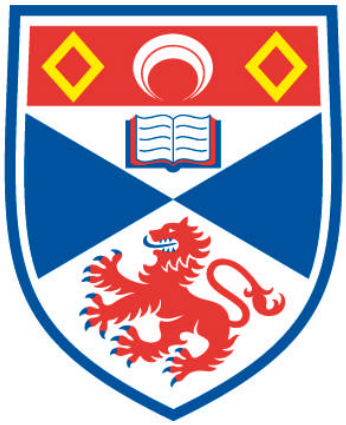
A COMPUTATIONAL APPROACH TO DISCOVERING P53 BINDING SITES IN THE HUMAN GENOME

Ji-Hyun Lim

A Thesis Submitted for the Degree of PhD
at the
University of St Andrews



2012

University of
St Andrews

School of Biology

PhD Thesis

A computational approach to discovering p53 binding sites in the human genome

by

Ji-Hyun Lim

July 2012

Abstract

The tumour suppressor p53 protein plays a central role in the DNA damage response/checkpoint pathways leading to DNA repair, cell cycle arrest, apoptosis and senescence. The activation of p53-mediated pathways is primarily facilitated by the binding of tetrameric p53 to two 'half-sites', each consisting of a decameric p53 response element (RE). Functional REs are directly adjacent or separated by a small number of 1-13 'spacer' base pairs (bp). The p53 RE is detected by exact or inexact matches to the palindromic sequence represented by the regular expression [AG][AG][AG]C[AT][TA]G[TC][TC][TC] or a position weight matrix (PWM). The use of matrix-based and regular expression pattern-matching techniques, however, leads to an overwhelming number of false positives. A more specific model, which combines multiple factors known to influence p53-dependent transcription, is required for accurate detection of the binding sites.

In this thesis, we present a logistic regression based model which integrates sequence information and epigenetic information to predict human p53 binding sites. Sequence information includes the PWM score and the spacer length between the two half-sites of the observed binding site. To integrate epigenetic information, we analyzed the surrounding region of the binding site for the presence of mono- and trimethylation patterns of histone H3 lysine 4 (H3K4). Our model showed a high level of performance on both a high-resolution data set of functional p53 binding sites from the experimental literature (ChIP data) and the whole human genome. Comparing our model with a simpler sequence-only model, we demonstrated that the prediction accuracy of the sequence-only model could be improved by incorporating epigenetic information, such as the two histone modification marks H3K4me1 and H3K4me3.

Declarations

I, Ji-Hyun Lim, hereby certify that this thesis, which is approximately 26,600 words in length, has been written by me, that it is the record of work carried out by me and that it has not been submitted in any previous application for a higher degree.

I was admitted as a research student in August 2008 and as a candidate for the degree of PhD in August 2009; the higher study for which this is a record was carried out in the University of St Andrews between 2008 and 2012.

date _____ Signature of candidate _____

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of PhD in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree.

date _____ Signature of supervisor _____

In submitting this thesis to the University of St Andrews we understand that we are given permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. We also understand that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that my thesis will be electronically accessible for personal or research use unless exempt by award of an embargo as requested below, and that the library has the right to migrate my thesis into new electronic forms as required to ensure continued access to the thesis. We have obtained any third-party copyright permissions that may be required in order to allow such access and migration, or have requested the appropriate embargo below.

The following is an agreed request by candidate and supervisor regarding the electronic publication of this thesis: Access to printed copy and electronic publication of thesis through the University of St Andrews.

date _____ Signature of candidate _____

date _____ Signature of supervisor _____

Acknowledgement

I would like to thank my two excellent superviors, Daniel Barker and Richard Iggo, for their support and patient guidance.

I would also like to thank Dave Ferrier and Richard Abbott for helpful advice and feedback at review meetings.

I thank the School of Medicine and the BBSRC for providing the funding that gave me the opportunity to undertake this research.

Finally, I offer special thanks to my family and friends for their endless love, support and help.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

With the accumulation of large amounts of human genomic sequence data through significant improvements in high-throughput sequencing technology, bioinformatics is becoming increasingly important and useful for modern biological research. A very important, but challenging application of bioinformatics is the genome-wide identification of transcription factor binding sites. The comprehensive mapping of transcription factor binding sites is of great importance to better understand the complex mechanisms which regulate gene expression.

1.1 Regulation of eukaryotic transcription

The process of gene expression in eukaryotes involves several stages, including the two major processes of transcription and translation. During transcription, the genetic information in DNA is transferred to RNA. The process of transcription consists of the three main steps: initiation, elongation and termination. Transcription initiation begins with the binding of RNA polymerase, a key enzyme essential for carrying out transcription, to the DNA at a gene promoter. The appropriate DNA binding of the enzyme is achieved by specific proteins which form a multi-component complex called transcription initiation complex with the RNA polymerase. In eukaryotes, there are three types of RNA polymerases (RNA polymerase I, RNA polymerase II and RNA polymerase III) which activate the transcription of distinct classes of RNAs in the nucleus. RNA polymerase I is critical for the transcriptional activation of the genes coding for the 28S, 18S and 5.8S ribosomal RNAs (rRNAs). The genes which code for transfer RNAs and the 5S rRNA are transcribed by RNA polymerase III. RNA polymerase II, the most extensively studied enzyme of the three types, transcribes various small nuclear RNA genes and all the genes coding proteins (Latchman, 2008c). During elongation, the RNA polymerase attached to the promoter moves along the DNA,

unwinds the double-stranded DNA and begins to synthesize the RNA transcript by using one of two single-stranded DNA as a template until the termination sequence is reached. In the final step of termination, the complex of RNA polymerase dissociates and releases the synthesized RNA transcript from the template DNA.

The immediate RNA transcript produced from the DNA, called the primary transcript, must undergo post-transcriptional modifications, such as 5′ capping, 3′ polyadenylation and RNA splicing, to form a mature messenger RNA (mRNA) molecule. The process of 5′ capping happens shortly after the start of transcription and involves the addition of a 7-methylguanosine residue to the 5′ end of the transcript which is essential to protect the 5′ end from attack by 5′-exonucleases and thus to stabilize the transcript. After transcription, the initial RNA product is modified by polyadenylating its 3′ end. A poly-A tail consisting of up to 250 adenine nucleotides is added at the 3′ end to protect the transcript from enzymatic degradation. Furthermore, internal non-coding sequences, called introns, are removed and the remaining exons are joined together to form the mature, functional mRNA molecule during the splicing process. Transported to the cytoplasm, the mRNA is then translated into the final protein.

Although the genome is the same within each cell, there are many different types of cells in the human body. The set of activated and inactivated genes is different in every cell. The expression of genes in a particular cell type or in response to a particular signal is controlled by regulatory mechanisms involving transcription factors and their DNA binding sequences, chromatin structure and histone modifications. Gene expression is known to be regulated at any stage from RNA synthesis and RNA translation to protein degradation and the various processes and factors involved in the regulation are interconnected (Lackner and Bähler, 2008). A particularly extensively studied area in gene expression control is that of the regulation at the level of transcription.

1.1.1 Transcription factors

The transcription process is controlled by trans-regulatory proteins called transcription factors. Many transcription factors are able to bind to DNA and influence the rate of transcription either positively (activator) or negatively (repressor). There are three major types of transcription factors: basal factors, also called general transcription factors, activators/repressors and coactivators/corepressors (Lewin, 2004). General transcription factors, such as TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH, are essential to form the transcription initiation complex with RNA polymerase II (Campbell and Farrell, 2007). The stability

and activity of the complex can be regulated by the binding of other proteins to sites in the promoter or in regions further away from the transcriptional start site. Proteins which positively influence the transcription by increasing the rate of transcription so that more RNA transcripts are synthesized are called activators. In contrast, proteins exerting an inhibitory effect on transcript are known as repressors. Several mechanisms have been identified by which transcription factors achieve their role as a repressor. A repressor may interfere with the action of a positively acting factor by preventing it binding to DNA and activating gene expresssion. Other mechanisms involve direct repression by regulating the transcription initiation complex with the RNA polymerase. In many cases, this requires specific coregulators acting as coactivators or corepressors which serve to link the activator or repressor protein to the complex to either stimulate or inhibit transcription.

1.1.2 DNA sequence elements involved in transcription and its regulation

Transcription factors recognize their DNA binding sites by short conserved sequence elements called motifs. A typical eukaryotic gene consists of multiple distinct cis-acting elements which are essential for the basic process of transcription or involved in transcription regulation. Such control elements have been observed in promoters, upstream of promoters, and even in regions which are at a great distance away from the transcriptional start site of genes.

Core elements in the promoter region (Smale and Kadonaga, 2003) bind the general transcription factors and RNA polymerase II. One of the best characterized core promoter element is the AT-rich TATA box (Lifton et al., 1978) found in many eukaryotic genes. The TATA box is located approximately 25-30 bp upstream of the transcription start site and binds the TATA box binding protein (TBP), a component of TFIID (Burley and Roeder, 1996). It determines the location at which the general transcription factors together with RNA polymerase II assemble to form the initiation complex. In genes lacking the TATA box, this function is assigned to the initiator element (Inr).

Further control elements called upstream promoter elements (UPE) have been found upstream of the core elements. UPEs, such as the GC (Sp1) box and the CAAT box, can modulate the activity of a promoter in conjunction with other regulatory elements which are also located close to the promoter.

A variety of sequence elements have been observed in enhancers located upstream, downstream or within a gene. Enhancers contain multiple binding sites for transcription factors which interact together to form a functional multi-protein complex known as the

enhanceosome. The enhanceosome may mediate the binding of other proteins which are required to loose the tightly packed chromatin structure to facilitate the binding of several activators to their binding sites. Activators bound at distal enhancers can increase the activity of a promoter by direct interaction with components of the initiation complex through DNA looping. In some cases, the activator does not directly interact with the proteins at the promoter. A cofactor is required which acts as a bridge so that the activator can regulate the activity of the promoter. In contrast to enhancers, there are sequences called silencers which act in a negative manner to inhibit promoter function and thus decrease the rate of transcription. Silencers contain binding sites for repressors which inhibit transcription by altering chromatin structure or by directly interacting with the proteins bound at the promoter to negatively influence promoter activity. In some cases, the activity of enhancers or silencers has to be limited so that only a certain gene is affected and not the genes in adjacent regions. This is achieved by regulatory elements called insulators.

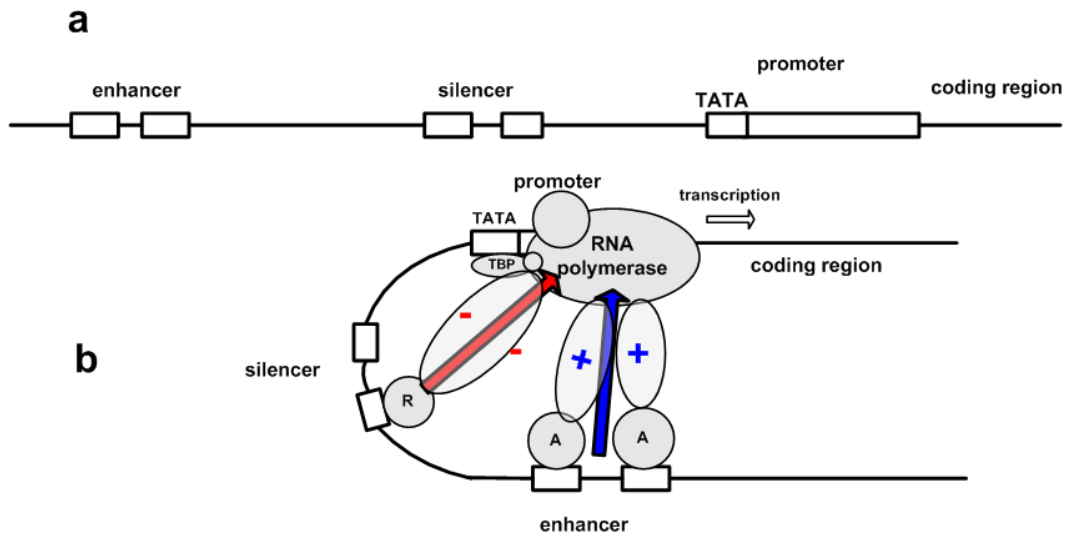


Figure 1.1 Structure of a typical eukaryotic gene with its regulatory elements. (a) Linear representation of a gene structure. (b) Interaction of transcription factors with the initiation complex via DNA looping. The TATA box binding protein (TBP) binds to the TATA box and forms a multi-protein complex called transcription initiation complex with the RNA polymerase and general transcription factors. Activators (A) and repressors (R) can influence transcription initiation of a particular gene by binding to enhancers and silencers, respectively, and interacting with the initiation complex.

1.1.3 Epigenetic factors

The ability of the DNA sequence elements to regulate gene expression together with their associated proteins depends on their accessibility. The chromatin structure plays a critical role in the transcription and regulation of genes. In eukaryotic cells, the DNA is wrapped twice around a complex of eight proteins called histones ($(\mathrm{H2A} \cdot \mathrm{H2B})_2(\mathrm{H3} \cdot \mathrm{H4})_2$) forming nucleosomes which are bundled into chromatin fibres. The binding affinity of a DNA binding protein to its binding sites can be affected by the chromatin architecture of the binding site. To make the DNA region accessible to the regulatory protein, alterations in chromatin structure by chromatin remodelling factors, such as the SWI/SNF complex, are required (Latchman, 2008a). In addition to the chromatin remodelling factors, histone modifications mediating chromatin changes play an important role in the regulation of gene expression. Histones are subject to multiple post-translational modifications, such as acetylation, methylation, phosphorylation and ubiquitination which can interact with each other. Acetylation of histone H3 on lysine residues (Roh et al., 2005) and phosphorylation of serine residues (Nowak and Corces, 2004) have been shown to produce a more open chromatin structure and thus to be critical for gene activation. Methylation of lysine residues, however, can produce both a more open or a more closed chromatin structure (Martin and Zhang, 2005).

1.2 Focus on p53

The gene encoding p53 (TP53) was initially described as an oncogene in 1979 (Deleo et al., 1979; Kress et al., 1979; Lane and Crawford, 1979; Linzer and Levine, 1979; Melero et al., 1979; Smith et al., 1979). In the late 80's, Bert Vogelstein et al. observed in human tumours that TP53 was mutated or lost and demonstrated that p53 might function as a tumour suppressor (Baker et al., 1989).

In its role as a transcription factor, the tumour suppressor p53 exerts its function by binding to DNA and activating expression of specific genes which are critical to inhibit the growth and proliferation of damaged cells. Many of these genes have been revealed to be involved in the regulation of cell cycle progression and in the distinct processes of cell death. In more than half of the human tumours, TP53 has been found to be mutated causing single residue changes (missense mutation) in the DNA binding domain of the protein with resultant alteration in p53 function (Sigal and Rotter, 2000).

1.2.1 Structural features of p53

The human p53 protein with a length of 393 amino acids consists of three main functional regions: the amino-(N)-terminal region, the central core and the carboxyl-(C)-terminal region.

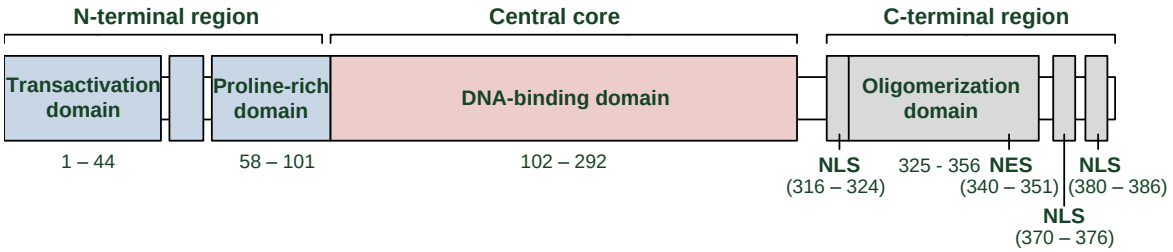


Figure 1.2 Functional domains of human p53. Numbering below the functional domains indicates residue number on human p53. Residue numbers were obtained from the p53 Knowledgebase (`http://p53.bii.a-star.edu.sg/aboutp53/protseq/index.php`).

The acidic amino-(N)-terminal region contains the transactivation domain and a proline-rich domain. The transactivation domain plays a critical role in transcriptional activation and regulation of p53. It is essential for post-translational modifications and interactions with numerous transcription factors, acetyltransferases, such as p300 and CBP, and the MDM2 ubiquitin ligase, a negative regulator of p53 (Espinosa and Emerson, 2001; Gu et al., 1997; Stommel and Wahl, 2004). The proline-rich domain has been shown to be important for efficient growth suppression which may be resulted from apoptosis (Baptiste et al., 2002; Bergamaschi et al., 2006; Millau et al., 2009; Sakamuro et al., 1997; Venot et al., 1998; Walker and Levine, 1996; Zhu et al., 1999). Furthermore, a negative regulatory domain has been identified within the proline-rich domain which negatively influences the sequence-specific DNA binding ability of p53 (Müller-Tiemann et al., 1998).

The sequence-specific DNA binding domain in the central core is essential for DNA binding of p53 and interacts with various proteins, such as the apoptosis stimulating p53 proteins (ASPPs) and the family members p63 and p73 (Flores et al., 2002; Patel et al., 2008; Samuels-Lev et al., 2001). It represents the region of p53 most affected by missense mutations.

The carboxyl-(C)-terminal region is able to bind DNA non-specifically (Bayle et al., 1995; Wang et al., 1993). It contains three nuclear localization signals (NLS), the oligomerization domain with a nuclear export signal (NES) and a negative regulatory domain at the very

end of the region. The nuclear localization signals are essential for the migration of p53 into the cell nucleus and the oligomerization domain for both dimer and tetramer formation of the protein.

1.2.2 Activation of p53 and cellular response

The activity of p53 is restrained under normal conditions. In unstressed cells, p53 is usually a protein with a short half-life of several minutes and is present at low levels. A carefully controlled balance between protein synthesis and degradation is required to keep the levels of p53 low. This is achieved by several distinct positive and negative autoregulatory feedback loops. Each of the loops creates a circuit in which p53-induced proteins regulate the activity and stability of the p53 protein in a positive (positive feedback loop) or negative manner (negative feedback loop) (Harris and Levine, 2005). Most of the known feedback loops act through the MDM2 ubiquitin ligase, a p53-responsive gene product which inhibits the transcriptional activity of p53. Bound to the transactivation domain of p53, MDM2 ubiquitylates several Lys residues at the C-terminus to trigger degradation of p53 by proteasomes (Rodriguez et al., 2000).

A wide range of cellular stresses, including DNA damage, oncogene activation, hypoxia, nutrient deprivation, nucleotide imbalances, ROS level and heat shock, lead to the induction and activation of p53 (Lavin and Gueven, 2006; Levine et al., 2006; Meek, 1999; Millau et al., 2009; Miyakoda et al., 2002) through multiple mechanisms involving post-translational modifications of the protein (Figure 1.3). Once the cell senses stress, signal mediators, such as ATM (ataxia-telangiectasia mutated), ATR (ataxia-telangiectasia mutated and Rad3-related) and Chk2 (checkpoint kinase 2) activate and stabilize the p53 protein by phosphorylating residues in the N-terminal transactivation domain of p53 (Jenkins et al., 2012). N-terminal phosphorylations have been shown to mediate the dissociation of MDM2-p53 complex causing an increase in p53 levels. Stress-induced phosphorylations have also been observed in the C-terminal region of p53. These phosphorylations have been shown to increase the sequence-specific DNA binding of p53. The p53 protein can be further stabilized through acetylation and methylation of C-terminal lysine residues by several acetyltransferases, such as CBP, p300 and PCAF, and the lysine-specific methyltransferase Set9, respectively. Once activated, the tetrameric p53 protein binds to its DNA response elements and recruits various cofactors such as histone modifying enzymes, chromatin remodeling factors and components of the transcription initiation complex (Beckerman and Prives, 2010) to regulate the transcription of its target genes. The p53 protein is known to transactivate genes which

encode proteins involved in DNA repair, cell-cycle arrest, apoptosis and senescence. But the list of target genes and functions regulated by p53 increases regularly as more information becomes available (Vousden, 2002).

