilde{\mathit{2}}\ \tilde{\mathit{1}}$) also performs on par with the above models, although it contains roughly half the degrees of freedom. These results suggest that there exist correlations between the positions in all four of the motifs above for the p53-binding site, although the correlations within the palindromic motif are the strongest. Furthermore, it can be seen that training the insert state emissions per corresponding position also boosts the predictive power of all the models (see Figures 3.4 and 3.5). In addition, the more correspondence placed between the synonymous positions during each training iteration, the better the resulting classifier at that point in the training (results not shown). For this training set, all the palindromic models with fixed correspondence factors between $c = 0.4$ and $c = 1.0$ eventually converged to the same predictive model, although lower correspondence factors required more iterations to do so. All the models converged on correspondence factors between $c = .98$ and $c = .999$ when training for optimum correspondence. Therefore the best predictive model completely corresponds (ties) the two half-sites in a palindromic structure during each iteration of the training. Our published p53HMM algorithm is this best predictive model: trained on the dataset of

160 functional p53 REs, fully corresponding the data per position based on the palindromic structural motif, and training the insert state emissions.

Validation of the p53HMM algorithm. The new p53HMM algorithm was used to screen for putative p53-binding sites in the endosomal compartment genes, which led to the discovery of a functional p53 site and a new p53-regulated gene, CHMP4C [287]. The putative p53RE sequence AAACAAGCCC agtagcagcagctgctcc GAGCTTGCCC was predicted in the promoter region (-497 to -460bp) of the CHMP4C gene. The data from the chromatin immunoprecipitation and the luciferase reporter assays showed that p53 protein can bind to this sequence and induce CHMP4C gene expression. Additionally, analysis by p53HMM found an alternative putative p53 binding site in the LIF gene that corresponds to a 5 bp shift to the right relative to the recently published putative site in intron 1 [106]. The p53HMM algorithm predicted the site GGACATGTCGGGACA–GCTC, which matches the consensus RRRCWWGYYYRRRCWWGYYY perfectly except for the gap ("–", deletion) at position 16. A PSSM approach predicted the shifted site AAcCAgGaCatGtCggGaCa, which is the best "gap-less" p53 site in the region conferring p53 regulation, but it still matches the consensus very poorly (consensus mismatches are in lowercase)[106]. A few genes in the dataset of 160 functional p53 binding sites have a deletion relative to the consensus exactly between the well-conserved C and G as seen above, including the genes: EGFR, TYRP1, EEF1A1, HSP90AB1, and BAI1. This discovery of an alternative p53-binding site that better matches known functional sites, by modeling for observed insertions and deletions, highlights some of the advantages of the new p53HMM algorithm.

Special considerations for the p53HMM algorithm. Although the spacer within a p53 RE has been shown to greatly affect the binding affinity for p53 protein, the ability to properly quantify this effect for all possible spacers of lengths 0-21 base pairs has been elusive. Therefore like previous algorithms, we have chosen to initially ignore the spacers of the training set and putative REs [101]. We are able to ignore arbitrary-length spacers by inserting a no-cost *Free Insertion Module* (FIM) between the two half-sites of the single-site PHMM [107, 10]. Similarly, we can ignore spacers with lengths between 1 and N base pairs by inserting a no-cost *Finite Emission Module* (FEM-N) between the two half-sites (see Figure 3.6). A prior p53 RE search algorithm (p53MH) was based upon a PSSM approach and a novel filtering matrix [101]. Unfortunately, the tables were not symmetric and the filtering table over-fit the available data at the time. The combined result was that the p53MH method completely rejects 58 of the 160 experimentally validated sites to date (receiving a score of 0 out of 100, where 100 represented the maximum relative binding affinity). Additionally, some sites received very high scores approaching 100, while the reverse-complement received a score of 0, and vice-versa. Due to these observations, we have purposely designed the p53HMM algorithm to be symmetric, so as to give identical scores for putative sites and their reverse complements. Secondly, we chose to abandon the filtering matrix to avoid over-fitting the available data. A feature that we preserved from p53MH is the normalizing of scores by the highest possible affinity for the motif (×100), so that the highest possible normalized score is 100.

Modeling dependencies between positions. PSSMs assume that all nucleotide positions within the motif contribute independently to the binding affinity of the binding site, which has been shown experimentally to not always be the case [239]. Recent research has focused on modeling dependencies between positions in protein-DNA binding sites [8, 294]. Typically *Tree Bayesian Networks* and *Mixtures* of trees have been used to attempt to model these dependencies between positions, which have been shown through cross validation to increase the predictive power of these models [8]. Our PHMM models do not attempt to model dependencies between the positions, however they can be extended to do so by using higher-order Profile Hidden Markov Models. Unfortunately, the ability to train for positional dependencies, and boost predictive power, is dependent upon the sampling size of the training set and requires larger training sets to train the extra parameters.

3.5 Modeling p53 cluster-sites

Binding affinity measurements have been obtained for certain p53 cluster-sites of different lengths by mutating or truncating known p53 cluster-sites in the genes: DDB2, TP53i3, CKM, IGFBP3, and RGC (See Table 3.1 and Figure 3.1) [246, 41, 20, 122]. Based on the relative binding affinities of these p53 cluster-sites, we propose a new p53 cluster-site algorithm that utilizes the trained PHMM to calculate and sum up the relative estimated binding-affinities, above a certain threshold, of all viable full-sites in the cluster with a spacer of 14bp or less (See Table 3.1). This model predicts a linear increase in p53 binding affinity dependent upon the number of half-sites in the cluster-site and the length of spacers between them. For example, for p53 cluster-sites with 2, 3, 4, 5, or 6 adjacent p53 half-sites, the number of possible full-sites with spacer-lengths $\leq$ 14bp would be 1, 3, 5, 7, and 9, respectively. Let N be the number of half-sites in the cluster-site, then the number of full-sites (to calculate binding affinities for and sum up) is given by the expression $2N - 3$ ($N \geq 2$). Although there exist functional sites with spacers $\geq$ 15bp, experiments suggest that their contribution to the overall binding affinity within a cluster-site is negligible.

These p53 cluster-site scores are attained through a two step process. The first step uses the cluster-site model which contains a generalized p53 half-site PHMM and a back-transition that limits any spacer between two half-sites to no more than 14bp (see part e of Figure 3.6). The dynamic programming *Viterbi* algorithm is used to find the highest scoring p53 half-sites in the sequence (that are separated by no more then 14bp). The second step then parses the state-path generated from step 1 and generates viable p53 full-sites with any spacers removed, while conserving the property that the half-sites in the cluster-site were not separated by more than 14bp. Now we use the more flexible p53 single-site model to score these viable full-sites using the *Viterbi* algorithm (see part d of Figure 3.6). We maintain a running sum of the log-odds scores of the candidate full-sites that are above a certain threshold. The log-odds score threshold and spacer-length limit (14bp) are chosen so as to best fit the experimental data (See Figure 3.7). Additionally, this p53 cluster-site model follows statistical mechanics, in that the overall binding affinity for the complete RE is proportional to the probability of any p53 protein binding to any of the allowed motifs found in the cluster-site.

3.5.1 Details of the p53 cluster-site algorithm

The p53 cluster-site algorithm is a two step process designed to sum the estimated relative binding affinities of all viable full-sites within a cluster-site. The first step uses the cluster-site model that contains a generalized p53 half-site PHMM and a back-transition through a no-cost FEM-14 module (see part e of Figure 3.6). The no-cost Finite Emission Module (FEM) of length 14 can match any sequence of length $\leq$ 14bp with no contribution to the over-all score. We score the entire putative cluster-site using the p53 cluster-site model and the *Viterbi* algorithm to find the best-supported path through the cluster-site. This path provides the strongest affinity half-sites that are not separated by more than 14bp. If

Table 3.1: Normalized Experimental Affinity of Cluster-sites

	Number of Half-sites													
	2	**3**	**4**	**5**	**5.5**	**6**	**7**	**7.5**	**8**	**8.5**	**9**	**10**	**11**	**12**
Cluster Site	**Relative Binding Affinity**													
DDB2	1		5											
TP53I3		3			6			10		12			16	
Theoretical Affinity Approximations														
# of Full-sites with spacers ≤ 14bp	1	3	5	7	8	9	11	12	13	14	15	17	19	21
# of Full-sites with spacers ≤ 24bp	1	3	6	9	10.5	12	15	16.5	18	19.5	21	24	27	30
# of Full-sites with any size spacer	1	3	6	10		15	21		28		36	45	55	66

This table contains the normalized experimental affinities of different cluster-sites dependent upon the number of half-sites contained in the RE. These affinity measurements were obtained by mutating or truncating p53 cluster-sites in the genes DDB2, and TP53i3 [246, 41]. These two p53 cluster-sites are chosen because they match the assumption of the theoretical models that no spacer sequences are present between the half-sites. All affinities are normalized by the two half-site (full-site) affinity respective of the RE. The theoretical models assume that all the half-sites in each cluster-site are identical, which is not the case for either of the two cluster-sites. Experimental results support a linear affinity growth model based upon the number of full-sites with spacers no longer than 14bp (in yellow).

we use the notation "[14]" for any spacer sequence of length 0 to 14 and H for a half-site sequence, then we can represent the cluster-site sequence path as:

$$H_1[14]H_2[14]H_3[14]....[14]H_N \quad \text{(where } N = \text{number of half-sites in the path)}$$

Step 2 now parses the cluster-site sequence path and generates a list of all viable full-sites, which are concatenations of any two half-sites such that they are not separated by more than 14bp:

$$\text{Set of viable full-sites } = \{H_1H_2, H_1H_3, H_2H_3, ...\}$$

Now we use the more flexible (and more accurate) single-site model with the Viterbi algorithm to estimate the relative binding affinity of all the viable full-sites in the cluster-site. The cluster-site affinity score is the sum of all viable full-site scores that exceed a certain threshold. If F denotes a viable full-site then:

$$\begin{aligned}
\text{cluster-site affinity score} &= \sum_{F}^{\{H_1H_2,H_1H_3,H_2H_3,....\}} \text{contribution}(F) \\
\text{contribution}(F) &= \begin{cases} Viterbi(F) & \text{if} \quad Viterbi(F) \geq \text{threshold;} \\ 0 & \text{if} \quad Viterbi(F) < \text{threshold.} \end{cases}
\end{aligned} \tag{3.29}$$

The spacer-length upper bound and the affinity-score lower bound were fit to best match the experimental results. In the case for p53-binding sites, the best fit is a spacer-length of no more than 14bp and a log-odds score of at least 27.5 (see Figure 3.7 on page 86).

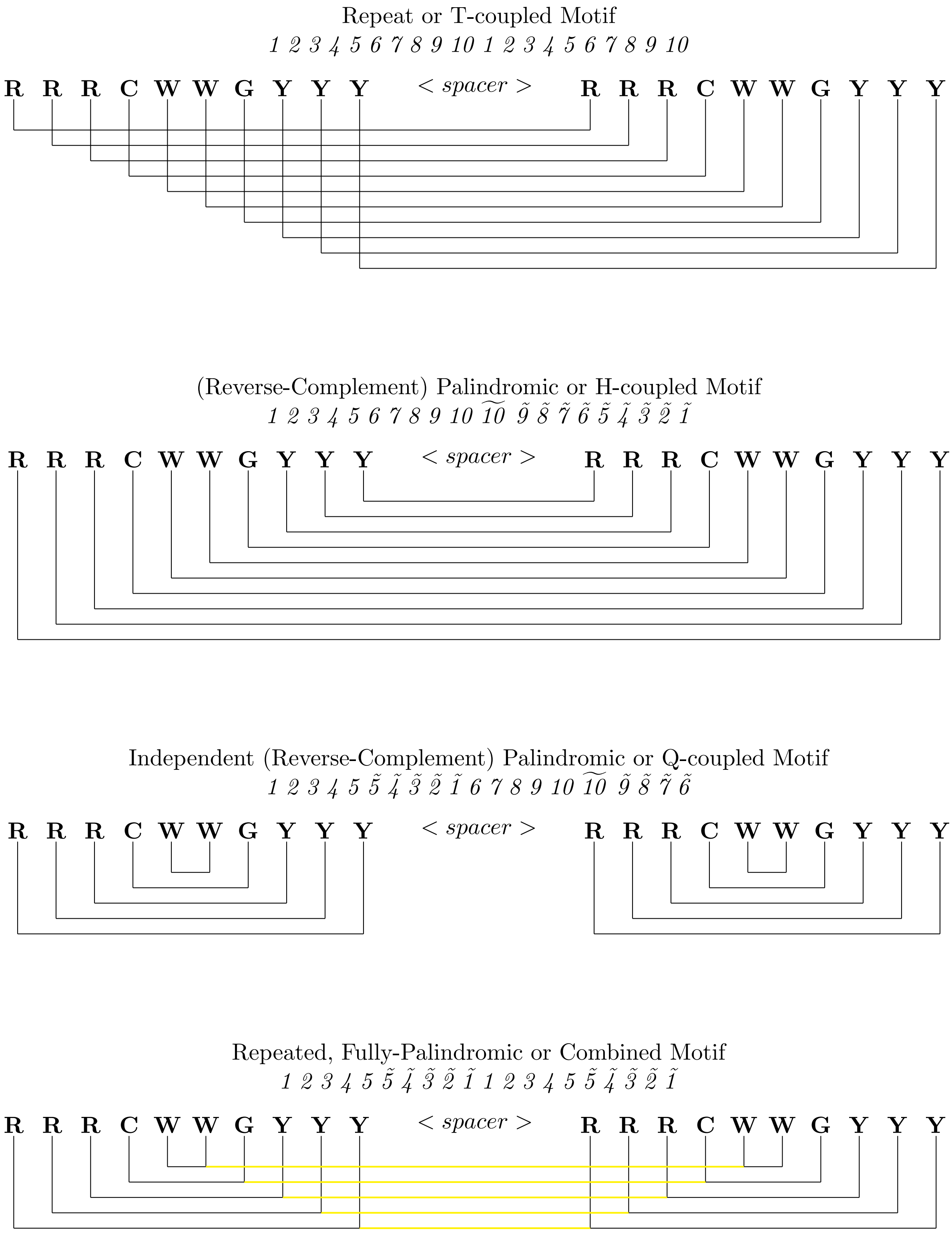

Repeat or T-coupled Motif
1 2 3 4 5 6 7 8 9 10 1 2 3 4 5 6 7 8 9 10
R R R C W W G Y Y Y < spacer > R R R C W W G Y Y Y
(Reverse-Complement) Palindromic or H-coupled Motif
1 2 3 4 5 6 7 8 9 10 $\widetilde{10}$ $\tilde{9}$ $\tilde{8}$ $\tilde{7}$ $\tilde{6}$ $\tilde{5}$ $\tilde{4}$ $\tilde{3}$ $\tilde{2}$ $\tilde{1}$
R R R C W W G Y Y Y < spacer > R R R C W W G Y Y Y
Independent (Reverse-Complement) Palindromic or Q-coupled Motif
1 2 3 4 5 $\tilde{5}$ $\tilde{4}$ $\tilde{3}$ $\tilde{2}$ $\tilde{1}$ 6 7 8 9 10 $\widetilde{10}$ $\tilde{9}$ $\tilde{8}$ $\tilde{7}$ $\tilde{6}$
R R R C W W G Y Y Y < spacer > R R R C W W G Y Y Y
Repeated, Fully-Palindromic or Combined Motif
1 2 3 4 5 $\tilde{5}$ $\tilde{4}$ $\tilde{3}$ $\tilde{2}$ $\tilde{1}$ 1 2 3 4 5 $\tilde{5}$ $\tilde{4}$ $\tilde{3}$ $\tilde{2}$ $\tilde{1}$
R R R C W W G Y Y Y < spacer > R R R C W W G Y Y Y

Figure 3.2: **The Four p53 Correlation Motifs.** The four correspondence motifs for the repeated, palindromic p53 RE are graphically represented. In the top three motifs, each line signifies correspondence between two synonymous positions. In the bottom motif, the previously independent half-sites are made “corresponding” (tied) by the yellow connecting lines so that now four synonymous positions are corresponded. (R = A or G, W = A or T, and Y = C or T. Position $\tilde{a}$ has the complement nucleotide emission distribution of a.)

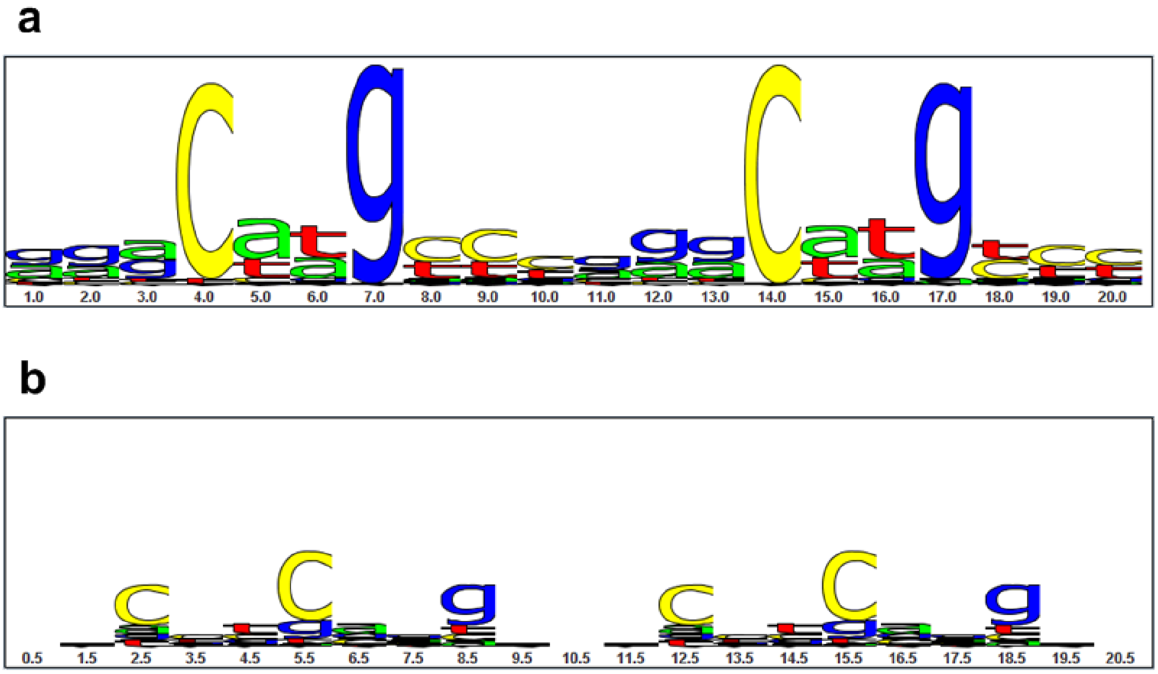


Figure 3.3: **(a)** The match-state sequence logo for the palindromic p53 motif: $\mathit{1}\ \mathit{2}\ \mathit{3}\ \mathit{4}\ \mathit{5}\ \mathit{6}\ \mathit{7}\ \mathit{8}\ \mathit{9}\ \mathit{10}\ \widetilde{\mathit{10}}\ \tilde{\mathit{9}}\ \tilde{\mathit{8}}\ \tilde{\mathit{7}}\ \tilde{\mathit{6}}\ \tilde{\mathit{5}}\ \tilde{\mathit{4}}\ \tilde{\mathit{3}}\ \tilde{\mathit{2}}\ \tilde{\mathit{1}}$. (Motif position $\tilde{a}$ has the complement nucleotide-emission distribution of a.) The height of each letter is made proportional to its frequency at each position, and the letters are sorted in descending frequency order. The height of the entire stack at each position is then adjusted to signify the information content (in bits) of that position [218]. The match-state nucleotide positions 4, 7, 14, and 17 (motif positions $\mathit{4}$, $\mathit{7}$, $\tilde{\mathit{7}}$, and $\tilde{\mathit{4}}$ respectively) are the most conserved and are the main points of contact with the p53 protein. **(b)** The insert-state sequence logo for the combined-palindromic p53-model: $\mathit{1}\ \mathit{2}\ \mathit{3}\ \mathit{4}\ \mathit{5}\ \tilde{\mathit{5}}\ \tilde{\mathit{4}}\ \tilde{\mathit{3}}\ \tilde{\mathit{2}}\ \tilde{\mathit{1}}\ \mathit{1}\ \mathit{2}\ \mathit{3}\ \mathit{4}\ \mathit{5}\ \tilde{\mathit{5}}\ \tilde{\mathit{4}}\ \tilde{\mathit{3}}\ \tilde{\mathit{2}}\ \tilde{\mathit{1}}$. These nucleotide insertions occur in-between the nucleotide positions shown in part **a**. The specificity motif of the insert-state emissions is different from that of the match-state emissions.

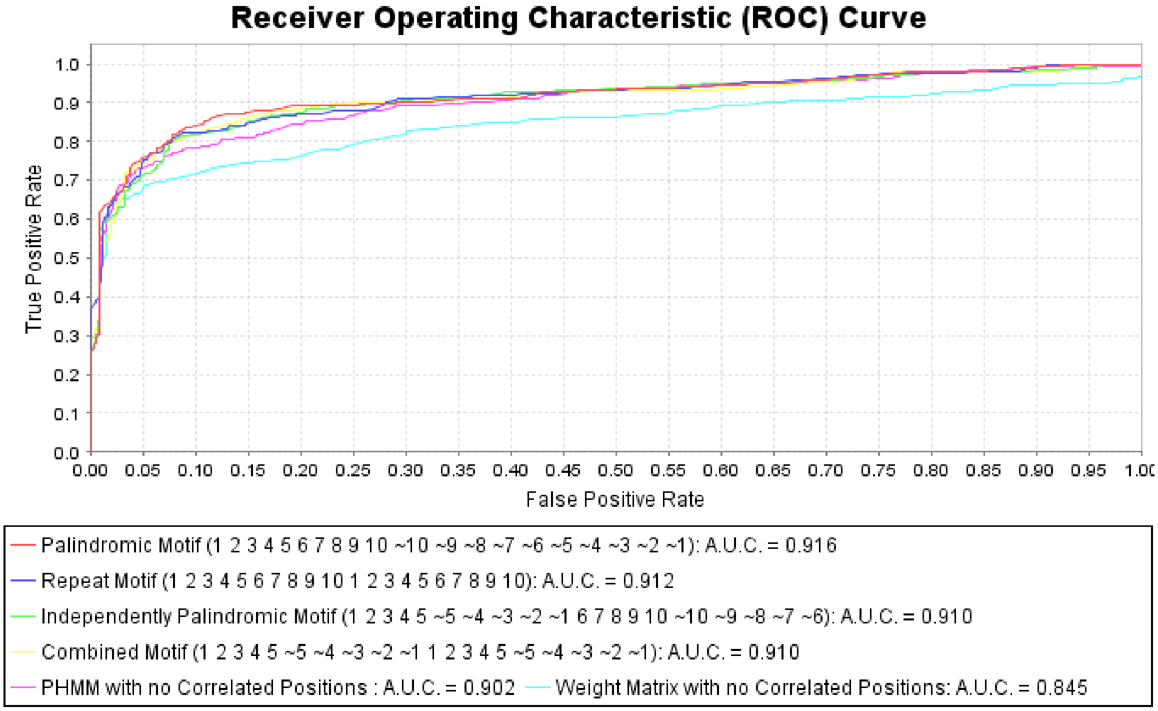


Figure 3.4: **Cross Validation with Receiver Operating Characteristic (ROC) curves reveals increased predictive power over weight matrices.** 1000 iterations of 10-fold random-split cross validation reveal that the most predictive model is the palindromic structure. The positive set contains 160 experimentally validated p53 binding sites, and the negative set contains 40bp random samples from the mononucleotide content of the training set. The true positive and false positive rates are calculated and plotted for all possible threshold values for each model. The predictive measure for comparing the curves is the AUC (Area Under the Curve). In all the PHMM models the insert state emissions are fixed to the A, G, C, T nucleotide distribution of the training set. The best classifier uses the palindromic training motif. (Position $\sim a$ has the complement nucleotide emission distribution of a).

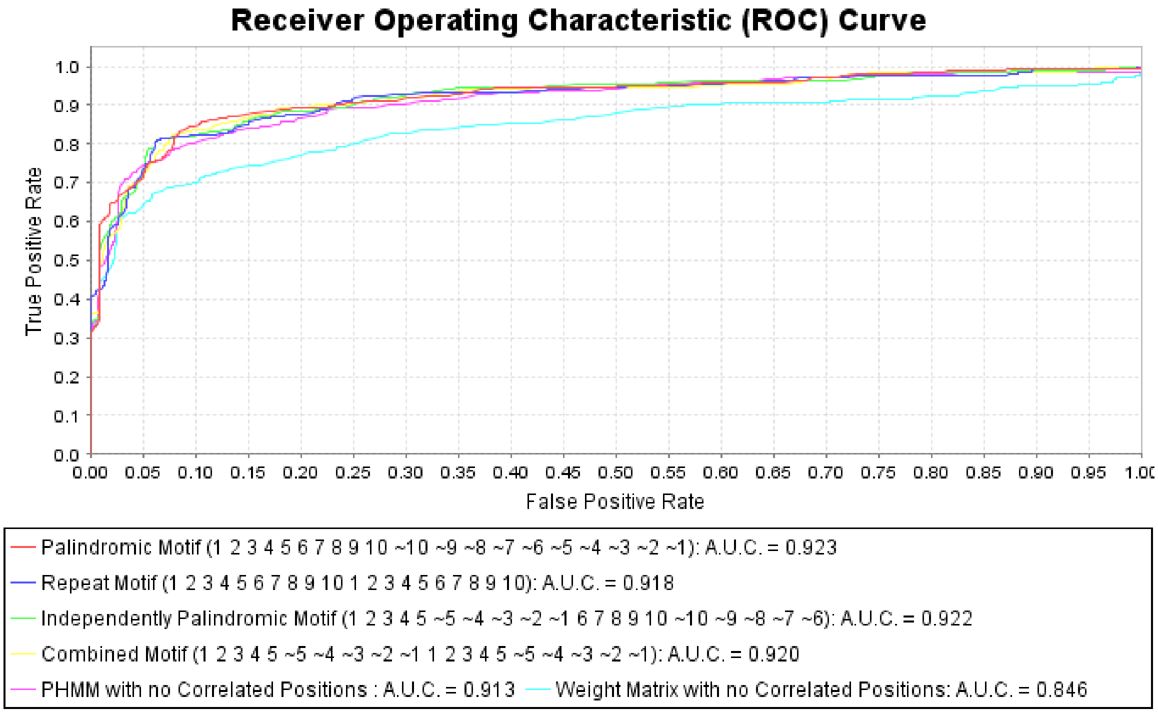


Figure 3.5: **Cross Validation with Receiver Operating Characteristic (ROC) curves reveals increased predictive power when training insert state emissions.** All the PHMM models in this comparison train the insert emission distributions based on positional insertions occurring in the training set. Again, 1000 iterations of 10-fold random-split cross validation reveal that the most predictive model is the palindromic structure. The positive set contains 160 experimentally validated p53 binding sites, and the negative set contains 40bp random samples from the mononucleotide content of the training set. The true positive and false positive rates are calculated and plotted for all possible threshold values for each model. The predictive measure for comparing the curves is the AUC (Area Under the Curve). The AUC values improve for all the PHMM models compared to Figure 3.4, but not for the weight-matrix model (which does not use the insert states). The best classifier (with the palindromic training motif) was used for the p53HMM algorithm. (Position $\sim a$ has the complement nucleotide emission distribution of a).

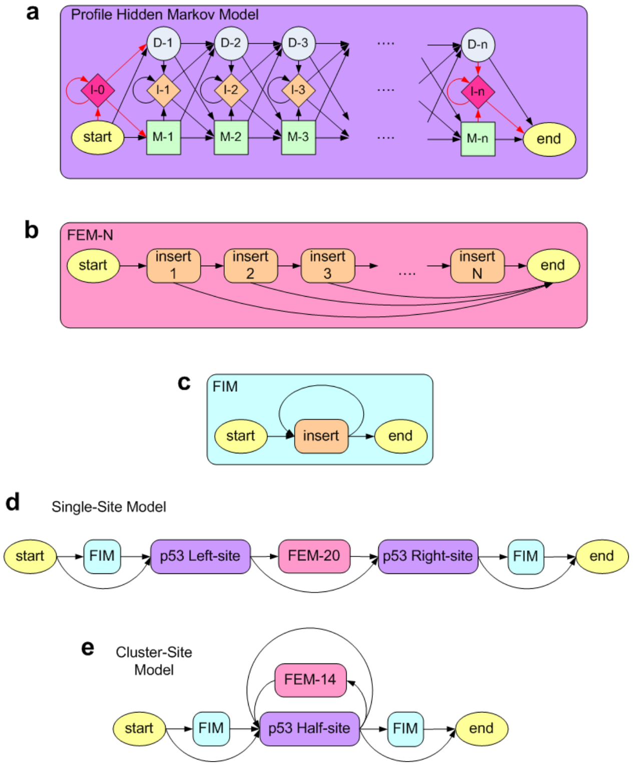


Figure 3.6: **The Topologies of p53 Single-site and Cluster-site Models. (a)** A Profile Hidden Markov Model (PHMM) contains three hidden states for each position in a sequence motif of length n: a match state (green squares), an insertion state (orange diamonds), and a delete state (gray circles). The arrows represent allowed transitions between states and have associated probabilities. The match and insertion states also have associated nucleotide emission probabilities. The first and last insertion states (I-0 and I-n) and associated transitions (in red) are shown for completeness. However, they are not present in the p53 models since they are replaced by FIM and FEM models. **(b)** The topology of the Finite Emission Module (FEM) of length N allows the ability to model any distribution of spacer-lengths between 1 and N. For the p53 models, the model and background probabilities within the FEM modules are identically uniform so that there is no-cost for spacer-lengths between 1 and N, and are referred to as “no-cost FEMs”. **(c)** The topology of the Free Insertion Module (FIM) allows for the ability to model an exponentially decaying distribution of spacer-lengths. However, by setting the model and background probabilities to identically uniform, the FIM can model any sequence of infinite length with no associated cost to the overall score (hence the word “Free”). **(d)** The main components of the p53 single-site model are the left and right half-site PHMMs, which potentially contain corresponding positions between them. These two half-site models are separated by a no-cost FEM model that limits the length of any intervening spacer sequence to 20bp. The half-site models are also wrapped by two FIMs that allow the Viterbi algorithm to find the best matching motifs anywhere in the candidate sequences. **(e)** The topology of the p53 cluster-site model consists of a single PHMM that models a general half-site, and two back-transitions that allow for modeling an infinite number of half-sites within the cluster-site. The back-transition through the no-cost FEM-14 model limits the spacer-sequence between the half-sites to lengths $\leq$ 14bp.