1.2.3 Transcriptional Regulation by p53 activity

The activation of p53-mediated pathways is usually facilitated by the binding of tetrameric p53 to two 'half-sites', each consisting of a decameric p53 response element (p53 RE). The p53 binding motif can be described by the palindromic regular expression pattern [AG][AG][AG]C[AT][TA]G[TC][TC][TC] (El-Deiry et al., 1992), where '[AG]', for example, matches the two bases 'A' and 'G'. Functional REs are likely to be either directly adjacent or separated by a small number of 'spacer' base pairs (bp). The length of the spacer region varies from 1 to 13 bp.

Many factors are known to influence p53-dependent transcription. In addition to the p53 REs, the sequence specific binding of p53 to its DNA binding sites can be affected by various trans-acting factors which play their roles in promoting post-translational modifications of p53. The p53 protein and its associated proteins are subject to numerous post-translational modifications. Post-translational modifications, such as phosphorylation, methylation and acetylation, have been shown to influence the stability and DNA binding affinity of p53 (Bode and Dong, 2004; Chuikov et al., 2004; Gu et al., 2004; Luo et al., 2004; Riley et al., 2008; Vousden, 2002). In response to UV-induced DNA damage, the histone acetyltransferase pCAF, for example, associated with the p300/CBP protein, acetylates the C-terminal residues of p53 to activate sequence-specific DNA binding of p53 (Chan and Thangue, 2001; Liu et al., 1999; Pawlak and Deckert, 2007). Some of the post-translational modifications are critical for optimal protein-protein interactions with other cellular factors associated with p53. N-terminal phosphorylations of Ser15, Thr18 and Ser20 in p53 have been reported to increase the association of p53 with p300/CBP (Dornan et al., 2003; Dumaz and Meek, 1999; Finlan and Hupp, 2004; Lambert et al., 1998) and additionally to block the interaction of p53 with the negative regulator MDM2 (Böttger et al., 1999; Chehab et al., 1999; Craig et al., 1999; Dumaz et al., 2001, 1999; Meek and Anderson, 2009; Sakaguchi et al., 2000; Schon et al., 2002; Shieh et al., 1997; Unger et al., 1999).

Another important factor which is known to have an influence on the transcriptional activity of p53 is the chromatin architecture encompassing p53 target genes. Specific histone modifying enyzmes recruited by p53 mediate post-translational modifications of histone tails (histone modifications), such as phosphorylation, acetylation, methylation and

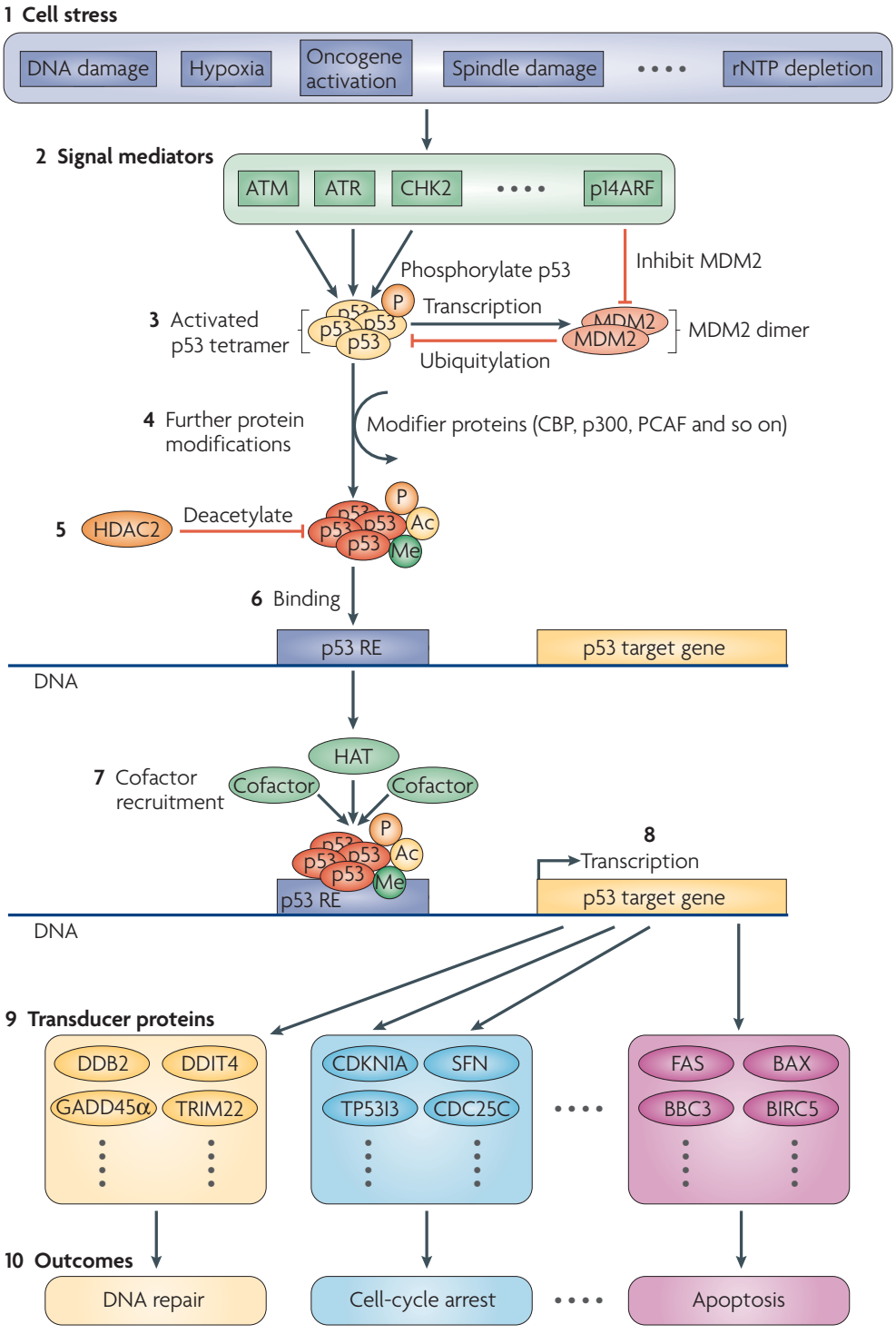


Figure 1.3 Activation of the p53 pathway (source taken from Riley et al. (2008)).

ubiquitination. These histone modifications have been reported to affect gene expression (Zhang and Reinberg, 2001). A correlation between histone methylation and gene activation has been detected by An et al. (2004). The two histone arginine methyltransferases CARM1 and PRMT1 which methylate Arg2, Arg17 and Arg26 of histone H3 and Arg3 of histone H4 have been shown to act as coactivators by directly interacting with p53 and its associated cofactor p300 (Fingerman and Briggs, 2004).

1.3 Experimental approaches to identifying TFBSs

Significant progress has been made in the development and improvement of high-throughput methods which enable genome-wide mapping of protein-DNA interactions in living cells (in vivo). A commonly used technique to identify and analyze protein-DNA interactions for DNA binding proteins is chromatin immunoprecipitation (ChIP) (Latchman, 2008b). The ChIP technique uses formaldehyde to cross-link protein to specific DNA sequence elements. After chromatin isolation and DNA fragmentation by sonication, the protein-bound fragments are immunoprecipitated using an antibody which targets the protein of interest. The cross-linking of the protein and DNA complexes is reversed so that the DNA can be separated from the proteins and purified. The isolated DNA can then be amplified and quantified by PCR.

The DNA enriched by ChIP can be analyzed and characterized in a number of ways. In ChIP-on-chip (Lee et al., 2006), chromatin immunoprecipitaion is combined with DNA microarrays allowing the mapping of all the binding sites for a particular protein in regions of the genome which are covered by the array. The enriched DNA fragments are fluorescently labeled and hybridized over the DNA microarray which contains single-stranded DNA sequences (probes) from the genomic portion of interest. The labeled fragments are hybridized to complementary probes and the hybridization can be determined by illuminating the microarray with fluorescent light and measuring the light intensity and color of each DNA fragment bound to the array. Initially used in yeasts (Blat and Kleckner, 1999; Hearnes et al., 2005; Ren et al., 2000), the ChIP-on-chip method now has broad application to various organisms, including higher organisms such as mammals. ChIP-on-chip is commonly used to determine and characterize transcription factor binding sites. Smeenk et al. (2008) and Kaneshiro et al. (2007) used a ChIP-on-chip approach to identify human p53 binding sites. Lupien et al. (2008) combined chromatin immunoprecipitation and DNA microarray to perform positional analyses of human FoxA1. In some cases, the method is used to

investigate the distributions of histone modifications. Heintzman et al. (2007), for example, described the chromatin architecture along a 30-Mb portion of the human genome (Encode consortium) using ChIP-on-chip. They found that active promoter and enhancer regions were characterized by specific patterns of histone modifications. Based on this information, Heintzman et al. (2007) developed a model to predict regulatory regions in the human genome.

Alternatively, in addition to DNA microarrays, ChIP can be combined with massively parallel sequencing technology (ChIP-seq). The ChIP-seq method has several advantages over ChIP-on-chip. ChIP-seq is a more cost-effective method and enables the precise mapping of protein-DNA interactions on a genome-wide scale (Ho et al., 2011). Recently, Shen et al. (2012) created a map of the cis-regulatory sequences identified in the mouse genome of different tissues and cell types based on ChIP-seq data. The authors performed ChIP-seq experiments to identify regions of the mouse genome bound by RNA polymerase II (polII) and the insulator-binding protein CCCTC-binding factor (CTCF) and to determine the localization of histone H3 lysine 4 monomethylation (H3K4me1), H3 lysine 4 trimethylation (H3K4me3) and H3 lysine 27 acetylation (H3K27ac) in different tissues and cell types. Several thousand binding sites for the human SRF, GABP and NRSF proteins were identified by Valouev et al. (2008) with ChIP-seq. The generated ChIP-seq data was used to evaluate their developed ChIP-seq data analysis method called QuEST (quantitative enrichment of sequence tags). A special form of ChIP-seq was used by Wei et al. (2006) who combined chromatin immunoprecipitation with the paired-end ditag sequencing strategy (ChIP-PET) to identify p53 binding sites in the human genome. Mapping transcription factor binding regions by the ChIP-PET method is a useful way of increasing the DNA sequencing efficiency while reducing the costs. In ChIP-PET, short sequences from both ends of the enriched ChIP DNA fragment are concatenated (PETs), sequenced and mapped to the genome to define the boundaries of each ChIP fragment. By identifying overlapping PET sequences, the binding sites for the particular DNA binding protein can be determined. The ChIP-PET method will be described in more detail later in Chapter 2.

1.4 Computational approaches to predicting binding sites

Experimentally derived data from high-througput methods such as ChIP can be used for computational prediction of transcription factor binding sites. Various attempts in the field of computational biology have been made to design practical strategies for predicting such

binding sites.

A wide range of computational methods identify binding sites of transcription factors by searching for common short nucleotide sequence patterns called motifs. Most of them use a position frequency matrix (PFM) or a position weight matrix (PWM), also known as position specific scoring matrix (PSSM), to represent the motif (Table 1.1). The PFM is derived from a multiple alignment of experimentally verified binding sites and contains the absolute or relative frequencies of observed nucleotides for each position in the alignment. The PWM represents a log-odds normalized version of PFM. The PWMs of many transcription factor binding sites are available in the TRANSFAC (Matys et al., 2003) and JASPAR (Sandelin et al., 2004) databases. There are a number of motif finding tools which are based on PWM. NestedMICA (Down and Hubbard, 2005), for example, is a motif finding tool which identifies significantly over-represented motif patterns in a set of sequences by optimizing a probabilistic matrix. Other well known methods for discovering motifs include MEME (Bailey and Elkan, 1994), AlignACE (Roth et al., 1998), CONSENSUS (Hertz and Stormo, 1999), BioProspector (Liu et al., 2001), MotifSampler (Thijs et al., 2001), MDScan (Liu et al., 2002), DWE (Smith et al., 2005) and MUSA (Mendes et al., 2006). A more complex model is described by Fu et al. (2009) who extended the PWM based search by some additional features, including conventional features such as transcription factor binding site sequence specificity and state transition probability, evolutionary features, structural and epigenetic features. The model called DISCOVER was developed for motif discovery in metazoan genomes.

Other methods successively use hidden Markov models (HMMs) to identify motifs in genomic sequences. Levkovitz et al. (2010) developed a model called DEMON which is based on HMM to detect enriched transcription factor binding sites in promoter sequences. Using DEMON, a strong enrichment of the binding sites for RUNX3 was observed in pancreatic adenocarcinoma (PAC) related genes, suggesting a possible role of RUNX3 as a potential target in pancreatic cancer biology. Riley et al. (2009) and Huang and Li (2005) developed computational methods based on profile HMMs to identify target genes of the p53 protein. Riley et al. (2009) showed that their model called p53HMM had better predictive abilities than PWMs.

Schneider (1997) and Lyakhov et al. (2008) used information theory to identify transcription factor binding sites. Information theory is generally used to measure overall sequence conservation in sets of nucleotide or protein sequences (Pierce, 1980; Shannon, 1948). The individual information (R_i) technique described by Schneider (1997) allows

Table 1.1 Position frequency matrix (PFM) and position weight matrix (PWM). If we assume equal background frequencies for each of the four nucleotides (0.25), the PWM value for nucleotide A in position 1 is given by the logarithm of the ratio ((relative frequency of A in position 1)/(background frequency of A)) called log odds, $\log_2(0.41/0.25) = 0.71$. For example, suppose our sequence of interest is ATACATGGCC. The PWM score of the decameric sequence can be determined by summing the log odds scores of the corresponding nucleotides (framed in (c)). Sequence ATACATGGCC has a total PWM score of 5.42. The higher the PWM score, the more likely the sequence of interest represents a binding site.

(a) Count matrix

	1	2	3	4	5	6	7	8	9	10
A	720	274	1064	14	1263	227	48	154	246	243
C	52	60	7	1700	42	41	30	759	834	691
G	726	1214	627	9	58	106	1664	27	69	268
T	260	210	60	35	395	1384	16	818	609	564

(b) PFM

	1	2	3	4	5	6	7	8	9	10
A	0.41	0.16	0.61	0.01	0.72	0.13	0.03	0.09	0.14	0.14
C	0.03	0.03	0.00	0.97	0.02	0.02	0.02	0.43	0.47	0.39
G	0.41	0.69	0.36	0.00	0.03	0.06	0.95	0.02	0.04	0.15
T	0.15	0.12	0.03	0.02	0.23	0.79	0.00	0.46	0.35	0.32

(c) PWM with equal background frequencies (0.25) for each nucleotide

	1	2	3	4	5	6	7	8	9	10
A	0.71	-0.64	0.78	-4.64	1.53	-0.94	-3.06	-1.47	-0.84	-0.84
C	-3.06	-3.06	-5.97	1.96	-3.64	-3.64	-3.64	0.78	0.91	0.64
G	0.71	1.46	0.53	-5.61	-3.06	-2.06	1.93	-3.64	-2.64	-0.74
T	-0.74	-1.06	-3.06	-3.64	-0.12	1.66	-4.78	0.88	0.49	0.36

investigating individual sequence conservations. An individual information weight matrix is first generated from the PFM of aligned sequences. The weight matrix can then be used to determine the information content of each individual sequence in the alignment which is measured in bits. All individual information contents are added together and divided by the number of aligned sequences to calculate the overall information content of the sequences. Binding sites can now be identified by comparing their individual information contents to the determined overall information content. The individual information method was used by Schneider (1997) to measure individual information distributions for *E. coli* ribosome binding sites, bacterial Fi binding sites and human splice acceptor and donor binding sites and by Lyakhov et al. (2008) to identify p53 binding sites from sequences near the transcriptional start sites in human chromosomes 1 and 2.

Another widely used approach integrates comparative genomics. Horvath et al. (2007) examined evolutionary conservation of experimentally verified human p53 binding sites across mouse, rabbit, rat and dog. Many of them were not significantly conserved across the four mammalian species. For a number of p53 binding sites, however, they observed differences in evolutionary conservation among p53 response elements and p53 related pathways. Xie et al. (2005) searched for conserved regulatory motifs in promoter regions and 3' UTRs of protein coding genes. Using a comparative genomics approach across human, mouse, rat and dog genomes, they successively found 174 candidate motifs in promoter regions and 106 motifs, which are likely to be involved in post-transcriptional regulation, in 3' UTRs.

Works on machine learning algorithms such as support vector machine (SVM) approaches for the prediction of transcription factor binding sites have been published by Sinha et al. (2007) and Jiang et al. (2007). Sinha et al. (2007) examined the flanking regions of experimentally verified human p53 binding sites for occurrences of functional motifs of other transcription factors to develop a SVM classifier based on this information for detecting p53 binding sites. Jiang et al. (2007) described a SVM based model called OSCAR which integrated multiple factors to identify regulatory motifs, such as information about positional preferences of the binding sites relative to the transcriptional start sites of the relevant genes and about occurrences of other motifs in their flanking regions. Their model was tested on binding sites of GATA and NHF transcription factor families.

Like OSCAR, there are several other approaches which integrate multiple information sources for the prediction of transcription factor binding sites. Ernst et al. (2010) described a logistic regression based method with 29 input features which was combined with motif

information for certain transcription factors. Based on 29 features, including, amongst other things, the distance to nearest transcriptional start site, information on conservation and levels of histone modifications, the logistic regression based model was used to determine the so-called general binding preference (GBP) score for a specific base location in the human genome. In addition, the motif score for the base in the specific location was determined using PWM and a zero-order background model. For a binding site of a specific length, an average GBP score was first determined over each base position in the binding site. This average score combined with the motif score associated with the base at a specific position represents the combined score for the particular base position in the binding site. To score the entire binding site region, either the maximum or the average value of the combined score over each base position was taken. Testing on new experimentally derived sequences, the method by Ernst et al. (2010) accurately predicted true binding sites. Won et al. (2010) developed a HMM based model called Chromia which combined sequence information with ChIP-seq signals of histone modifications at promoter and enhancer regions to detect functional sequence patterns.

1.5 Outline of the thesis

The thesis describes the analysis of human transcription factor binding sites for the p53 tumour suppressor. Central to the focus of our research is the prediction of these sites by integrating multiple features which are known to affect the DNA binding specificity and function of p53.

The first step in building a model for the prediction of transcription factor binding sites involves data collection. To develop a highly sophisticated prediction model, we used functional binding sites from two published ChIP data for the p53 protein (Smeenk et al., 2008; Wei et al., 2006). Chapter 2 extensively deals with these ChIP-based binding sites. Specific procedures for collecting and preparing the data are discussed and some exploratory analysis results are demonstrated. In our work, we performed various analyses, including basic analyses examining the positional distributions and spacer lengths between the two half-sites of the binding sites and Gene Ontology (GO) and KEGG pathway enrichment analyses to examine their possible biological roles.

Having explored the data, we turn our attention to the main work of the thesis which is presented in Chapter 3. Chapter 3 introduces our prediction model with detailed coverage of its building and application. It describes a multiple logistic regression approach which combines sequence information with epigenetic information to produce a reliable model for predicting human p53 binding sites.

In Chapter 4, we present a practical application of our combined evidence model which is based on logistic regression. A genome-wide analysis was performed using our model to identify all possible p53 binding sites in the whole human genome. The binding sites predicted by our model were extensively examined and compared with some experimental binding data published in the literature.

We compared our model to a more simple model which predicts the binding sites by searching for the specific p53 motif. The sequence-only model, described in Chapter 5, is applied on a genome-wide scale and its results analyzed and compared to those obtained from the logistic regression based model.

Finally, Chapter 6, briefly summarizes the contributions we made in this research work and discusses future directions and next steps to be undertaken which will lead to a better understanding of the process by which p53 recognizes its DNA binding sites.

Chapter 2

Data set

To develop and test a more specific model for the prediction of DNA binding sites preferred by p53, we obtained functional p53 binding sites from the experimental literature (Smeenk et al., 2008; Wei et al., 2006) which used the chromatin immunoprecipitaion (ChIP) technique. In this chapter, our main focus will be on these binding site data. We will first give details about the sources of the binding sites and explain them briefly. We will then present the procedures and methods which were carried out to collect the required data and finally examine and characterize them by exploratory analysis.

2.1 Introduction

ChIP data were obtained from two different studies. The first study by Wei et al. (2006) used ChIP in combination with a paired-end ditag (ChIP-PET) sequencing strategy (Fullwood et al., 2009; Hudson and Snyder, 2006) to map binding sites of p53. The binding site regions were identified in HCT116 human colon cancer cells with treatment of 5-fluorouracil which is known to stabilize the p53 protein (Sun et al., 2007). In the PET sequencing method, short sequences of length 36 bp, called tags or PETs, were extracted from cloned ChIP DNA fragments. A PET consisted of two ends, the 5′ and 3′ ends, of the DNA fragment, each having a length of 18 bp. The PETs were concatenated together to form longer PET sequences and cloned into a plasmid vector to construct the final PET libraries. The concatenated PETs were then sequenced for mapping. By using PETs which provided information about both ends of the ChIP DNA fragments, the location of the binding regions could be identified. The ChIP DNA fragments represented by PETs, which resulted from the same enriched binding region, were expected to overlap with each other, while those from background regions were randomly spread over the genome (PET singletons). From a set of 65572 unqiue ChIP DNA fragments (PETs) including PET singletons, 4302 were enriched for p53

binding by the ChIP-PET experiment. Among the 4302 PETs, 2886 were assigned to the PET clusters with two overlapping PETs (PET-2) and 1416 to the clusters with three or more overlapping PETs (PET-3+). 1451 and 327 distinct PET-2 and PET-3+ clusters were identified, each representing a p53 binding site region. A PET-13 cluster was presented within the PET-3+ clusters which matched the promoter region of the cyclin-dependent kinase inhibitor 1A (CDKN1A), a well known p53 target gene. In addition to CDKN1A, the PET-3+ clusters could be mapped to many other known p53 targets providing strong evidence that the 327 PET-3+ clusters represented high-confidence p53 binding sites.

In contrast, Smeenk et al. (2008) performed ChIP combined with whole genome tiling arrays (ChIP-on-chip) in human osteosarcoma U2OS cells (see Chapter 1 for a brief description of the ChIP-on-chip technique). The U2OS cells were treated with Actinomycin D to activate p53-dependent transcription (Choong et al., 2009). In this study, a total number of 1546 p53 binding sites were identified by ChIP-on-chip.

2.2 Methods

2.2.1 Data collection

In addition to the two ChIP-based binding data sets, we retrieved the sequences of numerous validated p53 binding sites (Horvath et al., 2007) which differed from the known p53 consensus sequence shown in Figure 2.1. These binding sites with consensus-poor REs were required to be present in our data set to avoid problems related to complete or quasi-complete separation in the later step when we defined our prediction model based on binary logistic regression (see Chapter 4). The sequences used in the study by Horvath et al. (2007) along with their coordinates from the human genome assembly hg17 were extracted from the supplementary material of the referred publication. For the ChIP-PET data set by Wei et al. (2006), we extracted the hg17 coordinates of all the PET-3+ clusters with three or more overlappinng DNA fragments and obtained the DNA sequences by using human genome data downloaded from the Ensembl ftp site (Ensembl release 35, November 2005). The genome-wide ChIP-on-chip binding data were received directly from Smeenk's group upon request. Since these data were based on NCBI36 coordinates (corresponding to UCSC's release number hg18), we first BLASTed each binding sequence against the human genome assembly hg17 to get coordinates from the same build of the human genome as for the other binding sites. We used the 'blastall' function from the NCBI's standalone command line 'blast' package (Altschul et al., 1990) and performed 'blastn' search for DNA sequences

using our own database which contained the human genome data downloaded from the Ensembl ftp site (Ensembl release 35, November 2005). The sequence data were fetched using the Ensembl Perl API (Ensembl release 35, November 2005).

If there are 300 binding sites in the genome (Smeenk et al., 2008), then theoretically there is the possibility for the data sets to capture all of them, because each data set contains more than 300 regions. However, things do not work like that, because any method will have some false negatives. Comparison of both data sets obtained from ChIP-on-chip and ChIP-PET experiments, respectively, revealed appreciable agreement for the enriched p53 targets (38% for Wei data set), but also a notable number of targets that were uniquely identified by one of the experiments. This observation arises from the fact that cellular proteins that interact with p53 and genes regulated by p53 are likely to vary in different cell types and with different stimuli (Han and Kulesz-Martin, 1992). In addition to that, ChIP-on-chip and ChIP-PET technologies have been observed to show different abilities to identify transcription factor binding regions for targets with low signals (Euskirchen et al., 2007; Smeenk et al., 2008). Strategies for improving the performances of both methods are given by Euskirchen et al. (2007). In order to avoid redundancy, we deleted the 124 sites from the published Smeenk data set that were also present in the Wei data set.

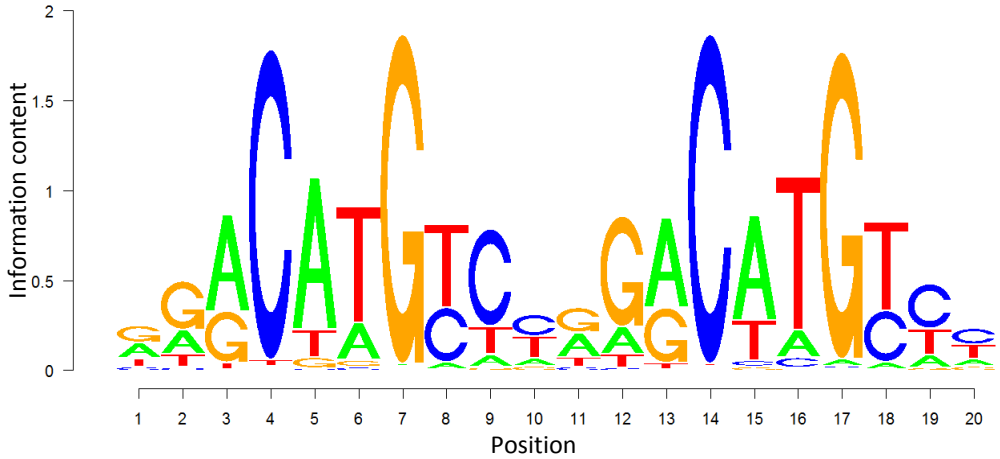


Figure 2.1 Sequence logo of the TRANSFAC PFM M01651 for p53, visualized using the seqLogo Bioconductor package in R (`www.bioconductor.org/packages/release/bioc/html/seqLogo.html`)

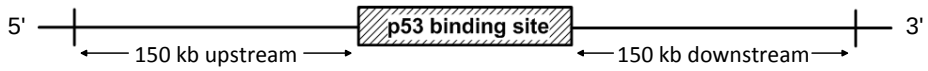
To determine the locations of the binding sites bound by p53 we scanned the ChIP sequences with FIMO (Grant et al., 2011), a motif-based sequence analysis tool available

from the MEME Suite motif finding tools (Bailey et al., 2009), using the M01651 motif (Figure 2.1) for p53 derived from the TRANSFAC (Matys et al., 2003) database. In order to allow spacing between the two half-sites in the p53 binding sites we divided the 20-mer TRANSFAC motif which was presented as a position frequency matrix (PFM) into two decamer half-sites. Occurrences of each individual decamer half-site were identified for both strands by running FIMO. Two decamer half-sites were joined together to a full binding site if they were directly adjacent or separated by a spacer of at most 13 base pairs (bp). The score of a full site was determined as the sum of the individual scores for the half-site occurrences. The full site with the highest score was selected to represent the p53 binding site in the ChIP sequence.

2.2.2 Exploratory data analysis

To analyze the functional annotation of the extracted p53 binding sites, we assigned the binding sites to their nearest genes by using the Ensembl Perl API (Ensembl release 35, November 2005).

Definition 1. (Nearest gene) *A gene is defined as a nearest gene of a binding site, if it is a protein coding gene, which lies within a distance of at most 150 kb up- or downstream to the binding site.*



If a binding site was located within the transcribed region of a gene, that gene was considered as the nearest gene of the binding site. If there was no such gene, we took one up- and one downstream genes, which overlapped the 150 kb flaking region at both the 5′ and 3′ ends of the binding site (the closer the better).

We analyzed the locations of the p53 binding sites in the human genome relative to Ensembl genes. All annotated single-exon genes, likely to be pseudogenes were excluded from the analysis. Locations of the binding sites were divided into intragenic (all introns and exons except the first exon and intron), TSS flanking (first intron, first exon and 5 kb upstream of TSS), 5 kb downstream (5 kb downstream of last exon), 5-25 kb downstream, 5-25 kb upstream and intergenic regions. This type of classification was also used by Smeenk et al. (2008) with the exception that the 5-25 kb downstream and the 5-25 kb upstream regions were merged together and treated as one classification group. The observed counts across the six genomic regions were then compared with the counts expected by chance by

using a G-test.

Definition 2. (G-statistic) *A G-statistic is also called a likelihood ratio test or a log-likelihood test. The observed value of G can be calculated as follows:*

$$G = 2\ln L = 2\sum^{c} O_c \ln(\frac{O_c}{E_c}),\ c \in C,$$

where O_c and E_c are the observed and expected counts of the classification group c, respectively. The observed value of G is compared with a χ^2-distribution with $|C| - 1$ degrees of freedom to compute the probability of getting that G value. (Sokal and Rohlf, 1995)

Table 2.1 Gene Ontology (GO) evidence codes and their reliability (Lee and Marcotte, 2009).

Evidence code	Description	Reliability
TAS	Traceable Author Statement	High
IDA	Inferred from Direct Assay	High
IMP	Inferred from Mutant Phenotype	High
IGI	Inferred from Genetic Interaction	High
IPI	Inferred from Physical Interaction	High
ISS	Inferred from Sequence or Structural Similarity	Low
IEP	Inferred from Expression Pattern	Low
NAS	Non-traceable Author Statement	Low
IEA	Inferred from Electronic Annotation	Low

To characterize the functions of the binding sites, we performed Gene Ontology (GO) (Ashburner et al., 2000) and gene enrichment analyses of the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2012) pathways on the list of their nearby genes with the web-based functional annotaion tool provided by DAVID (Huang et al., 2009a,b), the Database for Annotation, Visualisation and Integrated Discovery. For a given gene list, DAVID identified significantly enriched Gene Ontology (GO) terms associated with the genes in the list. In general, GO terms are categorized into biological process, molecular function and cellular component terms. Each GO term is associated with an evidence code which gives information on the source of the evidence and its reliability (Table 2.1). DAVID uses all GO evidence codes. The current version does not provide any options which allow filtering specific evidence codes.

2.3 Results

A total of 1757 functional p53 binding sites were gathered from the experimental literature.

Table 2.2 Binding sites of the p53 protein used in this study.

Total number of p53 binding sites	Published study
1422	(Smeenk et al., 2008)
327	(Wei et al., 2006)
8	(Horvath et al., 2007)

These binding sites were used as positive data for training and testing our prediction model in the later steps.

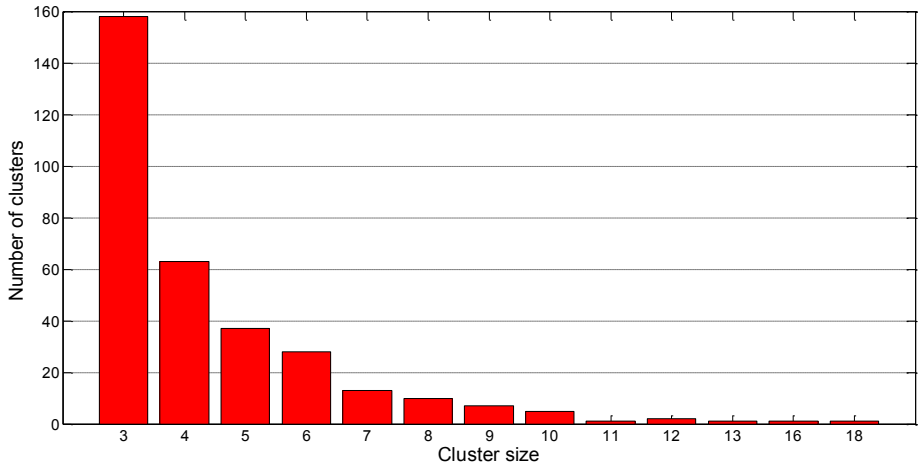


Figure 2.2 Number of p53 targets in Wei's PET clusters. Among the 13 different PET clusters the PET-3 cluster is the largest one with 158 (out of 327) targets. Clusters with only one p53 target are PET-11, PET-13, PET-16 and PET-18.

2.3.1 Exploratory data analysis

As shown in Figure 2.3, the majority of the p53 binding sites (69%) did not have a spacer between the two half-sites. For the remaining binding sites, the spacer lengths varied from 1 to 13 bp, whereas 1 bp and 10 bp spacers were found to be more frequent than others. This result was consistent with recent studies which examined the influence of spacers between the two half sites in the p53 binding site. An increase in the spacer length from 1 bp to 4 bp was reported to inhibit p53-mediated transcription, while a 10 bp spacing enhanced the transcriptional activity of p53 (Vukojevic et al., 2010).

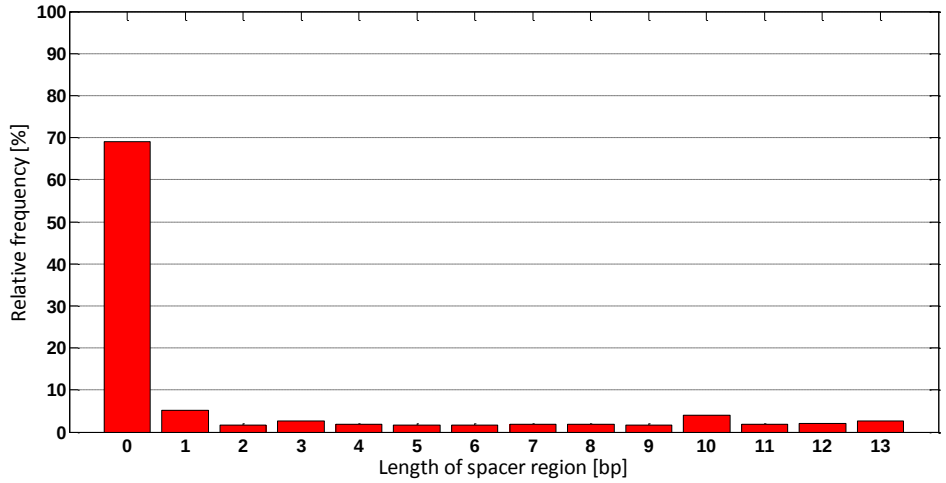


Figure 2.3 The spacer lengths between the two decameric half-sites of the 1757 p53 binding sites.

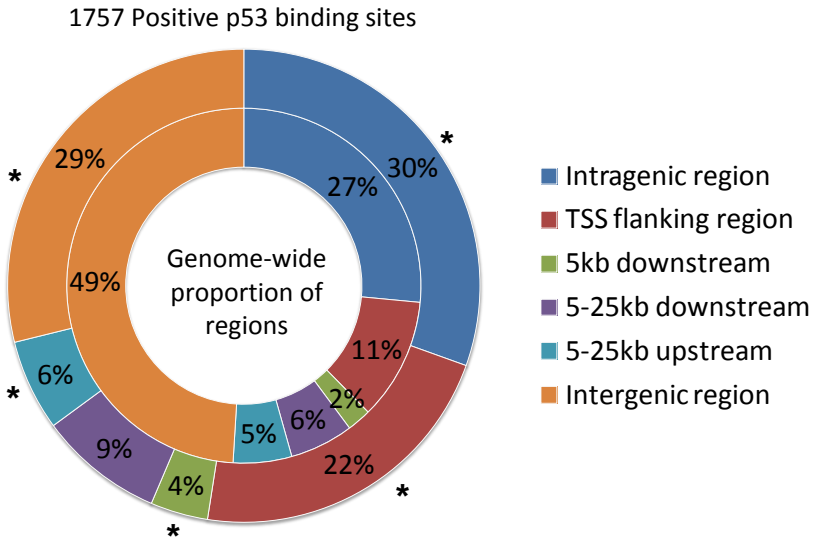


Figure 2.4 Distribution of the 1757 p53 binding sites in intragenic, TSS flanking, 5 kb downstream, 5-25 kb downstream, 5-25 kb upstream and intergenic regions relative to Ensembl genes (outer ring) compared to the genome-wide proportions of the six regions of interest (inner ring). Significantly enriched or under-represented regions (G-test, P<0.05) are marked with an asterisk (*).

The binding sites were distributed all over the genome (Figure 2.4). Out of 1757, 839 binding sites were found within a gene (intragenic) or near the transcriptional start site (TSS) of a gene (TSS flanking region). 54 sites were located within a distance of 5 kb downstream of a gene, 124 within 5-25 kb downstream, 120 within 5-25 kb upstream and

620 in intergenic regions. Comparing the observed counts across the six genomic regions with the counts we would expect by chance, we found that the result obtained by using the G-test was highly significant (G=167.19, df=5, $P\approx$0). Individual G-tests applied for testing each genomic region identified a number of significantly enriched regions for the p53 binding sites. The binding sites were significantly over-represented in intragenic (G=6.69, df=1, P=9.72$\times10^{-3}$), TSS flanking (G=75.16, df=1, $P\approx$0), 5 kb downstream (G=8.89, df=1, P=2.87$\times10^{-3}$) and 5-25 kb upstream (G=11.17, df=1, P=8.31$\times10^{-4}$) regions. Statistically significant under-representation was observed for the intergenic region (G=134.42, df=1, $P\approx$0).