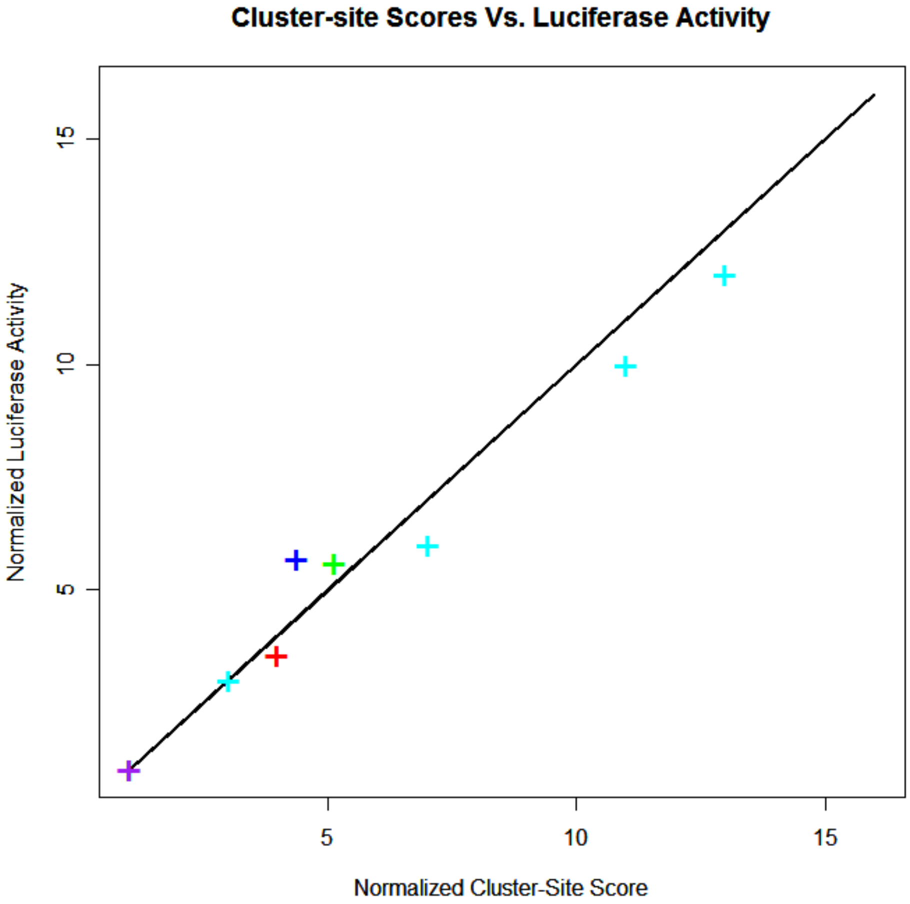


Figure 3.7: **Comparison of Cluster-site scores and Luciferase Activity** This graph compares the estimated relative binding affinity given by the cluster-site score to the luciferase activity from four experiments for four different p53 cluster-sites. The four cluster-sites regulate the genes DDB2 (blue), CKM (red), IGFBP3 (green), and TP53I3 (cyan). In all four experiments the luciferase activity of truncated mutants of the respective p53 cluster-site were compared to the luciferase activity of the full cluster-site. All cluster-site scores and activity measurements are normalized by the full-site (two half-sites) measurement. The cluster-site scores are attained by summing the estimated binding affinity of all viable full-sites in the cluster-site that have an affinity above a lower bound and spacer-lengths below an upper bound. The full-site affinity lower bound and spacer-length upper bound were chosen to best match the experimental data. The best fit was attained by enforcing that spacer-lengths not exceed 14bp and affinity scores exceed 27.5.

3.6 Conclusions

Profile Hidden Markov Models (PHMMs) can boost predictive power over weight matrices (PSSMs) when the binding motif is highly degenerative and tolerates insertions and/or deletions at various positions. The increase in predictive power for the p53-binding motif can be seen in Figures 3.4 and 3.5. When the RE has a known repeated and/or palindromic motif, this prior knowledge can be used to correspond parameters in the model to exploit the redundancy in the information in the motif. We propose a novel "Corresponded Baum-Welch" training algorithm that significantly boosts the predictive power of the p53-RE model, as seen in Figure 3.5. When the motif is not known, all possible motifs for the given size can be sampled and cross-validation techniques leveraged to infer the correct motif that maximizes predictive power. For example, Figure 3.5 reveals that the maximally predictive p53-binding motif corresponds the two half-sites in a palindromic structure.

Our algorithms demonstrate the best predictive capability to date in classifying putative p53-binding sites. One algorithm uses a novel "Corresponded Baum-Welch" training method that exploits the repeated palindromic structure of the p53 motif to train for allowed insertions and deletions relative to the consensus. The second algorithm properly models the relative increase in binding affinity for p53 cluster-sites (REs with ≥ 3 adjacent half-sites) by using a two step process that scores all viable full-sites in the cluster-site while restricting the spacer-length to 14bp. This new cluster-site algorithm best matches the experimental data (see Figure 3.1 on page 79).

The faithfulness within the p53 RE. Analysis of the 37 total p53-binding sites from El-Deiry *et al.*, 1992 and Funk *et al.*, 1992 showed that the left half-site appeared to be more faithful than the right half-site, graphically expressed as $\longrightarrow\longleftarrow_{spacer}\longrightarrow\longleftarrow$. It also appeared that the ⟵ motif was more faithful than the ⟶ motif within each half-site, graphically expressed as $\longrightarrow \longleftarrow {}_{spacer} \longrightarrow \longleftarrow$. However, these differences were not statistically significant [101]. Our findings with the current dataset of 160 p53-binding sites show no significant differences in the faithfulness between the quarter-sites. Additional evidence that the half-sites share the same binding properties is given by the fact that the best computational predictor in this analysis assumes and leverages that the two half-sites share the same binding preferences.

Chapter 4

Dynamic Acceptance Thresholds

4.1 Accepting affinity scores as a function of the distance from the TSS

An interesting finding from the analysis of our dataset of 160 functional p53-binding sites is that the low relative affinity scores from our model are significantly correlated with short distances from the Transcription Start Site (TSS). We find that low affinity sites exist only in a tight band around the TSS (see Figure 4.1 on the following page). Therefore a dynamic binding-affinity acceptance threshold, dependent upon the putative site's distance from the TSS, can greatly reduce the false positive rate of our classifier. With a dynamic acceptance threshold, putative sites will require higher calculated binding affinities as their distance from the TSS increases in order to be accepted as potentially functional.

For example, consider the linear dynamic acceptance threshold $65.16 + .00107\Delta X$ shown in Figure 4.1, with the additional restriction that the putative sites must be within 5,000bp upstream and 1,000bp downstream of the gene. Let the static acceptance threshold be all normalized scores above 70 with the same restriction that the putative sites must be within 5,000bp upstream and 1,000bp downstream of the gene. Even though the restricted dynamic threshold has a false negative rate of 22 out of 158 validated p53 sites (13.9%), and the restricted static threshold 32 out of 158 (20.3%), the restricted static threshold generates over 3.2 times as many positive hits when scoring all 39,288 isoforms of known genes in the human genome (hg18). Thus, the dynamic acceptance threshold has a lower

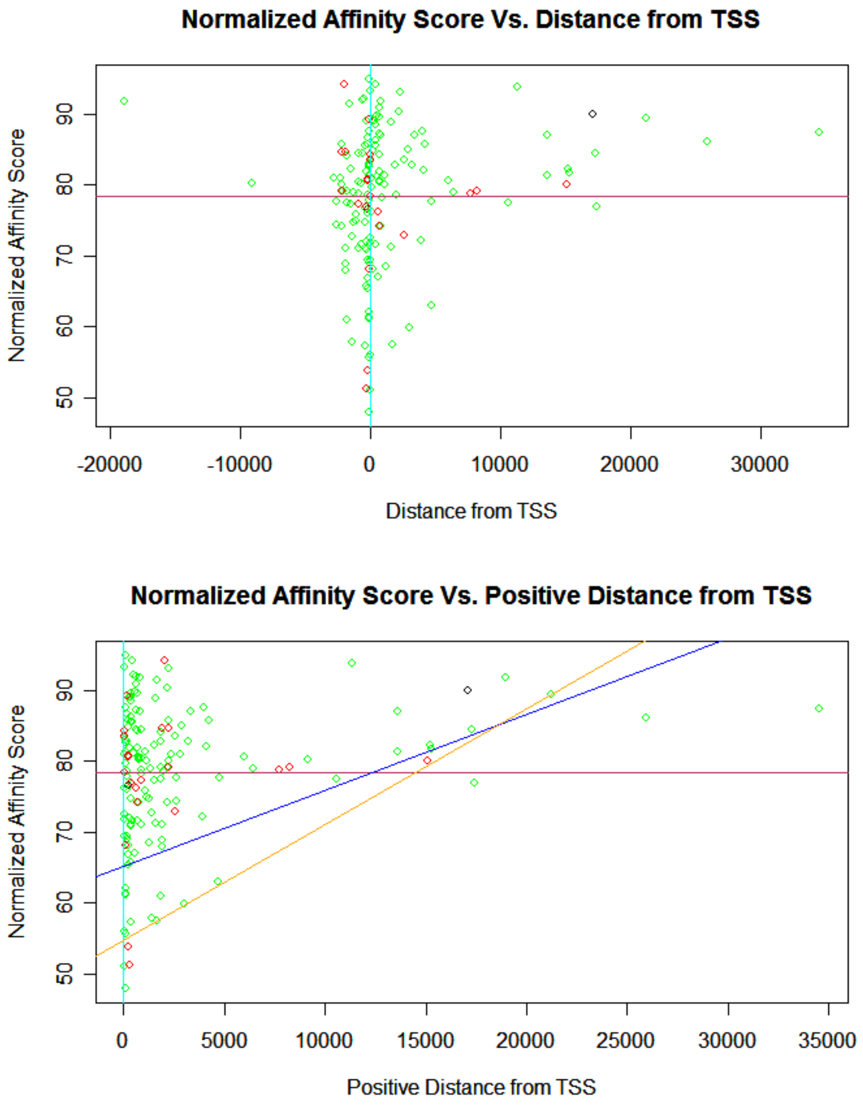


Figure 4.1: **Normalized affinity scores versus distances from the TSS. (Upper)** This plot presents the normalized affinity scores returned from the p53 single-site model versus the distance from the Transcription Start Site (TSS) for 158 experimentally validated p53-binding sites. Low affinity sites exist in a tight band around the TSS (cyan vertical line). p53 activation-sites are plotted in green, repression-sites in red, and both activation and repression in black. All sites $\geq$ 11Kb from the TSS have relative affinity scores above the average of $\approx$ 78 (purple horizontal line). **(Lower)** This plot presents the estimated normalized affinity scores versus the positive distance (absolute value) from the TSS. Two linear dynamic acceptance thresholds are shown for scoring for putative p53-binding sites. The orange threshold corresponds to the formula $54.69 + .00163\Delta X$ and has a false negative rate of 7 out of 158 validated p53 sites (4.4%). The blue threshold corresponds to the formula $65.16 + .00107\Delta X$ and has a false negative rate of 18 out of 158 validated p53 sites (11.4%) (ΔX = distance from TSS).

known false negative rate and a considerably lower false positive rate.

Functional low-affinity p53-sites only exist near the TSS. Therefore the binding affinity threshold for accepting a putative site should be dependent on the putative site's distance from the TSS. By this method, putative sites with relatively low calculated binding affinities that are near the TSS may be accepted, while those sites with equal scores but more distant from the TSS will be rejected. A dynamic threshold, as a function of the distance from the TSS, can greatly reduce the false positive rate when searching for putative p53-sites in genes.

Chapter 5

The Effects of Non-linear DNA conformations

5.1 Introduction

Experiments have been performed to directly or indirectly measure the binding affinities of certain p53-binding sites *in vitro*. One group measured the *in vitro* dissociation constants of 20 known p53 response elements using fluorescence anisotropy [275]. Another group measured the *in vitro* dissociation constants for the functional human p53-RE found in the DDB2 gene and 11 mutants using competitive EMSA (Electrophoretic Mobility Shift Assay) [246]. A third group indirectly measured p53-RE binding affinities through yeast luciferase-transactivation experiments for 22 known p53 response elements and four mutants [110]. Other than different p53 REs at the same positional location, the yeast strains were isogenic. All three groups reported that there was no correlation between their experimental results and relative binding affinity scores from available weight matrix models (PSSMs, PWMs).

Although the palindromic PHMM models more accurately classified known p53-REs compared to PWMs, simple linear models using the PHMM relative-affinity scores still failed to closely match experimental measurements (see Figure 5.1 on the next page). However, an analysis of the PHMM normalized affinity scores versus their distance from the TSS revealed that all the lowest affinity scores (as measured by sequence similarity) only exist within a relatively tight (3kb) band on either side of the TSS (see Figure 4.1 on page 90). It

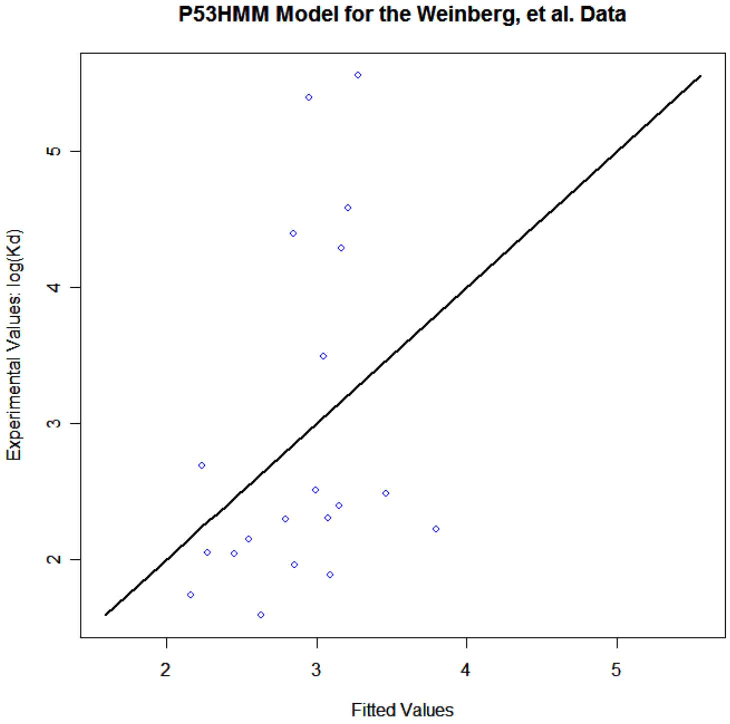


Figure 5.1: **Weinberg experimental measurements versus p53HMM scores.** A plot of the experimental p53 binding affinity measurements (y-axis) versus linear fitted values, where the only predictor variable is the PHMM relative binding affinity score using the palindromic p53 model (x-axis). The adjusted R-squared is 0.0657 and the F-statistic p-value is 0.1438.

appears that there may be some other affinity variable (besides similarity to the consensus site) that may be rescuing these poorly-matching p53 sites near the TSS.

5.2 p53 Binding to Holliday Junctions

Experiments have revealed that the p53 protein is pleiotropic, in that it performs multiple functions in the cell. The primary, well-known role is that of a transcription factor to activate cancer-suppression pathways, such as cell cycle arrest and apoptosis, after DNA damage events. The secondary, less well-known role is to bind to Holliday junctions with high affinity and recruit other proteins (including T4 endonuclease VII and T7 endonuclease

I) to form a complex that resolves (cleaves and separates) the DNA helices [135]. Holliday junctions are produced by homologous recombination events that occur spontaneously or during DNA damage repair. They are potentially lethal DNA structures if they pass unresolved across the G2/M phase check-point. Once in metaphase, during chromosome condensation and segregation, a Holliday junction can lead to chromosome loss, duplication, or breakage in one of the daughter cells.

Lee, et al. used electron microscopy (EM) and gel retardation assays to show that p53 tetramers bind with very high specificity to four-way Holliday junctions (and with $\approx$ 3-fold less affinity to three-way junctions) in a completely structure-dependent and sequence- independent manner [135]. In addition, the locus for binding to Holliday junctions is found in the C-terminal domain of the p53 protein, as opposed to the DNA-binding domain used for transcriptional regulation. Therefore, the p53 protein has two completely different mechanisms for binding to DNA. One DNA-binding mechanism, mediated through the Holliday junction-binding domain, is completely DNA-structure dependent and DNA-sequence independent. The other DNA-binding mechanism, mediated through the DNA-binding domain, appears completely DNA-sequence dependent and DNA-structure independent.

5.3 p53-REs readily form cruciform structures

As mentioned before, the p53-binding site consists of repeated palindromes, possibly separated by a spacer sequence. This motif is ideal for forming stable cruciform (double-hairpin) structures, although the activation barrier to go from linear B-DNA to a non-linear cruciform conformation can be prohibitive. However, we know that strand separation, mediated by enzymes, occurs near the TSS to form the transcription bubble. In addition, we know from experiments that many transcription initiation events fail to produce full-length mRNA transcripts [4]. Therefore, the region of DNA around the TSS of a gene can be a very dynamic environment with frequent strand separation events. Once the two complementary DNA strands separate, the activation barrier has been breached and local hairpin structures are very likely to occur along palindromic motifs, where base-pair complementarity facilitates local, single-stranded alpha-helix formations. Another important observation is that cruciform (double hairpin) DNA conformations locally mimic four-way Holliday junctions, the exact same DNA structure to which p53 binds with high specificity. Once these cruciform DNA conformations form, then the high activation barrier must be breached again in order for the DNA to re-assume a linear B-DNA conformation. Thus these cruciform DNA structures may have long residence times before they are resolved. (In addition, p53 may play a role in resolving these cruciform structures in a fashion similar to how it helps to resolve Holliday junctions.) Therefore, it's possible that DNA conformation may also play an important role in p53 binding to its response elements and regulating its transcriptional target genes.

Recent experiments support this theory by showing that p53 prefers to bind to p53-REs forced into non-linear conformations *in vitro*. Prives, et al. showed that p53 prefers to bind to bent, circularized p53 response elements over linear, B-DNA conformation sites [163]. Even more convincing, the Deppert Lab has shown through EMSA and electron microscopy (EM) that p53 prefers to bind to its response elements when one strand is forced into a hairpin structure [124, 82]. They were able to force one of the strands of the p53-RE into a hairpin (stem-loop) structure *in vitro* by annealing that strand with an opposite strand that is missing the p53-RE sequence. Kim, et. al also showed that p53 can bind to a single p53 half-site sequence with high affinity if that half-site is presented at the end of a hairpin DNA structure [124]. Both labs showed that modifications to the carboxy-terminal domain of the p53 protein greatly affect the preference for non-linear DNA sites versus linear B-DNA p53-binding sites. Experiments suggest that unmodified p53 tetramers prefer to bind to response elements in non-linear, cruciform structures. In fact, the unmodified C-terminal domain appears to hinder binding to linear conformation sites. However, p53 proteins with C-terminal modifications, such as phosphorylation or C-terminal binding to other proteins (like antibody PAb421), prefer to bind to linear B-DNA binding sites. Therefore C-terminal modifications of the p53 protein affect its mode of binding.

5.4 Using UNAFold to estimate p53-RE folding profiles

I used a computational approach to find whether the capability to fold into a cruciform (double hairpin) structure is a factor that affects the overall binding affinity of a p53-RE for the p53 protein. I used the UNAFold folding software to find the probabilities and minimum free energies of different possible cruciform conformations of the 160 known human p53-binding sites. Interestingly, the human p53-RE (single-site) with the highest known binding affinity is also an ideal cruciform folder(see Figure 5.2 on the next page) [275, 110]. This highest affinity site is the $5'$ p53-RE found in the CDKN1A (p21) promoter. Using this CDKN1A p53-RE as the template, we see that there are five stable structures possible. They are labeled as: the linear B-DNA conformation, the full cruciform, the left cruciform, the right cruciform, and the double cruciform (left and right together) (see Figure 5.2 on the following page).

It should be noted that adjacent $5'$ and $3'$ sequence surrounding a p53-RE, and the spacer-sequence, can all have a considerable effect on the folding capability of a p53-site, by either reinforcing a fold, weakening a fold, or affecting the activation barrier to separate the two DNA strands. For example, over-represented poly-A and poly-T signals near the

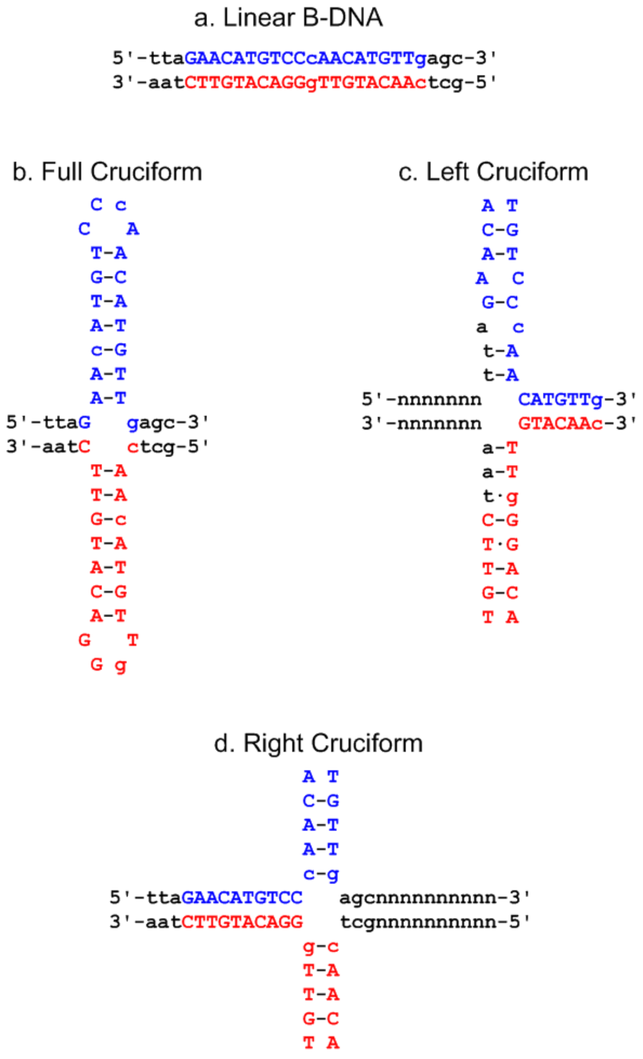


Figure 5.2: **Folding Conformations of a CDKN1A (p21) p53-RE** We present four of the five stable structures of the same CDKN1A p53-RE embedded in a DNA sequence. The coding-strand p53-RE is shown in blue, the template-strand p53-RE in red. Mismatches to the p53-RE consensus (RRRCWWGYYY) are displayed in lower-case. Watson-Crick base pairing is displayed as "-", and wobble base pairing as "·". **a.** The linear B-DNA conformation shows good similarity with only two nucleotide mismatches to the p53 consensus. **b.** The full cruciform structure is very stable with seven Watson-Crick base pairs per strand. **c.** The left cruciform is reinforced by the 5′ sequence that contributes two extra Watson-Crick base pairs per strand for a total of five per strand. The template strand also contains two wobble ($G \cdot T$) base pairs. **d.** The right cruciform structure contains only four Watson-Crick base pairings per hairpin. The double cruciform structure (not shown) is a combination of the left and right cruciforms put together, with possible interactions (stacking) between the stems and loops of the hairpins.

TSS are known to lower the activation barrier to separate the two strands to form the transcription envelope. After computing the folding capacity values, I created simple linear and non-linear models that predicted overall binding affinity based upon two sets of predictor variables: (1) sequence similarity to the consensus motif (the p53HMM score), and (2) the calculated folding probabilities and free energies of assuming any of the p53-preferred cruciform conformations (estimated by UNAFold).

It should be noted that p53-REs that closely match the consensus RRRCWWGYYY will also on average fold better. However, this property is not guaranteed for any one particular p53-RE. For example, the two hypothetical p53-REs GGGCATGCCCGGGCATGCCC and AAACATGCCCAAACATGCCC both match the p53-consensus perfectly. However, the former sequence is by far a better cruciform folder since it contains many possible Watson-Crick base-pairings while the latter possesses very few. Another perfect, hypothetical p53-RE GGGCATGTTTGGGCATGTTT would rank in between those two sequences in capability to cruciform-fold, since it contains many possible wobble base pairings (G-T). In addition, it is not necessary to match the p53 consensus at all in order to be a great cruciform-folder.

UNAFold is an extension of the Zuker (mfold) folding algorithm [155]. The software suite provides a unified format to predict hybridization and secondary structure (folding) of DNA and RNA molecules using equilibrium thermodynamic models and dynamic programming [299, 297, 298, 155, 47]. The methods assume an additive-free energy model where the overall free energy of a fold is the sum of the individual free energy contributions of atomic structures found in the fold (see Figure 5.3 on the next page). Individual free energy contributions include: base-pair stackings, hairpin loop lengths, bulge loop lengths, interior loop lengths, multi-branch loop lengths, and terminal mismatches of stems. The individual free energy contributions for both DNA and RNA structures are fitted to thermodynamic measurements of small DNA and RNA fragments [77, 259, 214]. UNAFold

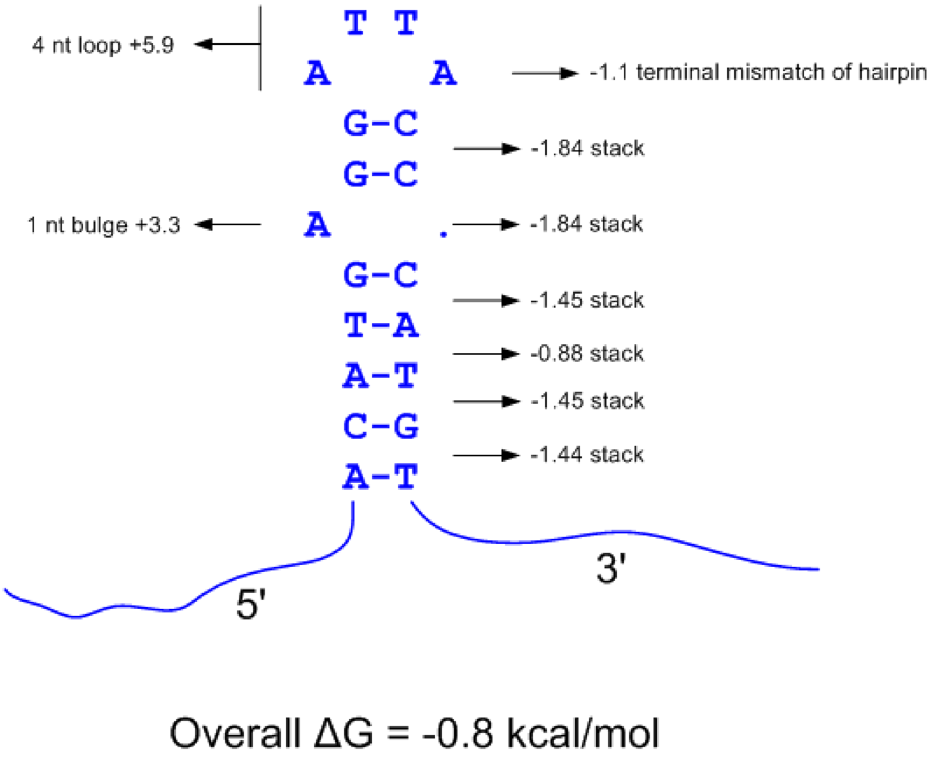


Figure 5.3: **UNAFold Free Energy Calculation** UNAFold uses dynamic programming algorithms to calculate the free energy of DNA and RNA secondary structures using a free-energy additive model. Free energy contributions include base-pair stackings, hairpin loop lengths, bulge loop lengths, interior loop lengths, multi-branch loop lengths, and terminal mismatches of stems. All the free energy parameters for both RNA and DNA calculations have been fitted to thermodynamic measurements of small DNA and RNA fragments. The DNA values shown here are from the SantaLucia tables [214]. The overall free energy of a secondary structure is the sum of the individual contributions. UNAFold is also able to calculate the minimum free energy fold, the partition function, and the probabilities of all possible base-pairings [299, 161, 155].

uses the dynamic-programming Zuker algorithm to efficiently find the minimum free energy structure. In addition, UNAFold uses the McCaskill algorithm to calculate the partition function and probabilities of all possible base pairs. The McCaskill algorithm converts free energies into probabilities by using the Gibbs-Boltzmann equation $e^{-\frac{\Delta G(s)}{RT}}$, and sums the probabilities of all possible structures instead of finding the minimum free energy structure.

Predicting RNA and DNA secondary structure can be approached as a stochastic context-free grammar (SCFG) problem [53]. SCFGs are ideal for modeling a palindromic language and finding nested palindromes in sequences. The RNA and DNA secondary structure problem is essentially equivalent to finding the nested stem-loops (palindromes) that give the minimum free energies, and thus can be viewed through a SCFG-context. The Zuker algorithm can be viewed as a variant of the Cocke-Younger-Kasami (CYK) algorithm that finds the optimum nested stem-loop (palindromic) structure with the minimum free energy. In addition, the McCaskill algorithm can be viewed as a variant of the *inside-outside* algorithm that sums over all possible structures (paths) to calculate the partition function of the sequence and sub-sequences [53]. SCFG models define rules that create nested, long-distance pairwise correlations between terminal symbols. However, SCFGs cannot model a copy language where long-distance pairwise correlations can cross over each other (and are no longer nested) [53]. Therefore, SCFGs (and the Zuker algorithm) cannot model RNA (or DNA) pseudoknots which contain inter-hairpin base-pairing that violates the nesting property [47, 48]. Because of this limitation, the UNAFold algorithm cannot model possible (though unlikely) base pairing between any of the hairpins found in a fold of a p53-RE (see Figure 5.2 on page 97). We can only look at DNA conformations where base pairing is restricted to occur within a particular hairpin.

5.5 Constructing models to predict binding-affinity measurements

I constructed simple linear models to predict the measured binding affinity (or luciferase activity) using the p53HMM score and calculations from folding both strands into four of the possible stable cruciform structures presented in Figure 5.2. We refer to these four possible folds, that are very stable for the functional human p53-RE with the highest known affinity, as the "p53-preferred conformations". I force all of the binding sites into these p53-preferred conformations by first aligning the sequences to the p53 consensus (using a

PHMM) and then using constraints to force base pairings that comply with the preferred fold. We fold both strands (independently) since their folding profiles are not necessarily the same (although they tend to be similar on average.) This difference in the folding profiles between the strands is partially due to the fact that the reverse complement of wobble base pairs (G-T) do not themselves base pair (e.g. G-T can possibly base pair but A-C cannot). I also calculate the probability of a strand folding into any non-linear structure (i.e. forming any base pairing) by taking the minimum of the probabilities of all possible base pairings. When measuring the similarity to the p53-consensus sequence, we need to calculate only one p53HMM score for both strands since the p53HMM algorithm is symmetric (gives the same score for the reverse complement of any putative site). The folding predictor-variables calculated for each DNA strand are: the minimum free energy of all possible folds, the probability of folding into the minimum energy structure, the probability of folding into any non-linear conformation, the probability of folding into each of the four p53-preferred conformations, the free energy of folding into each of the four p53-preferred conformations, the minimum free energy of all four p53-preferred conformations, the sum of the probabilities of folding into the four preferred p53-conformations, and the sum of the probabilities of all possible base-pairings. It should be mentioned that many of these predictor variables can be highly correlated. The goal is to find the smallest subset of predictor variables that generates the best predicting model.

5.6 Results

We refer to the three datasets under analysis as the Weinberg, Inga, and Tan datasets. The Weinberg and Tan datasets are similar in that they both contain dissociation measurements from *in vitro* experiments (fluorescence anisotropy and EMSA, respectively). The Inga dataset is different in that it contains relative luciferase-activity measurements from the *in vivo* context of isogenic yeast strains. Comparisons of the predictive models reveals interesting differences between the two *in vitro* and one *in vivo* datasets.

Analysis of the Weinberg dataset reveals two interesting properties. First, both (1) the calculated probability of the two strands folding into any of the four p53-preferred structures, and (2) the calculated probability of folding into any non-linear structure, were slightly better stand-alone predictor variables than the p53HMM measure of how well the sequence matches the p53 consensus. The model that predicts the measured binding affinity solely on the calculated probability of the two strands folding into any of the four p53-preferred structures has a Multiple R-squared of .1417 and an F-statistic p-value of 0.1019. The model that predicts the measured binding affinity solely on the calculated probability of folding into any non-linear structure has a Multiple R-squared of .1403 and an F-statistic p-value of 0.1037. Finally, the model that predicts the measured binding affinity solely on the p53HMM measure (of how well the sequence matches the p53 consensus) has a Multiple R-squared of .1149 and an F-statistic p-value of 0.1438. These results indicate that in the Weinberg *in vitro* experiments (performed in duplicate), the capability to fold into hairpin structures is a slightly more important determinant of p53 binding affinity than matching the p53 consensus. Second, the fact that the calculated probabilities of different folds are better predictor variables than the free-energy calculations suggests that the calculated equilibrium partition function sufficiently models the *in vitro* folding landscape in the Weinberg experiments. Using a forward step procedure and the AIC criterion for choosing predictor variables, we generate a combined binding-affinity model with the p53HMM score and the 12 best folding variables, with an adjusted R-squared of 0.8385 and an F-statistic p-value of 0.0074 (see Figure 5.4 on the next page).