Functional annotation and enrichment analysis

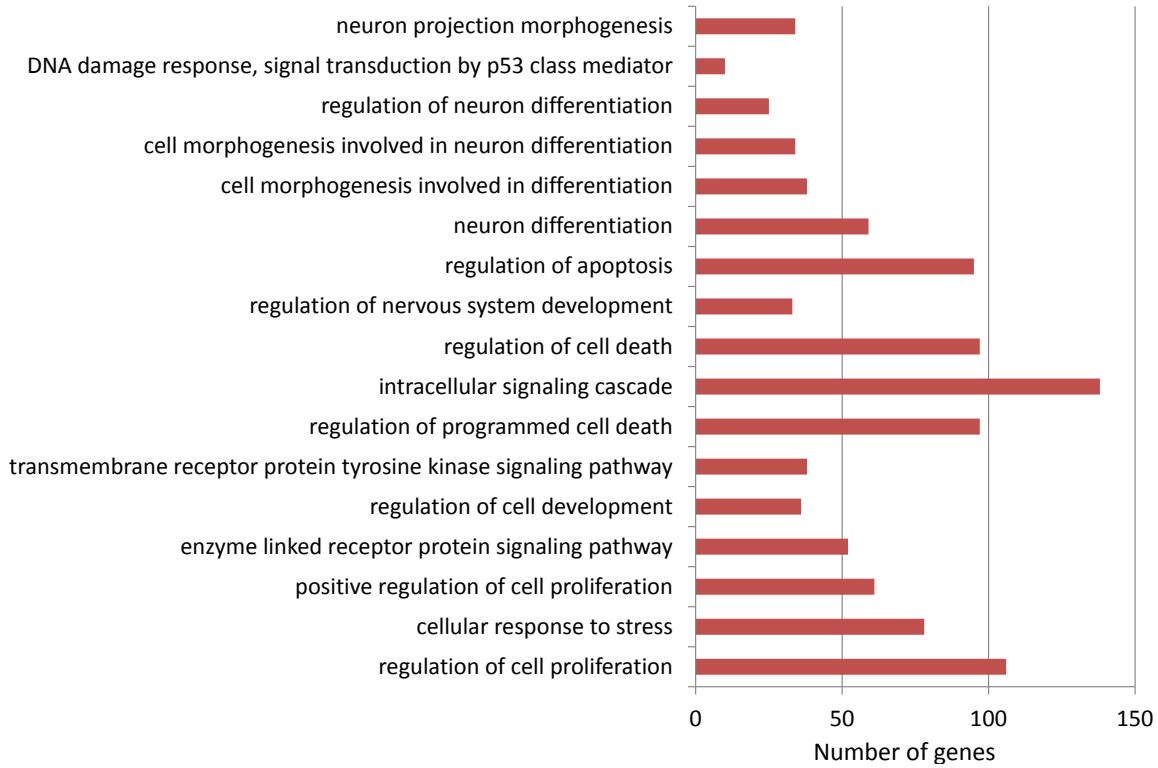


Figure 2.5 Biological process terms in the 'Gene Ontology FAT' annotation category found to be highly significantly enriched (P<0.0001) in our list of 1047 (out of 1509) genes.

In total, 1718 unique nearby genes were identified for 1533 (out of 1757) binding sites. To annotate the genes using GO terms we used DAVID and explored enriched GO terms and KEGG pathways in the given gene list. As a summary from the GO analysis using the 'GO FAT' annotation categories, significantly enriched terms (with low P values) associated with our gene list in the biological process and molecular function categories are shown

in Figures 2.5 and 2.6. A list of enriched cellular component GO terms can be found in Appendix C. The order of the displayed GO terms is from less significant to more significant (from top to bottom). The most statistically significant term can therefore be found at the bottom. All biological process and molecular function GO terms presented here have P values of less than 0.0001 and less than 0.05, respectively. DAVID uses a modified version of the Fisher's Exact test (EASE score) to compute the P values. The 'GO FAT' annotation category provided by DAVID represents a subset of the standard set of GO terms excluding the broadest (non-informative) terms so that more specific terms can be presented more clearly. The annotation category considers the top five levels from the full tree and filters out all 'higher' GO terms.

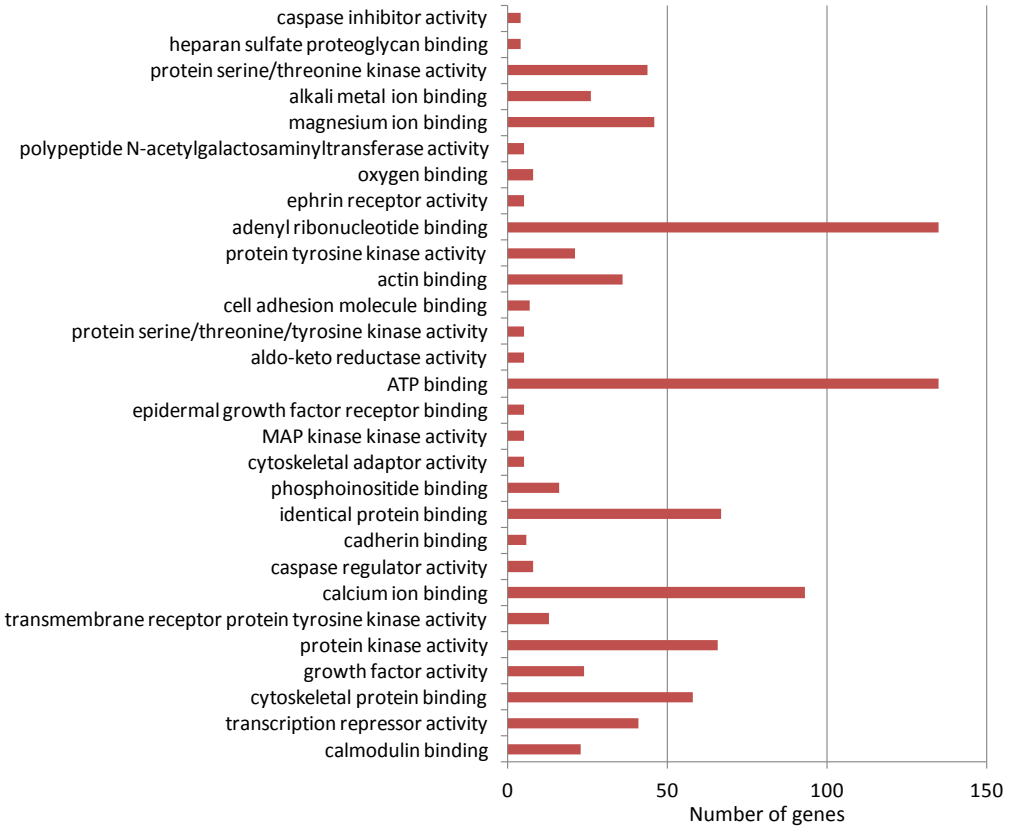


Figure 2.6 Molecular function terms in the 'Gene Ontology FAT' annotation category found to be highly significantly enriched (P<0.05) in our list of 1002 (out of 1509) genes.

We found numerous highly enriched GO terms related to the well-known anti-cancer functions of p53, such as apoptosis, the programmed cell death and cell cycle arrest. In addition, many of the 1047 genes involved in the biological process GO Fat annotation category were shown to be stress response genes and related to many various regulation processes (Figure 2.5). Among the molecular function terms, 'calmodulin binding' was

found to be the most statistically significant term (Figure 2.6). The most numerous term was 'ATP' binding. This finding is consistent with the fact that p53 can interact with ATP which has been shown to trigger changes in the configuration of p53 (Brain and Jenkins, 1994; Warnock and Raines, 2004).

The results of the KEGG pathway analysis involving 451 genes are shown in Figure 2.7. Pathways with P<0.05 were considered to be statistically significant. Not surprisingly, the most statistically significant pathway was represented by 'p53 signaling pathway' and the most numerous pathway by 'pathways in cancer'.

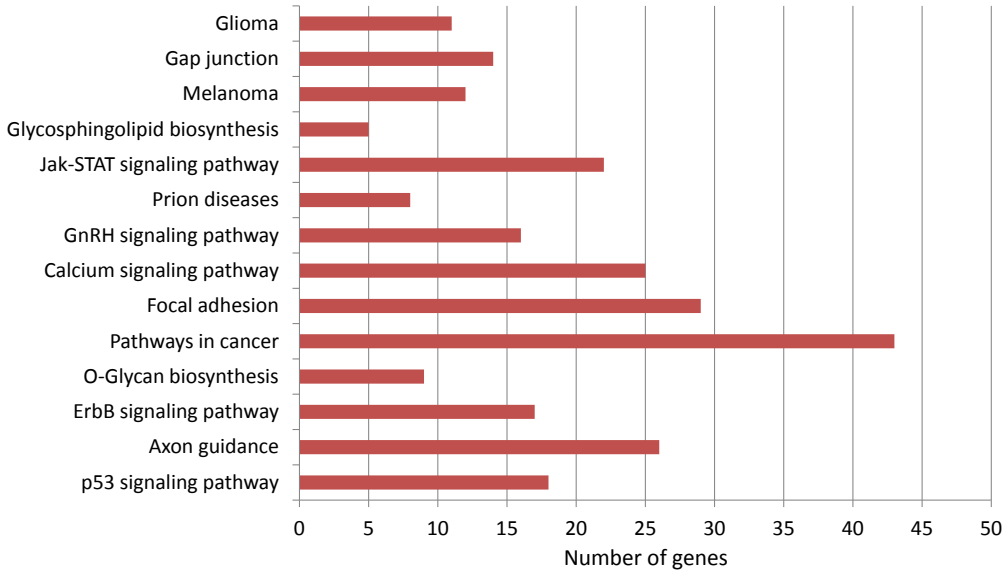


Figure 2.7 Enriched KEGG pathways (P<0.05) associated with 451 genes from our gene list. KEGG is a database resource comprising various fields of genomes, enzymatic pathways, and biological chemicals.

2.4 Discussion

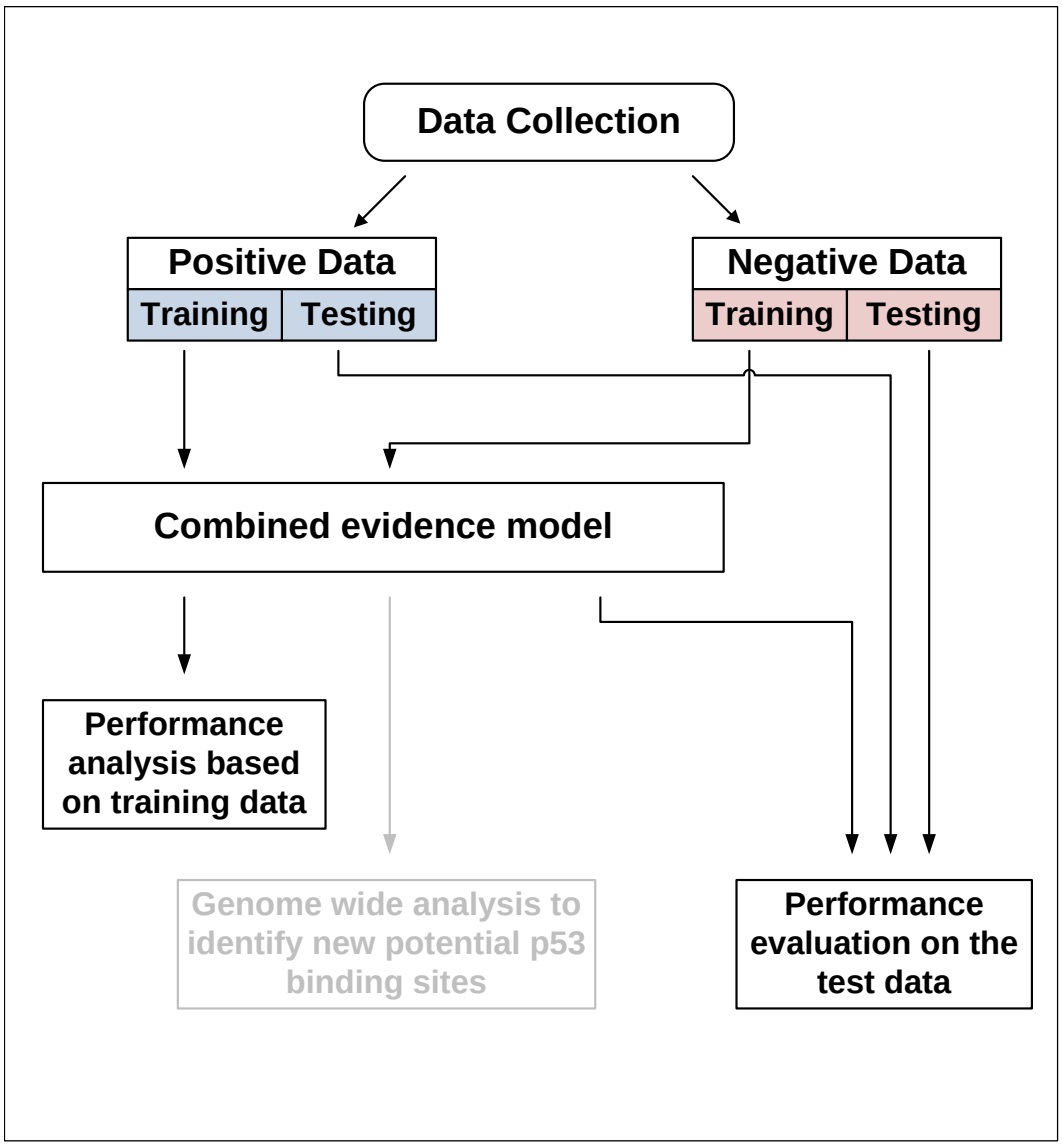
To create a set of positive samples (positive data set), we collected 1757 human p53 binding sites from the experimental literature (Horvath et al., 2007; Smeenk et al., 2008; Wei et al., 2006). The majority of the functional p53 binding sites consisted of two adjacent half-sites. Among the binding sites which contained a spacer region, spacer lengths of size 1 bp and 10 bp were observed more frequently than other spacers. The spacer separating the two decameric half-sites has been shown to play a critical role in regulating the DNA binding affinity and transactivation property of the p53 protein (Riley et al., 2008; Vukojevic et al., 2010). Several studies have suggested that optimal interactions with the p53 protein can

be established with adjacent half-sites of the binding site or half-sites separated by a 10 bp spacer (which corresponds to one turn of the helix) (Cook et al., 1995; Wang et al., 1995). The analysis of the functional p53 binding sites by gene region revealed that more than half (52%) of the binding sites were located in intragenic or TSS flanking regions. In addition, a significant portion was found distal to a transcription start site. These distant binding sites with their bound proteins have been shown to interact with the promoter and the transcription initiation complex through DNA looping (Riley et al., 2008).

A number of studies have suggested that p53 binding sites may be grouped on the basis of the activated pathway, such as apoptosis and cell cycle related control mechanisms. Distinct sequence conservation patterns have been observed among the p53 binding sites of apoptosis- and cell cycle-regulating genes (Horvath et al., 2007; Qian et al., 2002). Based on these observations, we performed clustering on the 1688 unique decameric half-sites of our 1757 positive binding sites to discover meaningful sequence patterns which might be biologically important (Appendix A). Some decamers may be more associated with particular stress response processes. For example, the p53 protein is known to be modified by phosphorylation which has been shown to subtly alter the sequence specificity of p53. Our hypothesis was that genes involved in those processes should be more likely to have both decamers within the same cluster, on the grounds that p53 would probably have both members of the dimer phosphorylated in the same way. To test the hypothesis that divergent sites might be biologically very different from the others, we clustered the decameric half-sites into subgroups of sites which were similar in sequence with Ward's (Ward, 1963) and unweighted pair group method with arithmetic mean (UPGMA) (Sokal and Michener, 1958) methods. The Ward's method gave nice balanced groups whereas the UPGMA method produced highly unbalanced groups which were essentially unusable (Figure A.1). Reasonable results were obtained by the Ward's method. Performing clustering analysis using the Ward's method, we could find distinct sequence patterns which were related to specific cellular processes. In contrast, the UPGMA clustering generated non-significant results.

Chapter 3

Prediction model



In this chapter, we will introduce our prediction model which is based on logistic regression. Logistic regression is one of many techniques which can be used to predict p53 binding sites. Other useful approaches are weight matrix-based methods, Markov chain models, hidden Markov models (HMMs), support vector machines (SVMs) and information theory, for which many attempts have been made. For detailed information, please see Chapter 1.

3.1 Introduction

3.1.1 Binary logistic regression

Logistic regression is a statistical method widely used in social and natural sciences. It is the most important model for categorical response data (Agresti, 2002). A logistic regression model which has two possible outcomes is called binary logistic regression model. For categorical response variables with more than two possible outcomes we refer to nominal or ordinal (ordered outcomes) logistic regression models. A special case of binary logistic regression is multiple logistic regression. Multiple logistic regression refers to methods for analyzing the relationship between a binary response variable and multiple predictor variables which can be categorical and/or continuous.

Definition 3. (Multiple logistic regression model) *For a binary response variable y and n independent predictor variables $x = x_1, ..., x_n$, let $P(y = 1)$ represent the probability that an event occurs given the n predictors with*

$$P(y = 1) = \frac{\exp(\beta_0 + \beta_1 x_1 + ... + \beta_n x_n)}{1 + \exp(\beta_0 + \beta_1 x_1 + ... + \beta_n x_n)},$$

where $\beta_0, \beta_1, ..., \beta_n$ are the parameters of the model.
The multiple logistic regression model is

$$\text{logit}[P(y = 1)] = \log[\frac{P(y = 1)}{1 - P(y = 1)}] = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n.$$

The formula for the logistic regression model uses the log *of the odds ratio,* $\log[\frac{P(y=1)}{1-P(y=1)}]$*, called the logistic transformation or logit (Agresti, 2002; Agresti and Finlay, 2007).*

Let a be the logistic transformation with $a = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n$. P achieves its minimum value at $\lim_{a \to -\infty} \frac{\exp(a)}{1+\exp(a)} = 0$ and its maximum value at $\lim_{a \to +\infty} \frac{\exp(a)}{1+\exp(a)} = 1$ (Larose, 2006).

Definition 4. (Likelihood function) *Let X be the observed data. The likelihood function of the parameters $\beta = \beta_0, \beta_1, ..., \beta_n$ is defined as*

$$L(\beta|X) = \prod_{i=1}^{m} P(y = 1|x_i)^{y_i}(1 - P(y = 1|x_i))^{1-y_i},$$

where $P(y = 1|x_i)$ is the probability of the i-th observation where the response variable is positive ($y = 1$) and $1 - P(y = 1|x_i)$ the probability of the i-th observation where the response is negative ($y = 0$) (Larose, 2006).

The parameters $\beta = \beta_0, \beta_1, ..., \beta_n$ are estimated using maximum likelihood (Eliason, 1993). The optimal values of the model parameters are those that maximize the likelihood $L(\beta|X)$ of observing the data X. The likelihood tends to become too small to be represented by any calculator or computer. Thus, we generally work with the log of the likelihood. Taking the natural logarithm, the log-likelihood function is

$$\ln(L(\beta|X)) = \sum_{i=1}^{m} y_i \ln(P(y = 1|x_i)) + (1 - y_i)\ln(1 - P(y = 1|x_i)).$$

3.1.2 Complete and quasi-complete separation in logistic regression

There may be situations where the outcome variable y separates the predictor x or a combination of predictors completely. In the following example data,

x =	x_1	...	-3.8	-2.1	-1.4	0.8	1.4	2.2	4.9	...	x_m
y =	0	...	0	0	0	1	1	1	1	...	1

there is just one predictor variable x. The greatest x value for the observations with $y = 0$ (negative response) is less than the smallest x value for the observations with $y = 1$ (positive response). There is a problem of complete separation of data points involving x. We say that the dependent variable y separates the independent variable x completely or x predicts y perfectly. In situations of quasi-complete separation, the greatest x for the observations with negative response is equal to or less than the smallest x for the observed values with positive response. When we run a logistic regression model on our example data,

$$P(y = 1) = \frac{\exp((x - a)b)}{1 + \exp((x - a)b)}$$

a perfect fit occurs, if we make b as large as possible, i.e. infinite, for $a = 0$. The larger b, which is the coefficient for x, the larger the likelihood. The maximum likelihood estimate does not exist here. The 'glm' function in R will produce the warning message "glm.fit: fitted probabilities numerically 0 or 1 occurred" due to extremely large and extremely small logits which are mapped to high probabilities near one and to low probabilities near zero, respectively.

One good technique for dealing with the problem of complete and quasi-complete separation in logistic regression is Firth logistic regression (Firth, 1993). Firth logistic regression uses a penalized likelihood estimation procedure which penalizes very large coefficients running to infinity. A detailed description of the penalized likelihood estimation procedure can be found in Heinze and Ploner (2003). In R, we can use the 'logistf' package to perform Firth logistic regression (Heinze and Schemper, 2002).

3.2 Methods

3.2.1 Data set: training and testing data sets

We split our positive set of 1757 p53 binding sites (Chapter 2) into training and testing sets. Half of the binding sites from each study were randomly taken as positive training sites and the remaining other half were used as positive testing set. In addition, we generated the same proportion of negative sites for training and testing. Since p53 binding sites are assumed to be rare in coding exons (Riley et al., 2008), we generated the negative sites by randomly selecting repeat-free regions from protein coding exons (that is translated into protein) of the human genome. We used the Ensembl Perl API (Ensembl release 35, November 2005) to retrieve the data from the human Ensembl Core database 'homo_sapiens_core_35_35h'. The length of spacers between the two half-sites of the negative sequences was chosen randomly and allowed to be up to 13 bp long.

3.2.2 Individual predictors used for building the prediction model

To predict p53 binding sites we used a number of potential factors which have been reported to influence p53-dependent transactivation. Our predictor variables included the individual match scores of the two decamer half-sites representing a full binding site, as well as the combined score of the full site, spacer length between the two half-sites, overlap with predicted enhancers and overlap with histone modification sites, such as Lys4 mono-, di-,

and trimethyl H3 sites. Because of the binary nature of the response variable which we defined as

$$Y = \begin{cases} 0, & \text{if given site is not a p53 binding site;} \\ 1, & \text{if given site is a p53 binding site,} \end{cases}$$

logistic regression was chosen as our prediction method.

Histone modification ChIP-seq data for H3K4me1, H3K4me2 and H3K4me3 in the HMEC, HUVEC, NHEK and NHLF cell lines and for H3K4me2 and H3K4me3 in the HEPG2 cell were obtained from the ENCODE/Broad Institute available at the UCSC Genome Browser (`http://genome.ucsc.edu`) (Fujita et al., 2011; Kent et al., 2002). All five cell lines are known to express wild-type p53. The coordinates of the 36589 enhancers recently predicted via chromatin signatures by Heintzman et al. (Heintzman et al., 2007) were downloaded from the Ren Lab web site. Since our p53 binding data used hg17 coordinates, the hg18 data of histone modifications, as well as those of the enhancers were converted to hg17 coordinates using the LiftOver tool accessible from the UCSC Genome Browser web site.

Using the training samples, including positive and negative sites, we collected data for the predictor variables listed in Table 3.1. We determined the motif scores *decamer1_score.cont*, *decamer2_score.cont* and *pair_score.cont* by using the M01651 motif for p53 from the TRANSFAC database within the TRANSFAC Suite (Figure 2.1) as described in Chapter 2. The score of the full site (*pair_score.cont*) was computed by the sum of the two individual half-site scores (*decamer1_score.cont* and *decamer2_score.cont*). The three score variables can be encoded either as continuous or binary predictors. We tested both encodings and chose between them based on their ability to improve prediction accuracy. To determine the *in_enhancer* and *in_XH3K4me1*, *in_XH3K4me2* and *in_XH3K4me3* predictors for the different cell lines we examined overlaps with any predicted enhancers and searched for H3K4me1, H3K4me2 and H3K4me3 signals at the training sites.

Table 3.1 Response and predictor variables included in the training models. The response variable is represented by *p53_bs_status*.

Description	Coding	Variable name
Status of being a p53 binding site	Binary: 0: is not a p53 binding site 1: is a p53 binding site	*p53_bs_status*

Continued on next page

Table 3.1 – continued from previous page

Description	Coding	Variable name
FIMO score for the occurrence of the first decamer motif	a) Continuous: floating point numbers b) Binary: 0: negative score 1: positive score	a) *decamer1_score.cont* b) *decamer1_score.positive*
FIMO score for the occurrence of the second decamer motif	a) Continuous: floating point numbers b) Binary: 0: negative score 1: positive score	a) *decamer2_score.cont* b) *decamer2_score.positive*
Maximum score for the occurrence of a decamer-decamer pair which is the sum of the two decamer scores	a) Continuous: floating point numbers b) Binary: 0: negative score 1: positive score	a) *pair_score.cont* b) *pair_score.positive*
Length of the spacer region between two decamer half-sites	Continuous: integers	*spacer.cont*
Overlaps with enhancer site	Binary: 0: does not overlap any enhancer site 1: overlaps an enhancer	*in_enhancer*
Overlaps with H3K4me1 site in HMEC cell line	Binary: 0: does not overlap any H3K4me1 site in HMEC cell line 1: overlaps a H3K4me1 site in HMEC cell line	*in_HmecH3K4me1*
Overlaps with H3K4me2 site	Binary: 0: does not overlap	*in_HmecH3K4me2*

Continued on next page

Table 3.1 – continued from previous page

Description	Coding	Variable name
in HMEC cell line	any H3K4me2 site in HMEC cell line 1: overlaps a H3K4me2 site in HMEC cell line	
Overlaps with H3K4me3 site in HMEC cell line	Binary: 0: does not overlap any H3K4me3 site in HMEC cell line 1: overlaps a H3K4me3 site in HMEC cell line	*in_HmecH3K4me3*
Overlaps with H3K4me1 site in NHLF cell line	Binary: 0: does not overlap any H3K4me1 site in NHLF cell line 1: overlaps a H3K4me1 site in NHLF cell line	*in_NhlfH3K4me1*
Overlaps with H3K4me2 site in NHLF cell line	Binary: 0: does not overlap any H3K4me2 site in NHLF cell line 1: overlaps a H3K4me2 site in NHLF cell line	*in_NhlfH3K4me2*
Overlaps with H3K4me3 site in NHLF cell line	Binary: 0: does not overlap any H3K4me3 site in NHLF cell line 1: overlaps a H3K4me3 site in NHLF cell line	*in_NhlfH3K4me3*
Overlaps with H3K4me1 site	Binary: 0: does not overlap	*in_NhekH3K4me1*

Continued on next page

Table 3.1 – continued from previous page

Description	Coding	Variable name
in NHEK cell line	any H3K4me1 site in NHEK cell line 1: overlaps a H3K4me1 site in NHEK cell line	
Overlaps with H3K4me2 site in NHEK cell line	Binary: 0: does not overlap any H3K4me2 site in NHEK cell line 1: overlaps a H3K4me2 site in NHEK cell line	*in_NhekH3K4me2*
Overlaps with H3K4me3 site in NHEK cell line	Binary: 0: does not overlap any H3K4me3 site in NHEK cell line 1: overlaps a H3K4me3 site in NHEK cell line	*in_NhekH3K4me3*
Overlaps with H3K4me1 site in HUVEC cell line	Binary: 0: does not overlap any H3K4me1 site in HUVEC cell line 1: overlaps a H3K4me1 site in HUVEC cell line	*in_HuvecH3K4me1*
Overlaps with H3K4me2 site in HUVEC cell line	Binary: 0: does not overlap any H3K4me2 site in HUVEC cell line 1: overlaps a H3K4me2 site in HUVEC cell line	*in_HuvecH3K4me2*
Overlaps with H3K4me3 site	Binary: 0: does not overlap	*in_HuvecH3K4me3*

Continued on next page

Table 3.1 – continued from previous page

Description	Coding	Variable name
in HUVEC cell line	any H3K4me3 site in HUVEC cell line 1: overlaps a H3K4me3 site in HUVEC cell line	
Overlaps with H3K4me2 site in HEPG2 cell line	Binary: 0: does not overlap any H3K4me2 site in HEPG2 cell line 1: overlaps a H3K4me2 site in HEPG2 cell line	*in_Hepg2H3K4me2*
Overlaps with H3K4me3 site in HEPG2 cell line	Binary: 0: does not overlap any H3K4me3 site in HEPG2 cell line 1: overlaps a H3K4me3 site in HEPG2 cell line	*in_Hepg2H3K4me3*

3.2.3 Model selection procedure

Given our data set, we were interested in finding a model that was complex enough to fit our data well, and that was simple to interpret at the same time. We first performed separate univariate analyses to identify important predictor variables. We fitted logistic regression models with one independent predictor variable and investigated their fits.

Based on the results obtained from the univariate analyses, we fitted a multiple logistic regression model using all the important predictor variables. Beginning from a complex model, we sequentially removed predictors by comparing the Akaike information criteria (AIC) (Akaike, 1973) and the area under the curve (AUC) values until no improvement was observed. This procedure is known as 'backward elimination'. We used the standard 'glm' (generalized linear model) function in R with 'family=binomial', as well as the 'logistf' function in the logistf package to perform Firth logistic regression.

Definition 5. (Akaike Information Criterion (AIC)) *The Akaike's information criterion (Akaike, 1973) judges the adequacy of a model. The model which minimizes AIC is generally considered as an optimal model (Kadane and Lazar, 2004). AIC is defined as*

$$\mathsf{AIC} = -2((\mathsf{argmax}_{\beta}\mathsf{ln}(L(\beta|X))) - N),$$

where $\mathsf{argmax}_{\beta}\mathsf{ln}(L(\beta|X))$ *is the maximized log likelihood and N the number of parameters in the model (Agresti, 2002). As the equation shows, AIC penalizes models with many parameters.*

3.2.4 Performance analysis using ROC curve

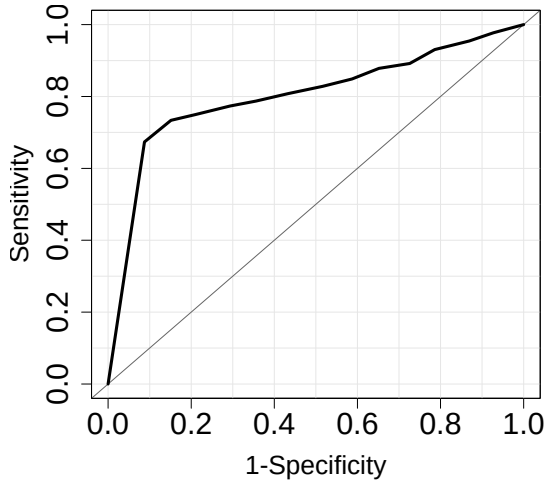


Figure 3.1 ROC curve of the sensitivity (y-axis) versus 1-specificity (x-axis) for hypothetical data along with a diagonal reference line (sensitivity=1-specificity).

To investigate the predictive ability of the logistic regression model, we performed a receiver operating characteristic (ROC) curve analysis. The ROC curve shown in Figure 3.1 is a plot of the sensitivity (true positive rate) of the model prediction against the complement of its specificity, 1-specificity (false positive rate), for a series of cut-off points. The closer the curve comes to the upper left corner (sensitivity of 1, specificity of 1), the higher is the overall accuracy of the model (Zweig and Campbell, 1993). In Figure 3.2, the sensitivity and specificity for a hypothetical model are plotted for a range of different cut-off points. The minimized difference threshold (MDT) is the point where sensitivity and specificity are equal. The sum of sensitivity and specificity is maximized at the maximized sum threshold (MST). The two threshold criteria of MDT and MST are commonly used to determine optimal cut-off values.

Definition 6. (True Positives (*TP*), False Positives (*FP*), True Negatives (*TN*) and False Negatives (*FN*) (Tompa et al., 2005)) *Suppose we have a data set with positive ($y = 1$) and negative ($y = 0$) observations. For a given cut-off point, the number of positive observations correctly classified as positive is called TP, and FN when those positives are classified as negative. TN is then the number of negative observations correctly classified as negative and FP the number of the negative observations classified as positive. The different counts can be organized into the following classification table called confusion matrix:*

	Observation		
Prediction	1	0	Total
1	*TP*	*FP*	*TP*+*FP*
0	*FN*	*TN*	*FN*+*TN*
Total	*TP*+*FN*	*FP*+*TN*	m=*TP*+*FN*+*FP*+*TN*

where m is the total number of observations.

Definition 7. (Sensitivity, specificity, precision and accuracy (Sinha et al., 2007; Tompa et al., 2005)) *The sensitivity is the proportion of positive observations correctly predicted as positive*

$$Sensitivity = \frac{TP}{TP + FN}$$

and the specificity is the proportion of negative observations correctly predicted as negative

$$Specificity = \frac{TN}{TN + FP}.$$

The precision, also known as the positive predictive value, and the accuracy can be determined as follows:

$$Precision = \frac{TP}{TP + FP}$$

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}.$$

Definition 8. (Area under the ROC curve (AUC) (Vittinghoff et al., 2011)) *The ROC curve defined as*

$$\mathsf{ROC} = \{(1 - \mathsf{Specificity}(c), \mathsf{Sensitivity}(c))\},\ c \in [0,1],$$

with cut-off c can be summarized by the AUC. The AUC is a single number ranging from zero to one which can be used to evaluate and to compare the performance of models. A good model is characterized by an AUC close to 1.

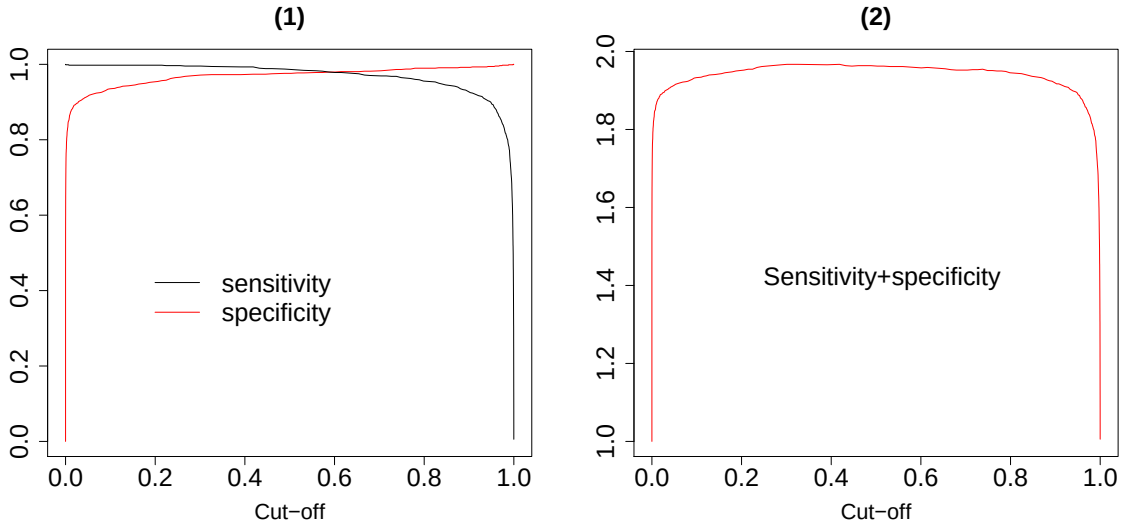


Figure 3.2 (1) Sensitivity and specificity for all possible cut-off points for a hypothetical model. The point where the two curves cross is called the minimzed difference threshold (MDT). The MDT represents the cut-off value at which sensitivity and specificity are equal. (2) The sum of sensitivity and specificity for different cut-off points. The highest point of the curve is called the maximized sum threshold (MST) and is the point on the ROC curve closest to the upper left corner. This is the cut-off point which maximizes the sum of sensitivity and specificity.

3.3 Results

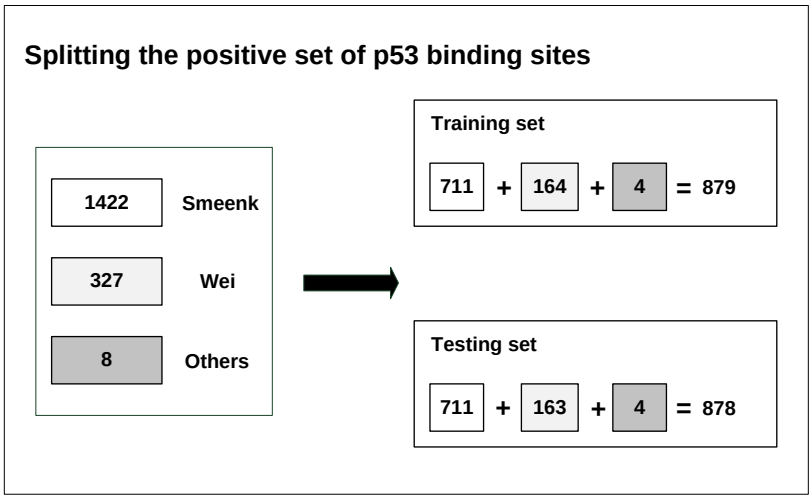


Figure 3.3 Positive training and testing sets. When randomly splitting the positive set of p53 binding sites, we made sure that each study was presented in the same proportion in the training and testing sets. In the end, 879 binding sites were selected as positive training and 878 as positive testing sites.

In total, we gathered 1758 sites for training and 1756 sites for testing including positive and negative sites.

3.3.1 Simple logistic regression for each predictor

Continuous vs. binary score predictors

Table 3.2 Descriptive statistics for the continuous *decamer1_score.cont*, *decamer2_score.cont* and *pair_score.cont* by *p53_bs_status*.

Predictor	*p53_bs_status*	Min.	1st Qu.	Median	Mean	3rd Qu.	Max
decamer1_score.cont	0	-41.08	-23.88	-18.30	-18.04	-12.91	9.03
	1	-16.81	6.15	9.03	8.39	11.36	14.52
decamer2_score.cont	0	-42.00	-23.4	-18.29	-17.66	-12.44	11.82
	1	-10.11	5.69	8.75	7.98	10.99	14.52
pair_score.cont	0	-72.57	-42.87	-36.13	-35.69	-28.04	2.17
	1	-6.66	13.14	17.41	16.37	20.30	28.10

The first independent variable to examine was the continuous *decamer1_score.cont* predictor. When looking at the scatterplot of the independent *decamer1_score.cont* variable and the response variable *p53_bs_status* given in Figure 3.4, the positive training sites (*p53_bs_status*=1) had considerably higher scores than the negative training sites (*p53_bs_status*=0) whose scores were mostly negative. Table 3.2 shows that only 25% of the positive training sites had scores smaller than 6.15, and 75% of the negative training sites fell below the value of -12.91.

To explore the relationship between *decamer1_score.cont* and *p53_bs_status*, we fitted a simple logistic regression model with the continuous *decamer1_score.cont* predictor. The logistic regression model linear in *decamer1_score.cont* as shown in Figure 3.4 resembled the lowess curve, indicating that a linear model was appropriate.

The continuous *decamer1_score.cont* predictor was statistically significant in the linear model (Table 3.3). To evaluate the overall fit of the logistic regression model, we looked at the null deviance (2437.11) and the residual deviance (201.13). We used the likelihood ratio test to compare the single-predictor logistic regression model to a null model with just an intercept (R^2 statistic). The χ^2 of 2235.98 (2437.11 − 201.13) with 1 degree of freedom yielded a P value close to zero. The logistic regression model as a whole fitted our data significantly better than the null model. The residual deviance was 201.13 on df=1756 indicating that the

fitted values were not significantly different from the observed values ($P \approx 1$).