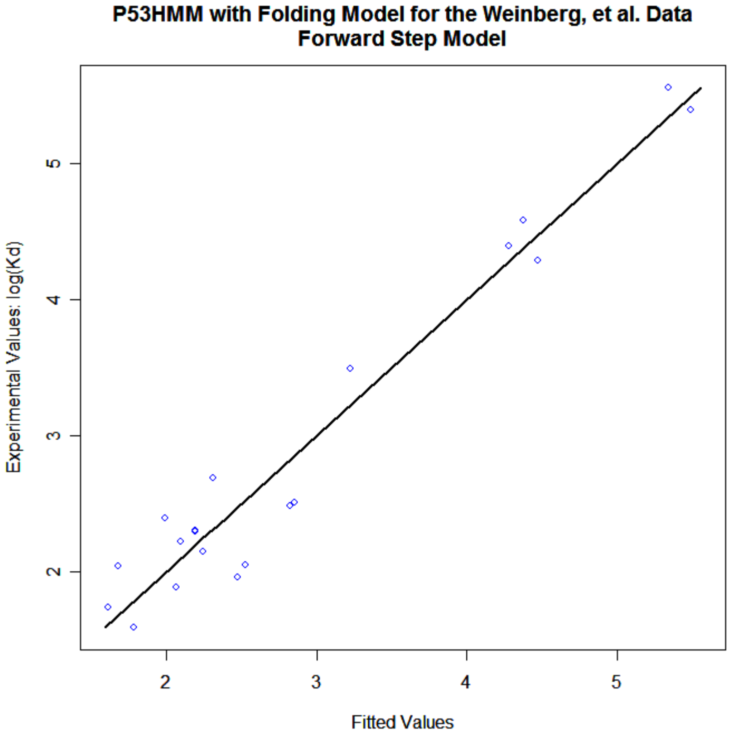


Figure 5.4: **The Weinberg Binding Affinity Model** Using a forward step procedure and the AIC criterion for choosing predictor variables, we generate a combined binding-affinity model with the p53HMM score and the 12 best folding variables. The adjusted R-squared is 0.8385 and an F-statistic p-value is 0.0074. We combine the p53HMM score with the 12 best folding variables to generate a linear binding-affinity model with an adjusted R-squared of 0.8385 and an F-statistic p-value of 0.0074. The folding variables include the calculated probabilities and free-energy calculations for all four p53-preferred cruciform conformations for both strands. The x-axis displays the fitted values while the y-axis displays the experimental measurements.

Analysis of the Tan dataset is even more surprising. All of the effective predictor variables are highly correlated. In addition, similar to the Weinberg results we see that some of the folding variables are better stand-alone predictors than the p53HMM sequence-similarity score in predicting the measured binding affinity. For example, the model that predicts the measured binding affinity solely on the calculated ΔG of the minimum free energy non-linear structure of the template-strand has a Multiple R-squared of .6324 and an F-statistic p-value of 0.001987. The model that predicts the measured binding affinity solely on the p53HMM measure (of how well the sequence matches the p53 consensus) has a Multiple

R-squared of .4369 and an F-statistic p-value of 0.01926. In fact, the best combination of any two predictor variables for predicting the measured dissociation constants are the ΔG calculations of the minimum free energy, non-linear folds of the two separate strands. The fact that the probability calculations were not very effective predictor variables suggests that our calculated equilibrium partition function does not fully match the *in vitro* folding landscape of the experiment. This suggests that the conditions of the Tan experiments were favorable for some of the binding sites to readily assume the minimum free energy, non-linear conformations, and that the affinity scores are mostly driven by the p53-affinity to non-linear structure in a sequence-independent manner. Using a forward step procedure and the AIC criterion for choosing predictor variables, we generate a binding-affinity model with only two structural predictor variables (the free energy calculations of the minimum free energy fold of both strands), with an adjusted R-squared of 0.6203 and an F-statistic p-value of 0.0052 (see Figure 5.5 on the following page).

Analysis of the *in vivo* luciferase (Inga) dataset shows surprising differences in comparison to the two *in vitro* experiments. First, the two best binding-affinity predictor variables were the free energy calculations of both strands in the right cruciform conformation (combined Multiple R-squared of .5257 and F-statistic p-value of 0.0001297). The model that predicts the measured binding affinity solely on the p53HMM measure (of how well the sequence matches the p53 consensus) has a Multiple R-squared of .2255 and an F-statistic p-value of 0.01233. This suggests two interesting properties of the Inga experiments. First, the folding properties of the two hairpins that make up the right palindrome (nearest the TSS) are most indicative of the measured luciferase activity. Although the reason for the disproportionate dependence on the folding profile of the right palindrome is not known, the reason may be that strand separation did not readily extend to include the left palindrome.

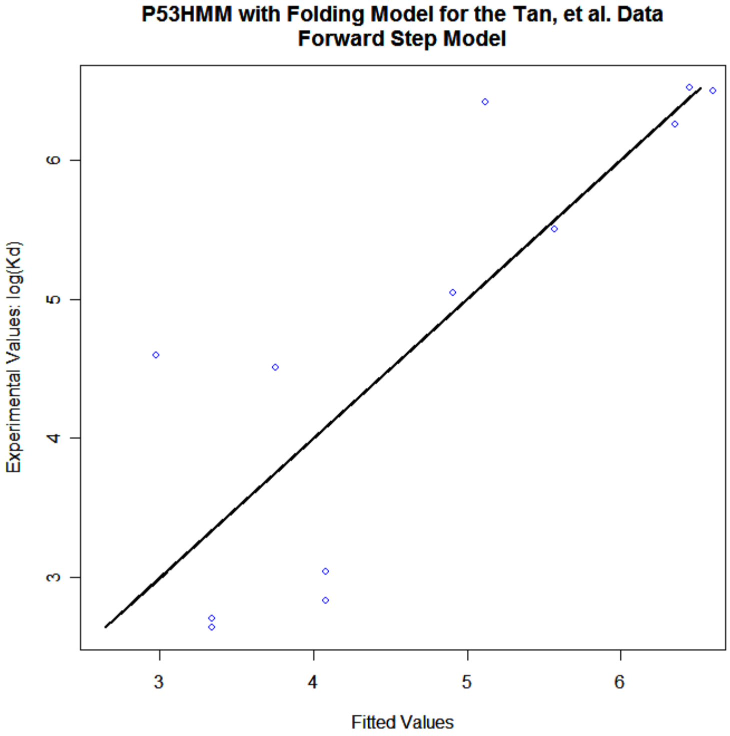


Figure 5.5: **The Tan Binding Affinity Model** Using a forward step procedure and the AIC criterion for choosing predictor variables, we generate a binding-affinity model with only two structural predictor variables (the free energy calculations of the minimum free energy fold of both strands), with an adjusted R-squared of 0.6203 and an F-statistic p-value of 0.0052 It's clear that the model is less accurate in predicting the low-affinity sites. The x-axis displays the fitted values while the y-axis displays the experimental measurements.

Second, the fact that the probability calculations for forming the right cruciform were not nearly as effective in predicting binding affinity as the free-energy calculations, suggests that our calculated equilibrium partition function does not completely model the *in vivo* folding landscape of the experiment. This is not surprising, since forming the transcription envelope, and transcription itself, *in vivo* are enzymatic and energy-driven processes (not equilibrium processes). Using a forward step procedure and the AIC criterion for choosing predictor variables, we generate a combined binding-affinity model with the p53HMM score and the 10 best folding variables, with an adjusted R-squared of 0.799 and an F-statistic p-value of $3.818e-05$ (see Figure 5.6 on the next page). The one slight outlier in the Inga

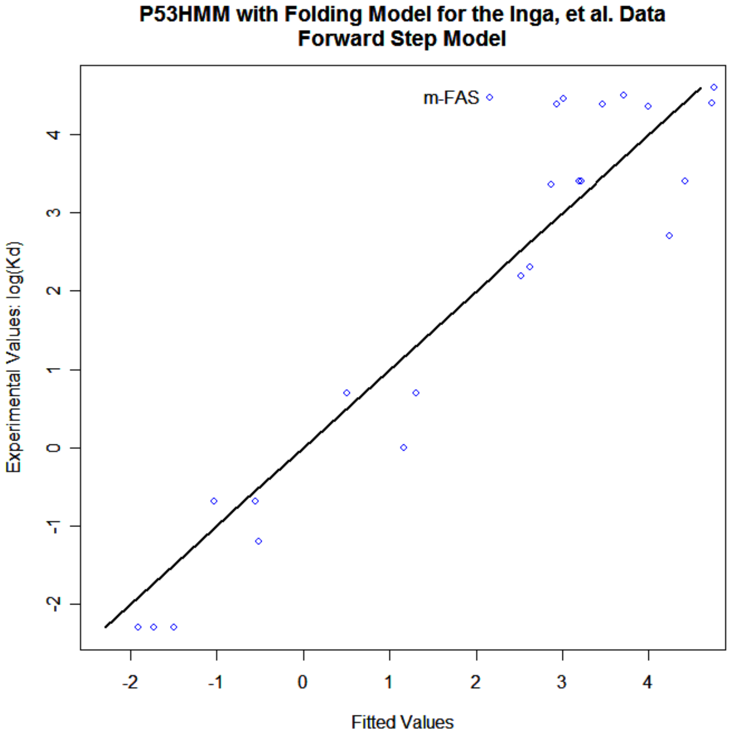


Figure 5.6: **The Inga Binding Affinity Model** Using a forward step procedure and the AIC criterion for choosing predictor variables, we generate a combined binding-affinity model with the p53HMM score and the 10 best folding variables (with an adjusted R-squared of 0.799 and an F-statistic p-value of $3.818e-05$). The folding variables include the calculated probabilities and free energy calculations for all four p53-preferred cruciform conformations for both strands. The one slight outlier in the Inga dataset is the m-FAS p53-RE. It's clear that the model is less accurate in predicting the high-affinity sites. The x-axis displays the fitted values while the y-axis displays the experimental measurements.

dataset is the m-FAS p53-RE, which exhibits luciferase activity higher than predicted. The accuracy of the model is somewhat surprising, considering the inherent noisiness found in luciferase expression assays.

5.7 Variable Selection

Our goal with variable selection is to find the "best", or at least a "good", subset of predictors in order to explain the binding-affinity measurements in the simplest way. In the process we wish to remove un-necessary or redundant predictors. If there are p predictor variables, then in order to find the "best" (optimum) model we would need to fit all the possible 2^p models and choose the best one according to some criteria. However, to generate our models for the three datasets, we utilized a forward step criterion-based procedure. The forward step method is a greedy procedure that iteratively adds predictor variables that minimize the criterion at each step. This method is not guaranteed to find the optimal set of predictor variables of any set size except one. However, the method tends to find "good" sets of predictors of any given size. Two popular criteria are the following:

$$\begin{aligned}
\text{Akaike Information Criterion (AIC)} &= n \log \frac{RSS}{n} + 2p \\
\text{Bayes Information Criterion (BIC)} &= n \log \frac{RSS}{n} + p \log n \\
RSS &= \text{Residual Sum of Squares} \\
n &= \text{Number of observations} \\
p &= \text{Number of parameters}
\end{aligned} \tag{5.1}$$

We wish to minimize the AIC or BIC at every step. Notice that as models grow they will fit better and so have a smaller RSS, but they will also have more parameters. Therefore the iterative process will stop when the model has the proper balance of fit and model size. The BIC penalizes larger models more heavily, and so will tend to generate smaller models compared to those generated by the AIC criterion. We used the AIC criterion to generate all our models.

5.8 Conclusion

By taking into account the capability of the p53 tetramer to bind to non-linear, cruciform DNA structures, I was able to create simple linear models with a high degree of accuracy in predicting experimental measurements. The differences between the models are indicative of the different *in vitro* and *in vivo* conditions of the experiments. The Weinberg and Inga models depend almost equally on both the degree of matching the p53 consensus sequence and the degree of folding into cruciform structures. However, the Tan model depends only on the degree of folding into cruciform structures. Similarity to the p53 consensus sequence has little bearing on the binding affinity measurements performed by Tan, et. al. Interestingly, the *in vivo* Inga model exhibits a very strong dependence on the folding capacities of the two strands of just the right half-site (the palindrome that is nearest to the TSS). It's possible that in this experimental setting of gene transcription in yeast, the strand separation did not (or could not) readily extend to allow for possible hairpin-folding of the left half-site. Analysis of the two *in vitro* models shows no preference for the folding profile of either half-site.

Unfortunately, our equilibrium partition function calculation may not always accurately model the folding landscape of either *in vitro* or *in vivo* conditions. Further experimentation is required to measure the actual amounts and probabilities of non-linear conformations in both *in vitro* and *in vivo* conditions. Then we can build better predictive models that can more properly model the folding landscapes and more accurately estimate binding affinities.

Chapter 6

Searching for Nearby Motifs

6.1 Introduction

Chromatin Immunoprecipitation microarray (ChIP-Chip) analysis has shown that not all putative DNA binding-sites that are found through sequence analysis can be bound by their corresponding transcription-factor proteins [273]. In fact, ChIP-chip analysis has shown that a small percentage of putative binding sites are found to bind to transcription factors for any given cell type [273]. A popular theory explaining this phenomenon is that most putative sites are not accessible to the TF-proteins due to nucleosome positioning and chromatin silencing [4]. In addition, TF-factors can be occluded from binding sites through competition with other proteins. Another important factor is that TF-binding alone to a promoter region is not sufficient to confer TF-regulation of a gene for a particular TF-protein [41, 252]. There are examples where adjacent co-factor sites are necessary in order to exhibit a p53-response [126, 251]. Also, it has been shown experimentally that the p53 response can change drastically (from activation to repression, and vice versa), dependent upon which co-factor sites are adjacent to the p53 site [126]. Finally, it has been shown that adjacent co-factor sites that effectively mediate DNA-looping can confer “functional-ness” to p53-binding sites very far from the transcription start site (TSS) of the regulated gene.

Due to these observations, it is apparent that the surrounding sequence of a putative p53-binding site is an important variable in determining whether a putative site can be functional, and the form and degree of that function (activating or repressing). Therefore, sequence analysis of the DNA segments around the set of 160 experimentally validated p53-binding sites may shed light into the co-factor binding sites (and other motifs) necessary to confer "functional-ness" of a site. If we can stratify the set of functional p53-binding sites by tissue-type and p53-response (apoptosis, DNA repair, etc.), we may also be able to find the "DNA footprints" that determine these behaviors as well.

The ability to determine "functional-ness" of a putative site is an open and challenging problem. In addition, the ability to discover the shared binding motif in the promoter and intronic regions of genes that are believed to be co-regulated by the same transcription factor is also an open and challenging problem. The difficulty is multi-fold: (1) DNA is not random and in general contains many over-represented motifs [248], (2) the DNA segments containing genes can vary greatly in their GC and AT contents (*isochores*), (3) the promoter, exonic, intronic, and downstream regions in and around genes differ considerably in oligonucleotide content, and (4) enhancer regions of genes can be extremely far from the genes they regulate and the problem of determining functional enhancer regions is still un-solved. Therefore the oligonucleotide content of DNA differs considerably, and is dependent upon nested contexts. In essence, the motif discovery problem boils down to attempting to find a *functional*, over-represented motif (a needle) in a haystack of many *non-functional*, over-represented motifs. Therefore, it is imperative to have a thorough model that correctly captures all of the non-functional motifs found in the background DNA, in order that the functional motif of interest stands out above the background. To properly model background DNA, a 3^{rd} *Order Markov Model* or higher, or a mixture of 2^{nd} *Order Markov Models* have been used (where each mixture member models a certain isochore type in the genome) [248]. All attempts at the *Motif Discovery* problem center on finding over-represented oligonucleotides that are present in the sequences of interest and are similar to each other by some distance metric [44]. The challenge lies in finding the over-represented motifs with respect to the proper expected background distribution of DNA oligonucleotides.

6.2 Motif Discovery Methods

Many algorithms have been proposed in order to find a commonly shared transcription factor binding site (TFBS) within a set of promoter sequences believed to be co-regulated, which is commonly known as the *Motif Discovery Problem* [256]. These sets of putatively co-regulated genes often come from micro-array experiments, where the genes in question share similar mRNA expression profiles [256]. The most common techniques to find shared motifs use iterative *Expectation Maximization* or *Gibbs Sampling* methods in order to maximize the log-likelihood or *information content* of the training sequences (or some other similar measure), where the training sequences are the members of the set of promoter sequences believed to be co-regulated [55]. Most *Motif Discovery* algorithms use PSSMs that assume independence between the nucleotide positions. Newer algorithms use HMM's or Bayesian Networks in attempts to learn possible dependencies between the positions. However, these more complex models require larger datasets in order to train the extra parameters [8].

The many methods proposed for discovering over-represented motifs in biological sequences generally fall into two major categories: probabilistic and deterministic [44]. Probabilistic models estimate probabilities for nucleotides at each position, using Bayesian inference or maximum-likelihood methods. These probabilistic models are not guaranteed to find the optimal solutions, but only locally optimal solutions. Examples of probabilistic motif finders are Gibbs sampling, expectation maximization (EM), and greedy methods.

Deterministic models typically rely on counting and comparing oligonucleotide (word) frequencies through exhaustive enumeration. Deterministic models typically guarantee global optimality, but finding long motifs through exhaustive enumeration has been computationally intractable.

Examples of published deterministic methods are Oligo/Dyad-Analysis, YMF, Moby-Dick, Brazma's regular expression method, Marsan and Sagot's suffix tree method, Weeder (a suffix tree method), MITRA (a suffix tree method), QuickScore, MaMF, MDScan, and WINNOWER [263, 265, 26, 255, 230, 21, 156, 192, 63, 200, 194, 103, 144]. Examples of published probabilistic methods are Consensus, Lawrence and Reilly's EM method, MEME (an EM method), Gibbs sampling, AlignACE (a Gibbs sampling method), MotifSampler (a Gibbs sampling method), BioProspector (a Gibbs sampling method), GibbsST (a Gibbs sampling method), GLAM (a Gibbs sampling method), and NestedMICA [97, 132, 6, 131, 207, 249, 143, 222, 78, 52] (It should be noted that MEME does have an exhaustively enumerative component that can iterate over all possible seeds.) Reviews and comparisons of these methods can be found in [273, 256, 104, 44]. The reviews of these methods revealed that the performance of these algorithms degrades significantly as the length of the analyzed sequences increases. The comparisons also revealed that the best performers were ensemble methods (like EMD) that use multiple algorithms and a voting scheme to leverage the inherent advantages of the different approaches [104, 105, 44].

I have approached the *Motif Discovery* problem around functional p53 binding sites using methods from information theory and statistical physics. In the process, I have developed a method called PURE (Patterns Using Relative Entropy) to find over- and under-represented motifs in different regions of human genes. PURE has a few advantages over current motif-finding methods. The method centers on finding meaningful, over- and under-represented DNA motifs that contribute most to the relative entropy (Kullback-Leibler pseudo-distance) between two real DNA sequences, or between a real DNA sequence and a randomly generated background sequence that preserves certain properties of the real sequence. A major advantage of the method is that by integrating simultaneous searches of both the coding and template strands, I can define measures to classify over- and under-represented motifs as either *strand-independent* or *strand-dependent*.

In most cases, this classification also labels a motif as either *transcriptional* or *post-transcriptional*. A *transcriptional* (strand-independent) motif is defined as being found with relatively equal frequency on both strands, while a *post-transcriptional* (strand-dependent) motif is found disproportionately only on the coding strand. For example, transcriptional motifs include TF-binding sites and fixed nucleosome positions; while post-transcriptional motifs include splice sites, exonic splice enhancers (ESEs), intronic splice enhancers (ISEs), translational target sites, microRNA binding sites, and $3'$ polyadenylation sites. The above classification almost always applies since our analysis shows that the only known transcriptional motif that appears to exhibit a strand preference is the TATA-box. Our analysis coincides with promoter analysis performed by [74]. However, all known post-transcriptional signals have considerable bias for occurring on the coding strand versus the template strand, unless they happen to be palindromes.

6.3 Using Relative Entropy to find over- and under-represented words

A Relative-Entropy Algorithm has been used to find the over- and under-represented oligonucleotides (words) of the coding regions of bacteria, independent of codon usage [203]. These codon-usage independent words were sufficient in establishing bacterial host-phage relationships [203]. In that analysis, the relative entropy algorithm ranks the oligonucleotides that contribute the most to the relative entropy (*Kullback-Leibler Distance*) between the observed distribution of oligonucleotides in the coding regions and the expected background distribution after sufficient random shuffling of the positions of synonymous codons [203, 204]. A crucial part of the method is the ability to re-scale out the contribution of the most-contributing word to the relative entropy at each step before searching for the next most over- or under-represented word (see Equations 6.15.1 on page 141). This is necessary since the frequency of words of different lengths are not independent (e.g. if TGAC is over-represented, then TGACA will be as well). Another Relative-Entropy algorithm has been used to rank the constraints and dependencies that contribute most to the relative entropy of $5'$ and $3'$ splice sites compared to an expected background distribution [282].

6.4 The PURE method: using Relative Entropy to find motifs

Motivated by the success of the relative-entropy approach, I augmented the approaches above by adding methods to systematically agglomerate over- and under-represented words into binding motifs, and methods that use information from both strands to maximize the ability to learn different types of DNA motifs.

The difference between two nucleotide sequences can be measured by their relative entropy, also known as Kullback-Leibler (KL) pseudo-distance D_{KL} (see Appendix). One of the nucleotide sequences is real DNA sequence from which we wish to find over- and under-represented motifs. The other nucleotide sequence is either a randomized version of real DNA or some different, real DNA sequence that represents background DNA. A measure that gives the contribution $S(w)$ of a word w to the overall Kullback-Leibler pseudo-distance between the two distributions is:

$$S(w) \quad = \quad P_R(w) \log \frac{P_R(w)}{P_B(w)} + [1 - P_R(w)] \log \left(\frac{1 - P_R(w)}{1 - P_B(w)} \right) \qquad (6.1)$$

where $P_R(w)$ and $P_B(w)$ represent the probabilities of w in the real and background distributions, respectively. $S(w)$ can be interpreted as the Kullback-Leibler pseudo-distance between the coarse-grained real and background distributions, where all we know is that a given word in the distribution is w or not. The advantageous property of $S(w)$ is that, unlike other methods, it allows for a fair comparison of words of different lengths [203, 204].

The method from Ref [203] finds the most significant word w_{max}, defined as that one with the highest $S(w)$. An essential part of the algorithm is that after the most significant word w_{max} is found, the background distribution is rescaled so that $S(w_{max}) = 0$. Then the whole procedure is repeated to find the next most-significant word, and the procedure is iterated. The rescaling is necessary since the frequencies of words of different lengths are not independent. For example, if TGAC is functional and over-represented, then (before rescaling) TGACA and GAC will also likely be over-represented, although neither may be functional (they are just being dragged in by the functional TGAC word). In such a case, after rescaling the relative entropy, contributions of TGACA and GAC will become negligible if they aren't over- or under-represented outside of the context of TGAC.

Experiments have shown that proteins and microRNAs typically do not bind to just one oligosequence, but usually binds to a family of similar oligosequences that form a motif. Often, the binding specificity of a protein or microRNA is represented using a weight matrix (PSSM or PWM) that quantifies the binding specificity for each position in the motif and assumes binding affinity independence between positions [239]. If we know that the binding motif of a TF-protein consists of only two words w_1 and w_2 of the same length, then we can represent the relative entropy contribution $S(m)$ of this motif $\{w_1, w_2\}$ as:

$$S(m) = S(w_1, w_2) \quad = \quad P_R(w_1) \log \frac{P_R(w_1)}{P_B(w_1)} + P_R(w_2) \log \frac{P_R(w_2)}{P_B(w_2)} + [1 - P_R(w_1) - P_R(w_2)] \log \left(\frac{1 - P_R(w_1) - P_R(w_2)}{1 - P_B(w_1) - P_B(w_2)} \right) \quad (6.2)$$

Similarly, we can calculate the relative entropy contribution $S(m)$ for any motif m that consists of a set of different words $\{w_1, w_2, ..., w_N\}$ of the same length:

$$S(m) = S(w_1, w_2, ..., w_N) \quad = \quad \sum_{i=1}^{N} \left(P_R(w_i) \log \frac{P_R(w_i)}{P_B(w_i)} \right) + \left[1 - \sum_{i=1}^{N} P_R(w_i) \right] \log \left(\frac{1 - \sum_{i=1}^{N} P_R(w_i)}{1 - \sum_{i=1}^{N} P_B(w_i)} \right) \quad (6.3)$$

Similar to equation (6.1), we can interpret equation (6.3) as the Kullback-Leibler pseudo-distance between the coarse-grained real and background distributions, where all we know is that a given word is in the motif m or not. If we wish to find the set of words $\{w_i\}$ that constitute an over- or under-represented motif m, then our goal is to find the set of words $\{w_i\}$ that maximize $S(m)$ with the constraint that $d(w_j, w_k) \leq C$ (a constant) for all w_j and w_k in the motif m, where $d(w_j, w_k)$ is a distance measure. This approach is central to all the methods used by PURE to find different types of DNA motifs.

In order to adequately measure the sequence-similarity distance between motifs and words, we associate a position-specific scoring matrix (PSSM, weight matrix, or PWM) to each motif $m = \{w_i\}$ and to each single word w. These PSSMs provide a statistical summary of the per-position nucleotide frequencies over all the words in a motif m, and are referred to as the $PSSM(m)$. A PSSM consists of a $4 \times N$ matrix where each column represents the probability distribution $P_m(n_i)$ at motif position $i = 1...N$ and where $n_i \in \{A, C, G, T\}$. The emission probabilities are assumed to be independent at each position, so that for any sequence $n_1 n_2 ... n_N$ the probability $P_m(n_1 n_2 ... n_N)$ of seeing the sequence in the $PSSM(m)$ is given by $P_m(n_1 n_2 ... n_N) = \prod_{i=1}^{N} P_m(n_i)$.

PURE uses a weighted, normalized Jensen-Shannon divergence (JS-entropy) as the distance measure between two PSSMs p and q that are associated with two motifs m_1 and m_2, respectively. The JS-entropy assumes that if the PSSMs p and q are similar, then they will be close to their weighted average. I normalize the JS-entropy by the length of the two motifs:

$$d(p,q) = \left[\frac{wt_1 \cdot D_{KL}(p||wt_1 \cdot p + wt_2 \cdot q) + wt_2 \cdot D_{KL}(q||wt_1 \cdot p + wt_2 \cdot q)}{N}\right] \tag{6.4}$$

where N is the number of columns in the PSSMs, and wt_1 and wt_2 are the weights that represent the combined probability (or occurrences) of seeing any of the words $w_i \in m_1$ or $w_j \in m_2$, respectively, over their combined sum (in the foreground sequences):

$$\begin{aligned}
wt_1 &= \frac{\sum\limits_{w_i \in m_1} P_R(w_i)}{\sum\limits_{w_i \in m_1} P_R(w_i) + \sum\limits_{w_j \in m_2} P_R(w_j)} \\
&= \frac{\sum\limits_{w_i \in m_1} N_R(w_i)}{\sum\limits_{w_i \in m_1} N_R(w_i) + \sum\limits_{w_j \in m_2} N_R(w_j)} \\
wt_2 &= \frac{\sum\limits_{w_j \in m_2} P_R(w_j)}{\sum\limits_{w_i \in m_1} P_R(w_i) + \sum\limits_{w_j \in m_2} P_R(w_j)} \\
&= \frac{\sum\limits_{w_j \in m_2} N_R(w_j)}{\sum\limits_{w_i \in m_1} N_R(w_i) + \sum\limits_{w_j \in m_2} N_R(w_j)} \\
N_R(w) &= \text{Number of occurrences of the word } w \text{ in the real sequences} \qquad (6.5)
\end{aligned}$$

In equation (6.4), $wt_1 \cdot p + wt_2 \cdot q$ represents the weighted average PSSM of p and q. Since we have that for any sequence $n_1 n_2 ... n_N$ the probability $P(n_1 n_2 ... n_N)$ of seeing the sequence in a $PSSM(m)$ is given by $P_m(n_1 n_2 ... n_N) = \prod_{i=1}^{N} P_m(n_i)$, then we have that:

$$\begin{aligned} D_{KL}(s||t) &= \sum_{i=1}^{N} \sum_{n^j}^{\{A,C,G,T\}} P_s(n_i^j) \log\left(\frac{P_s(n_i^j)}{P_t(n_i^j)}\right) \\ (wt_1 \cdot p + wt_2 \cdot q)_i^j &= wt_1 \cdot P_p(n_i^j) + wt_2 \cdot P_q(n_i^j) \quad \forall i = 1..N, \forall n^j \in \{A, C, G, T\} \qquad (6.6) \end{aligned}$$

The distance measure in PURE is similar to the ones used by MEDUSA and the agglomerative information bottleneck algorithm [165, 231]. A single word w can be associated to a $PSSM(w)$ by using 0 or 1 emission probabilities. If we believe that a motif m is an incomplete sampling of all the words that should be in the motif, then we can add pseudocounts as prior knowledge from the background distribution to make the $PSSM(m)$ less stringent. When two similarly over- or under-represented motifs m_1 and m_2 are close in distance $(d(PSSM(m_1), PSSM(m_2)) \leq C)$, then they are merged to form a new motif $m = m_1 \cup m_2$ that consists of the union of the words in m_1 and m_2. The new motif m has the combined relative entropy contribution $S(m_1, m_2)$, and the associated weight matrix $PSSM(m) = wt_1 \cdot PSSM(m_1) + wt_2 \cdot PSSM(m_2)$, which is the weighted average motif.

6.5 Finding Transcriptional (strand-independent) signals

PURE takes advantage of the fact that human DNA is double-stranded, where one strand is the reverse complement of the other. By analyzing both strands, PURE is able to use this apparently redundant information in order to boost the ability to find meaningful biological motifs. It has been previously noticed that TF-biding sites can be found on either strand [4]. Since experiments have shown that TF-proteins bind to double-stranded DNA (possibly in the presence of chromatin), this implies that most known TF-proteins do not exhibit a $5'$ to $3'$ or $3'$ to $5'$ directional bias with regards to its binding motif [4]. Our analysis of the promoter regions of all known isoforms of human genes suggests that known human TF-binding sites occur with equal over-representation on both strands (see Table 6.1 on page 144). The only known exception is the TATA-box, which is on average more over-represented on the coding strand. Therefore, in the case of transcriptional (strand-independent) signals, the over-represented word w and its reverse complement $\overline{w}$ are in fact the same motif. Furthermore, the combined tuple $(w, \overline{w})$ should be over- or under-represented if there is no biological pressure to prefer one over the other.