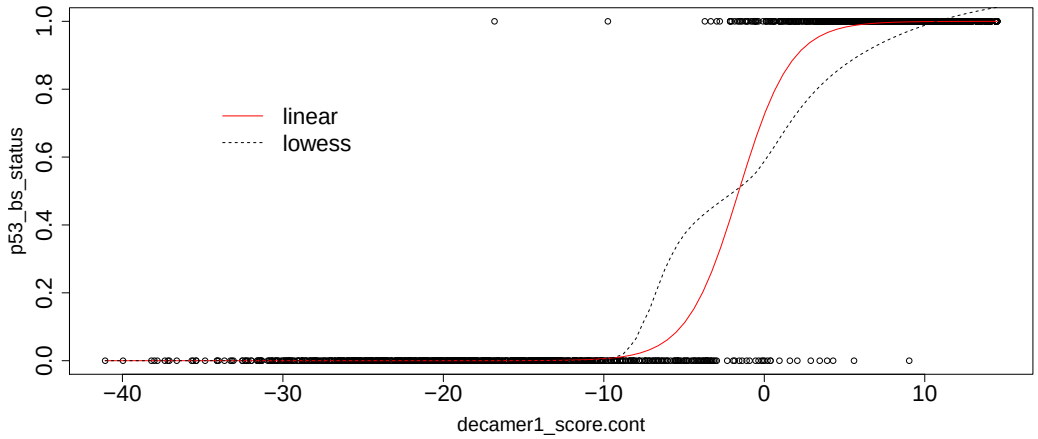


Figure 3.4 Scatterplot of *decamer1_score.cont* and *p53_bs_status* for the training sites with the single-predictor logistic regression model linear in *decamer1_score.cont* and a lowess curve displayed on a probability scale. The scatterplot clarifies the binary nature of the response variable *p53_bs_status*. All data points fall on one of the two horizontal lines representing the presence of p53 binding sites (*p53_bs_status*=1) and the absence of p53 binding sites (*p53_bs_status*=0).

Table 3.3 Logistic regression for *decamer1_score.cont* predictor

Predictor	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.9702	0.1911	5.08	3.82e-07 ***
decamer1_score.cont	0.6118	0.0512	11.95	< 2e-16 ***
(Intercept)	-3.4317	0.1920	-17.87	<2e-16 ***
decamer1_score.positive[T.TRUE]	7.6131	0.3391	22.45	<2e-16 ***

When evaluating the overall performance of the binary *decamer1_score.positive* predictor, the χ^2 of 2053.12 (2437.11-383.99) on 1 degree of freedom with a P value close to zero told us that the logistic regression model as a whole fitted significantly better than the null model (Table 3.3). Furthermore, the residual deviance of 383.99 on df=1756 ($P \approx 1$) indicated that the fitted values were not significantly different from the observed values.

Comparing the two logistic regression models based on AIC and AUC, the model with the continuous *decamer1_score.cont* predictor was to be preferred to the one with the binary *decamer1_score.positive* variable due to its smaller AIC and greater AUC values (Table 3.7).

Similar to *decamer1_score.cont*, the *decamer2_score.cont* predictor showed a strong effect on the outcome variable *p53_bs_status*.

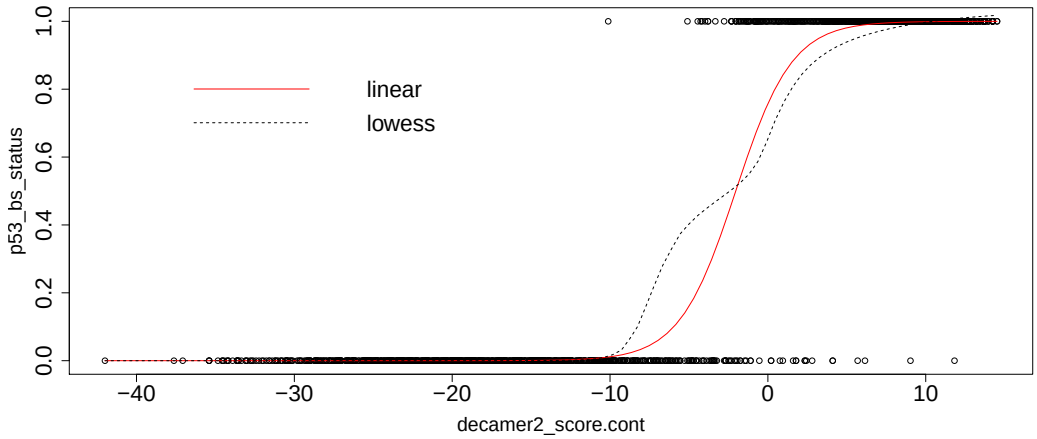


Figure 3.5 Scatterplot of *decamer2_score.cont* and *p53_bs_status* for the training sites with the single-predictor logistic regression models linear in *decamer2_score.cont* and a lowess curve displayed on a probability scale.

As shown in Figure 3.5, a linear term for the *decamer2_score.cont* predictor was appropriate when fitting a single-predictor logistic regression model to our data. The continuous *decamer2_score.cont* predictor was statistically significant (Table 3.4) and the logistic regression model as a whole fitted significantly better than the null model. When we performed a χ^2-test using the difference in deviance residuals between our logistic regression model and the null model, we obtained a χ^2 of 2179.66 with 1 degree of freedom, yielding a small P value which was very close to zero.

Replacing the *decamer2_score.cont* predictor by the binary *decamer2_score.positive* variable, an increase in AIC and a decrease in AUC were observed (Table 3.7). Thus, a continuous version for the *decamer2_score* predictor was preferred.

Table 3.4 Logistic regression for *decamer2_score* predictor.

Predictor	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.1388	0.1726	6.60	4.15e-11 ***
decamer2_score.cont	0.5597	0.0429	13.06	< 2e-16 ***
(Intercept)	-2.7703	0.1403	-19.75	<2e-16 ***
decamer2_score.positive[T.TRUE]	6.6524	0.2823	23.56	<2e-16 ***

We next analyzed the *pair_score* predictor. To determine the correct structural form of the continuous *pair_score.cont* predictor, we fitted two simple logistic regression models, one which was linear and another one which was quadratic in *pair_score.cont* (Figure 3.6).

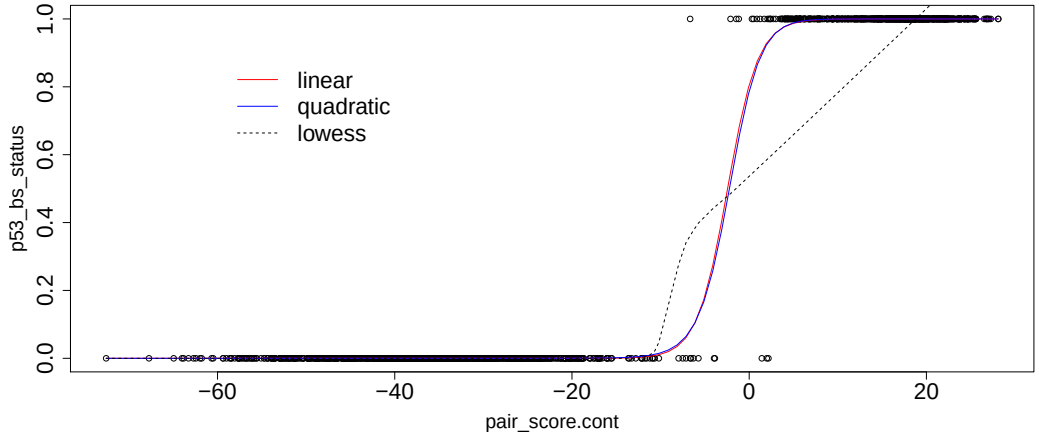


Figure 3.6 Scatterplot of *pair_score.cont* and *p53_bs_status* for the training sites with the single-predictor logistic regression model linear and quadratic in *pair_score.cont* and a lowess curve displayed on a probability scale.

As shown in Table 3.7, adding a quadratic term did not decrease the AIC. In addition, the *pair_score.cont* predictor was statistically significant, whereas the quadratic term was not (Table 3.5). Thus, only the linear term for *pair_score.cont* should be used in the model.

Table 3.5 Logistic regression for *pair_score* predictor

Predictor	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.4215	0.5165	2.75	0.00592 **
pair_score.cont	0.5830	0.1133	5.14	2.7e-07 ***
(Intercept)	1.3073	0.5663	2.31	0.021 *
pair_score.cont	0.5991	0.1242	4.82	1.41e-06 ***
I(pair_score.cont^2)	0.0057	0.0114	0.50	0.620
(Intercept)	-5.3891	0.5011	-10.75	<2e-16 ***
pair_score.positive[T.TRUE]	11.0647	0.7652	14.46	<2e-16 ***

When running the univariate logistic regression model using the standard 'glm' function in R, we obtained a warning saying "fitted probabilities numerically 0 or 1 occurred". This normally means that a perfect fit is possible within the parametrization of the model.

Table 3.6 Table of *pair_score.cont* by *p53_bs_status*.

p53_bs_status	<0	≥ 0
0	876	3
1	4	875

Here, in our case, however, this message seemed to be a warning about very small fitted probabilities. As shown in Table 3.6, there was no evidence that a problem of complete or quasi-complete separation occured. In addition, the parameter estimate for *pair_score.cont*, as well as the standard errors for the parameter estimation were not too large to worry about (Table 3.5). For more information on complete separation, see Subsection 3.1.2.

When we fitted a logistic regression model of *p53_bs_status* on *pair_score.positive* which was presented as a binary predictor, R didn't report a warning. The predictor was reported to be significant. The AIC was 95.20, and thus larger than the AIC for the model with the continuous *pair_score.cont* predictor (Table 3.7). The χ^2 of 2345.91 on 1 degree of freedom with a P value close to zero showed that the logistic regression model as a whole fitted significantly better than the null model. And the residual deviance of 91.195 on df=1756 showed that the fitted values were not significantly different from the observed values ($P \approx 1$).

Table 3.7 Simple logistic regression models of the score predictors with their AIC and AUC values.

Predictor	AIC	AUC
decamer1_score.cont	205.13	0.9973591
decamer1_score.positive	387.99	0.976678
decamer2_score.cont	261.45	0.9956196
decamer2_score.positive	580.85	0.9596132
pair_score.cont	34.11	0.9999573
pair_score.cont+I(*pair_score.cont*^{2})	35.95	0.9999573
pair_score.positive	95.20	0.9960182

Since *pair_score.cont* represented the sum of the two predictors *decamer1_score.cont* and *decamer2_score.cont*, it was worth examining correlation between the predictors. The correla-

tion between *pair_score.cont* with either *decamer1_score.cont* or *decamer2_score.cont* was quite high (Table 3.8). The two continuous predictors *decamer1_score.cont* and *decamer2_score.cont* were highly correlated, too. It was advisable not to use three of them at the same time in the model. Based on the findings, the predictor which best represented motif match was the continuous *pair_score.cont* predictor (Table 3.7).

Table 3.8 Correlations between the *decamer1_score.cont*, *decamer2_score.cont* and *pair_score.cont' predictors*.

	decamer1_score	*decamer2_score*	*pair_score*
decamer1_score	1		
decamer2_score	0.80	1	
pair_score	0.95	0.95	1

Continuous *spacer.cont* predictor

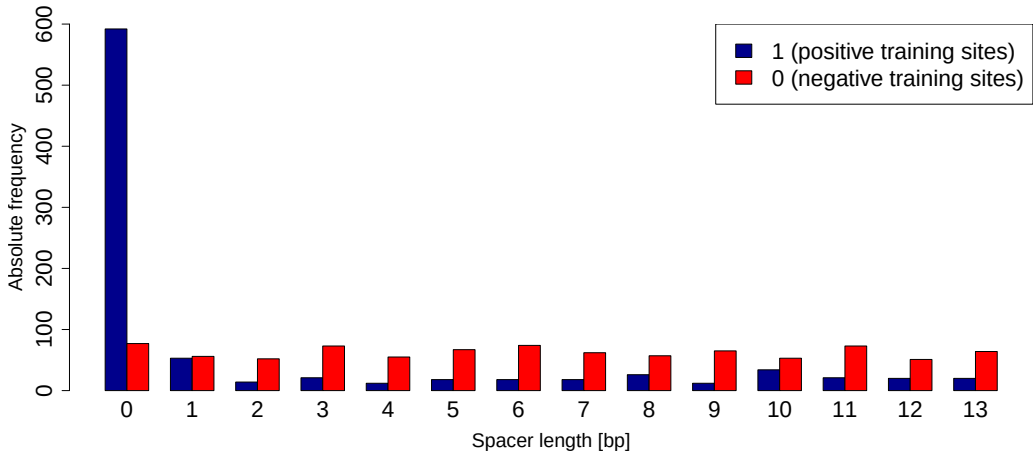


Figure 3.7 Spacer length distribution of the training sites by *p53_bs_status*.

As shown in Figure 3.7, the majority of the positive training sites (*p53_bs_status*=1) did not have a spacer between the two half-sites. For the remaining positive training sites, the length of spacers varied from 1 to our upper limit of 13 bp, whereas 1 bp and 10 bp spacers were present more frequently than others.

We fitted several simple logistic regression models to examine different structural forms (linear, quadratic and cubic) for the *spacer.cont* predictor (Figure 3.8). When we compared

the different parametric forms by using AIC and AUC, the quadratic or the cubic forms for *spacer.cont* seemed to be appropriate in the logistic regression models (Table 3.9). The continuous *spacer.cont* predictor was statistically significant in the three models. The χ^2 of 623.10 with df=2 for the quadratic model and the χ^2 of 717.30 with df=3 for the cubic model yielded both a P value close to zero. This indicated that the logistic regression models quadratic and cubic in *spacer.cont* as a whole fitted significantly better than the null model. Furthermore, the fitted values were not significantly different from the observed values for both models ($P(\chi^2_{1755}) > 1814 = 0.16$; $P(\chi^2_{1754}) > 1719.8 = 0.72$).

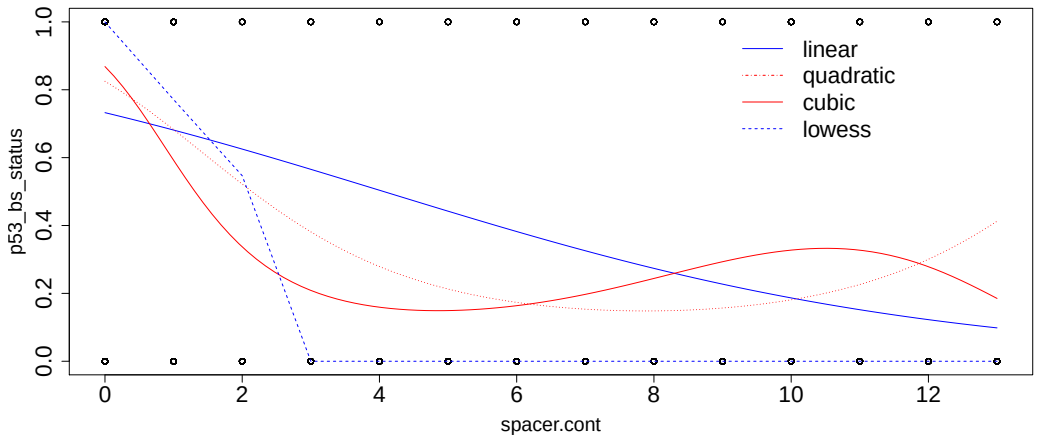


Figure 3.8 Scatterplot of *spacer.cont* and *p53_bs_status* for the training sites with the single-predictor logistic regression models linear, quadratic and cubic in *spacer.cont* and a lowess curve displayed on a probability scale.

Table 3.9 Logistic regression for the *spacer.cont* predictors.

Predictor	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.0075	0.0739	13.64	<2e-16 ***
spacer.cont	-0.2480	0.0136	-18.21	<2e-16 ***
(Intercept)	1.5502	0.0925	16.77	<2e-16 ***
spacer.cont	-0.8376	0.0468	-17.89	<2e-16 ***
I(spacer.cont^2)	0.0532	0.0038	13.82	<2e-16 ***
(Intercept)	1.8867	0.1081	17.45	<2e-16 ***
spacer.cont	-1.7679	0.1129	-15.66	<2e-16 ***
I(spacer.cont^2)	0.2663	0.0232	11.50	<2e-16 ***
I(spacer.cont^3)	-0.0116	0.0012	-9.39	<2e-16 ***

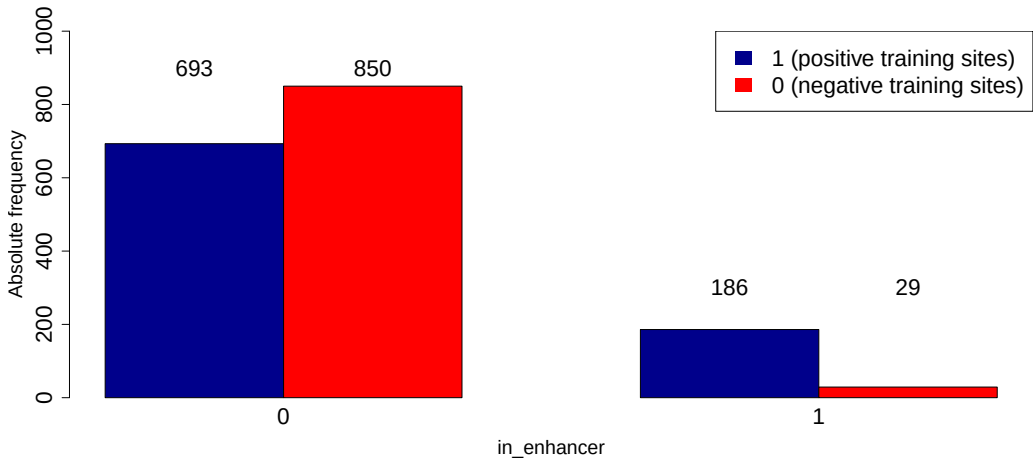


Figure 3.9 Absolute frequencies of overlaps with enhancers by *p53_bs_status*.

Investigating the *in_enhancer* predictor by *p53_bs_status*, the majority of the training sites did not overlap a predicted enhancer (Figure 3.9). However, the difference between observed (positive training) and expected (negative training) counts of sites which did overlap an enhancer and which did not was statistically significant (G=408.31, df=1, $P\approx 0$). We fitted a

Table 3.10 Logistic regression models for the binary *in_enhancer* predictor.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.2042	0.0512	-3.99	6.61e-05 ***
in_enhancer[T.1]	2.0627	0.2061	10.01	< 2e-16 ***

Null deviance: 2437.1 on 1757 degrees of freedom

Residual deviance: 2293.1 on 1756 degrees of freedom

AIC: 2297.1

AUC: 0.589306

simple logistic regression of *p53_bs_status* on the binary *in_enhancer* predictor. As reported in Table 3.10, the P value for the coefficient of the predictor variable was strongly significant ($P<2\times10^{-16}$). The odds of those p53 binding sites overlapping any known enhancer was 6.41 ($P(y=1|1)(1-P(y=1|1))=186/29$), while the odds of those p53 binding sites which did not overlap an enhancer was 0.815 ($P(y=1|0)(1-P(y=1|0))=693/850$). The resulting odds ratio of 7.8669 (6.41/0.815) indicated that the odds of sites overlapping an enhancer

are p53 binding sites were nearly 8 times the odds of sites which do not overlap an enhancer. The probability of sites which did overlap an enhancer are p53 binding sites was 0.8651 ($186/(186+29)$), and thus larger than the overall proportion of p53 binding sites in the data set which is 0.5. The estimated probability could also be determined by using the equation from Definition 3 in Subsection 3.1.1:

$$P = \frac{\exp(-0.2042 + 2.0627(1))}{1 + \exp(-0.2042 + 2.0627(1))} = 0.8651.$$

The probability of sites not overlapping an enhancer are p53 binding sites was 0.4491 ($693/(693+850)$ or $\exp(-0.2042)/(1+\exp(-0.2042))$). The corresponding likelihood ratio of 1.9263 (0.8651/0.4491) indicated that sites overlapping an enhancer were nearly 2 times more likely to be p53 binding sites than for those which did not overlap an enhancer.

The logistic regression model as a whole for the *in_enhancer* predictor fitted our data significantly better than the null model with just an intercept. The χ^2 of 144 with 1 degree of freedom yieled a P value which was very close to zero. The residual deviance of 2293.1 on 1756 degrees of freedom, however, indicated that the fitted values were significantly different from the observed values (P=1.110223$\times 10^{-16}$). The *in_enhancer* predictor alone was not able to make a significant contribution to the model for the prediction of p53 binding sites. There was room for improvement in the model.