PURE is able to take advantage of this additional information by searching for tuples $(w, \overline{w})$ that contribute most to the relative entropy contribution of a motif m. Remember that a word w can be either over-represented or under-represented relative to the expected background, and that the word's relative entropy contribution is a positive measure of the degree of over- or under-representation. (A word w that is neither over- nor under-represented (i.e. $P_R(w) = P_B(w)$) has a relative entropy contribution of 0.) Let us define:

$$\begin{aligned}
w &= \text{word on the coding strand} \\
\overline{w} &= \text{template-strand version of } w \text{ (reverse complement of } w) \\
m &= \{w_1, \overline{w}_1, w_2, \overline{w}_2, ..., w_N, \overline{w}_N\} \\
W_{over} &= \{w \in m \mid w \text{ is over-represented}\} \\
\overline{W}_{over} &= \{\overline{w} \in m \mid \overline{w} \text{ is over-represented}\} \\
W_{under} &= \{w \in m \mid w \text{ is under-represented}\} \\
\overline{W}_{under} &= \{\overline{w} \in m \mid \overline{w} \text{ is under-represented}\} \qquad (6.7)
\end{aligned}$$

Assuming statistical independence, word probabilities are multiplicative and entropies are additive. PURE uses these properties to define a new measure S^{over}_{txn} to find the transcriptional (strand-independent), over-represented motifs that are found on both strands:

$$S^{over}_{txn}(m) \quad = \quad S(W_{over} \cup \overline{W}_{over}) - S(W_{under} \cup \overline{W}_{under}) \tag{6.8}$$

Similarly, PURE defines a new measure S^{under}_{txn} to find the transcriptional (strand-independent), under-represented motifs that are absent from both strands:

$$S^{under}_{txn}(m) \quad = \quad S(W_{under} \cup \overline{W}_{under}) - S(W_{over} \cup \overline{W}_{over}) \tag{6.9}$$

6.6 Finding Post-transcriptional (strand-dependent) signals

Transcription of DNA results in a single-stranded messenger RNA (mRNA) or non-coding RNA (ncRNA) molecule. (Non-coding RNAs include tRNAs, microRNAs, ribosomal RNAs, etc.) Therefore, there is evolutionary pressure to maintain post-transcriptional signals on the coding strand, but not on the template strand. Thus, we expect that post-transcriptional signals will be over-represented on the coding strand with respect to both the template strand and the background DNA. Similar to finding transcriptional motifs, PURE uses the additional information found on the template strand to define two new measures $S^{over}_{post-txn}(m)$ and $S^{under}_{post-txn}(m)$ to find the over- and under-represented coding-strand motifs relative to both the template strand and background DNA, respectively:

$$\begin{aligned} S^{over}_{post-txn}(m) &= S(W_{over} \cup \overline{W}_{under}) - S(W_{under} \cup \overline{W}_{over}) \\ S^{under}_{post-txn}(m) &= S(W_{under} \cup \overline{W}_{over}) - S(W_{over} \cup \overline{W}_{under}) \end{aligned} \tag{6.10}$$

Some transcriptional signals are very strong and can have both relatively high transcriptional and relatively high post-transcriptional entropy contributions. This can happen if the coding-strand word w and the template-strand word $\overline{w}$ are both significantly over-represented, and w is more over-represented than $\overline{w}$. In these cases, post-transcriptional signals can be "drowned-out" by the transcriptional signals. PURE uses an entropy measure $R_{post-txn}(m)$ to find relatively weak post-transcriptional signals by finding the percentage contribution of the post-transcriptional measure in the sum of the transcriptional and post-transcriptional measures:

$$\begin{aligned} R^{over}_{post-txn}(m) &= \frac{S^{over}_{post-txn}(m)}{S^{over}_{txn}(m) + S^{over}_{post-txn}(m)} \\ R^{under}_{post-txn}(m) &= \frac{S^{under}_{post-txn}(m)}{S^{under}_{txn}(m) + S^{under}_{post-txn}(m)} \end{aligned} \tag{6.11}$$

with the additional constraint that $S_{post-txn}(w_i) \leq C$ (a constant) for all w_i in the motif m. The entropy measure $R_{post-txn}(m)$ is used to find the motifs that have the highest percentage of post-transcriptional entropy contribution relative to their combined transcriptional and post-transcriptional entropy contributions, independent of how the entropy contributions compare to others from words not in the motif. This entropy percentage measure has the property that $0 \leq R_{post-txn}(m) \leq 1$. The entropy measure $R_{post-txn}(m)$ is extremely useful in finding weak post-transcriptional motifs such as ESEs and ISEs.

6.7 The PURE Algorithm

PURE iterates through a two step process. The first step is to find the seed tuple $(w_{max}, \overline{w}_{max})$ that maximizes the particular entropy contribution measure $S(m)$ or $R(m)$ of interest. In the second step, PURE iterates through all possible tuples $(w, \overline{w})$ to find the words

$\{w_1, \overline{w}_1, w_2, \overline{w}_2, ..., w_N, \overline{w}_N\}$ that maximally increase the entropy contribution measure $S(m)$ or $R(m)$ at each step, until the increase in the entropy contribution measure is below a certain threshold. The additional constraint is that the newly added word w_{n+1} and the current motif $m = \{w_1, w_2, ..., w_n\}$ must be adequately similar, as determined by their normalized, weighted JS-entropy distance. After step 2, the newly discovered motif is compared to the previously inferred motifs (using the normalized, weighted JS-entropy distance) to weed-out duplicates. Step 1 is repeated again with the next highest seed tuple $(w_{max}, \overline{w}_{max})$ that maximizes the particular entropy contribution measure $S(m)$ or $R(m)$ of interest, after we re-scale out the last seed tuple from the background distribution (so that $S(w_{max}, \overline{w}_{max}) = 0$). This entire process is repeated until the entropy contribution of the seed tuple $(w_{max}, \overline{w}_{max})$ is below a given threshold, or all seed tuples are exhausted. PURE returns a list of the discovered over- or under-represented motifs, sorted by their entropy contribution measures $S(m)$ or $R(m)$, and the words included in each motif. (See Appendix for further details.)

6.8 Properties of the $S_{txn}(m)$ and $S_{post-txn}(m)$ Measures

By looking at the entropy contribution measures $S^{over}_{txn}(w, \overline{w})$ and $S^{under}_{txn}(w, \overline{w})$ for a single tuple $(w, \overline{w})$, we see some obvious but interesting properties:

$$\begin{aligned} S^{over}_{txn}(w, \overline{w}) &= -S^{under}_{txn}(w, \overline{w}) && (6.12) \\ S_{txn}(w, \overline{w}) &= S_{txn}(\overline{w}, w) && (6.13) \\ S_{txn}(w, \overline{w}) &= \text{sgn}(w) \cdot S(w) \quad \text{for} \quad w = \overline{w} && (6.14) \\ \text{where} \quad \text{sgn}(w) &= \text{sgn}\left(\frac{P_R(w)}{P_B(w)} - 1\right) \end{aligned}$$

(i.e., $\mathrm{sgn}(w) = 1$ if w is over-represented, $\mathrm{sgn}(w) = -1$ if w is under-represented, and $\mathrm{sgn}(w) = 0$ if $P_R(w) = P_B(w)$.) Property (6.12) allows us to efficiently find both over- and under-represented transcriptional motifs using the same method. Property (6.13) implies that for any tuple $(w, \overline{w}) \in m$, it does not matter which words we label as the coding-strand word and as the template-strand word when finding either over- or under-represented motifs. Property (6.14) implies that palindromes contribute equally to $S(m)$ and $S_{txn}(m)$ when finding either over- or under-represented motifs.

The entropy contribution measures $S^{over}_{post-txn}(w, \overline{w})$ and $S^{under}_{post-txn}(w, \overline{w})$ for a single tuple $(w, \overline{w})$ also have some obvious but interesting properties:

$$\begin{aligned}
S^{over}_{post-txn}(w, \overline{w}) &= -\, S^{under}_{post-txn}(w, \overline{w}) && (6.15) \\
S_{post-txn}(w, \overline{w}) &= 0 \quad \text{for} \quad w = \overline{w} && (6.16) \\
S_{post-txn}(w, \overline{w}) &= \mathrm{sgn}(w) \cdot S(w) \quad \text{for} \quad P_R(\overline{w}) = P_B(\overline{w}) && (6.17) \\
\text{where} \quad \mathrm{sgn}(w) &= \mathrm{sgn}\left(\frac{P_R(w)}{P_B(w)} - 1\right)
\end{aligned}$$

Property (6.15) allows us to efficiently find both over- and under-represented post-transcriptional motifs using the same method. Property (6.16) implies that we cannot detect palindromes as post-transcriptional (strand-dependent) signals. Property (6.17) implies that if w is a completely post-transcriptional (strand-dependent) signal (i.e., there is no signal on the template strand), then w contributes equally to $S_{post-txn}(w, \overline{w})$ and $S(w)$.

Finally, the entropy contribution ratios $R^{over}_{post-txn}(w, \overline{w})$ and $R^{under}_{post-txn}(w, \overline{w})$ for a single tuple $(w, \overline{w})$ also have some obvious but interesting properties:

$$R^{over}_{post-txn}(w, \overline{w}) \quad = \quad -\, R^{under}_{post-txn}(w, \overline{w}) \tag{6.18}$$

$$R_{post-txn}(w, \overline{w}) \quad = \quad 0 \quad \text{for} \quad w = \overline{w} \tag{6.19}$$

Property (6.18) allows us to efficiently find both weakly over- and under-represented post-transcriptional motifs using the same method. Property (6.19) implies that we cannot detect palindromes as post-transcriptional (strand-dependent) signals.

6.9 Safe Mode

There exists a *Motif Discovery* scenario that presents a difficult problem to motif-finding algorithms. Consider the hypothetical case where we have ten genes we believe are co-regulated, and we wish to find a commonly shared CIS-regulatory motif. If five of the promoter regions share highly over-represented motifs and the other five promoter regions share some other highly over-represented motifs, then the less over-represented motifs that all ten genes may share may be drowned out. Thus, we wish to discover the over-represented motifs that are maximally over-represented for each promoter region. PURE addresses this optimization problem by having a “safe-mode”, whereby we wish to maximize the relative entropy measure $S(m)$ while also minimizing the variance of $S(m)$ across the different sequences being analyzed:

$$SMS(m) \quad = \quad S(m) \times \frac{1}{[variance_{seqs}(S(m))]^k} \tag{6.20}$$

where k is a constant that controls the level of preference for less variance in the signal across the different sequences. Since $variance_{seqs}(S(m))$ is a fraction, the preference for less variance increases as k increases from 0. Although the safe-mode has advantageous characteristics, it is much more computationally cumbersome to calculate $S(m)$ for every sequence. I have found that the safe-mode is usually not necessary, but can be useful in certain circumstances similar to the hypothetical case above.

6.10 $S(m_1 \cup m_2)$ is Preferable to $S(m_1) + S(m_2)$

If we consider w_1 and w_2 to be similar words that are members of the same motif m, then we wish to merge the two motifs (to their weighted average), and calculate the relative entropy contribution of both words together. If we compare the calculations $S(w_1 \cup w_2)$ and $S(w_1) + S(w_2)$, we see that all the terms in the resulting sum are identical except for the last terms:

$$[1 - P_R(w_1) - P_R(w_2)] \log\left(\frac{1 - P_R(w_1) - P_R(w_2)}{1 - P_B(w_1) - P_B(w_2)}\right) \quad \neq \quad [1 - P_R(w_1)] \log\left(\frac{1 - P_R(w_1)}{1 - P_B(w_1)}\right) + [1 - P_R(w_2)] \log\left(\frac{1 - P_R(w_2)}{1 - P_B(w_2)}\right) \tag{6.21}$$

Where the left-hand side is the last summand of $S(w_1 \cup w_2)$, and the right-hand side contains the last two summands of $S(w_1) + S(w_2)$. Close analysis of the differences reveals that the calculation $S(w_1) + S(w_2)$ is problematic. The problem is that every word w_i should be represented only once in the summation of relative entropy contributions. However, in the calculation $S(w_1) + S(w_2)$ every word w_i which is not w_1 or w_2 is represented twice in the overall summation, which leads to "over-counting". By using induction over sets of words, we see that for any two motifs m_1 and m_2 that we wish to merge, the last terms of the calculations $S(m_1 \cup m_2)$ and $S(m_1) + S(m_2)$ are not equivalent:

$$\left[1 - \sum_{w_i}^{m_1 \cup m_2} P_R(w_i)\right] \log \left(\frac{1 - \sum_{w_i}^{m_1 \cup m_2} P_R(w_i)}{1 - \sum_{w_i}^{m_1 \cup m_2} P_B(w_i)} \right) \neq \left[1 - \sum_{w_i}^{m_1} P_R(w_i)\right] \log \left(\frac{1 - \sum_{w_i}^{m_1} P_R(w_i)}{1 - \sum_{w_i}^{m_1} P_B(w_i)} \right) + \left[1 - \sum_{w_i}^{m_2} P_R(w_i)\right] \log \left(\frac{1 - \sum_{w_i}^{m_2} P_R(w_i)}{1 - \sum_{w_i}^{m_2} P_B(w_i)} \right) \tag{6.22}$$

Similarly, the calculation $S(m_1 \cup m_2)$ is the preferable, exact calculation of $S(m)$ where $m = m_1 \cup m_2$. Furthermore, the calculation $S(m_1) + S(m_2)$ can be considered an inexact approximation to $S(m_1 \cup m_2)$.

6.11 Modeling Background DNA

I have used three different background models in our analysis. The first background model is for non-coding regions and it contains the expected counts of all oligonucleotides, assuming: (1) random DNA, and (2) the mononucleotide content of all the sequences under analysis. This background model is easy to calculate, but has a distinct disadvantage: the over- and under-represented motifs relative to this random background DNA may be common to all DNA segments, or at least common to some larger class of segments than the one of interest. The second background model is for coding regions and it contains the expected counts of all oligonucleotides, assuming: (1) random DNA, and (2) the codon usage of all the coding sequences under analysis. A method that models the codon usage in the background DNA is described in [203], whereby synonymous codons are randomly shuffled while still preserving the amino acid sequence. Both of the background DNA models above are maximum entropy distribution (MED) models subject to different constraints (mononucleotide content (no constraints) and codon-usage) [282, 204].

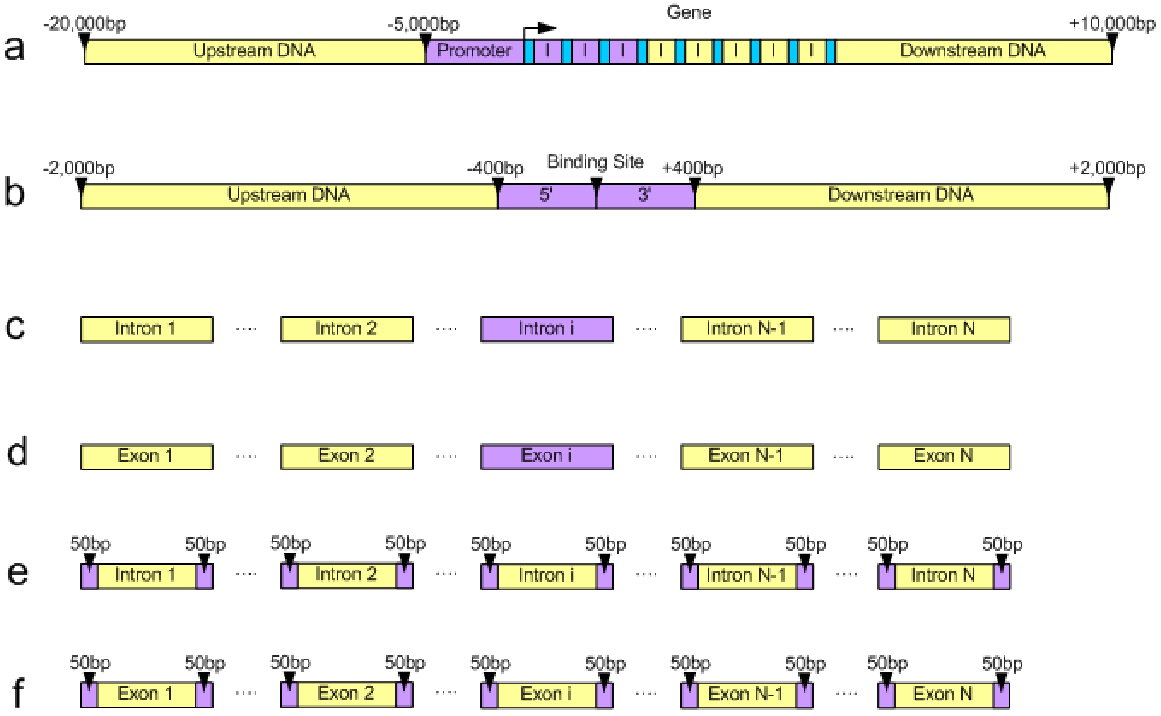


Figure 6.1: **Methods for Modeling Background DNA** Different models for obtaining expected background DNA sequences are presented. The proposed background sequences are shown in yellow. The foreground sequences from which we wish to find over- or under-represented motifs are in purple. Coding sequence is shown in cyan. **a.** This background/foreground (DNA) model is proposed when we wish to find over- or under-represented transcription factor binding sites (TFBSs) in the promoter regions and first three introns of human genes. The foreground/background boundary location can be moved depending on the organism (e.g., humans have longer promoter regions than yeast). **b.** This background/foreground model is used to find under or over-represented co-factor motifs around a set of known, functional TF-binding sites. **c.** This foreground/background model, which avoids coding sequence, is used when we wish to search in a particular intron for motifs. **d.** This foreground/background model, which avoids non-coding sequence, is used when we wish to search a particular exon for motifs. **e.** We use this foreground/background model to find splice-sites and intronic splice enhancers (ISEs) within the first 50 base pairs on the intron-side of exon/intron boundaries. **f.** We use this foreground/background model to find exonic splice enhancers (ESEs) within the first 50 base pairs on the exon-side of exon/intron boundaries.

The third background model uses real DNA from regions different from the numerator DNA, and directly counts the oligonucleotide content. If the goal is to find motifs found in a region of DNA that stand out, versus some other region, then the third method is preferable. For example, when searching for ESE signals in the 50-bp regions at either end of exons, the numerator sequences should be the 50-bp sequences from both ends, and the background DNA should be the exons themselves minus those 50-bp end-regions (see part f of Figure 6.1 on the preceding page). In this example, one would not want to use the mononucleotide-content background, since the list of over-represented motifs would include many motifs that are global to all regions of exons (like codons). Using the codon-shuffled exon background is also problematic, since we wish to find the ESE motifs that stand out versus the interior regions of the exons, not the maximum entropy distribution of all coding regions. I have found that using real background DNA, from either side of the DNA regions of interest, has been more effective in finding known biological motifs. By using real DNA from regions nearby the sequences being analyzed, the background DNA model takes into account regional sequence-biases that are not of interest (e.g., isochore biases). For example, in order to analyze promoter regions of genes, one should use real DNA just upstream of these promoter regions (see part a of Figure 6.1 on the previous page). Additionally, to analyze a certain exon or intron, one should use other exonic or intronic sequences from the same gene or nearby genes (see parts c and d of Figure 6.1 on the preceding page).

6.12 Properties of PURE

PURE has many desirable properties that are advantageous for discovering functional regulatory DNA motifs:

1. PURE assumes a TCM model (Two Component Mixture, zero or more occurrences per sequence), as opposed to a an OOPS model (One Occurrence Per Sequence) or ZOOPS model (Zero or One Occurrence Per Sequence) [6]. The TCM model is generally considered more appropriate for most motif discovery problems [195]. In addition, PURE includes a safe-mode where a parameter k controls a smooth change in

the search bias toward motifs with less variance in their over- or under-representation across the sequences. In effect, PURE allows for a configurable, smooth transition from a TCM model to a OMOPS model (One or More Occurrences Per Sequence).

2. PURE is a combination of a deterministic and a probabilistic model. PURE is deterministic in that it can iterate over all possible motifs to guarantee finding the globally optimal solution. PURE is probabilistic in that it generates probabilistic models (PWMs) of motifs based completely on the relative entropy contributions of the words that have been combined to form the motif.

3. We incorporate a high-order Markov-chain background model of real DNA from DNA regions within or near the sequences of interest. This method helps to find the over- and under-represented motifs relative to similar background sequences that contain general, un-important motifs.

4. PURE analyzes both strands in order to exploit all the available information in double-stranded DNA sequences to find biologically-relevant transcriptional and post-transcriptional motifs.

6.13 Results

In order to test the performance of PURE when applied to the *Motif Discovery Problem* in humans, I attempted to "re-discover" already known motifs in well-studied regions of the human genome: (1) the 50bp region upstream of the start sites of transcription, and (2) the 50bp region in both directions of the exon-intron boundary within genes. PURE found the known motifs in the first 50-bps upstream region of human transcription start sites

(TSSs): the *TATAA*-Box, *BRE*-box, *Poly-T* signal, *Poly-A* signal, *CpG*-signal, *CAAT*-box, *INR*-box, and others (see Table 6.2 on page 146). The method was also successful in finding the motifs known to exist on either side of the exon-intron boundary within human genes: $5'$ and $3'$ splice sites, and $5'$ and $3'$ *exonic splice enhancers* (ESEs) (see Figures 6.3 on page 147, 6.4 on page 149, 6.5 on page 151, and 6.6 on page 153, and Tables 6.3 on page 148, 6.4 on page 150, 6.5 on page 152, and 6.6 on page 153. The $5'$ and $3'$ splice sites are typically strong signals and are relied on heavily by modern gene-finders to classify *Open Reading Frames* (ORFs) [282]. The $5'$ and $3'$ ESEs are typically much weaker signals, but can enhance and rescue the splicing activity at nearby, weak splice sites *in vitro* [65, 66]. Also, *in vivo* experiments have shown that mutations in splice sites and ESEs can produce alternative splicing events that can greatly alter the structure and function of the translated protein [65]. This loss or reduction of protein function by alternative splicing can have severe effects on regulatory pathways, and can lead to disease and cancer [65]. The PURE algorithm also discovered other over-represented motifs in these regions that currently have no known function.

By looking at the reverse complements of the most under- and over-represented words in the 50-bp region upstream from human transcription start sites (TSSs), some very interesting features surfaced. First, the reverse complements of the top under- and over-represented words (all of which were TF-binding sites or *Poly-A* signals) had nearly equal estimated contributions to the relative entropy (see Table 6.1 on page 144). The only exception was the TATAA box, where the coding-strand TATAA signal was about ten times higher. This is in stark contrast to the splice-site and ESE motifs, which are known to be post-transcriptional, mRNA splicing signals. The estimated entropy contributions of the coding-strand splice-site and ESE motifs were hundreds or thousands of times higher than their reverse complement words.

Second, it is well-known that the CpG dinucleotide motif is under-represented in general in the human genome, but over-represented in the promoter regions of genes [273]. These "CpG islands" correlate with the locations of promoters to varying degrees depending on the organism ($\approx 60\%$ of human promoters co-locate with CpG islands) [273]. The prevailing theory for the observed difference in CpG content between intergenic and genic regions is that CpGs are lost over evolutionary time in intergenic regions due to the conversion of CpGs to TpGs by the process of methylation and deamination [4]. However, there appears to be evolutionary pressure to preserve the CpG content in promoters in order to facilitate the regulatory mechanism of gene inhibition through methylation [273]. Our analysis of the promoter regions of human genes suggests a slightly more complicated picture. It is more accurate to say that CpG dinucleotides are over-represented in human promoter regions only when surrounded by other cytidines (Cs) and guanines (Gs) (see Table 6.1 on page 144). In fact, the CpG dinucleotide motif is actually highly under-represented in human promoters when it is accompanied by adjacent adenosines (As) or thymidines (T's) on either side. Therefore, there appear to be two evolutionary pressures at work with respect to the CpG content of promoter regions. The first evolutionary pressure promotes over-representation of CpGs in the local context of surrounding Cs and Gs, presumably to facilitate regulatory gene-inhibition through methylation. The second evolutionary pressure promotes under-representation CpGs in the local context of immediately adjacent As and Ts. A proposed biological explanation for this second evolutionary pressure is found in the innate immune response. In the innate immune response, the toll-like receptor 9 (TLR9) protein recognizes unmethylated $NpWpCpGpWpN$ motifs as foreign DNA and activates the NFKB immunostimmulatory response [260]. Therefore, through high under-representation of the *WpCpGpW* motif, human genes with unmethylated promoters are able to label themselves as "self" and evade the innate immune response. In addition, recent analysis has revealed that viruses that successfully jump to humans quickly evolve to mimic the *WpCpGpW* under-representation found in human genes in order to evade the host's immune system [84].

After validating that the PURE *de novo* motif discovery method could successfully "rediscover" known motifs, I attempted to discover over-represented transcriptional motifs in the regions around experimentally validated p53-binding sites. I searched all 400bp segments centered around the 159 human, experimentally validated p53-sites, while using the regions from 500 to 2000bp on either side of the sites as background DNA (See part b of Figure 6.1 on page 127). In addition to finding the known p53 binding sites in the sequences, PURE was able to discern other over-represented transcriptional motifs (See Figure 6.7 on page 155 and Table 6.7 on page 155). One notable entry is the full-length motif of the GC box which binds SP1, a known co-factor of p53. Thus, the presence of the full SP1 binding motif (GGGCGGG) and its variants can help discern functional p53-binding sites.

In an attempt to validate our *de novo* motifs discovered by PURE, we used MAPPER to perform a known-motif search through the same 400bp-segment regions centered around the 159 human p53-binding sites. MAPPER uses PHMMs to model 1,079 TFBSs, using experimentally determined functional sites provided by the TRANSFAC and JASPAR databases [154, 160, 213]. MAPPER models are built from TFBS sequences from the human, mouse, fly, worm, and yeast genomes [154]. The top three motifs with the highest number of hits when searching for sequence matches to all 1,079 models, with an E-value cut-off of 25, are the human p50, p53, and SP1 binding motifs. The p50-binding site is very similar to the p53-binding motif, but lacks the high specificity at positions four and seven.

Therefore, there is a good correlation between the results from the over-represented *de novo* motif search and the most over-represented known-motif search. From these results we conclude that PURE can find biologically relevant, over- and under-represented transcriptional motifs.

6.14 Future Improvements to PURE

There has been a great deal of research, publications, and algorithms presented to solve the *Motif Discovery Problem* in DNA sequences. Here I present some improvements that can be applied to PURE to: (1) increase predictive performance, and (2) integrate other possibly available data. By integrating data from other sources, the goal is to weed-out the high number of false positives that plague motif-finding algorithms. Analysis has shown that with our current database of ≈ 700 known binding motifs, an exhaustive motif search of human DNA generates at least one hit (of some motif) every several base pairs [273]. It has been shown that many of these putative sites can bind to their respective TF-proteins *in vitro*, but not *in vivo* [257]. Therefore, *in vivo* conditions exist that determine a binding site's "functional-ness", that are not currently considered by our motif-finding methods. In fact, it has been estimated that only 0.1% of putative sites, predicted by models of individual binding sites (like PWMs), are actually functional. (This estimate has been labeled the "futility theorem") [273].)

1. **Integrating Phylogenetic Footprinting.** By assuming that functional motifs are more evolutionarily conserved than non-functional sequences, analyzing orthologous sequences of differing evolutionary distances has been very helpful in finding functional sites [38, 18, 13, 39, 270, 28, 44]. Example algorithms that perform motif discovery using phylogenetic footprinting include CONREAL, CLUSTAL W, PHYLONET, and PhyloScan. In addition, combining phylogenetic footprinting analysis with sequence analysis of co-regulated genes has improved searches for functional motifs even further [80, 162, 121, 195, 174, 228, 229]. These integrated methods combine motif over-representation and cross-species sequence conservation into one probabilistic score

to predict functional sites. Example algorithms of this integrated approach include OrthoMEME, PhyloCon, PhyME, EMnEm, PhyloGibbs, and Stubb. Evidence from these recent successes suggests that adding phylogenetic footprinting capability to PURE should improve its functional motif-finding capabilities as well.

2. **Search for spaced dyad (gapped) motifs.** Many TF-binding sites consist of two conserved half-sites separated by an un-conserved gap (spacer) region. This occurs when the TF-protein binds as a dimer or tetramer, and the complex makes two separate contacts with the DNA sequences. Some spaced dyad motifs have a variable length spacer while others have a fixed length [44]. Since spaced dyad motifs are so prevalent, it makes sense to augment PURE to look for over- and under-representation of those particular motifs. Algorithms that specifically consider the spaced dyad motif include MEDUSA, Oligo/Dyad Analysis, and BioProspector [165, 265, 143] Currently, PURE will recognize the conserved half-sites as separate motifs (or one identical motif if the two half-sites are identical).

3. **Integrating Expression Data.** Levels of mRNA expression from microarry analysis can also be used to help find functional motifs. Algorithms that integrate co-expression profiles with sequence analysis include REDUCE, matrixREDUCE, MEDUSA, MotifRegressor, KIMONO, cis/TF, and FIRE [27, 76, 165, 40, 102, 16, 60].

4. **Model Inserts and Deletions in a Motif.** Weight matrices (PSSMs or PWMs) are able to model the effects of mutations at each position in a motif, assuming independence between positions. However, they are not able to model insertions and deletions of nucleotides within the motif. Some TF-proteins are able to bind to variable length motifs, where deletions and insertions within the motif affect binding specificit but

are still tolerated [202]. Profile Hidden Markov Models (PHMMs) provide a consistent probabilistic model to include the binding-affinity effects of possible insertions and deletions of nucleotides within a motif [127, 55]. By augmenting PURE to utilize PHMMs (instead of PWMs), PURE will be able to find over- and under- represented motifs that model tolerated nucleotide insertions and deletions.

5. **Integrating Positional Clustering of Motifs.** Analysis has shown that functional binding sites tend to occur in tight clusters in the DNA sequence [120, 276, 198, 14, 221]. There is much experimental evidence to support that functional binding sites significantly coincide with regions of DNA with low nucleosome occupancy. These short open windows occur between long DNA segments that are subject to chromatin silencing. Therefore, looking for over-represented motifs that are spatially near other (or the same) over-represented motifs will help reduce the false positive rate of PURE. Examples of algorithms that use positional clustering to improve motif prediction are COMPEL, Stubb, and CREME [120, 229, 221].

6. **Integrating Nucleosome Occupancy Data.** New technologies are allowing for whole or partial genome-wide nucleosome occupancy maps [166]. This additional data can be merged with PURE to further reduce the false positive rates.

7. **Integrating Positional Bias Relative to the TSS.** Many functional binding sites *in vivo* exhibit a positional bias relative to the Transcription Start Site (TSS) [74, 235, 279, 202]. Algorithms that attempt to discover a possible positional bias relative to the TSS (in order to reduce the false positive rate) include: ITB, AlignACE, and PositionAnalysis [123, 207, 264].

6.15 Appendix

6.15.1 Appendix A - Relative Entropy Methods

Let:

$$
\begin{aligned}
W_N &= \text{Set of words of length } N \\
W_N^i &= \text{A word in the set } W_N \\
N_B(W_N^i) &= \text{The background count of the word } W_N^i \\
N_R(W_N^i) &= \text{The observed (real) count of the word } W_N^i \\
P_B(W_N^i) &= \text{The probability of } W_N^i \text{ in the (expected) background} \\
P_R(W_N^i) &= \text{The probability of } W_N^i \text{ in the observed (real) distribution} \\
w &= \text{any word of length 1 to } N \\
N_B(w) &= \text{Approximated background count of the word } w \\
N_R(w) &= \text{Approximated observed (real) count of the word } w \\
P_B(w) &= \text{Approximated probability of } w \text{ in the (expected) background} \\
P_R(w) &= \text{Approximated probability of } w \text{ in the observed (real) distribution} \\
C(W_N^i, w) &= \text{The number of times } w \text{ is contained in } W_N^i \\
L(w) &= \text{The length of the word } w \\
\Gamma &= \text{The overall length of the sequence}
\end{aligned}
$$

Maintaining and repetitively re-scaling the background distribution of all words of length 1 to N is very computationally intensive. In order to achieve gains in execution time and memory usage, we have implemented shortcuts in the algorithm, similar to [203], at the expense of complete accuracy . We actually maintain the background (and foreground) distribution of all words of only length N. Then we closely approximate the distributions of all shorter words with lengths less than N. In order to closely approximate the counts of a word w with length $< N$, we consider all the W_N^is of length N that include it, and there are $(N+1) - L(w)$ such W's, assuming we are not near an edge of a sequence. The time performance gain during re-scaling is considerable, while the accuracy loss do to "edge effects" is minimal. The inaccuracies are introduced at the very beginning and end regions (of length N) that butt against an "edge" of a sequence. The accuracy loss due to edge effects decreases as the length of the separate sequences in the background and foreground increase. In effect, the method presented here to approximate the probability of observing a word shorter than N is completely accurate only in the hypothetical case of infinite sequence length.

Then we have that for any word w of length 1 to N:

$$\begin{aligned}
N_B(w) &= \sum_{i=0}^{4^N-1} \frac{N_B(W_N^i) \times C(W_N^i, w)}{(N+1) - L(w)} \\
N_R(w) &= \sum_{i=0}^{4^N-1} \frac{N_R(W_N^i) \times C(W_N^i, w)}{(N+1) - L(w)} \\
P_B(w) &= N_B(w)/\ \Gamma \\
P_R(w) &= N_R(w)/\ \Gamma
\end{aligned}$$

The Kullback-Leibler pseudo-distance between the observed (real) and background probability distributions (P_R and P_B) is given by:

$$D_{KL}(P_R \| P_B) \quad = \quad \sum_{i=0}^{4^N-1} P_R(W_N^i) \log \frac{P_R(W_N^i)}{P_B(W_N^i)}$$

The Kullback-Leibler divergence gives a distance measure between two distributions. However, it is not a true metric, since it is not symmetric and does not satisfy the triangle inequality:

$$D_{KL}(D_1 \| D_2) \neq D_{KL}(D_2 \| D_1)$$

$$D_{KL}(D_1 \| D_3) \nleq D_{KL}(D_1 \| D_2) + D_{KL}(D_2 \| D_3)$$

However, there is a symmetric version known as Jeffreys' J-Divergence:

$$J_{Div}(D_1, D_2) \quad = \quad D_{KL}(D_1 \| D_2) + D_{KL}(D_2 \| D_1)$$

Both of the versions are always non-negative and are zero if and only if the two distributions are identical:

$$\begin{aligned} D_{KL}(D_1 \| D_2) &\geq 0 \\ J_{Div}(D_1, D_2) &\geq 0 \\ D_{KL}(D_1 \| D_2) = 0 &\iff D_1 = D_2 \\ J_{Div}(D_1, D_2) = 0 &\iff D_1 = D_2 \end{aligned}$$

The Jensen-Shannon divergence gives the mean of the relative entropy of each distribution to the mean distribution of D_1 and D_2:

$$JS(D_1, D_2) \quad = \quad \frac{1}{2}\left(D_{KL}(D_1\|\frac{1}{2}(D_1 + D_2)) + D_{KL}(D_2\|\frac{1}{2}(D_1 + D_2))\right)$$

While both the Kullback-Leibler divergence and Jeffreys' J-divergence range between zero and positive infinity, the Jensen-Shannon divergence ranges from zero and $\ln 2$ (i.e. 1 bit). It can be shown that Jeffrey's J-divergence J_{Div} and the Jensen-Shannon divergence JS are related by the inequality [43]:

$$JS(D_1, D_2) \quad \leq \quad \ln\left(\frac{2}{1 + \exp\left(-\frac{1}{2}J_{Div}(D_1, D_2)\right)}\right)$$

A measure that gives the contribution of a word w, of length 1 to N, to the overall Kullback-Leibler pseudo-distance between the two distributions is:

$$S(w) \quad = \quad P_R(w)\log\frac{P_R(w)}{P_B(w)} + [1 - P_R(w)]\log\left(\frac{1 - P_R(w)}{1 - P_B(w)}\right)$$

$S(w)$ can also be interpreted as the Kullback-Leibler pseudo-distance between the coarse-grained real and background distributions, where all we know is that a given word in the distribution is w or not.

In order to find the most over- and under-represented words, we want to find the words with the highest $S(w)$. An important property of these words is that their contributions $S(w)$ are not independent [203]. For example, if TGAC is over-represented, then TGACA (probably) will be as well. If we believe that only one of the words in an over-represented, nested cluster has biological significance, and that all the other words are being dragged along, then we would want to somehow remove the effects of the significant word from the nested cluster. A proposed method is to re-scale the background after each iteration of finding the $w_{max} = \max(S(w_i))\forall w_i$, and removing the effects of w_{max} on the contributions $S(w_i)$ for all other words w_i [203]. With this method we will attain a sorted list of the words that *independently* contribute most to the Kullback-Leibler pseudo-distance between the two distributions.

Rescaling the background as presented in [203]

We want to remove the contribution of w, $S(w)$, to the D_{KL} by making the contribution of w identical in the observed and background distributions. For the rescaling to be minimal, we re-scale all of the words W_N^i with the same $C(W_N^i, w)$ by the same factor. Therefore, we partition the set W_N into disjoint subsets, where each element in a given subset contains w an equal number of times. The sets are thus defined:

$$\begin{aligned} K_J(w) &= \left\{W_N^i \mid C(W_N^i, w) = J\right\} \quad J = 1, 2, .., N \\ K_0 \cup ... \cup K_N &= W_N \end{aligned}$$

Since we want to re-scale these disjoint subsets so that the probabilities of being in a given subset are equal for both the observed and background distributions, we calculate:

$$\begin{aligned} Q_R(K_J) &= \sum_{W_N^i \in K_J} P_R\left(W_N^i\right) \\ Q_B(K_J) &= \sum_{W_N^i \in K_J} P_B\left(W_N^i\right) \end{aligned}$$

Then we use them to re-scale all the background counts in the containment set K_J:

$$N_B(W_N^i) \longrightarrow \frac{Q_R(K_J)}{Q_B(K_J)} N_B(W_N^i) \qquad \forall \; W_N^i \in K_J$$

Then the contribution of w to the D_{KL} has been removed since now we have:

$$S_{rescaled}(w) \;=\; 0$$

6.15.2 Appendix B - Discovered Motifs using Relative Entropy Methods

Ranked List of the most over/under-represented words in the Human Promoter Regions (-50,0)
against the mononucleotide background and with Re-scaling

Rank	w	+/-	$S(w)$	$\bar{w}$	$S(w) - S(\bar{w})$
1	TTT	+	0.00669622	AAA	0.00136639
2	AAA	+	0.00560457	TTT	0.00560457
3	TCG	-	0.00463774	CGA	0.000162366
4	ACG	-	0.0051526	CGT	0.00171207
5	TA	-	0.00294945	TA	0
6	CGA	-	0.00331754	TCG	0.00326504
7	CGT	-	0.00301892	ACG	0.00297245
8	GGGCGGG	+	0.00225123	CCCGCCC	0.000493133
9	GGT	-	0.00190124	ACC	3.50562e-05
10	ACC	-	0.00213444	GGT	0.00213444
11	CCCGCCC	+	0.00163478	GGGCGGG	0.0016275
12	GGGGCGG	+	0.00157568	CCGCCCC	0.000141837
13	GGCGGGG	+	0.00153286	CCCCGCC	0.000150364
14	ATG	-	0.00151373	CAT	0.000746087
15	CCGCCCC	+	0.0013982	GGGGCGG	0.00139753
16	GGAGG	+	0.00135808	CCTCC	0.000175457
17	CCCCGCC	+	0.00135835	GGCGGGG	0.00135802
18	GCGGCGG	+	0.00133634	CCGCCGC	0.000628948
19	GGCGGCG	+	0.00125388	CGCCGCC	0.000630252
20	CCTCC	+	0.00121439	GGAGG	0.00121436
21	GCGGGGC	+	0.00106887	GCCCCGC	0.000426889
22	CGGCGGC	+	0.00105062	GCCGCCG	0.000553378
23	GGAAG	+	0.000919067	CTTCC	0.000229621
24	GAG	+	0.000896704	CTC	0.000331838
25	ATAAA	+	0.000876449	TTTAT	0.000635312
26	CAA	-	0.000782654	TTG	2.93757e-05
27	TTG	-	0.000838625	CAA	0.000838625
28	ATC	-	0.000788353	GAT	0.000283002
29	CCGCCGC	+	0.000690018	GCGGCGG	0.000689678
30	TCT	+	0.000669254	AGA	0.000495494
31	GCCCCGC	+	0.000632698	GCGGGGC	0.000632593
32	CTTCC	+	0.000631985	GGAAG	0.000631198
33	CGCCGCC	+	0.000626285	GGCGGCG	0.000626282
34	TGACGT	+	0.000600157	ACGTCA	0.000178186
35	GACGTCA	+	0.000523123	TGACGTC	0.000474148
36	GCCGCCG	+	0.000493553	CGGCGGC	0.000493527
37	TTAA	+	0.000460568	TTAA	0
38	AGAA	+	0.000462243	TTCT	0.000348979
39	ACACACA	+	0.00041932	TGTGTGT	0.000205603
40	CTCCCTC	+	0.000390886	GAGGGAG	0.000297069
41	TTATT	+	0.000383781	AATAA	0.000223584
42	GTCCG	-	0.000362307	CGGAC	0.000105473
43	CTCCTC	+	0.000357151	GAGGAG	0.000302533
44	GCGCG	+	0.000357186	CGCGC	5.50704e-05
45	CGGGGCG	+	0.000356865	CGCCCCG	0.000160819
46	CGCCCCC	+	0.000346089	GGGGGCG	4.4485e-05
47	CACGTG	+	0.000344168	CACGTG	0

Table 6.1: This ranked list gives the most over- and under-represented words in the first 50-bp region just upstream of the transcription start site, relative to the mononucleotide background, for all human genes. The degree of over- and under-representation (+ or -) is measured by the words' contribution $S(w)$ to the relative entropy distance between the promoter sequences' oligo distribution and the expected background distribution. Notice that words and their reverse complements appear close together in the list. For example, the first two motifs (in yellow) are reverse complements of each other: TTT and AAA. Interestingly, we see from entries three, four, six, and seven (in cyan), that the word $WpCpG$ and its reverse complement $CpGpW$ are highly under-represented in the promoter regions of human genes (W represents an A or T). However, we see from the green entries that CpG dinucleotides that are surrounded by other cytidines (Cs) and guanines (Gs) are over-represented.

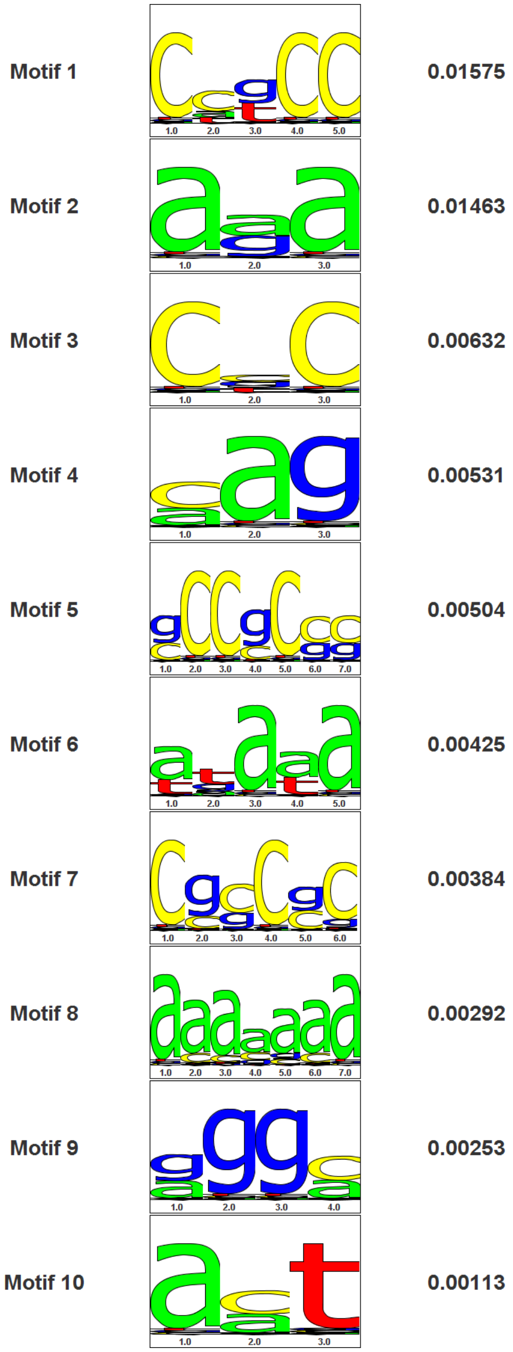


Figure 6.2: **The Top 10 Over-represented, Transcriptional Promoter-Region Motifs.** This list provides the top 10 most over-represented motifs in the Promoter Regions (-50,0) just upstream of the TSS of all known human genes. These motifs consist of sets of tuples $m = \{w_i, \bar{w}_i\}$ that maximize the transcriptional entropy-contribution measure $S_{txn}^{over}(m)$, against the mononucleotide background and with Re-scaling. When using the transcriptional entropy-contribution measure $S_{txn}^{over}(m)$, words w_i and their reverse complements $\bar{w}_i$ are considered to be equivalent motifs on different strands. Motifs 1, 5, 7, and 9 match the GC box (or parts thereof). Motif 6 matches the TATA box. Motifs 2 and 8 match the poly-A signals that facilitate melting (strand separation). Motif 4 matches part of the Myo D site. Motif 7 matches the BRE Box. Motif 10 matches part of the MTE Box.

Ranked List of the most over-represented tuple-motifs $m = (w, \bar{w})$ in the Promoter Regions (-50,0) of human genes that maximize the transcriptional entropy contribution measure $S_{txn}^{over}(m)$
against the mononucleotide background and with Re-scaling

Rank	$w/\bar{w}$	+/- / +/-	$S_{txn}^{over}(m)$	Motif
1	AAA/TTT	(+/+)	0.012142963710993378	Melting
2	CCGCC/GGCGG	(+/+)	0.005390318878979367	GC Box
3	CTCC/GGAG	(+/+)	0.005026509691252731	INR Box
4	CAG/CTG	(+/+)	0.003968616965861579	MTE Box
5	CTTCC/GGAAG	(+/+)	0.003621381090718398	INR Box
6	CCC/GGG	(+/+)	0.003426268046064107	GC Box
7	AGAG/CTCT	(+/+)	0.0028801831501160547	INR Box
8	GCGC/GCGC	(+/+)	0.0030939789817518125	BRE Box
9	AGAA/TTCT	(+/+)	0.002250242063324839	INR Box
10	CGCCGC/GCGGCG	(+/+)	0.0019246152501235865	BRE Box
11	GCCCCGC/GCGGGGC	(+/+)	0.001932352334181332	BRE Box
12	AA/TT	(+/+)	0.0015966934925324798	Melting
13	CCTC/GAGG	(+/+)	0.0018545477188168947	INR Box
14	GC/GC	(+/+)	0.0014391322584564985	BRE Box, GC Box
15	CCCGCCC/GGGCGGG	(+/+)	0.0012773504502755781	GC Box
16	TATAA/TTATA	(+/+)	0.0012251911453985663	TATA Box
17	ACACACA/TGTGTGT	(+/+)	0.0011046525156472357	
18	AGGAA/TTCCT	(+/+)	9.5014580515938E-4	
19	CGCCCCC/GGGGGCG	(+/+)	9.445259312556126E-4	GC Box
20	CCGCCCC/GGGGCGG	(+/+)	9.418208908182862E-4	
21	CACACAC/GTGTGTG	(+/+)	9.071697357169304E-4	
22	CCCCGCC/GGCGGGG	(+/+)	8.963002122745756E-4	GC Box
23	AGGAGGA/TCCTCCT	(+/+)	8.214503794956535E-4	
24	CCCGGCC/GGCCGGG	(+/+)	7.761861997016208E-4	
25	CGCCCCG/CGGGGCG	(+/+)	7.704076592416819E-4	
26	TCA/TGA	(+/+)	7.577933165127584E-4	
27	TATATA/TATATA	(+/+)	7.709336565496528E-4	TATA Box
28	CGCCCGC/GCGGGCG	(+/+)	7.428765635318227E-4	
29	CCGG/CCGG	(+/+)	7.386923586819713E-4	
30	AG/CT	(+/+)	7.517047729841807E-4	
31	CGGCGGC/GCCGCCG	(+/+)	9.000636832637893E-4	BRE Box
32	CGGCCGC/GCGGCCG	(+/+)	7.84350369174976E-4	BRE Box
33	AAT/ATT	(+/+)	7.687319863343943E-4	
34	CC/GG	(+/+)	9.324056539551373E-4	
35	ACA/TGT	(+/+)	7.911599972864622E-4	
36	CCGCG/CGCGG	(+/+)	8.569034102541627E-4	BRE Box
37	AAG/CTT	(+/+)	7.040562865587774E-4	
38	ATAAA/TTTAT	(+/+)	6.544494394750478E-4	TATA Box
39	GACGTCA/TGACGTC	(+/+)	6.382596575954781E-4	
40	GCC/GGC	(+/+)	5.911944240261525E-4	GC Box
41	CCACC/GGTGG	(+/+)	5.989664730584919E-4	
42	TTAA/TTAA	(+/+)	5.962556290507563E-4	
43	ATAT/ATAT	(+/+)	5.598002690177426E-4	TATA Box
44	CCGGAA/TTCCGG	(+/+)	4.988653407326678E-4	
45	GGGA/TCCC	(+/+)	5.059598094653579E-4	
46	CA/TG	(+/+)	5.225650828928757E-4	
47	AAATA/TATTT	(+/+)	4.724946860629666E-4	TATA Box
48	ACTTCCG/CGGAAGT	(+/+)	4.711535261116658E-4	
49	CACTTC/GAAGTG	(+/+)	4.6060928779331897E-4	
50	GAGC/GCTC	(+/+)	4.3514651677507756E-4	

Table 6.2: This ranked list gives the most over-represented words (of lengths 1-7 bp) on both strands in the first 50-bp region just upstream of the transcription start site (TSS), relative to the mononucleotide background, for all known human genes. Words and their reverse complements are considered to be equivalent motifs on different strands. This list shows individual tuples that help construct motifs like those shown in Figure 6.2 on the preceding page.

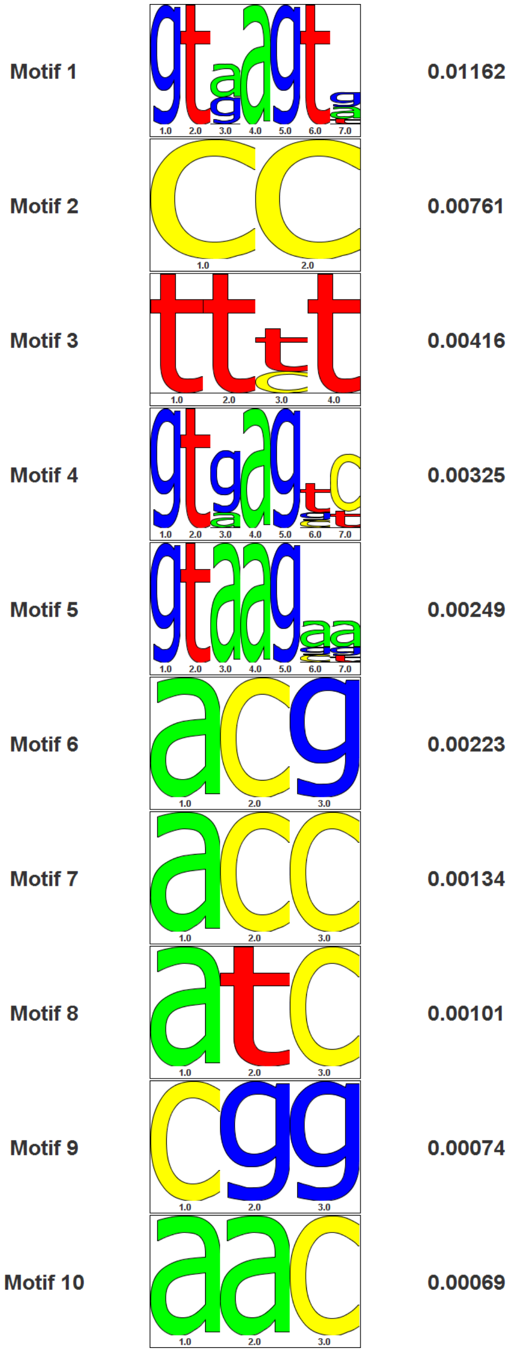


Figure 6.3: **The Top 10 Over-represented, Post-Transcriptional Motifs in the beginning regions of human introns.** This list provides the top 10 most over-represented, post-transcriptional motifs (of lengths 1-7 bp) in the first 50-bp region of all introns in all known human genes. These motifs consist of sets of tuples $m = \{w_i, \bar{w}_i\}$ that maximize the post-transcriptional entropy-contribution measure $S^{over}_{post-txn}(m)$, against the mononucleotide background and with Re-scaling. The $5'$ splice site motifs are $AG|GTRAGT$.

Ranked List of the most over-represented tuple-motifs $m = (w, \bar{w})$ in the intron(0,50) regions of human genes that maximize the post-transcriptional entropy contribution measure $S^{over}_{post-txn}(m)$ against the mononucleotide background and with Re-scaling

The $5'$ splice site motifs are $AG|GTRAGT$.

Rank	$w/\bar{w}$	+/- / +/-	$S^{over}_{post-txn}(m)$
1	CC/GG	(+/+)	0.007606766746999127
2	GTGAGTG/CACTCAC	(+/+)	0.0035345230403304667
3	TTTT/AAAA	(+/+)	0.0035166662597895604
4	GTAAGTA/TACTTAC	(+/+)	0.0028037376152889337
5	ACG/CGT	(-/-)	0.0022861839759536215
6	GTAAGTG/CACTTAC	(+/-)	0.0019748315901672994
7	GTGAGTA/TACTCAC	(+/-)	0.0016464615904590617
8	GTAAGTT/AACTTAC	(+/+)	0.0015163907155177042
9	ACC/GGT	(-/-)	0.0014581593422086166
10	ATC/GAT	(-/-)	0.0010697489964987617
11	GTAAGAA/TTCTTAC	(+/+)	9.978881792188908E-4
12	GTGAGTC/GACTCAC	(+/-)	9.13115114341284E-4
13	CGG/CCG	(-/-)	7.628159639648325E-4
14	ACTA/TAGT	(-/-)	7.166235003200293E-4
15	GTAAGTC/GACTTAC	(+/-)	6.204129270896352E-4
16	CTGG/CCAG	(+/+)	5.87422155379456E-4
17	GTGAGGC/GCCTCAC	(+/+)	5.813557402606899E-4
18	TTTATTT/AAATAAA	(+/+)	5.417887188773696E-4
19	TTTCTTT/AAAGAAA	(+/+)	5.391521969078973E-4
20	TGAGTGG/CCACTCA	(+/+)	5.317088409123245E-4
21	GTGAGTT/AACTCAC	(+/+)	5.283389406390586E-4
22	CGAA/TTCG	(-/-)	5.205673026696406E-4
23	GTAAGGA/TCCTTAC	(+/-)	5.020747725017928E-4
24	TAAGTAT/ATACTTA	(+/-)	5.003749691899603E-4
25	GTAAGAG/CTCTTAC	(+/-)	4.917754758013632E-4
26	AAC/GTT	(-/-)	4.4460502523030693E-4
27	TCT/AGA	(+/+)	4.4554307457174815E-4
28	GTAAGCA/TGCTTAC	(+/-)	4.358782978838502E-4
29	GTGAGCA/TGCTCAC	(+/+)	4.325548972385005E-4
30	TGAGTGC/GCACTCA	(+/-)	4.0612873794146294E-4
31	GTGAGCC/GGCTCAC	(+/+)	4.043522744980639E-4
32	GTGTGTG/CACACAC	(+/+)	4.0248045177062404E-4
33	TAAGTAA/TTACTTA	(+/+)	3.904893238429955E-4
34	GTAAGAT/ATCTTAC	(+/-)	3.864091183130103E-4
35	GTGAGGG/CCCTCAC	(+/+)	3.7965824723958097E-4
36	TTTGTTT/AAACAAA	(+/+)	3.785126414799403E-4
37	GTGAGGA/TCCTCAC	(+/+)	3.702997718645142E-4
38	TAAGTTT/AAACTTA	(+/+)	3.691004242378078E-4
39	CAGG/CCTG	(+/+)	3.5533079297743377E-4
40	TGTGTGT/ACACACA	(+/+)	3.594304946610601E-4
41	GGC/GCC	(+/+)	3.1281568559348707E-4
42	TTC/GAA	(+/+)	3.7324310338131164E-4
43	GTAGGTG/CACCTAC	(+/+)	3.2439785516657355E-4
44	GTGGGTG/CACCCAC	(+/+)	3.027167350972143E-4
45	TGAGTGT/ACACTCA	(+/+)	2.956203475809266E-4
46	CTCG/CGAG	(-/-)	2.890462277561072E-4
47	AGCG/CGCT	(-/-)	2.920001328187347E-4
48	TGAGTGA/TCACTCA	(+/+)	2.802466362977953E-4
49	GTGAGAG/CTCTCAC	(+/+)	2.358937466940281E-4
50	GTACGTG/CACGTAC	(+/+)	2.338330144353744E-4

Table 6.3: This ranked list gives the most over-represented words (of lengths 1-7 bp) on the coding strand in the first 50-bp region of all introns, relative to the mononucleotide background, for all known human genes. The motifs in yellow match known $5'$ splicing sites. This list shows individual tuples that help construct motifs like those shown in Figure 6.3 on the preceding page.

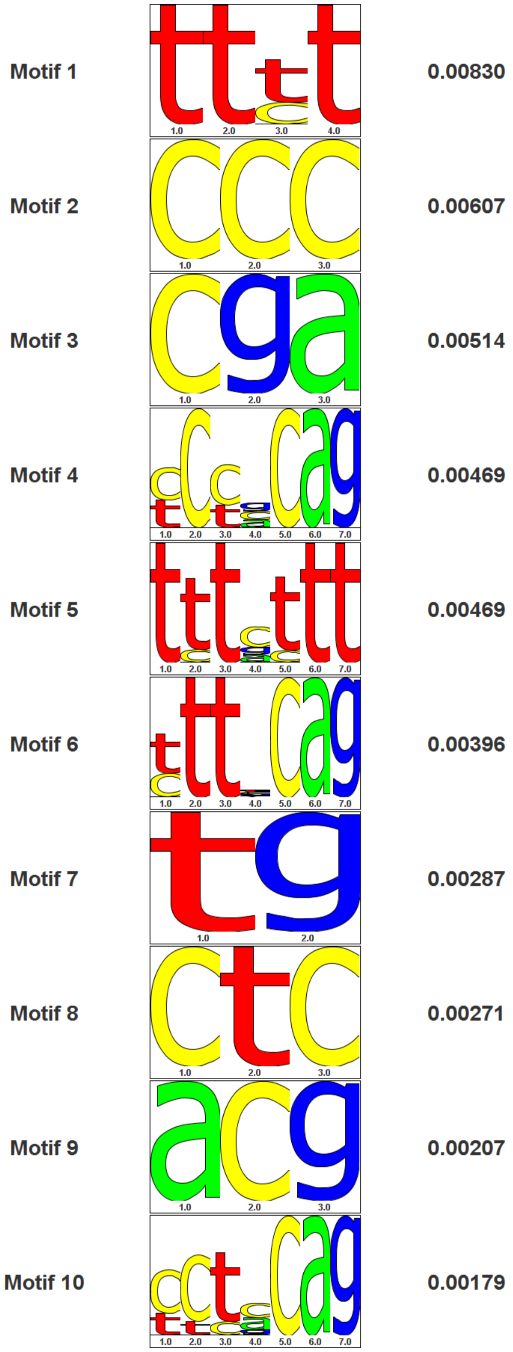


Figure 6.4: **The Top 10 Over-represented, Post-Transcriptional Motifs in the end-regions of human introns.** This list provides the top 10 most over-represented, post-transcriptional motifs (of lengths 1-7 bp) in the last 50-bp region of all introns in all known human genes. These motifs consist of sets of tuples $m = \{w_i, \bar{w}_i\}$ that maximize the post-transcriptional entropy-contribution measure $S^{over}_{post-txn}(m)$, against the mononucleotide background and with Re-scaling. The $3'$ splice site motifs are $YYYYYYYYNCAG|G$.

Ranked List of the most over-represented tuple-motifs $m = (w, \bar{w})$ in the intron(-50,0) regions of human genes that maximize the post-transcriptional entropy contribution measure $S^{over}_{post-txn}(m)$ against the mononucleotide background and with Re-scaling

The $3'$ splice site motifs are $YYYYYYYYNCAG|G$.

Rank	$w/\bar{w}$	+/- / +/-	$S^{over}_{post-txn}(m)$
1	TTTT/AAAA	(+/+)	0.00679548011286107
2	CGA/TCG	(-/-)	0.005732036100534097
3	CCC/GGG	(+/+)	0.004801817827452967
4	CTC/GAG	(+/+)	0.0030133726199413156
5	CTG/CAG	(+/+)	0.0028841926440741307
6	TTTCTTT/AAAGAAA	(+/+)	0.002206934799593214
7	GTA/TAC	(-/-)	0.002078496830381361
8	ACG/CGT	(-/-)	0.0021178845193865897
9	TTTGTTT/AAACAAA	(+/+)	0.0014631306022040065
10	ATAG/CTAT	(-/-)	0.0013129866927489248
11	CTA/TAG	(-/-)	0.0013169423530145089
12	GAT/ATC	(-/-)	0.001066940407302762
13	CGGA/TCCG	(-/-)	8.892641383403599E-4
14	AACT/AGTT	(-/-)	8.858653381139908E-4
15	CCCACAG/CTGTGGG	(+/+)	8.567071054391215E-4
16	TGTGT/ACACA	(+/+)	8.432170953661942E-4
17	CCTGCAG/CTGCAGG	(+/-)	7.939848600523447E-4
18	AGCG/CGCT	(-/-)	7.397867332593134E-4
19	AA/TT	(+/+)	6.65538458253628E-4
20	TTG/CAA	(-/-)	6.855802020798427E-4
21	CGCA/TGCG	(-/-)	6.395493626828244E-4
22	GACT/AGTC	(-/-)	5.840965262054558E-4
23	CTTCC/GGAAG	(+/+)	5.27498012899531E-4
24	GAGT/ACTC	(-/-)	4.672858869555445E-4
25	CTTGCAG/CTGCAAG	(+/-)	4.621166221933004E-4
26	ACC/GGT	(-/-)	4.5649342983299047E-4
27	CTGA/TCAG	(-/-)	4.457877088462007E-4
28	TTTGCAG/CTGCAAA	(+/-)	4.4143917635133497E-4
29	TCCACAG/CTGTGGA	(+/-)	4.0850681770532234E-4
30	TTTCTAG/CTAGAAA	(+/+)	3.999312728056908E-4
31	AATA/TATT	(+/+)	3.7271414529173395E-4
32	TTTATTT/AAATAAA	(+/+)	3.802204823007458E-4
33	GTG/CAC	(+/+)	3.77520798316819E-4
34	CTTTCAG/CTGAAAG	(+/-)	3.762032213717009E-4
35	TTTTCTT/AAGAAAA	(+/+)	3.621841552421502E-4
36	TTTTCAG/CTGAAAA	(+/-)	3.608805217530901E-4
37	TCTTTC/GAAAGA	(+/+)	3.4657828887014726E-4
38	TGAC/GTCA	(+/-)	3.4290848962098444E-4
39	TTTACAG/CTGTAAA	(+/-)	3.4113537391334224E-4
40	CCCGCAG/CTGCGGG	(+/+)	3.4119463052312087E-4
41	TTTGTAG/CTACAAA	(+/+)	3.25728667958095E-4
42	AATG/CATT	(+/-)	3.214057651093297E-4
43	TTGTTTT/AAAACAA	(+/+)	3.1572096818088455E-4
44	CGGGA/TCCCG	(-/-)	2.931151350453779E-4
45	TTTCCTT/AAGGAAA	(+/+)	2.750504788257974E-4
46	CTTACAG/CTGTAAG	(+/-)	2.707506609703922E-4
47	TTTATAG/CTATAAA	(+/+)	2.675940705690651E-4
48	ATTTCAG/CTGAAAT	(+/-)	2.5307405487761373E-4
49	AAAGAAA/TTTCTTT	(-/-)	2.483977819827864E-4
50	AAAAGGG/CCCTTTT	(-/-)	2.4428483026164036E-4

Table 6.4: This ranked list gives the most over-represented words (of lengths 1-7 bp) on the coding strand in the last 50-bp region of all introns, relative to the mononucleotide background, for all known human genes. The motifs in yellow match known $3'$ splicing sites. This list shows individual tuples that help construct motifs like those shown in Figure 6.4 on the preceding page.