Binary *in_H3K4me1*, *in_H3K4me2* and *in_H3K4me3* predictors

The difference between observed (positive training) and expected (negative training) counts of sites which do and do not overlap histone Lys4 methylation regions was highly statistically significant at the conventional 0.05 level for most of the *in_H3K4me1*, *in_H3K4me2* and *in_H3K4me3* predictors according to the G-tests. The G-tests for the *in_HuvecH3K4me2* and *in_NhlfH3K4me3* predictors, however, resulted in G=1.33, df=1, P= 0.25 and G=2.58, df=1, P=0.11, respectively. For those two out of the 14 predictors shown in Figure 3.10, no significant results were obtained. This observation is consistent with the results reported by logistic regression. As shown in Table 3.12 and Table 3.13, the two single predictor logistic regression models with *in_HuvecH3K4me2* and *in_NhlfH3K4me3*, respectively, had a poor fit (see R^2 statistic). The small χ^2 values and the resulting P values larger than 0.05 indicated that the models as a whole did not fit the data significantly better than the null model. In addition, when we used the AIC and AUC values for model comparison, the logistic regression models with *in_HuvecH3K4me2* and *in_NhlfH3K4me3* had the largest AIC (AIC(*in_HuvecH3K4me2*)=2440.4; AIC(*in_NhlfH3K4me3*)=2439.8) and the lowest AUC

(AUC(*in_HuvecH3K4me2*)=0.51; AUC(*in_NhlfH3K4me3*)=0.51) values.

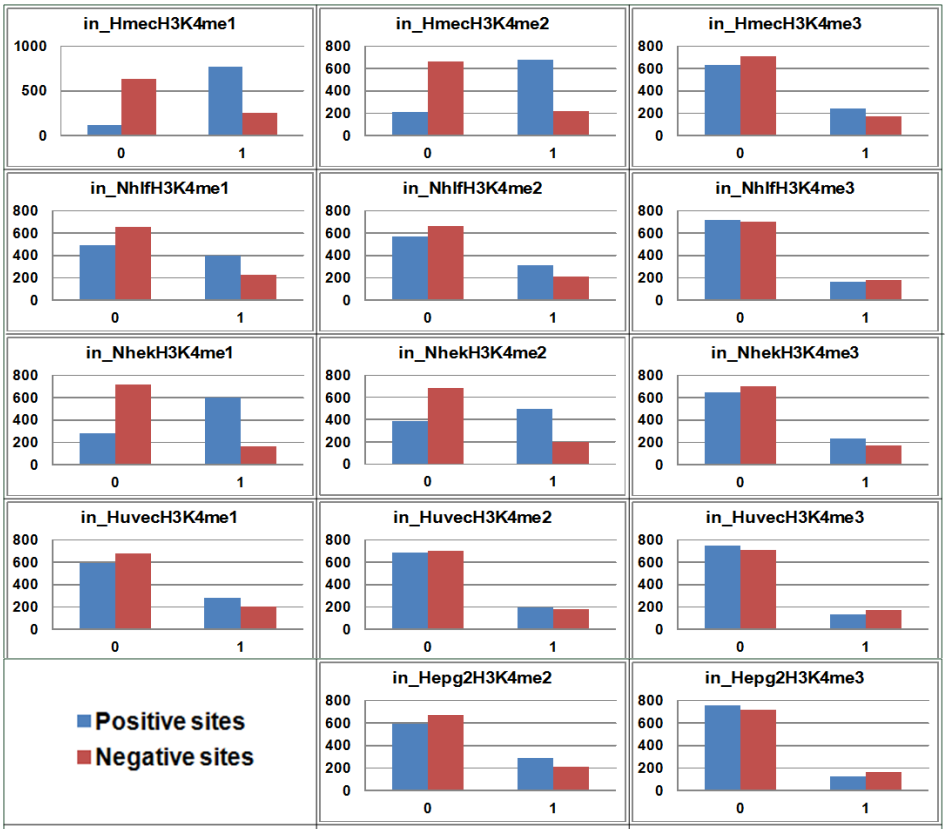


Figure 3.10 Absolute frequencies of observations *in_H3K4me1*, *in_H3K4me2* and *in_H3K4me3* for the cell lines 'HMEC', 'NHLF', 'NHEK', 'HUVEC' and 'HEPG2' classified by *p53_bs_status*.

In general, the differences between observed and fitted values were statistically significant for most of the 14 models indicating a very poor fit to the data. The binary *in_H3K4me1*, *in_H3K4me2* and *in_H3K4me3* predictors alone did not seem to be sufficient to explain the observed outcome and to correctly predict p53 binding sites. The only exception to this was the model with the *in_HmecH3K4me1* predictor. The χ^2 of 666.9 on 1 degree of freedom with a P value close to zero showed that the logistic regression model based on *in_HmecH3K4me1* as a whole fitted significantly better than the null model. And the residual deviance of 1770.2 on 1756 degrees of freedom indicated that the fitted values were not significantly different from the observed values ($P(\chi^2_{1756}) > 1770.2 = 0.4012$). However, here as well, there was room for improvement in the model.

Table 3.11 Results of the logistic regression models for the binary *in_H3K4me1* predictors.

Cell line	AIC	AUC	Deviance χ^2 test	R^2 statistic
HMEC	1774.2	0.79	$P(\chi^2_{1756}) > 1770.2 = 0.40$	$P(\chi^2_1) > 666.9 \approx 0$
NHLF	2368.2	0.60	$P(\chi^2_{1756}) > 2364.2 \approx 0$	$P(\chi^2_1) > 72.9 \approx 0$
NHEK	1979.1	0.75	$P(\chi^2_{1756}) > 1975.1 = 1.00 \times 10^{-4}$	$P(\chi^2_1) > 462 \approx 0$
HUVEC	2422.3	0.55	$P(\chi^2_{1756}) > 2418.3 \approx 0$	$P(\chi^2_1) > 18.8 = 1.45 \times 10^{-5}$

Table 3.12 Results of the logistic regression models for the binary *in_H3K4me2* predictors.

Cell line	AIC	AUC	Deviance χ^2 test	R^2 statistic
HMEC	1946.2	0.76	$P(\chi^2_{1756}) > 1942.2 = 1.00 \times 10^{-3}$	$P(\chi^2_1) > 494.9 \approx 0$
NHLF	2415.9	0.55	$P(\chi^2_{1756}) > 2411.9 \approx 0$	$P(\chi^2_1) > 25.2 = 5.17 \times 10^{-7}$
NHEK	2225.4	0.67	$P(\chi^2_{1756}) > 2221.4 = 1.92 \times 10^{-13}$	$P(\chi^2_1) > 215.7 \approx 0$
HUVEC	2440.4	0.51	$P(\chi^2_{1756}) > 2436.4 \approx 0$	$P(\chi^2_1) > 0.7 = 0.40$
HEPG2	2425.2	0.54	$P(\chi^2_{1756}) > 2421.2 \approx 0$	$P(\chi^2_1) > 15.9 = 6.68 \times 10^{-5}$

Table 3.13 Results of the logistic regression models for the binary *in_H3K4me3* predictors.

Cell line	AIC	AUC	Deviance χ^2 test	R^2 statistic
HMEC	2422.8	0.54	$P(\chi^2_{1756}) > 2418.8 \approx 0$	$P(\chi^2_1) > 18.3 = 1.45 \times 10^{-5}$
NHLF	2439.8	0.51	$P(\chi^2_{1756}) > 2435.8 \approx 0$	$P(\chi^2_1) > 1.3 = 0.25$
NHEK	2428.4	0.54	$P(\chi^2_{1756}) > 2424.4 \approx 0$	$P(\chi^2_1) > 12.7 = 4.0 \times 10^{-4}$
HUVEC	2435.6	0.52	$P(\chi^2_{1756}) > 2431.6 \approx 0$	$P(\chi^2_1) > 5.5 = 1.9 \times 10^{-2}$
HEPG2	2435.4	0.52	$P(\chi^2_{1756}) > 2431.4 \approx 0$	$P(\chi^2_1) > 5.7 = 1.7 \times 10^{-2}$

3.3.2 Model selection using backward elimination

Table 3.14 Predictor names and their abbreviations. For simplicity, we abbreviated the names of the predictors. These abbreviations are used in the tables which present the multiple logistic regression models.

Predictor	Abbr.	Predictor	Abbr.
decamer1_score.cont	*D1.c*	*in_NhlfH3K4me1*	*Nhl1*
decamer1_score.positive	*D1.p*	*in_NhlfH3K4me2*	*Nhl2*
decamer2_score.cont	*D2.c*	*in_NhlfH3K4me3*	*Nhl3*
decamer2_score.positive	*D2.p*	*in_NhekH3K4me1*	*Nhe1*
pair_score.cont	*P.c*	*in_NhekH3K4me2*	*Nhe2*
pair_score.positive	*P.p*	*in_NhekH3K4me3*	*Nhe3*
spacer.cont	*S.c*	*in_HuvecH3K4me1*	*Hu1*
in_enhancer	*E*	*in_HuvecH3K4me2*	*Hu2*
in_HmecH3K4me1	*Hm1*	*in_HuvecH3K4me3*	*Hu3*
in_HmecH3K4me2	*Hm2*	*in_Hepg2H3K4me2*	*He2*
in_HmecH3K4me3	*Hm3*	*in_Hepg2H3K4me3*	*He3*

GLM models

We first built multiple logistic regression models using the standard 'glm' function in R. We started by fitting a full model containing 17 predictor variables to predict p53 binding sites (Model 1 shown in Table 3.15). In addition, quadratic and cubic terms for the continuous *spacer.cont* predictor were added to the full model. We used the binary *pair_score.positive* instead of the continuous *pair_score.cont* due to convergence problems reported by R. Other multiple logistic regression models with the highly correlated continuous *decamer1_score.cont* and *decamer2_score.cont* predictors were tested separately and compared to the models that included *pair_score.cont*. Models including *pair_score.cont* showed a better performance in terms of AIC and AUC.

Due to lack of significance for individual effects in the full model, we dropped those predictors with the largest P values for which their removal had a significant effect on the model. The likelihood-ratio statistic comparing the full model (model 1) to the reduced model 2 which did not include the *in_HmecH3K4me2*, *in_HmecH3K4me3*, *in_NhekH3K4me3*

and *in_HuvecH3K4me2* predictors equaled 0.12 ($df = 4$) indicating that the four predictors were not necessary ($P = 9.98 \times 10^{-1}$) and did not significantly improve the fit in terms of increasing the likelihood or decreasing the deviance.

Table 3.15 Results of fitting several logistic regression models using the standard glm function. A list of explanation of used abbreviations for the predictors can be found in Table 3.14.

Model	Predictors	AIC	AUC
1	$P.p + S.c + I(S.c^2) + I(S.c^3) + E$ $+Hm1 + Hm2 + Hm3 + Nhl1 + Nhl2 + Nhl3$ $+Nhe1 + Nhe2 + Nhe3 + Hu1 + Hu2 + Hu3$ $+He2 + He3$	99.46	0.999492
2	$P.p + S.c + I(S.c^2) + I(S.c^3) + E$ $+Hm1 + Nhl1 + Nhl2 + Nhl3 + Nhe1 + Nhe2$ $+Hu1 + Hu3 + He2 + He3$	91.57	0.9995011
3	$P.p + S.c + I(S.c^2) + I(S.c^3) + Hm1$ $+Nhl1 + Nhl2 + Nhl3 + He2 + He3$	82.09	0.9994978
4	$P.p + S.c + I(S.c^2) + I(S.c^3) + Hm1$ $+Nhl2$	79.29	0.9993574
5	$P.p + S.c + I(S.c^2) + Hm1 + Nhl2$	78.87	0.9992772
6.1	$P.p + S.c + I(S.c^2) + Hm1$	79.02	0.9988358
6.2	$P.p + S.c + I(S.c^2) + Nhl2$	79.51	0.9992526

We next considered a model which removed *in_enhancer*, *in_NhekH3K4me1*, *in_NhekH3K4me2*, *in_HuvecH3K4me1* and *in_HuvecH3K4me3* (model 3). The likelihood ratio test comparing model 2 to the reduced model 3 resulted in an increased deviance of 0.52 on df=5 with P=0.99. The reduced model 3 was significantly better than model 2. Model 4, which in comparison to model 5 did not include *in_NhlfH3K4me1*, *in_NhlfH3K4me3*, *in_Hepg2H3K4me2* and *in_Hepg2H3K4me3* had an increased deviance of 5.20 on df=4 resulting in P=0.27. The next likelihood ratio test comparing model 4 and model 5 suggested that the cubic term I($spacer.cont^3$) included in model 4, but not in model 5 is unnecessary (deviance.difference = 1.58, df=1, P=0.21). It was possible to further reduce the model by dropping *in_NhlfH3K4me2* (model 6.1) or by removing *in_HmecH3K4me1* (model 6.2). In both cases, the removal was

highly significant. Further simplification resulted in a significantly poorer fit and was therefore not recommended. We also analyzed models with all types of interaction terms between the predictors, but none of them were found to be significant.

Model 5, model 6.1 and model 6.2, seemed reasonable for a good prediction model of p53 binding sites. Comparing the three models using AIC, AIC was smallest for model 5 which included *pair_score.positive*, *spacer.cont*, I($spacer.cont^2$), *in_HmecH3K4me1* and *in_NhlfH3K4me2*. Thus, based on AIC, model 5 would be the model which rates best.

Logistf models

Based on the results obtained from the univariate analyses, AIC was smallest for the single predictor logistic regression model with *pair_score.cont* (AIC $= 34.11$). This demonstrates strong evidence of an effect of *pair_score.cont* on predicting p53 binding sites. The continuous predictor seems to be an extremely good predictor. Since the standard 'glm' function in R reported convergence problems that may have been caused by complete or quasi-complete separation of data points when using *pair_score.cont* in combination with other predictor variables, we used the 'logistf' function included in Heinze's 'logistf' package to perform Firth logistic regression (Heinze and Schemper, 2002). Unlike 'glm', the 'logistf' function does not report an AIC statistic, but a likelihood ratio test statistic. Thus, AIC cannot be used to compare models.

Table 3.16 shows the results of the multiple logistic regression models fitted by the 'logistf' function. We started with a full model with 17 predictor variables, including a quadratic term for *pair_score.cont*. We next dropped all predictor variables whose estimated coefficients resulted in large P values. When we compared the reduced model 2 to the full model (model 1), the AUC value was smaller for the full model, even though much more predictors were included in the full model indicating that the reduced model was the preferred model. Continuing the simplification process, we resulted in models 3, 4 and 5 with slightly different AUC values.

Among the five models listed in Table 3.16 which all as a whole fitted well to our data (see likelihood ratio test statistics), the AUC was largest for model 2. However, the other three models, model 3, model 4 and model 5, do also seem reasonable, because the difference in AUC was very minimal. Model 1 is not preferable due to a deep lack of significance for individual effects.

Table 3.16 Results of fitting several logistic regression models using the 'logistf' function. A list of explanation of used abbreviations for the predictors can be found in Table 3.14.

Model	Predictors	AUC	Likelihood ratio test (test statistic,df,P value)
1	$P.c + S.c + I(S.c^2) + E + Hm1 + Hm2 + Hm3 + Nhl1 + Nhl2 + Nhl3 + Nhe1 + Nhe2 + Nhe3 + Hu1 + Hu2 + Hu3 + He2 + He3$	0.9999961	(2333.19,18,$P \approx 0$)
2	$P.c + S.c + I(S.c^2) + Nhl1 + Nhe1 + Nhe3 + Hu3$	0.9999974	(2379.46,7,$P \approx 0$)
3	$P.c + S.c + I(S.c^2) + Nhe1 + Nhe3 + Hu3$	0.9999871	(2380.69,6,$P \approx 0$)
4	$P.c + S.c + I(S.c^2) + Nhe1 + Nhe3$	0.9999832	(2383.30,5,$P \approx 0$)
5	$P.c + S.c + I(S.c^2) + Nhe1$	0.9999819	(2387.52,4,$P \approx 0$)

3.3.3 Performance analysis on the training data

The seven top models we selected as potential models for predicting p53 binding sites were: glm.model5, glm.model6.1, glm.model6.2, logistf.model2, logistf.model3, logistf.model4 and logistf.model5.

We evaluated the prediction ability of the chosen models to discriminate p53 binding sites from non-p53 binding sites by determining specific performing measures, such as sensitivity and specificity, and using ROC curve analysis based on the training data set. Figure 3.11

shows the sum of sensitivity and specificity at various cut-off points (left plot) and the ROC curve (right plot) for model 'glm.model5'. The cut-off value which provided the largest sum of sensitivity and specificity was MST=0.1074 with a sensitivity of 0.9954 and a specificity of 0.9966. The corresponding confusion (classification) matrix for MST as probability cut-off was:

	observation	
prediction	1	0
1	875	3
0	4	876

with TP=875, FP=3, FN=4 and TN=876. The value for the area under the ROC curve (AUC) on the right plot was reported to be AUC=0.9993.

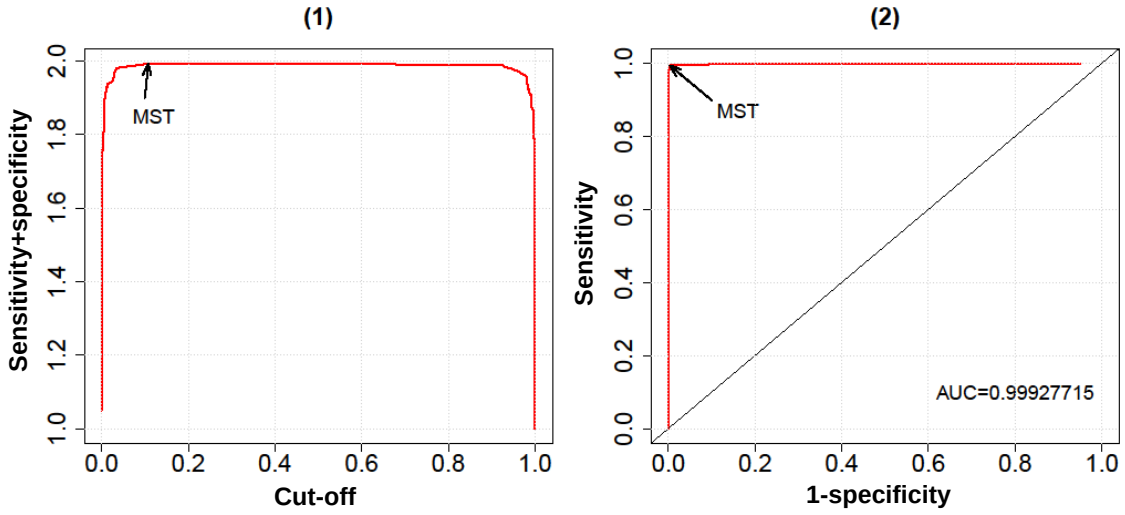


Figure 3.11 (glm.model5) (1) The sum of sensitivity and specificity for different cut-off points. The highest point of the curve is called MST. The MST of the 'glm' model is equal to 0.1074 resulting in a sum of 1.9920, where the values for sensitivity and specificity are 0.9954 and 0.9966, respectively. (2) ROC curve when plotting sensitivity against 1-specificity. The area under the ROC curve is called AUC.

For model 'glm.model6.1', the value of MST was reported to be 0.0789 resulting in the same sensitivity and specificity of 0.9954 and 0.9966 as model 'glm.model5' (Figure 3.12). The corresponding confusion matrix was the same, too. The AUC of 0.9988, however, was smaller than that of the previous model.