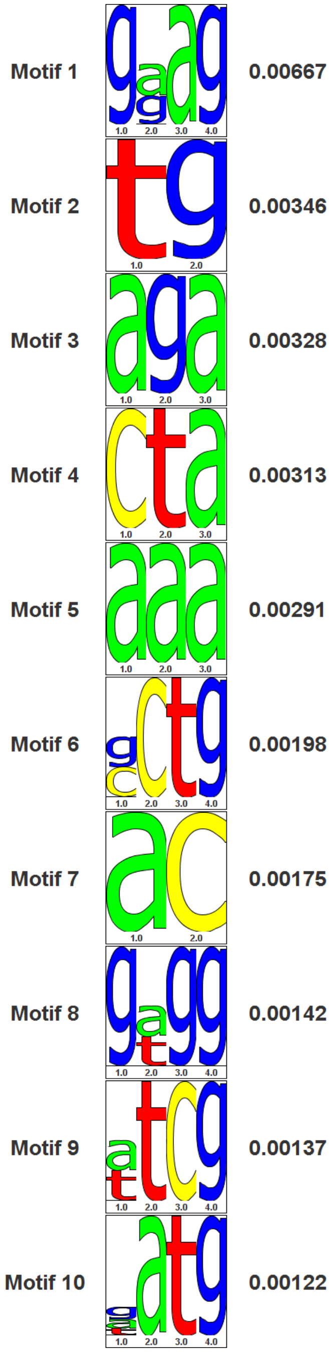


Figure 6.5: **The Top 10 Over-represented, Post-Transcriptional Motifs in the beginning regions of human exons.** This list provides the top 10 most over-represented, post-transcriptional motifs (of lengths 1-7 bp) in the first 50-bp region of all exons in all known human genes. These motifs consist of sets of tuples $m = \{w_i, \bar{w}_i\}$ that maximize the post-transcriptional entropy-contribution measure $S^{over}_{post-txn}(m)$, against the interior Exon(+50, -50) background and with Re-scaling. (See part f of Figure 6.1 on page 127). All the motifs except six, eight, and ten are known 5′ ESEs, or parts thereof.

Ranked List of the most over-represented tuple-motifs $m = (w, \bar{w})$ in the exon(0,50) regions of human genes that maximize the post-transcriptional entropy contribution measure $S^{over}_{post-txn}(m)$ against the interior Exon(+50, -50) background and with Re-scaling

Rank	$w/\bar{w}$	+/- / +/-	$S^{over}_{post-txn}(m)$
1	GAAG/CTTC	(+/+)	0.004048695844903189
2	CTA/TAG	(-/-)	0.0031626782674147574
3	AGA/TCT	(+/+)	0.003597415670708517
4	TG/CA	(+/+)	0.0030390368605560762
5	GGA/TCC	(+/+)	0.0030209950050174607
6	AAA/TTT	(+/+)	0.0027024617258089552
7	CCTG/CAGG	(+/+)	0.002000831841100705
8	TAC/GTA	(-/-)	0.001166299895974904
9	CTAC/GTAG	(+/+)	9.232782370311172E-4
10	AG/CT	(+/+)	8.87224053352953E-4
11	CCC/GGG	(+/+)	0.0010670541402720561
12	CTTCG/CGAAG	(-/-)	8.821105770158426E-4
13	CGA/TCG	(-/-)	6.012246432176632E-4
14	AGCAG/CTGCT	(+/+)	5.13578689611971E-4
15	ATGA/TCAT	(+/+)	5.263563383366894E-4
16	GCTG/CAGC	(+/+)	5.104968410438596E-4
17	GAAA/TTTC	(+/+)	5.437925316063626E-4
18	GTGG/CCAC	(+/+)	5.451328746207387E-4
19	GAGGAG/CTCCTC	(+/+)	4.610834473725388E-4
20	TATG/CATA	(-/-)	3.878252196743954E-4
21	GCG/CGC	(-/-)	3.7824814203015356E-4
22	CAAG/CTTG	(+/-)	3.3903049323965055E-4
23	CTTCC/GGAAG	(-/-)	3.372971442782766E-4
24	TTA/TAA	(-/-)	3.3589605432150695E-4
25	GAC/GTC	(-/-)	3.569341051760807E-4
26	AACG/CGTT	(-/-)	3.112534217600703E-4
27	GCGGC/GCCGC	(+/+)	2.881812697542727E-4
28	CGAG/CTCG	(+/+)	2.364466390925624E-4
29	GTG/CAC	(-/-)	2.1951130672369872E-4
30	ACCA/TGGT	(+/+)	2.2188110414978076E-4

Table 6.5: This ranked list gives the most over-represented words (of lengths 1-7 bp) on the coding strand in the first 50-bp region of all exons, relative to the mononucleotide background, for all known human genes. The motifs in yellow match known 5′ Exonic Splice Enhancer Sites (ESE) sites, or parts thereof. This list shows individual tuples that help construct motifs like those shown in Figure 6.5 on the preceding page.

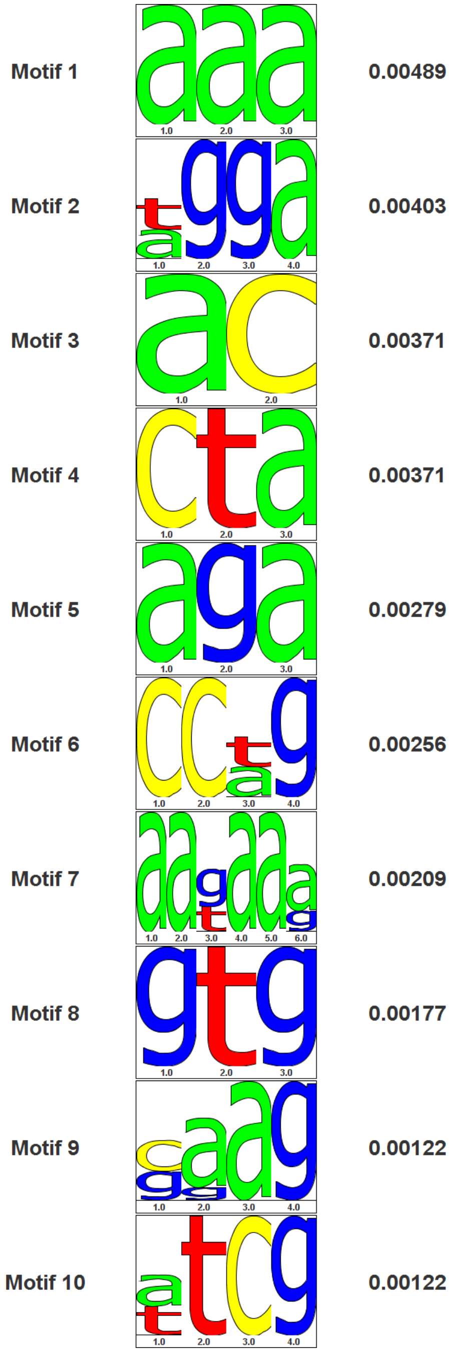


Figure 6.6: **The Top 10 Over-represented, Post-Transcriptional Motifs in the end-regions of human exons.** This list provides the top 10 most over-represented, post-transcriptional motifs (of lengths 1-7 bp) in the last 50-bp region of all exons in all known human genes. These motifs consist of sets of tuples $m = \{w_i, \bar{w}_i\}$ that maximize the post-transcriptional entropy-contribution measure $S^{over}_{post-txn}(m)$, against the interior Exon(+50, -50) background and with Re-scaling. (See part f of Figure 6.1 on page 127). All motifs except four, six, and ten are known $3'$ ESEs, or parts thereof.

Ranked List of the most over-represented tuple-motifs $m = (w, \bar{w})$ in the exon(0,50) regions of human genes that maximize the post-transcriptional entropy contribution measure $S^{over}_{post-txn}(m)$ against the interior Exon(+50, -50) background and with Re-scaling

Rank	$w/\bar{w}$	+/- / +/-	$S^{over}_{post-txn}(m)$
1	AAA/TTT	(+/+)	0.004890672357847098
2	AC/GT	(-/-)	0.0037109233655847444
3	TGGA/TCCA	(+/+)	0.0026383706247931284
4	AG/CT	(+/+)	0.002748678133590105
5	CTA/TAG	(-/-)	0.004323377196141551
6	CCTG/CAGG	(+/+)	0.0018340860030188527
7	GTG/CAC	(+/+)	0.0015307591979683748
8	AGAA/TTCT	(+/+)	0.0016197696903002797
9	AATAAA/TTTATT	(+/+)	0.0012238205650436298
10	AGGA/TCCT	(+/+)	8.716431556973791E-4
11	ATCG/CGAT	(-/-)	7.153304205296418E-4
12	TTCG/CGAA	(-/-)	6.510006830113184E-4
13	CTTA/TAAG	(-/-)	6.596903937702932E-4
14	GAAGA/TCTTC	(+/+)	6.085246758769655E-4
15	CGAG/CTCG	(-/-)	5.767545886029121E-4
16	CGT/ACG	(-/-)	5.654842220159089E-4
17	CTAC/GTAG	(+/+)	6.25154370798946E-4
18	TATG/CATA	(-/-)	6.111898071316576E-4
19	GCTG/CAGC	(+/+)	5.427852699728579E-4
20	CCA/TGG	(+/+)	5.371913774101849E-4
21	TACC/GGTA	(-/-)	4.447578973180086E-4
22	ATGAA/TTCAT	(+/+)	4.2831695444750387E-4
23	CAAG/CTTG	(+/-)	3.325868100797346E-4
24	GTTA/TAAC	(-/-)	3.012283020482866E-4
25	AGATG/CATCT	(+/+)	2.7674563598791273E-4
26	GATT/AATC	(-/-)	2.588528691282514E-4
27	CAACA/TGTTG	(+/+)	2.3843637374700683E-4
28	TTTG/CAAA	(+/-)	2.3231056828609673E-4
29	TCA/TGA	(+/+)	2.3012803300291942E-4

Table 6.6: This ranked list gives the most over-represented words on the coding strand in the last 50-bp region of all exons, relative to the mononucleotide background, for all known human genes. The motifs in yellow match known $3'$ Exonic Splice Enhancer Sites (ESE) sites, or parts thereof. The motif in cyan matches the known polyadenylation signal ($AATAAA$), which is present in the last 50bp of the last exon of most genes. This list shows individual tuples that help construct motifs like those shown in Figure 6.6 on the preceding page.

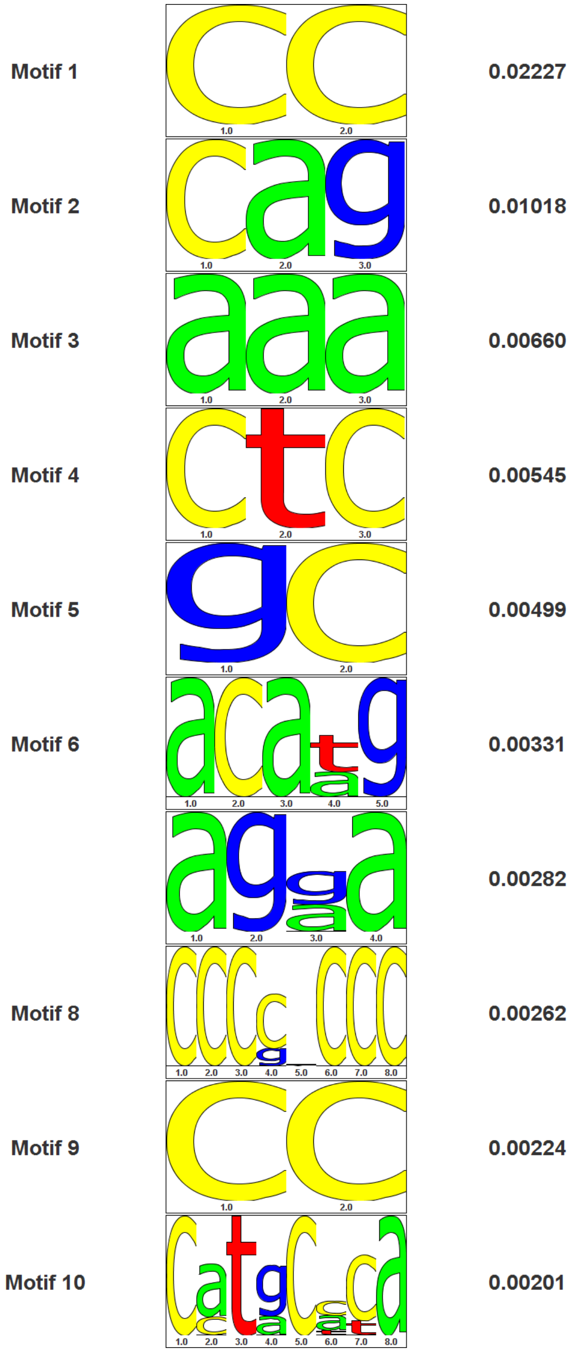


Figure 6.7: **The Top 10 Over-represented, Transcriptional Motifs within a 400bp window centered around the 160 functional human p53 REs.** This list provides the top 10 most over-represented, transcriptional motifs (of lengths 1-8 bp) within a 400bp sequence-window centered around 160 functional human p53 REs. These motifs consist of sets of tuples $m = \{w_i, \bar{w_i}\}$ that maximize the transcriptional entropy-contribution measure $S_{txn}^{over}(m)$, against the background DNA from the regions from 500 to 2000bp on either side of the site (see part b of Figure 6.1 on page 127), and with Re-scaling. All motifs except two, seven, and eight are known conserved regions of p53 REs. The most prevalent (and best binding) p53-RE is RRRCATGYYY. Motif eight is an obvious GC Box (Binds SP1) and is a known co-factor of p53.

Ranked List of the most over-represented tuple-motifs $m = (w, \bar{w})$ in the 400bp regions around 160 functional human p53 REs
that maximize the transcriptional entropy contribution measure $S^{over}_{txn}(m)$
against the background of surrounding DNA and with Re-scaling

Rank	$w/\bar{w}$	+/- / +/-	$S^{over}_{post-txn}(m)$
1	CC/GG	(+/+)	0.022271277247304003
2	CAG/CTG	(+/+)	0.01017801589079955
3	AAA/TTT	(+/+)	0.006598455938214537
4	CTC/GAG	(+/+)	0.005445456500169858
5	GC/GC	(+/+)	0.004990412898294265
6	ACATG/CATGT	(+/+)	0.0024903889414493027
7	AAG/CTT	(+/+)	0.0022252200341660363
8	CC/GG	(+/+)	0.0021687592658422294
9	CA/TG	(+/+)	0.0019961710418497057
10	AGGA/TCCT	(+/+)	0.0014510752955463587
11	AAA/TTT	(+/+)	0.00147698496112573
12	AGAA/TTCT	(+/+)	0.0012327432433121054
13	CCCCGCCC/GGGCGGGG	(+/+)	0.0012104230074030968
14	CCGCGG/CCGCGG	(+/+)	8.782962712328938E-4
15	GCC/GGC	(+/+)	8.519916066560516E-4
16	CTCC/GGAG	(+/+)	8.583415478638848E-4
17	CAC/GTG	(+/+)	8.934930894683394E-4
18	GGAA/TTCC	(+/+)	9.860706055785018E-4
19	AGA/TCT	(+/+)	8.464009430879868E-4
20	CCCAGGC/GCCTGGG	(+/+)	7.623347087374656E-4
21	CATACACA/TGTGTATG	(+/+)	7.457168267390651E-4
22	CCCGGG/CCCGGG	(+/+)	7.390237189098721E-4
23	ATGCATAC/GTATGCAT	(-/+)	7.028478871238439E-4
24	ATACACAA/TTGTGTAT	(-/+)	6.947631924228929E-4
25	AATATT/AATATT	(+/+)	6.499039639025234E-4
26	CCGCCGC/GCGGCGG	(+/+)	6.497484416861103E-4
27	TAATCCCA/TGGGATTA	(+/+)	6.464885086453086E-4
28	CCCGCCCC/GGGGCGGG	(+/+)	6.37243590059867E-4
29	GCATACAC/GTGTATGC	(+/+)	6.301308888700846E-4
30	GCCCCGC/GCGGGGC	(+/+)	6.298752695265208E-4
31	GAATTGAA/TTCAATTC	(+/+)	6.222284901025285E-4
32	TGAA/TTCA	(+/+)	5.939564765509615E-4
33	TGCATACA/TGTATGCA	(+/+)	5.887318248210412E-4
34	AGGCATG/CATGCCT	(+/+)	5.671330125574255E-4
35	AGCCCAGG/CCTGGGCT	(+/+)	5.6733052098333E-4
36	CATGCATA/TATGCATG	(+/+)	5.656501252690313E-4
37	CGGCCGC/GCGGCCG	(+/+)	5.613866839799184E-4
38	ATGTACAT/ATGTACAT	(+/+)	5.209388827425665E-4
39	CCGGGCA/TGCCCGG	(+/+)	5.156724545478054E-4
40	ACAAGC/GCTTGT	(+/+)	4.953541242437797E-4
41	AATAATAA/TTATTATT	(-/+)	4.905815461607241E-4
42	CTGTAATC/GATTACAG	(+/+)	4.8515599235118254E-4
43	GCCCAGAC/GTCTGGGC	(+/+)	4.521413353905173E-4
44	AACAAAA/TTTTGTT	(+/+)	4.5117534304659726E-4
45	ATTACAGG/CCTGTAAT	(+/+)	4.5075000153572227E-4
46	ATGCTCAC/GTGAGCAT	(+/+)	4.4835242259512464E-4
47	CGCCGCCG/CGGCGGCG	(+/+)	4.413389977119478E-4
48	AGGCTGAG/CTCAGCCT	(+/+)	4.331792008631744E-4
49	AAACTAG/CTAGTTT	(+/+)	4.32894597860302E-4
50	CCCCACCC/GGGTGGGG	(+/+)	4.2336074017180407E-4

Table 6.7: This ranked list gives the most over-represented words (of lengths 1-8 bp) on both strands within a 400bp window centered around the 160 functional human p53 REs, relative to surrounding DNA. Words and their reverse complements are considered to be equivalent motifs on different strands. The motifs in yellow match regions of the p53 consensus binding-site RRRCWWGYYY. The motifs in cyan match possible GC boxes that bind the SP1 transcription factor, a known co-factor of p53. This list shows individual tuples that help construct motifs like those shown in Figure 6.7 on the preceding page.

References

[1] Katsuya Adachi, Minoru Toyota, Yasushi Sasaki, Toshiharu Yamashita, Setsuko Ishida, Mutsumi Ohe-Toyota, Reo Maruyama, Yuji Hinoda, Tsuyoshi Saito, Kohzoh Imai, Ryuichi Kudo, and Takashi Tokino, *Identification of SCN3B as a novel p53-inducible proapoptotic gene.*, Oncogene **23** (2004), no. 47, 7791–7798.

[2] S. N. Agoff, J. Hou, D. I. Linzer, and B. Wu, *Regulation of the human hsp70 promoter by p53.*, Science **259** (1993), no. 5091, 84–87.

[3] Silvia Di Agostino, Sabrina Strano, Velia Emiliozzi, Valentina Zerbini, Marcella Mottolese, Ada Sacchi, Giovanni Blandino, and Giulia Piaggio, *Gain of function of mutant p53: the mutant p53/NF-Y protein complex reveals an aberrant transcriptional mechanism of cell cycle regulation.*, Cancer Cell **10** (2006), no. 3, 191–202.

[4] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J. D. Watson, *Molecular Biology of the Cell*, 4*th* ed., Garland Publishing, 2002.

[5] S. F. Altschul, R. J. Carroll, and D. J. Lipman, *Weights for data related by a tree.*, J Mol Biol **207** (1989), no. 4, 647–653.

[6] T. L. Bailey and C. Elkan, *The value of prior knowledge in discovering motifs with MEME.*, Proc Int Conf Intell Syst Mol Biol **3** (1995), 21–29.

[7] P. Balagurumoorthy, Stuart M Lindsay, and Rodney E Harrington, *Atomic force microscopy reveals kinks in the p53 response element DNA.*, Biophys Chem **101-102** (2002), 611–623.

[8] Yoseph Barash, Gal Elidan, Nir Friedman, and Tommy Kaplan, *Modeling dependencies in protein-DNA binding sites*, RECOMB '03: Proceedings of the seventh annual international conference on Research in computational molecular biology (New York, NY, USA), ACM Press, 2003, pp. 28–37.

[9] N. A. Barlev, L. Liu, N. H. Chehab, K. Mansfield, K. G. Harris, T. D. Halazonetis, and S. L. Berger, *Acetylation of p53 activates transcription through recruitment of coactivators/histone acetyltransferases.*, Mol Cell **8** (2001), no. 6, 1243–1254.

[10] Christian Barrett, Richard Hughey, and Kevin Karplus, *Scoring hidden Markov models*, Comput. Appl. Biosci. **13** (1997), no. 2, 191–199.

[11] Valentina Basile, Roberto Mantovani, and Carol Imbriano, *DNA damage promotes histone deacetylase 4 nuclear localization and repression of G2/M promoters, via p53 C-terminal lysines.*, J Biol Chem **281** (2006), no. 4, 2347–2357.

[12] V. Benoit, A. C. Hellin, S. Huygen, J. Gielen, V. Bours, and M. P. Merville, *Additive effect between NF-kappaB subunits and p53 protein for transcriptional activation of human p53 promoter.*, Oncogene **19** (2000), no. 41, 4787–4794.

[13] Eugene Berezikov, Victor Guryev, Ronald H A Plasterk, and Edwin Cuppen, *CONREAL: conserved regulatory elements anchored alignment algorithm for identification of transcription factor binding sites by phylogenetic footprinting.*, Genome Res **14** (2004), no. 1, 170–178.

[14] Benjamin P Berman, Yutaka Nibu, Barret D Pfeiffer, Pavel Tomancak, Susan E Celniker, Michael Levine, Gerald M Rubin, and Michael B Eisen, *Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome.*, Proc Natl Acad Sci U S A **99** (2002), no. 2, 757–762.

[15] J. Bian and Y. Sun, *Transcriptional activation by p53 of the human type IV collagenase (gelatinase A or matrix metalloproteinase 2) promoter.*, Mol Cell Biol **17** (1997), no. 11, 6330–6338.

[16] K. Birnbaum, P. N. Benfey, and D. E. Shasha, *cis element/transcription factor analysis (cis/TF): a method for discovering transcription factor/cis element relationships.*, Genome Res **11** (2001), no. 9, 1567–1573.

[17] A. Bist, C. J. Fielding, and P. E. Fielding, *p53 regulates caveolin gene transcription, cell cholesterol, and growth by a novel mechanism.*, Biochemistry **39** (2000), no. 8, 1966–1972.

[18] Mathieu Blanchette and Martin Tompa, *Discovery of regulatory elements by a computational method for phylogenetic footprinting.*, Genome Res **12** (2002), no. 5, 739–748.

[19] Gareth L Bond, Wenwei Hu, Elisabeth E Bond, Harlan Robins, Stuart G Lutzker, Nicoleta C Arva, Jill Bargonetti, Frank Bartel, Helge Taubert, Peter Wuerl, Kenan Onel, Linwah Yip, Shih-Jen Hwang, Louise C Strong, Guillermina Lozano, and Arnold J Levine, *A single nucleotide polymorphism in the MDM2 promoter attenuates the p53 tumor suppressor pathway and accelerates tumor formation in humans.*, Cell **119** (2004), no. 5, 591–602.

[20] J. C. Bourdon, V. Deguin-Chambon, J. C. Lelong, P. Dessen, P. May, B. Debuire, and E. May, *Further characterisation of the p53 responsive element–identification of new candidate genes for trans-activation by p53.*, Oncogene **14** (1997), no. 1, 85–94.

[21] A. Brazma, I. Jonassen, J. Vilo, and E. Ukkonen, *Predicting gene regulatory elements in silico on a genomic scale.*, Genome Res **8** (1998), no. 11, 1202–1215.

[22] C. Britschgi, M. Rizzi, T. J. Grob, M. P. Tschan, B. Hgli, V. A. Reddy, A-C. Andres, B. E. Torbett, A. Tobler, and M. F. Fey, *Identification of the p53 family-responsive element in the promoter region of the tumor suppressor gene hypermethylated in cancer 1.*, Oncogene **25** (2006), no. 14, 2030–2039.

[23] L. Buckbinder, R. Talbott, S. Velasco-Miguel, I. Takenaka, B. Faha, B. R. Seizinger, and N. Kley, *Induction of the growth inhibitor IGF-binding protein 3 by p53.*, Nature **377** (1995), no. 6550, 646–649.

[24] V. Budhram-Mahadeo, P. J. Morris, M. D. Smith, C. A. Midgley, L. M. Boxer, and D. S. Latchman, *p53 suppresses the activation of the Bcl-2 promoter by the Brn-3a POU family transcription factor.*, J Biol Chem **274** (1999), no. 21, 15237–15244.

[25] Timothy F Burns, Peiwen Fei, Kimberly A Scata, David T Dicker, and Wafik S El-Deiry, *Silencing of the novel p53 target gene Snk/Plk2 leads to mitotic catastrophe in paclitaxel (taxol)-exposed cells.*, Mol Cell Biol **23** (2003), no. 16, 5556–5571.

[26] H. J. Bussemaker, H. Li, and E. D. Siggia, *Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis.*, Proc Natl Acad Sci U S A **97** (2000), no. 18, 10096–10100.

[27] ______, *Regulatory element detection using correlation with expression.*, Nat Genet **27** (2001), no. 2, 167–171.

[28] C. Steven Carmack, Lee Ann McCue, Lee A Newberg, and Charles E Lawrence, *PhyloScan: identification of transcription factor binding sites using cross-species evidence.*, Algorithms Mol Biol **2** (2007), 1.

[29] Michele Ceribelli, Myriam Alcalay, Maria Alessandra Vigan, and Roberto Mantovani, *Repression of new p53 targets revealed by ChIP on chip experiments.*, Cell Cycle **5** (2006), no. 10, 1102–1110.

[30] Hee Don Chae, Jeanho Yun, and Deug Y Shi, *Transcription repression of a CCAAT-binding transcription factor CBF/HSP70 by p53.*, Exp Mol Med **37** (2005), no. 5, 488–491.

[31] Jiguo Chen and Ivan Sadowski, *Identification of the mismatch repair genes PMS2 and MLH1 as p53 target genes by using serial analysis of binding elements.*, Proc Natl Acad Sci U S A **102** (2005), no. 13, 4813–4818.

[32] X. Chen, Y. Zheng, J. Zhu, J. Jiang, and J. Wang, *p73 is transcriptionally regulated by DNA damage, p53, and p73.*, Oncogene **20** (2001), no. 6, 769–774.

[33] Y. Cho, S. Gorina, P. D. Jeffrey, and N. P. Pavletich, *Crystal structure of a p53 tumor suppressor-DNA complex: understanding tumorigenic mutations.*, Science **265** (1994), no. 5170, 346–355.

[34] Sergei Chuikov, Julia K Kurash, Jonathan R Wilson, Bing Xiao, Neil Justin, Gleb S Ivanov, Kristine McKinney, Paul Tempst, Carol Prives, Steven J Gamblin, Nickolai A Barlev, and Danny Reinberg, *Regulation of p53 activity through lysine methylation.*, Nature **432** (2004), no. 7015, 353–360.

[35] Abel C S Chun and Dong-Yan Jin, *Transcriptional regulation of mitotic checkpoint gene MAD1 by p53.*, J Biol Chem **278** (2003), no. 39, 37439–37450.

[36] Selvon St Clair, Luciana Giono, Shohreh Varmeh-Ziaie, Lois Resnick-Silverman, Wen-Jun Liu, Abhilash Padi, Jayasri Dastidar, Andrea DaCosta, Melissa Mattia, and James J Manfredi, *DNA damage-induced downregulation of Cdc25C is mediated by p53 via two independent mechanisms: one involves direct binding to the cdc25C promoter.*, Mol Cell **16** (2004), no. 5, 725–736.

[37] Michael J Clemens, *Targets and mechanisms for the regulation of translation in malignant transformation.*, Oncogene **23** (2004), no. 18, 3180–3188.

[38] P. F. Cliften, L. W. Hillier, L. Fulton, T. Graves, T. Miner, W. R. Gish, R. H. Waterston, and M. Johnston, *Surveying Saccharomyces genomes to identify functional elements by comparative DNA sequence analysis.*, Genome Res **11** (2001), no. 7, 1175–1186.

[39] Paul Cliften, Priya Sudarsanam, Ashwin Desikan, Lucinda Fulton, Bob Fulton, John Majors, Robert Waterston, Barak A Cohen, and Mark Johnston, *Finding functional features in Saccharomyces genomes by phylogenetic footprinting.*, Science **301** (2003), no. 5629, 71–76.

[40] Erin M Conlon, X. Shirley Liu, Jason D Lieb, and Jun S Liu, *Integrating regulatory motif discovery and genome-wide expression analysis.*, Proc Natl Acad Sci U S A **100** (2003), no. 6, 3339–3344.

[41] Ana Contente, Alexandra Dittmer, Manuela C Koch, Judith Roth, and Matthias Dobbelstein, *A polymorphic microsatellite that mediates induction of PIG3 by p53.*, Nat Genet **30** (2002), no. 3, 315–320.

[42] J. L. Cook, R. N. R, J. F. Giardina, F. E. Fontenot, D. Y. Cheng, and J. Alam, *Distance constraints and stereospecific alignment requirements characteristic of p53 DNA-binding consensus sequence homologies.*, Oncogene **11** (1995), no. 4, 723–733.

[43] Gavin E. Crooks, *Inequalities between the Jenson-Shannon and Jeffreys divergences.*, Tech. Report 004, Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California, 94720, USA, April 2008.

[44] Modan K Das and Ho-Kwok Dai, *A survey of DNA motif finding algorithms.*, BMC Bioinformatics **8 Suppl 7** (2007), S21.

[45] Elisa de Stanchina, Emmanuelle Querido, Masako Narita, Ramana V Davuluri, Pier Paolo Pandolfi, Gerardo Ferbeyre, and Scott W Lowe, *PML is a direct p53 target that modulates p53 effector functions.*, Mol Cell **13** (2004), no. 4, 523–535.

[46] V. Deguin-Chambon, M. Vacher, M. Jullien, E. May, and J. C. Bourdon, *Direct transactivation of c-Ha-Ras gene by p53: evidence for its involvement in p53 transactivation activity and p53-mediated apoptosis.*, Oncogene **19** (2000), no. 51, 5831–5841.

[47] Roumen A Dimitrov and Michael Zuker, *Prediction of hybridization and melting for double-stranded nucleic acids.*, Biophys J **87** (2004), no. 1, 215–226.

[48] Robert M Dirks, Milo Lin, Erik Winfree, and Niles A Pierce, *Paradigms for computational nucleic acid design.*, Nucleic Acids Res **32** (2004), no. 4, 1392–1403.

[49] Marko Djordjevic, Anirvan M. Sengupta, and Boris I. Shraiman, *A Biophysical Approach to Transcription Factor Binding Site Discovery*, Genome Res. **13** (2003), no. 11, 2381–2390.

[50] M. Dohn, J. Jiang, and X. Chen, *Receptor tyrosine kinase EphA2 is regulated by p53-family proteins and induces apoptosis.*, Oncogene **20** (2001), no. 45, 6503–6515.

[51] David Dornan, Ingrid Wertz, Harumi Shimizu, David Arnott, Gretchen D Frantz, Patrick Dowd, Karen O'Rourke, Hartmut Koeppen, and Vishva M Dixit, *The ubiquitin ligase COP1 is a critical negative regulator of p53.*, Nature **429** (2004), no. 6987, 86–92.

[52] Thomas A Down and Tim J P Hubbard, *NestedMICA: sensitive inference of over-represented motifs in nucleic acid sequence.*, Nucleic Acids Res **33** (2005), no. 5, 1445–1453.

[53] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological sequence analysis*, 1^{st} ed., Cambridge University Press, 1998.

[54] Cyril Duriez, Nicole Falette, Carole Audoynaud, Caroline Moyret-Lalle, Karim Bensaad, Stphanie Courtois, Qing Wang, Thierry Soussi, and Alain Puisieux, *The human BTG2/TIS21/PC3 gene: genomic structure, transcriptional regulation and evaluation as a candidate tumor suppressor gene.*, Gene **282** (2002), no. 1-2, 207–214.

[55] S. R. Eddy, *Profile hidden Markov models.*, Bioinformatics **14** (1998), no. 9, 755–763.

[56] S. R. Eddy, G. Mitchison, and R. Durbin, *Maximum discrimination hidden Markov models of sequence consensus.*, J Comput Biol **2** (1995), no. 1, 9–23.

[57] W. S. el Deiry, S. E. Kern, J. A. Pietenpol, K. W. Kinzler, and B. Vogelstein, *Definition of a consensus binding site for p53.*, Nat Genet **1** (1992), no. 1, 45–49.

[58] W. S. el Deiry, T. Tokino, T. Waldman, J. D. Oliner, V. E. Velculescu, M. Burrell, D. E. Hill, E. Healy, J. L. Rees, and S. R. Hamilton, *Topological control of p21WAF1/CIP1 expression in normal and neoplastic tissues.*, Cancer Res **55** (1995), no. 13, 2910–2919.

[59] S. Abou Elela and R. N. Nazar, *The ribosomal 5.8S RNA as a target site for p53 protein in cell differentiation and oncogenesis.*, Cancer Lett **117** (1997), no. 1, 23–28.

[60] Olivier Elemento, Noam Slonim, and Saeed Tavazoie, *A universal framework for regulatory element discovery across all genomes and data types.*, Mol Cell **28** (2007), no. 2, 337–350.

[61] Leif W Ellisen, Kate D Ramsayer, Cory M Johannessen, Annie Yang, Hideyuki Beppu, Karolina Minda, Jonathan D Oliner, Frank McKeon, and Daniel A Haber, *REDD1, a developmentally regulated transcriptional target of p63 and p53, links p63 to regulation of reactive oxygen species.*, Mol Cell **10** (2002), no. 5, 995–1005.

[62] Y. Endo, T. Fujita, K. Tamura, H. Tsuruga, and H. Nojima, *Structure and chromosomal assignment of the human cyclin G gene.*, Genomics **38** (1996), no. 1, 92–95.

[63] Eleazar Eskin and Pavel A Pevzner, *Finding composite regulatory patterns in DNA sequences.*, Bioinformatics **18 Suppl 1** (2002), S354–S363.

[64] J. M. Espinosa and B. M. Emerson, *Transcriptional regulation by p53 through intrinsic DNA/chromatin binding and site-directed cofactor recruitment.*, Mol Cell **8** (2001), no. 1, 57–69.

[65] William G Fairbrother, Ru-Fang Yeh, Phillip A Sharp, and Christopher B Burge, *Predictive identification of exonic splicing enhancers in human genes.*, Science **297** (2002), no. 5583, 1007–1013.

[66] William G. Fairbrother, Gene W. Yeo, Rufang Yeh, Paul Goldstein, Matthew Mawson, Phillip A. Sharp, and Christopher B. Burge, *RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons*, Nucl. Acids Res. **32** (2004), 187–190.

[67] G. Farmer, J. Colgan, Y. Nakatani, J. L. Manley, and C. Prives, *Functional interaction between p53, the TATA-binding protein (TBP), andTBP-associated factors in vivo.*, Mol Cell Biol **16** (1996), no. 8, 4295–4304.

[68] G. Farmer, P. Friedlander, J. Colgan, J. L. Manley, and C. Prives, *Transcriptional repression by p53 involves molecular interactions distinct from those with the TATA box binding protein.*, Nucleic Acids Res **24** (1996), no. 21, 4281–4288.

[69] Peiwen Fei, Wenge Wang, Seok hyun Kim, Shulin Wang, Timothy F Burns, Joanna K Sax, Monica Buzzai, David T Dicker, W. Gillies McKenna, Eric J Bernhard, and Wafik S El-Deiry, *Bnip3L is induced by p53 under hypoxia, and its knockdown promotes tumor growth.*, Cancer Cell **6** (2004), no. 6, 597–609.

[70] Z. Feng, S. Jin, A. Zupnick, J. Hoh, E. de Stanchina, S. Lowe, C. Prives, and A. J. Levine, *p53 tumor suppressor protein regulates the levels of huntingtin gene expression.*, Oncogene **25** (2006), no. 1, 1–7.

[71] Zhaohui Feng, Wenwei Hu, Elisa de Stanchina, Angelika K Teresky, Shengkan Jin, Scott Lowe, and Arnold J Levine, *The regulation of AMPK beta1, TSC2, and PTEN expression by p53: stress, cell and tissue specificity, and the role of these gene products in modulating the IGF-1-AKT-mTOR pathways.*, Cancer Res **67** (2007), no. 7, 3043–3053.

[72] Zhaohui Feng, Haiyan Zhang, Arnold J Levine, and Shengkan Jin, *The coordinate regulation of the p53 and mTOR pathways in cells.*, Proc Natl Acad Sci U S A **102** (2005), no. 23, 8204–8209.

[73] D.S. Fields, Y-y. He, A.Y. Al-Uzri, and G.D. Stormo, *Quantitative Specificity of the Mnt Repressor*, Journal of Molecular Biology **271** (August 1997), 178–194(17).

[74] Peter C FitzGerald, Andrey Shlyakhtenko, Alain A Mir, and Charles Vinson, *Clustering of DNA sequences in human promoters.*, Genome Res **14** (2004), no. 8, 1562–1574.

[75] Elsa R Flores, Kenneth Y Tsai, Denise Crowley, Shomit Sengupta, Annie Yang, Frank McKeon, and Tyler Jacks, *p63 and p73 are required for p53-dependent apoptosis in response to DNA damage.*, Nature **416** (2002), no. 6880, 560–564.

[76] Barrett C Foat, Alexandre V Morozov, and Harmen J Bussemaker, *Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE.*, Bioinformatics **22** (2006), no. 14, e141–e149.

[77] S. M. Freier, R. Kierzek, J. A. Jaeger, N. Sugimoto, M. H. Caruthers, T. Neilson, and D. H. Turner, *Improved free-energy parameters for predictions of RNA duplex stability.*, Proc Natl Acad Sci U S A **83** (1986), no. 24, 9373–9377.

[78] Martin C Frith, Ulla Hansen, John L Spouge, and Zhiping Weng, *Finding functional sequence elements by multiple local alignment.*, Nucleic Acids Res **32** (2004), no. 1, 189–200.

[79] W. D. Funk, D. T. Pak, R. H. Karas, W. E. Wright, and J. W. Shay, *A transcriptionally active DNA-binding site for human p53 protein complexes.*, Mol Cell Biol **12** (1992), no. 6, 2866–2871.

[80] M. S. Gelfand, E. V. Koonin, and A. A. Mironov, *Prediction of transcription regulatory sites in Archaea by a comparative genomic approach.*, Nucleic Acids Res **28** (2000), no. 3, 695–705.

[81] M. Gerstein, E. L. Sonnhammer, and C. Chothia, *Volume changes in protein evolution.*, J Mol Biol **236** (1994), no. 4, 1067–1078.

[82] Thomas Ghler, Maurice Reimann, Dimitry Cherny, Korden Walter, Gabriele Warnecke, Ella Kim, and Wolfgang Deppert, *Specific interaction of p53 with target binding sites is determined by DNA conformation and is regulated by the C-terminal domain.*, J Biol Chem **277** (2002), no. 43, 41192–41203.

[83] Vita Golubovskaya, Aparna Kaur, and William Cance, *Cloning and characterization of the promoter region of human focal adhesion kinase gene: nuclear factor kappa B and p53 binding sites.*, Biochim Biophys Acta **1678** (2004), no. 2-3, 111–125.

[84] Benjamin D. Greenbaum, Arnold J. Levine, Gyan Bhanot, and Raul Rabadan, *Patterns of Evolution and Host Gene Mimicry in Influenza and Other RNA Viruses*, PLoS Pathog **4** (2008), no. 6, e1000079.

[85] W. Gu and R. G. Roeder, *Activation of p53 sequence-specific DNA binding by acetylation of the p53 C-terminal domain.*, Cell **90** (1997), no. 4, 595–606.

[86] W. Gu, X. L. Shi, and R. G. Roeder, *Synergistic activation of transcription by CBP and p53.*, Nature **387** (1997), no. 6635, 819–823.

[87] Wei Gu, Jianyuan Luo, Chris L Brooks, Anatoly Y Nikolaev, and Muyang Li, *Dynamics of the p53 acetylation pathway.*, Novartis Found Symp **259** (2004), 197–205; discussion 205–7, 223–5.

[88] S. Gupta, V. Radha, Y. Furukawa, and G. Swarup, *Direct transcriptional activation of human caspase-1 by tumor suppressor p53.*, J Biol Chem **276** (2001), no. 14, 10585–10588.

[89] T. D. Halazonetis, L. J. Davis, and A. N. Kandil, *Wild-type p53 adopts a 'mutant'-like conformation when bound to DNA.*, EMBO J **12** (1993), no. 3, 1021–1028.

[90] H. J. Han, T. Tokino, and Y. Nakamura, *CSR, a scavenger receptor-like protein with a protective role against cellular damage causedby UV irradiation and oxidative stress.*, Hum Mol Genet **7** (1998), no. 6, 1039–1046.

[91] David Christopher Harmes, Edward Bresnick, Emma A Lubin, Julie K Watson, Kelly E Heim, Joshua C Curtin, Anne M Suskind, Justin Lamb, and James DiRenzo, *Positive and negative regulation of deltaN-p63 promoter activity by p53 and deltaN-p63-alpha contributes to differential regulation of p53 target genes.*, Oncogene **22** (2003), no. 48, 7607–7616.

[92] Kelly Lynn Harms and Xinbin Chen, *Histone deacetylase 2 modulates p53 transcriptional activities through regulation of p53-DNA binding activity.*, Cancer Res **67** (2007), no. 7, 3145–3152.

[93] Sandra L Harris and Arnold J Levine, *The p53 pathway: positive and negative feedback loops.*, Oncogene **24** (2005), no. 17, 2899–2908.

[94] Jamie M Hearnes, Deborah J Mays, Kristy L Schavolt, Luojia Tang, Xin Jiang, and Jennifer A Pietenpol, *Chromatin immunoprecipitation-based screen to identify functional genomic binding sites for sequence-specific transactivators.*, Mol Cell Biol **25** (2005), no. 22, 10148–10158.

[95] S. Henikoff and J. G. Henikoff, *Position-based sequence weights.*, J Mol Biol **243** (1994), no. 4, 574–578.

[96] H. Hermeking, C. Lengauer, K. Polyak, T. C. He, L. Zhang, S. Thiagalingam, K. W. Kinzler, and B. Vogelstein, *14-3-3 sigma is a p53-regulated inhibitor of G2/M progression.*, Mol Cell **1** (1997), no. 1, 3–11.

[97] G. Z. Hertz and G. D. Stormo, *Identifying DNA and protein patterns with statistically significant alignments of multiple sequences.*, Bioinformatics **15** (1999), no. 7-8, 563–577.

[98] Kerstin Herzer, Christine S Falk, Jens Encke, Sren T Eichhorst, Axel Ulsenheimer, Barbara Seliger, and Peter H Krammer, *Upregulation of major histocompatibility complex class I on liver cells by hepatitis C virus core protein via p53 and TAP1 impairs natural killer cell cytotoxicity.*, J Virol **77** (2003), no. 15, 8299–8309.

[99] J. M. Heumann, A. S. Lapedes, and G. D. Stormo, *Neural networks for determining protein specificity and multiple alignment of binding sites.*, Proc Int Conf Intell Syst Mol Biol **2** (1994), 188–194.

[100] William H Hoffman, Siham Biade, Jack T Zilfou, Jiandong Chen, and Maureen Murphy, *Transcriptional repression of the anti-apoptotic survivin gene by wild type p53.*, J Biol Chem **277** (2002), no. 5, 3247–3257.

[101] J. Hoh, S. Jin, T. Parrado, J. Edington, A. J. Levine, and J. Ott, *The p53MH algorithm and its application in detecting p53-responsive genes.*, Proc Natl Acad Sci U S A **99** (2002), no. 13, 8467–8472.

[102] I. Holmes and W. J. Bruno, *Finding regulatory elements using joint likelihoods for sequence and expression profile data.*, Proc Int Conf Intell Syst Mol Biol **8** (2000), 202–210.

[103] Lawrence S Hon and Ajay N Jain, *A deterministic motif finding algorithm with application to the human genome.*, Bioinformatics **22** (2006), no. 9, 1047–1054.

[104] Jianjun Hu, Bin Li, and Daisuke Kihara, *Limitations and potentials of current motif discovery algorithms.*, Nucleic Acids Res **33** (2005), no. 15, 4899–4913.

[105] Jianjun Hu, Yifeng D Yang, and Daisuke Kihara, *EMD: an ensemble algorithm for discovering regulatory motifs in DNA sequences.*, BMC Bioinformatics **7** (2006), 342.

[106] W. Hu, Z. Feng, AK Teresky, and A. J. Levine, *p53 regulates maternal reproduction through LIF*, Nature **in press** (2007), in press.

[107] R. Hughey and A. Krogh, *Hidden Markov models for sequence analysis: extension and analysis of the basic method.*, Comput Appl Biosci **12** (1996), no. 2, 95–107.

[108] Hee-Jeong Im, Mark R Pittelkow, and Rajiv Kumar, *Divergent regulation of the growth-promoting gene IEX-1 by the p53 tumor suppressor and Sp1.*, J Biol Chem **277** (2002), no. 17, 14612–14621.

[109] Carol Imbriano, Aymone Gurtner, Fabienne Cocchiarella, Silvia Di Agostino, Valentina Basile, Monica Gostissa, Matthias Dobbelstein, Giannino Del Sal, Giulia Piaggio, and Roberto Mantovani, *Direct p53 transcriptional repression: in vivo analysis of CCAAT-containing G2/M promoters.*, Mol Cell Biol **25** (2005), no. 9, 3737–3751.

[110] Alberto Inga, Francesca Storici, Thomas A Darden, and Michael A Resnick, *Differential transactivation by the p53 transcription factor is highly dependent on p53 level and promoter target sequence.*, Mol Cell Biol **22** (2002), no. 24, 8612–8625.

[111] Steven A Innocente and Jonathan M Lee, *p53 is a NF-Y- and p21-independent, Sp1-dependent repressor of cyclin B1 transcription.*, FEBS Lett **579** (2005), no. 5, 1001–1007.

[112] V. Iotsova and D. Stehelin, *Down-regulation of fibronectin gene expression by the p53 tumor suppressor protein.*, Cell Growth Differ **7** (1996), no. 5, 629–634.

[113] A. S. Jaiswal and S. Narayan, *p53-dependent transcriptional regulation of the APC promoter in colon cancer cells treated with DNA alkylating agents.*, J Biol Chem **276** (2001), no. 21, 18193–18199.

[114] Kuang-Yu Jen and Vivian G Cheung, *Identification of novel p53 target genes in ionizing radiation response.*, Cancer Res **65** (2005), no. 17, 7666–7673.

[115] R. A. Johnson, T. A. Ince, and K. W. Scotto, *Transcriptional repression by p53 through direct binding to a novel DNA element.*, J Biol Chem **276** (2001), no. 29, 27716–27720.

[116] M. S. Jung, J. Yun, H. D. Chae, J. M. Kim, S. C. Kim, T. S. Choi, and D. Y. Shin, *p53 and its homologues, p63 and p73, induce a replicative senescence through inactivation of NF-Y transcription factor.*, Oncogene **20** (2001), no. 41, 5818–5825.

[117] T. Kanaya, S. Kyo, K. Hamada, M. Takakura, Y. Kitagawa, H. Harada, and M. Inoue, *Adenoviral expression of p53 represses telomerase activity through down-regulation of human telomerase reverse transcriptase transcription.*, Clin Cancer Res **6** (2000), no. 4, 1239–1247.

[118] K. Kannan, N. Amariglio, G. Rechavi, and D. Givol, *Profile of gene expression regulated by induced p53: connection to the TGF-beta family.*, FEBS Lett **470** (2000), no. 1, 77–82.

[119] M. V. Kato, H. Sato, M. Nagayoshi, and Y. Ikawa, *Upregulation of the elongation factor-1alpha gene by p53 in association with death of an erythroleukemic cell line.*, Blood **90** (1997), no. 4, 1373–1378.

[120] O. V. Kel-Margoulis, A. G. Romashchenko, N. A. Kolchanov, E. Wingender, and A. E. Kel, *COMPEL: a database on composite regulatory elements providing combinatorial transcriptional regulation.*, Nucleic Acids Res **28** (2000), no. 1, 311–315.

[121] Manolis Kellis, Nick Patterson, Matthew Endrizzi, Bruce Birren, and Eric S Lander, *Sequencing and comparison of yeast species to identify genes and regulatory elements.*, Nature **423** (2003), no. 6937, 241–254.

[122] S. E. Kern, J. A. Pietenpol, S. Thiagalingam, A. Seymour, K. W. Kinzler, and B. Vogelstein, *Oncogenic forms of p53 inhibit p53-regulated gene expression.*, Science **256** (1992), no. 5058, 827–830.

[123] S. M. Kielbasa, J. O. Korbel, D. Beule, J. Schuchhardt, and H. Herzel, *Combining frequency and positional information to predict transcription factor binding sites.*, Bioinformatics **17** (2001), no. 11, 1019–1026.

[124] E. Kim, N. Albrechtsen, and W. Deppert, *DNA-conformation is an important determinant of sequence-specific DNA binding by tumor suppressor p53.*, Oncogene **15** (1997), no. 7, 857–869.

[125] Y. Kimura, T. Furuhata, T. Urano, K. Hirata, Y. Nakamura, and T. Tokino, *Genomic structure and chromosomal localization of GML (GPI-anchored molecule-like protein), a gene induced by p53.*, Genomics **41** (1997), no. 3, 477–480.

[126] George Koutsodontis, Eleftheria Vasilaki, Wan-Chih Chou, Paraskevi Papakosta, and Dimitris Kardassis, *Physical and functional interactions between members of the tumour suppressor p53 and the Sp families of transcription factors: importance for the regulation of genes involved in cell-cycle arrest and apoptosis.*, Biochem J **389** (2005), no. Pt 2, 443–455.

[127] A. Krogh, M. Brown, I. S. Mian, K. Sjölander, and D. Haussler, *Hidden Markov models in computational biology. Applications to protein modeling.*, J Mol Biol **235** (1994), no. 5, 1501–1531.

[128] A. Krogh and G. Mitchison, *Maximum entropy weighting of aligned sequences of proteins or DNA.*, Proc Int Conf Intell Syst Mol Biol **3** (1995), 215–221.

[129] S. Kubicka, F. Khnel, L. Zender, K. L. Rudolph, J. Plmpe, M. Manns, and C. Trautwein, *p53 represses CAAT enhancer-binding protein (C/EBP)-dependent transcription of the albumin gene. A molecular mechanism involved in viral liver infection with implications for hepatocarcinogenesis.*, J Biol Chem **274** (1999), no. 45, 32137–32144.

[130] C. Kunz, S. Pebler, J. Otte, and D. von der Ahe, *Differential regulation of plasminogen activator and inhibitor gene transcription by the tumor suppressor p53.*, Nucleic Acids Res **23** (1995), no. 18, 3710–3717.

[131] C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton, *Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment.*, Science **262** (1993), no. 5131, 208–214.

[132] C. E. Lawrence and A. A. Reilly, *An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences.*, Proteins **7** (1990), no. 1, 41–51.

[133] J. Lee, J. Kim, and J. Kim, *Data-driven design of hmm topology for on-line handwriting recognition*, 2000.

[134] K. C. Lee, A. J. Crowe, and M. C. Barton, *p53-mediated repression of alpha-fetoprotein gene expression by specific DNA binding.*, Mol Cell Biol **19** (1999), no. 2, 1279–1288.

[135] S. Lee, L. Cavallo, and J. Griffith, *Human p53 binds Holliday junctions strongly and facilitates their cleavage.*, J Biol Chem **272** (1997), no. 11, 7532–7539.

[136] A. J. Levine, *p53, the cellular gatekeeper for growth and division.*, Cell **88** (1997), no. 3, 323–331.

[137] A. J. Levine, W. Hu, and Z. Feng, *The P53 pathway: what questions remain to be explored?*, Cell Death Differ **13** (2006), no. 6, 1027–1036.

[138] Kristina Lhr, Constanze Mritz, Ana Contente, and Matthias Dobbelstein, *p21/CDKN1A mediates negative regulation of transcription by p53.*, J Biol Chem **278** (2003), no. 35, 32507–32516.

[139] Maoxiang Li, Jun-Ying Zhou, Yubin Ge, Larry H Matherly, and Gen Sheng Wu, *The phosphatase MKP1 is a transcriptional target of p53 involved in cell cycle regulation.*, J Biol Chem **278** (2003), no. 42, 41059–41068.

[140] W. Liebetrau, A. Budde, A. Savoia, F. Grummt, and H. Hoehn, *p53 activates Fanconi anemia group C gene expression.*, Hum Mol Genet **6** (1997), no. 2, 277–283.

[141] Y. Lin, W. Ma, and S. Benchimol, *Pidd, a new death-domain-containing protein, is induced by p53 and promotes apoptosis.*, Nat Genet **26** (2000), no. 1, 122–127.

[142] Gang Liu and Xinbin Chen, *The ferredoxin reductase gene is regulated by the p53 family and sensitizes cells to oxidative stress-induced apoptosis.*, Oncogene **21** (2002), no. 47, 7195–7204.

[143] X. Liu, D. L. Brutlag, and J. S. Liu, *BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes.*, Pac Symp Biocomput (2001), 127–138.

[144] X. Shirley Liu, Douglas L Brutlag, and Jun S Liu, *An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments.*, Nat Biotechnol **20** (2002), no. 8, 835–839.

[145] Xiangguo Liu, Ping Yue, Fadlo R Khuri, and Shi-Yong Sun, *p53 upregulates death receptor 4 expression through an intronic p53 binding site.*, Cancer Res **64** (2004), no. 15, 5078–5083.

[146] ———, *Decoy receptor 2 (DcR2) is a p53 target gene and regulates chemosensitivity.*, Cancer Res **65** (2005), no. 20, 9169–9175.

[147] J. H. Ludes-Meyers, M. A. Subler, C. V. Shivakumar, R. M. Munoz, P. Jiang, J. E. Bigger, D. R. Brown, S. P. Deb, and S. Deb, *Transcriptional activation of the human epidermal growth factor receptor promoter by human p53.*, Mol Cell Biol **16** (1996), no. 11, 6009–6019.

[148] Jianyuan Luo, Muyang Li, Yi Tang, Monika Laszkowska, Robert G Roeder, and Wei Gu, *Acetylation of p53 augments its site-specific DNA binding both in vitro and in vivo.*, Proc Natl Acad Sci U S A **101** (2004), no. 8, 2259–2264.

[149] Buyong Ma, Yongping Pan, Jie Zheng, Arnold J Levine, and Ruth Nussinov, *Sequence analysis of p53 response-elements suggests multiple binding modes of the p53 tetramer to DNA targets.*, Nucleic Acids Res **35** (2007), no. 9, 2986–3001.

[150] Kaiwen Ma, Keigo Araki, Solachuddin J A Ichwan, Tamaki Suganuma, Mimi Tamamori-Adachi, and Masa-Aki Ikeda, *E2FBP1/DRIL1, an AT-rich interaction domain-family transcription factor, is regulated by p53.*, Mol Cancer Res **1** (2003), no. 6, 438–444.

[151] Timothy K MacLachlan and Wafik S El-Deiry, *Apoptotic threshold is lowered by p53 transactivation of caspase-6.*, Proc Natl Acad Sci U S A **99** (2002), no. 14, 9492–9497.

[152] James J Manfredi, *p53 and apoptosis: it's not just in the nucleus anymore.*, Mol Cell **11** (2003), no. 3, 552–554.

[153] Ofer Margalit, Hila Amram, Ninette Amariglio, Amos J Simon, Sigal Shaklai, Galit Granot, Neri Minsky, Avichai Shimoni, Alon Harmelin, David Givol, Mordechai Shohat, Moshe Oren, and Gideon Rechavi, *BCL6 is regulated by p53 through a response element frequently disrupted in B-cell non-Hodgkin lymphoma.*, Blood **107** (2006), no. 4, 1599–1607.

[154] Voichita D Marinescu, Isaac S Kohane, and Alberto Riva, *MAPPER: a search engine for the computational identification of putative transcription factor binding sites in multiple genomes.*, BMC Bioinformatics **6** (2005), 79.

[155] Nicholas R Markham and Michael Zuker, *DINAMelt web server for nucleic acid melting prediction.*, Nucleic Acids Res **33** (2005), no. Web Server issue, W577–W581.

[156] L. Marsan and M. F. Sagot, *Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification.*, J Comput Biol **7** (2000), no. 3-4, 345–362.

[157] Reo Maruyama, Fumio Aoki, Minoru Toyota, Yasushi Sasaki, Hirofumi Akashi, Hiroaki Mita, Hiromu Suzuki, Kimishige Akino, Mutsumi Ohe-Toyota, Yumiko Maruyama, Haruyuki Tatsumi, Kohzoh Imai, Yasuhisa Shinomura, and Takashi Tokino, *Comparative genome analysis identifies the vitamin D receptor gene as a direct target of p53-mediated transcriptional activation.*, Cancer Res **66** (2006), no. 9, 4574–4583.

[158] T. Mashimo, M. Watabe, S. Hirota, S. Hosobe, K. Miura, P. J. Tegtmeyer, C. W. Rinker-Shaeffer, and K. Watabe, *The expression of the KAI1 gene, a tumor metastasis suppressor, is directly activated by p53.*, Proc Natl Acad Sci U S A **95** (1998), no. 19, 11307–11311.

[159] Taido Matsui, Yuko Katsuno, Tomoharu Inoue, Fumitaka Fujita, Takashi Joh, Hiroyuki Niida, Hiroshi Murakami, Makoto Itoh, and Makoto Nakanishi, *Negative regulation of Chk2 expression by p53 is dependent on the CCAAT-binding transcription factor NF-Y.*, J Biol Chem **279** (2004), no. 24, 25093–25100.

[160] V. Matys, E. Fricke, R. Geffers, E. Gssling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, D-U. Kloos, S. Land, B. Lewicki-Potapov,

H. Michael, R. Mnch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, and E. Wingender, *TRANSFAC: transcriptional regulation, from patterns to profiles.*, Nucleic Acids Res **31** (2003), no. 1, 374–378.

[161] J. S. McCaskill, *The equilibrium partition function and base pair binding probabilities for RNA secondary structure.*, Biopolymers **29** (1990), no. 6-7, 1105–1119.

[162] A. M. McGuire, J. D. Hughes, and G. M. Church, *Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes.*, Genome Res **10** (2000), no. 6, 744–757.

[163] Kristine McKinney and Carol Prives, *Efficient specific DNA binding by p53 requires both its central and C-terminal domains as revealed by studies with high-mobility group 1 protein.*, Mol Cell Biol **22** (2002), no. 19, 6797–6808.

[164] A. M. Metcalfe, R. M. Dixon, and G. K. Radda, *Wild-type but not mutant p53 activates the hepatocyte growth factor/scatter factor promoter.*, Nucleic Acids Res **25** (1997), no. 5, 983–986.

[165] Manuel Middendorf, Anshul Kundaje, Mihir Shah, Yoav Freund, Chris H. Wiggins, and Christina Leslie, *Motif Discovery Through Predictive Modeling of Gene Regulation*, Research in Computational Molecular Biology (New York, NY), Springer Berlin, 2005, pp. 538–252.

[166] Tarjei S Mikkelsen, Manching Ku, David B Jaffe, Biju Issac, Erez Lieberman, Georgia Giannoukos, Pablo Alvarez, William Brockman, Tae-Kyung Kim, Richard P Koche, William Lee, Eric Mendenhall, Aisling O'Donovan, Aviva Presser, Carsten Russ, Xiaohui Xie, Alexander Meissner, Marius Wernig, Rudolf Jaenisch, Chad Nusbaum, Eric S Lander, and Bradley E Bernstein, *Genome-wide maps of chromatin state in pluripotent and lineage-committed cells.*, Nature **448** (2007), no. 7153, 553–560.

[167] Chaouki Miled, Marco Pontoglio, Serge Garbay, Moshe Yaniv, and Jonathan B Weitzman, *A genomic map of p53 binding sites identifies novel p53 targets involved in an apoptotic network.*, Cancer Res **65** (2005), no. 12, 5096–5104.

[168] Saverio Minucci and Pier Giuseppe Pelicci, *Histone deacetylase inhibitors and the promise of epigenetic (and more) treatments for cancer.*, Nat Rev Cancer **6** (2006), no. 1, 38–51.

[169] Asra Mirza, Qun Wu, Luquan Wang, Terri McClanahan, W. Robert Bishop, Ferdous Gheyas, Wei Ding, Beth Hutchins, Tish Hockenberry, Paul Kirschmeier, Jonathan R Greene, and Suxing Liu, *Global transcriptional program of p53 target genes during the process of apoptosis and cell cycle progression.*, Oncogene **22** (2003), no. 23, 3645–3654.

[170] M. Mller, S. Wilder, D. Bannasch, D. Israeli, K. Lehlbach, M. Li-Weber, S. L. Friedman, P. R. Galle, W. Stremmel, M. Oren, and P. H. Krammer, *p53 activates the CD95 (APO-1/Fas) gene in response to DNA damage by anticancer drugs.*, J Exp Med **188** (1998), no. 11, 2033–2045.

[171] Toshiki Mori, Yoshio Anazawa, Megumi Iiizumi, Seisuke Fukuda, Yusuke Nakamura, and Hirofumi Arakawa, *Identification of the interferon regulatory factor 5 gene (IRF-5) as a direct target for p53.*, Oncogene **21** (2002), no. 18, 2914–2918.

[172] G. F. Morris, J. R. Bischoff, and M. B. Mathews, *Transcriptional activation of the human proliferating-cell nuclear antigen promoter by p53.*, Proc Natl Acad Sci U S A **93** (1996), no. 2, 895–899.

[173] K. Mortensen, J. Skouv, D. M. Hougaard, and L. I. Larsson, *Endogenous endothelial cell nitric-oxide synthase modulates apoptosis in cultured breast cancer cells and is transcriptionally regulated by p53.*, J Biol Chem **274** (1999), no. 53, 37679–37684.

[174] Alan M Moses, Derek Y Chiang, Daniel A Pollard, Venky N Iyer, and Michael B Eisen, *MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model.*, Genome Biol **5** (2004), no. 12, R98.

[175] T. Mukhopadhyay and J. A. Roth, *p53 involvement in activation of the cytokeratin 8 gene in tumor cell lines.*, Anticancer Res **16** (1996), no. 1, 105–112.

[176] M. Murphy, J. Ahn, K. K. Walker, W. H. Hoffman, R. M. Evans, A. J. Levine, and D. L. George, *Transcriptional repression by wild-type p53 utilizes histone deacetylases, mediated by interaction with mSin3a.*, Genes Dev **13** (1999), no. 19, 2490–2501.

[177] N. Nagy, M. Takahara, J. Nishikawa, J. C. Bourdon, L. L. Kis, G. Klein, and E. Klein, *Wild-type p53 activates SAP expression in lymphoid cells.*, Oncogene **23** (2004), no. 53, 8563–8570.

[178] Takahito Nakagawa, Masato Takahashi, Toshinori Ozaki, Ken ichi Watanabe Ki, Satoru Todo, Hiroyuki Mizuguchi, Takao Hayakawa, and Akira Nakagawara, *Autoinhibitory regulation of p73 by Delta Np73 to modulate cell survival and death through a p73-specific target element within the Delta Np73 promoter.*, Mol Cell Biol **22** (2002), no. 8, 2575–2585.

[179] K. Nakano and K. H. Vousden, *PUMA, a novel proapoptotic gene, is induced by p53.*, Mol Cell **7** (2001), no. 3, 683–694.

[180] Ching-Ching Ng, Hirofumi Arakawa, Seisuke Fukuda, Hisato Kondoh, and Yusuke Nakamura, *p53RFP, a p53-inducible RING-finger protein, regulates the stability of p21WAF1.*, Oncogene **22** (2003), no. 28, 4449–4458.

[181] K. Nylander, J. C. Bourdon, S. E. Bray, N. K. Gibbs, R. Kay, I. Hart, and P. A. Hall, *Transcriptional activation of tyrosinase and TRP-1 by p53 links UV irradiation to the protective tanning response.*, J Pathol **190** (2000), no. 1, 39–46.

[182] Susanna Obad, Hans Brunnstrm, Johan Vallon-Christersson, Ake Borg, Kristina Drott, and Urban Gullberg, *Staf50 is a novel p53 target gene conferring reduced clonogenic growth of leukemic U-937 cells.*, Oncogene **23** (2004), no. 23, 4050–4059.

[183] K. Oda, H. Arakawa, T. Tanaka, K. Matsuda, C. Tanikawa, T. Mori, H. Nishimori, K. Tamai, T. Tokino, Y. Nakamura, and Y. Taya, *p53AIP1, a potential mediator of p53-dependent apoptosis, and its regulation by Ser-46-phosphorylated p53.*, Cell **102** (2000), no. 6, 849–862.

[184] C. Ohlsson, N. Kley, H. Werner, and D. LeRoith, *p53 regulates insulin-like growth factor-I (IGF-I) receptor expression and IGF-I-induced tyrosine phosphorylation in an*

osteosarcoma cell line: interaction between p53 and Sp1., Endocrinology **139** (1998), no. 3, 1101–1107.

[185] Takao Ohtsuka, Hoon Ryu, Yohji A Minamishima, Salvador Macip, Junji Sagara, Keiichi I Nakayama, Stuart A Aaronson, and Sam W Lee, *ASC is a Bax adaptor and regulates the p53-Bax mitochondrial apoptosis pathway.*, Nat Cell Biol **6** (2004), no. 2, 121–128.

[186] S. Okamura, H. Arakawa, T. Tanaka, H. Nakanishi, C. C. Ng, Y. Taya, M. Monden, and Y. Nakamura, *p53DINP1, a p53-inducible gene, regulates p53-dependent apoptosis.*, Mol Cell **8** (2001), no. 1, 85–94.

[187] W. K. Olson, A. A. Gorin, X. J. Lu, L. M. Hock, and V. B. Zhurkin, *DNA sequence-dependent deformability deduced from protein-DNA crystal complexes.*, Proc Natl Acad Sci U S A **95** (1998), no. 19, 11163–11168.

[188] A. Ori, A. Zauberman, G. Doitsh, N. Paran, M. Oren, and Y. Shaul, *p53 binds and represses the HBV enhancer: an adjacent enhancer element can reverse the transcription effect of p53.*, EMBO J **17** (1998), no. 2, 544–553.

[189] Motonobu Osada, Hannah Lui Park, Yuichi Nagakawa, Keishi Yamashita, Alexey Fomenkov, Myoung Sook Kim, Guojun Wu, Shuji Nomoto, Barry Trink, and David Sidransky, *Differential recognition of response elements determines target gene specificity for p53 and p63.*, Mol Cell Biol **25** (2005), no. 14, 6077–6089.

[190] Woong-Ryeon Park and Yusuke Nakamura, *p53CSV, a novel p53-inducible gene involved in the p53-dependent cell-survival pathway.*, Cancer Res **65** (2005), no. 4, 1197–1206.

[191] Brent J Passer, Vanessa Nancy-Portebois, Nathalie Amzallag, Sylvie Prieur, Christophe Cans, Aude Roborel de Climens, Giusy Fiucci, Veronique Bouvard, Marcel Tuynder, Laurent Susini, Stphanie Morchoisne, Virginie Crible, Alexandra Lespagnol, Jean Dausset, Moshe Oren, Robert Amson, and Adam Telerman, *The p53-inducible TSAP6 gene product regulates apoptosis and the cell cycle and interacts with Nix and the Myt1 kinase.*, Proc Natl Acad Sci U S A **100** (2003), no. 5, 2284–2289.

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

[193] Erica J Peterson, Oliver Bgler, and Shirley M Taylor, *p53-mediated repression of DNA methyltransferase 1 expression by specific DNA binding.*, Cancer Res **63** (2003), no. 20, 6579–6582.

[194] P. A. Pevzner and S. H. Sze, *Combinatorial approaches to finding subtle signals in DNA sequences.*, Proc Int Conf Intell Syst Mol Biol **8** (2000), 269–278.

[195] A. Prakash, M. Blanchette, S. Sinha, and M. Tompa, *Motif discovery in heterogeneous sequence data.*, Pac Symp Biocomput (2004), 348–359.

[196] Anne-Catherine Prats and Herv Prats, *Translational control of gene expression: role of IRESs and consequences for cell transformation and angiogenesis.*, Prog Nucleic Acid Res Mol Biol **72** (2002), 367–413.

[197] J. Raimond, F. Rouleux, M. Monsigny, and A. Legrand, *The second intron of the human galectin-3 gene has a strong promoter activity down-regulated by p53.*, FEBS Lett **363** (1995), no. 1-2, 165–169.

[198] Nikolaus Rajewsky, Massimo Vergassola, Ulrike Gaul, and Eric D Siggia, *Computational detection of genomic cis-regulatory modules applied to body patterning in the early Drosophila embryo.*, BMC Bioinformatics **3** (2002), 30.

[199] Elizabeth E Reczek, Elsa R Flores, Alice S Tsay, Laura D Attardi, and Tyler Jacks, *Multiple response elements and differential p53 binding control Perp expression during apoptosis.*, Mol Cancer Res **1** (2003), no. 14, 1048–1057.

[200] Mireille Regnier and Alain Denise, *Rare events and Conditional Events on random strings*, DMTCS **6** (2004), no. 2, 191214.

[201] Bart Rikhof, Paul G Corn, and Wafik S El-Deiry, *Caspase 10 levels are increased following DNA damage in a p53-dependent manner.*, Cancer Biol Ther **2** (2003), no. 6, 707–712.

[202] Todd Riley, Eduardo Sontag, Patricia Chen, and Arnold Levine, *Transcriptional control of human p53-regulated genes.*, Nat Rev Mol Cell Biol **9** (2008), no. 5, 402–412.

[203] Harlan Robins, Michael Krasnitz, Hagar Barak, and Arnold J. Levine, *A Relative-Entropy Algorithm for Genomic Fingerprinting Captures Host-Phage Similarities*, J. Bacteriol. **187** (2005), no. 24, 8370–8374.

[204] Harlan Robins, Michael Krasnitz, and Arnold J Levine, *The computational detection of functional nucleotide sequence motifs in the coding regions of organisms.*, Exp Biol Med (Maywood) **233** (2008), no. 6, 665–673.

[205] A. I. Robles, N. A. Bemmels, A. B. Foraker, and C. C. Harris, *APAF-1 is a transcriptional target of p53 in DNA damage-induced apoptosis.*, Cancer Res **61** (2001), no. 18, 6660–6664.

[206] F. M. Rollwagen, Z. Y. Yu, Y. Y. Li, and N. D. Pacheco, *IL-6 rescues enterocytes from hemorrhage induced apoptosis in vivo and in vitro by a bcl-2 mediated mechanism.*, Clin Immunol Immunopathol **89** (1998), no. 3, 205–213.

[207] F. P. Roth, J. D. Hughes, P. W. Estep, and G. M. Church, *Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation.*, Nat Biotechnol **16** (1998), no. 10, 939–945.

[208] Galit Rozenfeld-Granot, Janakiraman Krishnamurthy, Karuppiah Kannan, Amos Toren, Ninette Amariglio, David Givol, and Gideon Rechavi, *A positive feedback mechanism in the transcriptional activation of Apaf-1 by p53 and the coactivator Zac-1.*, Oncogene **21** (2002), no. 10, 1469–1476.

[209] Subhashini Sadasivam, Sanjeev Gupta, Vegesna Radha, Kiran Batta, Tapas K Kundu, and Ghanshyam Swarup, *Caspase-1 activator Ipaf is a p53-inducible gene involved in apoptosis.*, Oncogene **24** (2005), no. 4, 627–636.

[210] Z. Saifudeen, H. Du, S. Dipp, and S. S. El-Dahr, *The bradykinin type 2 receptor is a target for p53-mediated transcriptional activation.*, J Biol Chem **275** (2000), no. 20, 15557–15562.

[211] K. Saigusa, I. Imoto, C. Tanikawa, M. Aoyagi, K. Ohno, Y. Nakamura, and J. Inazawa, *RGC32, a novel p53-inducible gene, is located on centrosomes during mitosis and results in G2/M arrest.*, Oncogene **26** (2007), no. 8, 1110–1121.

[212] S. Sakuma, H. Saya, M. Tada, M. Nakao, T. Fujiwara, J. A. Roth, Y. Sawamura, Y. Shinohe, and H. Abe, *Receptor protein tyrosine kinase DDR is up-regulated by p53 protein.*, FEBS Lett **398** (1996), no. 2-3, 165–169.

[213] Albin Sandelin, Wynand Alkema, Pr Engstrm, Wyeth W Wasserman, and Boris Lenhard, *JASPAR: an open-access database for eukaryotic transcription factor binding profiles.*, Nucleic Acids Res **32** (2004), no. Database issue, D91–D94.

[214] J. SantaLucia, *A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics.*, Proc Natl Acad Sci U S A **95** (1998), no. 4, 1460–1465.

[215] Anna Saramki, Claire M Banwell, Moray J Campbell, and Carsten Carlberg, *Regulation of the human p21(waf1/cip1) gene promoter via multiple binding sites for p53 and the vitamin D3 receptor.*, Nucleic Acids Res **34** (2006), no. 2, 543–554.

[216] Joanna K Sax, Peiwen Fei, Maureen E Murphy, Eric Bernhard, Stanley J Korsmeyer, and Wafik S El-Deiry, *BID regulation by p53 contributes to chemosensitivity.*, Nat Cell Biol **4** (2002), no. 11, 842–849.

[217] S. J. Scherer, S. M. Maier, M. Seifert, R. G. Hanselmann, K. D. Zang, H. K. Muller-Hermelink, P. Angel, C. Welter, and M. Schartl, *p53 and c-Jun functionally synergize in the regulation of the DNA repair gene hMSH2 in response to UV.*, J Biol Chem **275** (2000), no. 48, 37469–37473.

[218] T. D. Schneider and R. M. Stephens, *Sequence logos: a new way to display consensus sequences.*, Nucleic Acids Res **18** (1990), no. 20, 6097–6100.

[219] D. W. Seol, Q. Chen, M. L. Smith, and R. Zarnegar, *Regulation of the c-met proto-oncogene promoter by p53.*, J Biol Chem **274** (1999), no. 6, 3565–3572.

[220] E. Seto, A. Usheva, G. P. Zambetti, J. Momand, N. Horikoshi, R. Weinmann, A. J. Levine, and T. Shenk, *Wild-type p53 binds to the TATA-binding protein and represses transcription.*, Proc Natl Acad Sci U S A **89** (1992), no. 24, 12028–12032.

[221] Roded Sharan, Asa Ben-Hur, Gabriela G Loots, and Ivan Ovcharenko, *CREME: Cis-Regulatory Module Explorer for the human genome.*, Nucleic Acids Res **32** (2004), no. Web Server issue, W253–W256.

[222] Kazuhito Shida, *GibbsST: a Gibbs sampling method for motif discovery with enhanced resistance to local optima.*, BMC Bioinformatics **7** (2006), 486.

[223] Y. Shiio, T. Yamamoto, and N. Yamaguchi, *Negative regulation of Rb expression by the p53 gene product.*, Proc Natl Acad Sci U S A **89** (1992), no. 12, 5206–5210.

[224] T. H. Shin, A. J. Paterson, and J. E. Kudlow, *p53 stimulates transcription from the human transforming growth factor alpha promoter: a potential growth-stimulatory role for p53.*, Mol Cell Biol **15** (1995), no. 9, 4694–4701.

[225] K. Shiraishi, S. Fukuda, T. Mori, K. Matsuda, T. Yamaguchi, C. Tanikawa, M. Ogawa, Y. Nakamura, and H. Arakawa, *Identification of fractalkine, a CX3C-type chemokine, as a direct target of p53.*, Cancer Res **60** (2000), no. 14, 3722–3726.

[226] Jiang Shou, Francis Ali-Osman, Asha S Multani, Sen Pathak, Paolo Fedi, and Kalkunte S Srivenugopal, *Human Dkk-1, a gene encoding a Wnt antagonist, responds to DNA damage and its overexpression sensitizes brain tumor cells to apoptosis following alkylation damage of DNA.*, Oncogene **21** (2002), no. 6, 878–889.

[227] P. R. Sibbald and P. Argos, *Weighting aligned protein or nucleic acid sequences to correct for unequal representation.*, J Mol Biol **216** (1990), no. 4, 813–818.

[228] Rahul Siddharthan, Eric D Siggia, and Erik van Nimwegen, *PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny.*, PLoS Comput Biol **1** (2005), no. 7, e67.

[229] Saurabh Sinha, *Discriminative motifs.*, J Comput Biol **10** (2003), no. 3-4, 599–615.

[230] Saurabh Sinha and Martin Tompa, *YMF: A program for discovery of novel transcription factor binding sites by statistical overrepresentation.*, Nucleic Acids Res **31** (2003), no. 13, 3586–3588.

[231] N. Slonim, N. Friedman, and N. Tishby, *Unsupervised document classification using sequential information maximization*, In Proceeding of SIGIR'02, 25th ACM intermational Conference on Research and Development of Information Retireval, Tampere, Finland, 2002. ACM Press, New York, USA, 2002.

[232] M. L. Smith, I. T. Chen, Q. Zhan, I. Bae, C. Y. Chen, T. M. Gilmer, M. B. Kastan, P. M. O'Connor, and A. J. Fornace, *Interaction of the p53-regulated protein Gadd45 with proliferating cell nuclear antigen.*, Science **266** (1994), no. 5189, 1376–1380.

[233] T. Soussi, *p53 alterations in human cancer: more questions than answers.*, Oncogene **26** (2007), no. 15, 2145–2156.

[234] V. Stambolic, D. MacPherson, D. Sas, Y. Lin, B. Snow, Y. Jang, S. Benchimol, and T. W. Mak, *Regulation of PTEN transcription by p53.*, Mol Cell **8** (2001), no. 2, 317–325.

[235] Alexander Stark, Michael F Lin, Pouya Kheradpour, Jakob S Pedersen, Leopold Parts, Joseph W Carlson, Madeline A Crosby, Matthew D Rasmussen, Sushmita Roy, Ameya N Deoras, J. Graham Ruby, Julius Brennecke, Harvard FlyBase curators, Berkeley Drosophila Genome Project, Emily Hodges, Angie S Hinrichs, Anat Caspi, Benedict Paten, Seung-Won Park, Mira V Han, Morgan L Maeder, Benjamin J Polansky, Bryanne E Robson, Stein Aerts, Jacques van Helden, Bassem Hassan, Donald G Gilbert, Deborah A Eastman, Michael Rice, Michael Weir, Matthew W Hahn, Yongkyu Park, Colin N Dewey, Lior Pachter, W. James Kent, David Haussler, Eric C Lai, David P Bartel, Gregory J Hannon, Thomas C Kaufman, Michael B Eisen, Andrew G Clark, Douglas Smith, Susan E Celniker, William M Gelbart, and Manolis

Kellis, *Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures.*, Nature **450** (2007), no. 7167, 219–232.

[236] Susanne Stein, Emily K Thomas, Birger Herzog, Matthew D Westfall, Jonathan V Rocheleau, Roger S Jackson, Mai Wang, and Peng Liang, *NDRG1 is necessary for p53-dependent apoptosis.*, J Biol Chem **279** (2004), no. 47, 48930–48940.

[237] J. E. Stenger, P. Tegtmeyer, G. A. Mayr, M. Reed, Y. Wang, P. Wang, P. V. Hough, and I. A. Mastrangelo, *p53 oligomerization and DNA looping are linked with transcriptional activation.*, EMBO J **13** (1994), no. 24, 6011–6020.

[238] Gary D. Stormo, *DNA binding sites: representation and discovery*, Bioinformatics **16** (2000), no. 1, 16–23.

[239] G.D. Stormo and D.S. Fields, *Specificity, free energy and information content in protein-DNA interactions*, Trends in Biochemical Sciences **23** (1 March 1998), 109–113(5).

[240] K. Subbaramaiah, N. Altorki, W. J. Chung, J. R. Mestre, A. Sampat, and A. J. Dannenberg, *Inhibition of cyclooxygenase-2 gene expression by p53.*, J Biol Chem **274** (1999), no. 16, 10911–10915.

[241] Yubo Sun, Xiao-Rong Zeng, Leonor Wenger, Gary S Firestein, and Herman S Cheung, *P53 down-regulates matrix metalloproteinase-1 by targeting the communications between AP-1 and the basal transcription complex.*, J Cell Biochem **92** (2004), no. 2, 258–269.

[242] Mori T., Anazawa Y., Matsui K., Fukuda S., Nakamura Y., Arakawa H., and Correspondence, *Cyclin K as a Direct Transcriptional Target of the p53 Tumor Suppressor*, Neoplasia **4** (1 May 2002), 268–274(7).

[243] R. Takimoto and W. S. El-Deiry, *Wild-type p53 transactivates the KILLER/DR5 gene through an intronic sequence-specific DNA-binding site.*, Oncogene **19** (2000), no. 14, 1735–1743.

[244] M. Tan, C. W. Heizmann, K. Guan, B. W. Schafer, and Y. Sun, *Transcriptional activation of the human S100A2 promoter by wild-type p53.*, FEBS Lett **445** (1999), no. 2-3, 265–268.

[245] M. Tan, Y. Wang, K. Guan, and Y. Sun, *PTGF-beta, a type beta transforming growth factor (TGF-beta) superfamily member, is a p53 target gene that inhibits tumor cell growth via TGF-beta signaling pathway.*, Proc Natl Acad Sci U S A **97** (2000), no. 1, 109–114.

[246] Thomas Tan and Gilbert Chu, *p53 Binds and activates the xeroderma pigmentosum DDB2 gene in humans but not mice.*, Mol Cell Biol **22** (2002), no. 10, 3247–3254.

[247] H. Tanaka, H. Arakawa, T. Yamaguchi, K. Shiraishi, S. Fukuda, K. Matsui, Y. Takei, and Y. Nakamura, *A ribonucleotide reductase gene involved in a p53-dependent cell-cycle checkpoint for DNA damage.*, Nature **404** (2000), no. 6773, 42–49.

[248] Gert Thijs, Magali Lescot, Kathleen Marchal, Stephane Rombauts, Bart De Moor, Pierre Rouze, and Yves Moreau, *A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling*, Bioinformatics **17** (December 2001), 1113–1122(10).

[249] Gert Thijs, Kathleen Marchal, Magali Lescot, Stephane Rombauts, Bart De Moor, Pierre Rouz, and Yves Moreau, *A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes.*, J Comput Biol **9** (2002), no. 2, 447–464.

[250] J. D. Thompson, D. G. Higgins, and T. J. Gibson, *Improved sensitivity of profile searches through the use of sequence weights and gap excision.*, Comput Appl Biosci **10** (1994), no. 1, 19–29.

[251] E. C. Thornborrow and J. J. Manfredi, *The tumor suppressor protein p53 requires a cofactor to activate transcriptionally the human BAX promoter.*, J Biol Chem **276** (2001), no. 19, 15598–15608.

[252] Edward C Thornborrow, Sejal Patel, Anthony E Mastropietro, Elissa M Schwartzfarb, and James J Manfredi, *A conserved intronic response element mediates direct p53-dependent transcriptional activation of both the human and murine bax genes.*, Oncogene **21** (2002), no. 7, 990–999.

[253] C. J. Thut, J. L. Chen, R. Klemm, and R. Tjian, *p53 transcriptional activation mediated by coactivators TAFII40 and TAFII60.*, Science **267** (1995), no. 5194, 100–104.

[254] T. Tokino, S. Thiagalingam, W. S. el Deiry, T. Waldman, K. W. Kinzler, and B. Vogelstein, *p53 tagged sites from human genomic DNA.*, Hum Mol Genet **3** (1994), no. 9, 1537–1542.

[255] M. Tompa, *An exact method for finding short motifs in sequences, with application to the ribosome binding site problem.*, Proc Int Conf Intell Syst Mol Biol (1999), 262–271.

[256] Martin Tompa, Nan Li, Timothy L Bailey, George M Church, Bart De Moor, Eleazar Eskin, Alexander V Favorov, Martin C Frith, Yutao Fu, W. James Kent, Vsevolod J Makeev, Andrei A Mironov, William Stafford Noble, Giulio Pavesi, Graziano Pesole, Mireille Rgnier, Nicolas Simonis, Saurabh Sinha, Gert Thijs, Jacques van Helden, Mathias Vandenbogaert, Zhiping Weng, Christopher Workman, Chun Ye, and Zhou Zhu, *Assessing computational tools for the discovery of transcription factor binding sites.*, Nat Biotechnol **23** (2005), no. 1, 137–144.

[257] F. Tronche, F. Ringeisen, M. Blumenfeld, M. Yaniv, and M. Pontoglio, *Analysis of the distribution of binding sites for a tissue-specific transcription factor in the vertebrate genome.*, J Mol Biol **266** (1997), no. 2, 231–245.

[258] R. Truant, H. Xiao, C. J. Ingles, and J. Greenblatt, *Direct interaction between the transcriptional activation domain of human p53 and the TATA box-binding protein.*, J Biol Chem **268** (1993), no. 4, 2284–2287.

[259] D. H. Turner, N. Sugimoto, J. A. Jaeger, C. E. Longfellow, S. M. Freier, and R. Kierzek, *Improved parameters for prediction of RNA structure.*, Cold Spring Harb Symp Quant Biol **52** (1987), 123–133.

[260] J. Van Uden and E. Raz, *Introduction to immunostimulatory DNA sequences.*, Springer Semin Immunopathol **22** (2000), no. 1-2, 1–9.

[261] Koji Ueda, Hirofumi Arakawa, and Yusuke Nakamura, *Dual-specificity phosphatase 5 (DUSP5) as a direct transcriptional target of tumor suppressor p53.*, Oncogene **22** (2003), no. 36, 5586–5591.

[262] T. Urano, H. Nishimori, H. Han, T. Furuhata, Y. Kimura, Y. Nakamura, and T. Tokino, *Cloning of P2XM, a novel human P2X receptor gene regulated by p53.*, Cancer Res **57** (1997), no. 15, 3281–3287.

[263] J. van Helden, B. Andr, and J. Collado-Vides, *Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies.*, J Mol Biol **281** (1998), no. 5, 827–842.

[264] J. van Helden, M. del Olmo, and J. E. Prez-Ortn, *Statistical analysis of yeast genomic downstream sequences reveals putative polyadenylation signals.*, Nucleic Acids Res **28** (2000), no. 4, 1000–1010.

[265] J. van Helden, A. F. Rios, and J. Collado-Vides, *Discovering regulatory elements in non-coding sequences by analysis of spaced dyads.*, Nucleic Acids Res **28** (2000), no. 8, 1808–1818.

[266] S. Velasco-Miguel, L. Buckbinder, P. Jean, L. Gelbert, R. Talbott, J. Laidlaw, B. Seizinger, and N. Kley, *PA26, a novel target of the p53 tumor suppressor and member of the GADD family of DNA damage and growth arrest inducible genes.*, Oncogene **18** (1999), no. 1, 127–137.

[267] B. Vogelstein, D. Lane, and A. J. Levine, *Surfing the p53 network.*, Nature **408** (2000), no. 6810, 307–310.

[268] Karen H Vousden, *Activation of the p53 tumor suppressor protein.*, Biochim Biophys Acta **1602** (2002), no. 1, 47–59.

[269] Luquan Wang, Qun Wu, Ping Qiu, Asra Mirza, Marnie McGuirk, Paul Kirschmeier, Jonathan R. Greene, Yaolin Wang, Cecil B. Pickett, and Suxing Liu, *Analyses of p53 Target Genes in the Human Genome by Bioinformatic and Microarray Approaches*, J. Biol. Chem. **276** (2001), no. 47, 43604–43610.

[270] Ting Wang and Gary D Stormo, *Identifying the conserved network of cis-regulatory sites of a eukaryotic genome.*, Proc Natl Acad Sci U S A **102** (2005), no. 48, 17400–17405.

[271] Y. Wang, J. F. Schwedes, D. Parks, K. Mann, and P. Tegtmeyer, *Interaction of p53 with its consensus DNA-binding site.*, Mol Cell Biol **15** (1995), no. 4, 2157–2165.

[272] C. T. Warnick, B. Dabbas, C. D. Ford, and K. A. Strait, *Identification of a p53 response element in the promoter region of the hMSH2 gene required for expression in A2780 ovarian cancer cells.*, J Biol Chem **276** (2001), no. 29, 27363–27370.

[273] Wyeth W Wasserman and Albin Sandelin, *Applied bioinformatics for the identification of regulatory elements.*, Nat Rev Genet **5** (2004), no. 4, 276–287.

[274] Chia-Lin Wei, Qiang Wu, Vinsensius B Vega, Kuo Ping Chiu, Patrick Ng, Tao Zhang, Atif Shahab, How Choong Yong, YuTao Fu, Zhiping Weng, JianJun Liu, Xiao Dong Zhao, Joon-Lin Chew, Yen Ling Lee, Vladimir A Kuznetsov, Wing-Kin Sung, Lance D Miller, Bing Lim, Edison T Liu, Qiang Yu, Huck-Hui Ng, and Yijun Ruan, *A global map of p53 transcription-factor binding sites in the human genome.*, Cell **124** (2006), no. 1, 207–219.

[275] Richard L Weinberg, Dmitry B Veprintsev, Mark Bycroft, and Alan R Fersht, *Comparative binding of p53 to its promoter and DNA recognition elements.*, J Mol Biol **348** (2005), no. 3, 589–596.

[276] T. Werner, *Models for prediction and recognition of eukaryotic promoters.*, Mamm Genome **10** (1999), no. 2, 168–175.

[277] G. S. Wu, P. Saftig, C. Peters, and W. S. El-Deiry, *Potential role for cathepsin D in p53-dependent tumor suppression and chemosensitivity.*, Oncogene **16** (1998), no. 17, 2177–2183.

[278] Min Wu, Liang-Guo Xu, Tian Su, Yang Tian, Zhonghe Zhai, and Hong-Bing Shu, *AMID is a p53-inducible gene downregulated in tumors.*, Oncogene **23** (2004), no. 40, 6815–6819.

[279] Xiaohui Xie, Tarjei S Mikkelsen, Andreas Gnirke, Kerstin Lindblad-Toh, Manolis Kellis, and Eric S Lander, *Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites.*, Proc Natl Acad Sci U S A **104** (2007), no. 17, 7145–7150.

[280] Junli Yan, Jianming Jiang, Ching Aeng Lim, Qiang Wu, Huck-Hui Ng, and Keh-Chuang Chin, *BLIMP1 regulates cell growth through repression of p53 transcription.*, Proc Natl Acad Sci U S A **104** (2007), no. 6, 1841–1846.

[281] Gong Yang, Daniel G Rosen, Zhihong Zhang, Robert C Bast, Gordon B Mills, Justin A Colacino, Imelda Mercado-Uribe, and Jinsong Liu, *The chemokine growth-regulated oncogene 1 (Gro-1) links RAS signaling to the senescence of stromal fibroblasts and ovarian tumorigenesis.*, Proc Natl Acad Sci U S A **103** (2006), no. 44, 16472–16477.

[282] Gene Yeo and Christopher B Burge, *Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals.*, J Comput Biol **11** (2004), no. 2-3, 377–394.

[283] Xia Yin, Beatriz M A Fontoura, Takashi Morimoto, and Robert B Carroll, *Cytoplasmic complex of p53 and eEF2.*, J Cell Physiol **196** (2003), no. 3, 474–482.

[284] E. Yonish-Rouach, D. Resnitzky, J. Lotem, L. Sachs, A. Kimchi, and M. Oren, *Wild-type p53 induces apoptosis of myeloid leukaemic cells that is inhibited by interleukin-6.*, Nature **352** (1991), no. 6333, 345–347.

[285] Heejei Yoon, Sandya Liyanarachchi, Fred A Wright, Ramana Davuluri, Janet C Lockman, Albert de la Chapelle, and Natalia S Pellegata, *Gene expression profiling of isogenic cells with different TP53 gene dosage reveals numerous genes that are affected by TP53 dosage and identifies CSPG2 as a direct target of p53.*, Proc Natl Acad Sci U S A **99** (2002), no. 24, 15632–15637.

[286] Xin Yu, Sandra L Harris, and Arnold J Levine, *The regulation of exosome secretion: a novel function of the p53 protein.*, Cancer Res **66** (2006), no. 9, 4795–4801.

[287] Xin Yu and A. J. Levine, *p53 regulation of the genes in the endosome compartment and autophagy*, 2007.

[288] J. Yun, H. D. Chae, H. E. Choy, J. Chung, H. S. Yoo, M. H. Han, and D. Y. Shin, *p53 negatively regulates cdc2 transcription via the CCAAT-binding NF-Y transcription factor.*, J Biol Chem **274** (1999), no. 42, 29677–29682.

[289] A. Zauberman, D. Flusberg, Y. Haupt, Y. Barak, and M. Oren, *A functional p53-responsive intronic promoter is contained within the human mdm2 gene.*, Nucleic Acids Res **23** (1995), no. 14, 2584–2592.

[290] Chun Zhang, Choungfeng Gao, Junya Kawauchi, Yoshinori Hashimoto, Nobuo Tsuchida, and Shigetaka Kitajima, *Transcriptional activation of the human stress-inducible transcriptional repressor ATF3 gene promoter by p53.*, Biochem Biophys Res Commun **297** (2002), no. 5, 1302–1310.

[291] D. W. Zhang, K-T. Jeang, and C. G L Lee, *p53 negatively regulates the expression of FAT10, a gene upregulated in various cancers.*, Oncogene **25** (2006), no. 16, 2318–2327.

[292] Ye Zhang, Jin-Shan Wang, Li-Ling Chen, Yong Zhang, Xiao-Kuan Cheng, Feng-Yan Heng, Ning-Hua Wu, and Yu-Fei Shen, *Repression of hsp90beta gene by p53 in UV irradiation-induced apoptosis of Jurkat cells.*, J Biol Chem **279** (2004), no. 41, 42545–42551.

[293] R. Zhao, K. Gish, M. Murphy, Y. Yin, D. Notterman, W. H. Hoffman, E. Tom, D. H. Mack, and A. J. Levine, *Analysis of p53-regulated gene expression patterns using oligonucleotide arrays.*, Genes Dev **14** (2000), no. 8, 981–993.

[294] Qing Zhou and Jun S Liu, *Modeling within-motif dependence for transcription factor binding site predictions.*, Bioinformatics **20** (2004), no. 6, 909–916.

[295] J. Zhu and X. Chen, *MCG10, a novel p53 target gene that encodes a KH domain RNA-binding protein, is capable of inducing apoptosis and cell cycle arrest in G(2)-M.*, Mol Cell Biol **20** (2000), no. 15, 5602–5618.

[296] Z. Zou, C. Gao, A. K. Nagaich, T. Connell, S. Saito, J. W. Moul, P. Seth, E. Appella, and S. Srivastava, *p53 regulates the expression of the tumor suppressor gene maspin.*, J Biol Chem **275** (2000), no. 9, 6051–6054.

[297] M. Zuker, *Computer prediction of RNA structure.*, Methods Enzymol **180** (1989), 262–288.

[298] ———, *On finding all suboptimal foldings of an RNA molecule.*, Science **244** (1989), no. 4900, 48–52.

[299] M. Zuker and P. Stiegler, *Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information.*, Nucleic Acids Res **9** (1981), no. 1, 133–148.

Curriculum Vita

Todd Robert Riley

2008 Ph.D. in Computational Biology and Molecular BioPhysics, Rutgers University

2004 M. Sc. in Applied Mathematics, University of North Carolina, Chapel Hill

1992 B. Sc. in Computer Science/Mathematics, Carnegie Mellon University

2005-08 Research Fellow, Simons Center for Systems Biology, Institute for Advanced Study

2005 Graduate Research Assistant, Rutgers University

2004-05 GAANN Fellowship, Rutgers University

2008 T. Riley, E. Sontag, P. Chen, A. Levine: The Transcriptional Control of Human p53-Regulated Genes. *Nature Reviews Molecular Cell Biology.* 9, 402-412