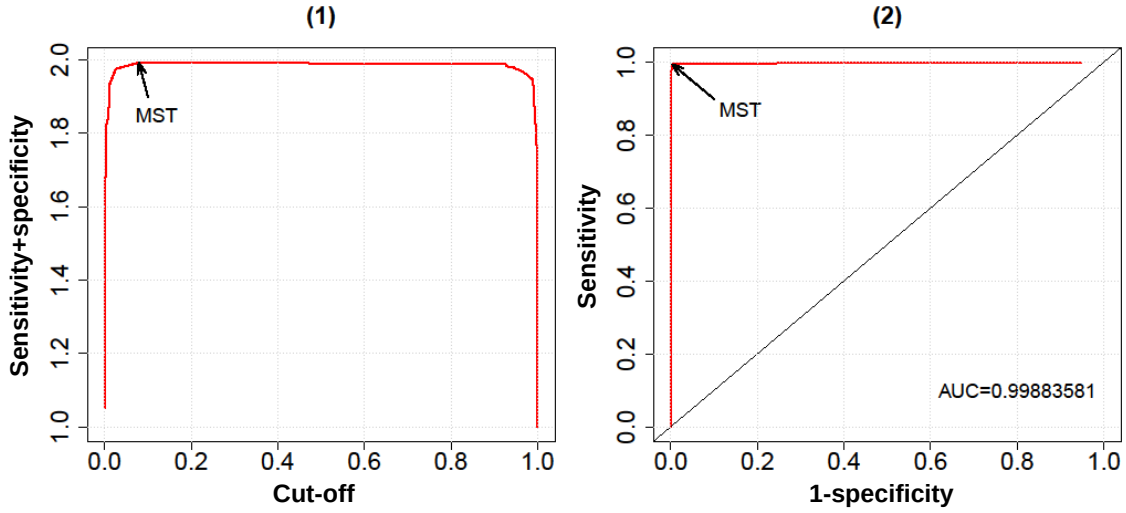


Figure 3.12 (glm.model6.1) (1) The sum of sensitivity and specificity for different cut-off points. The highest point of the curve is called MST. The MST of the 'glm' model is equal to 0.0789 resulting in a sum of 1.9920, where the values for sensitivity and specificity are 0.9954 and 0.9966, respectively. (2) ROC curve when plotting sensitivity against 1-specificity.

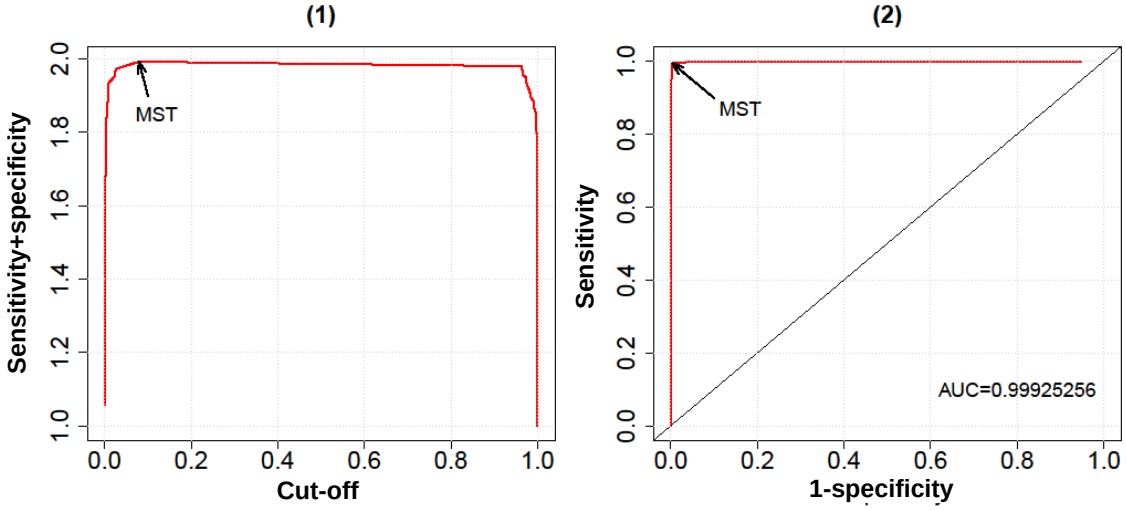


Figure 3.13 (glm.model6.2) (1) The sum of sensitivity and specificity for different cut-off points. The highest point of the curve is called MST. The MST of the 'glm' model is equal to 0.0793 resulting in a sum of 1.9920, where the values for sensitivity and specificity are 0.9954 and 0.9966, respectively. (2) ROC curve when plotting sensitivity against 1-specificity.

The value of MST for model 'glm.model6.2' was MST=0.0793 yielding the same confusion matrix, and thus the same sensitivity and specificity as for the other 'glm' models. The

reported AUC of 0.9993 was larger than 'glm.model6.1', but slightly less than 'glm.model5'.

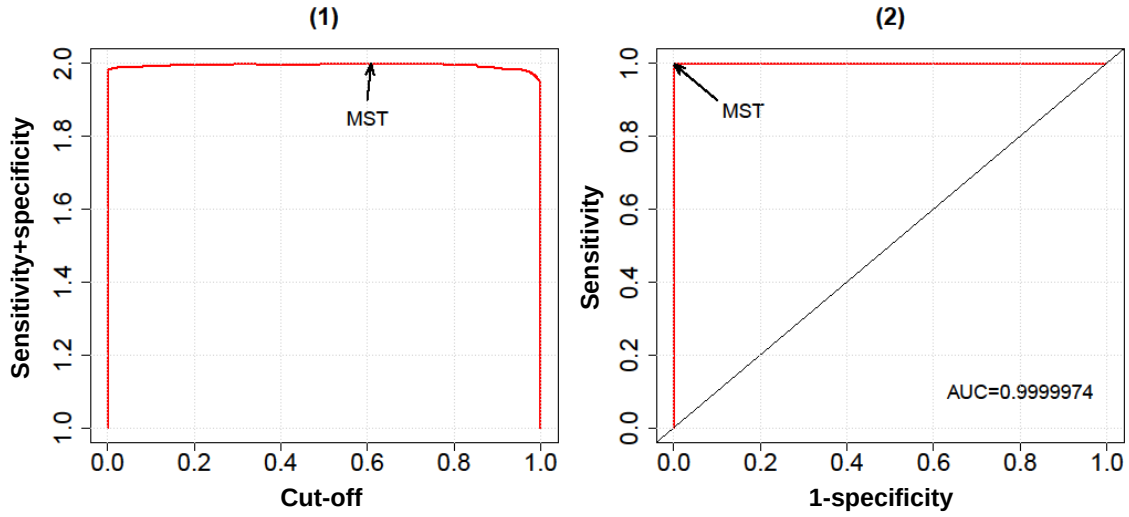


Figure 3.14 (logistf.model2) (1) The sum of sensitivity and specificity for different cut-off points. The highest point of the curve is called MST. The MST of the 'logistf' model is equal to 0.6088 resulting in a sum of 1.9989, where the values for sensitivity and specificity are 0.9989 and 1, respectively. (2) ROC curve when plotting sensitivity against 1-specificity.

As shown in Figure 3.14, the MST of the 'logistf.model2' model (MST=0.6088) was much greater than in the 'glm' models. Using a cut-off which was equal to the MST, we obtained the following confusion matrix:

	observation	
prediction	1	0
1	878	0
0	1	879.

Due to false positive fraction of zero the resulted value for specificity was equal to 1. The sensitivity was 0.9989. Adding both measures together, we obtained a sum of 1.9989. The area under the ROC curve was reported to be equal to AUC=0.9999974 which was much greater than those of the glm models.

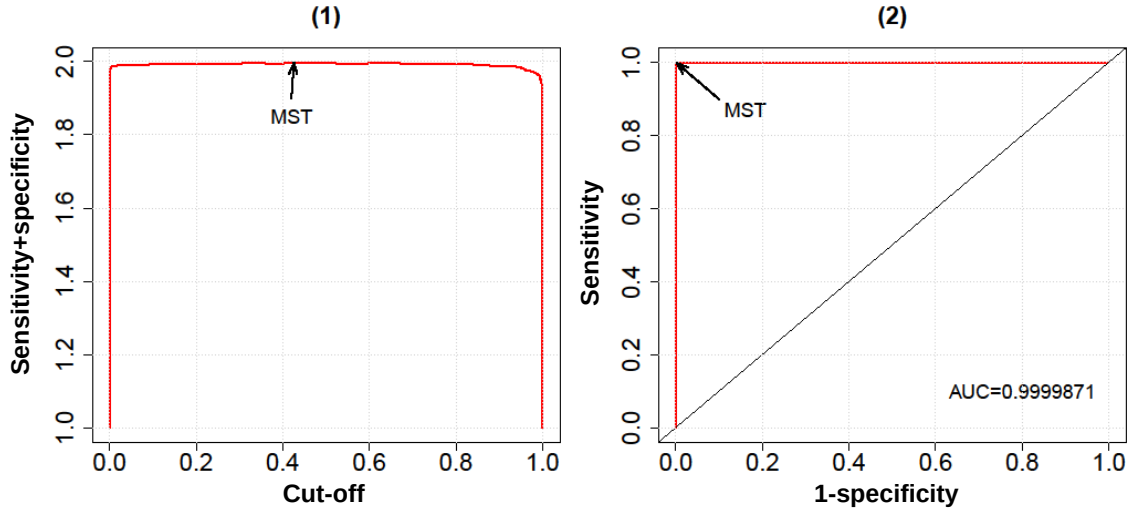


Figure 3.15 (logistf.model3) (1) The sum of sensitivity and specificity for different cut-off points. The highest point of the curve is called MST. The MST of the 'logistf' model is equal to 0.4259 resulting in a sum of 1.9966, where the values for sensitivity and specificity are 0.9989 and 0.9977, respectively. (2) ROC curve when plotting sensitivity against 1-specificity.

Figure 3.15 shows the performance result of model 'logistf.model3'. The MST of the model was equal to 0.4259 which produced the confusion matrix

	observation	
prediction	1	0
1	878	2
0	1	877.

The corresponding values for sensitivity and specificity were 0.9989 and 0.9977, respectively and the sum of both measures was 1.9966.

The logistf model 'logistf.model4' had two MST values, MST1=0.4014 and MST2=0.7882. MST1 produced the confusion matrix

	observation	
prediction	1	0
1	878	3
0	1	876.

with sensitivity=0.9989 and specificity=0.9966, and MST2 created the confusion matrix

	observation	
prediction	1	0
1	875	0
0	4	879

sensitivity=0.9954 and specificity=1. For both MST values, we obtained a sum of sensitivity and specificity which was equal to 1.9954.

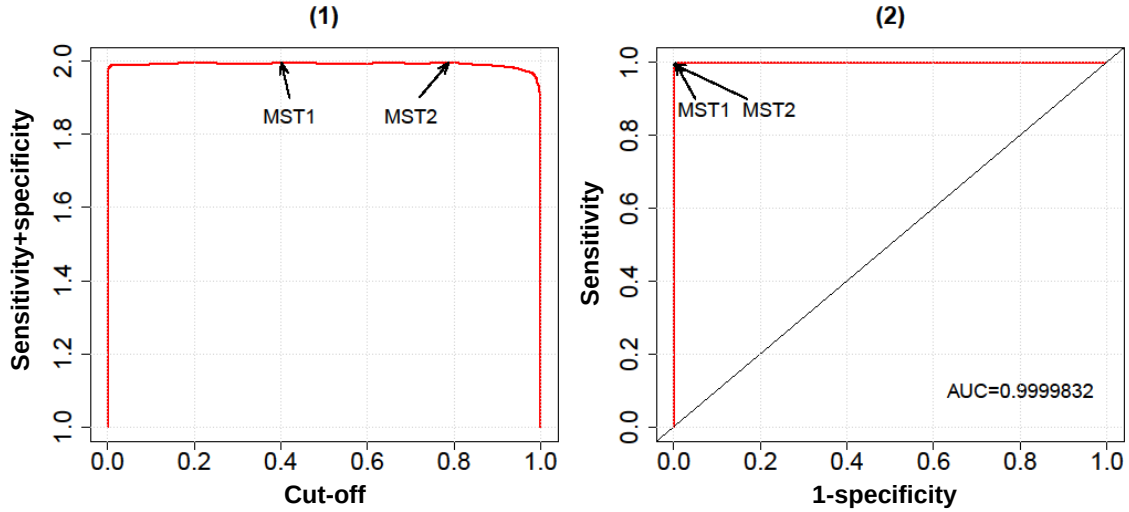


Figure 3.16 (logistf.model4) (1) The sum of sensitivity and specificity for different cut-off points. The points indicated by the two arrows on the plot represent the place where the sum of sensitivity and specificity is maximized. The values for the two MST values are MST1=0.4014 and MST2=0.7882, both resulting in a sum of 1.9954. MST1 results in a sensitivity of 0.9989 and a specificity of 0.9966 and MST2 in a sensitivity of 0.9954 and a specificity of 1. (2) ROC curve when plotting sensitivity against 1-specificity.

Model 'logistf.model5' also reported two MST values (MST1=0.2861, MST2=0.7911). The corresponding confusion matrices were

	observation	
prediction	1	0
1	879	4
0	0	875

with sensitivity=1 and specificity=0.9954 for MST1 and

	observation	
prediction	1	0
1	875	0
0	4	879

sensitivity=0.9954 and specificity=1 for MST2 resulting in a sum of 1.9954.

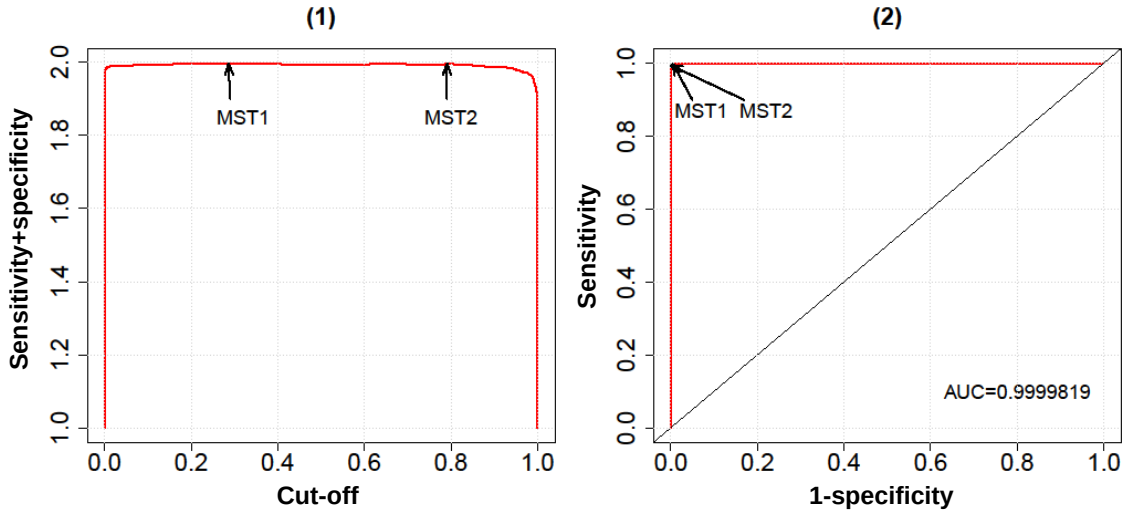


Figure 3.17 (logistf.model5) (1) The sum of sensitivity and specificity for different cut-off points. The points indicated by the two arrows on the plot represent the place where the sum of sensitivity and specificity is maximized. The values for the two MST values are MST1=0.2861 and MST2=0.7911, both resulting in a sum of 1.9954. MST1 results in a sensitivity of 1 and a specificity of 0.9954 and MST2 in a sensitivity of 0.9954 and a specificity of 1. (2) ROC curve when plotting sensitivity against 1-specificity.

3.3.4 Model evaluation using the testing data

Among the seven models, 'glm.model5', 'glm.model6.1', 'glm.model6.2', 'logistf.model2', 'logistf.model3', 'logistf.model4' and 'logistf.model5', the logistf model 'logistf.model2' produced the largest sum of sensitivity and specificity for the maximized sum threshold (MST) and the highest AUC associated with our training data. According to those two measures, 'logistf.model2' was deemed the best. Thus, we chose the combined evidence 'logistf.model2' model as our prediction model.

We evaluated the 'logistf.model2' model on the testing set with the MST from the previous subsection which gave the best sensitivity (0.999) and specificity (1) for detecting p53 binding sites in the training set. Using a probability threshold of 0.6088, we obatined the following

confusion matrix:

	observation	
prediction	1	0
1	873	6
0	5	872.

The values of 0.994 and 0.993 for the sensitivity and specificity were slightly less than those for the training set, but still high enough to produce a good prediction. We also plotted ROC curves based on the training and testing data sets (Figure 3.18). The AUC for the testing data was reported to be equal to AUC=0.9994 (Table 3.17).

Table 3.17 Comparison of performance of the combined evidence 'logistf.model2' model for the training and testing data. We used a probability threshold of 0.60879375 to distinguish between p53 and non-p53 binding sites.

Data set	Accuracy	Precision	Sensitivity	Specificity	AUC
Training set	0.9994	1	0.9989	1	0.9999
Testing set	0.9937	0.9932	0.9943	0.9932	0.9994

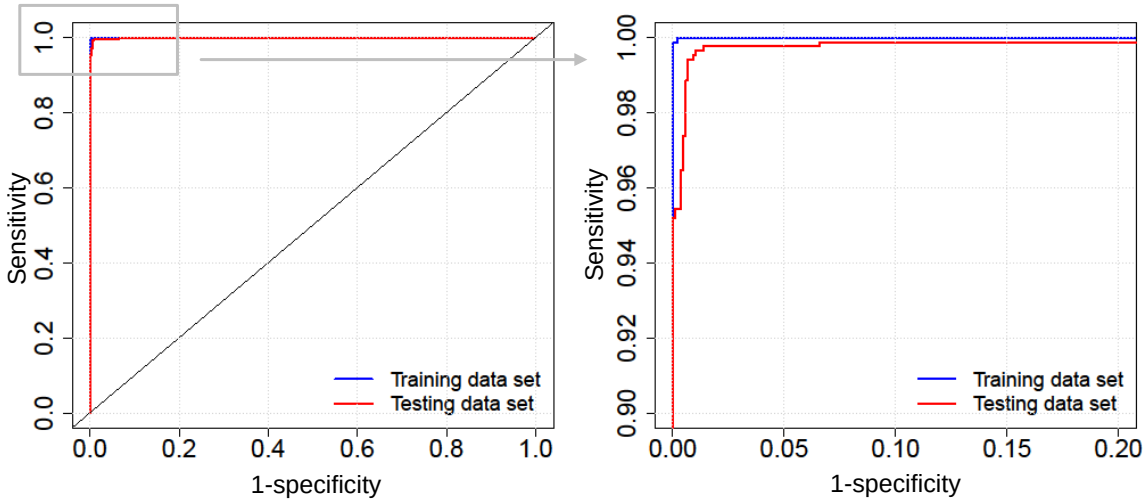


Figure 3.18 ROC curves of the combined evidence 'logistf.model2' model for training and testing data sets. The second graph on the right hand side is an enlarged version of the highlighted region in the first graph.

3.4 Discussion

Using the training data, the 'logistf.model2' model was the best performing model in terms of AUC, sensitivity and specificity. It is important to mention, however, that the differences in AUC and the other statistical measures were very minimal. The difference in AUC, for example, is roughly between one one-thousandth and one one-hundred-thousandth.

When we compare the 'glm' models with the 'logistf' models, the main difference between them is that the 'glm' models have to use a binary predictor variable for the pair score, while the same predictor in the 'logistf' models is treated as continuous. With independent variables that are all categorical in the model, only a limited choice of combinations exists for the set of predictor values. This also is the case for our 'glm' models despite the continuous *spacer.cont* variable which, however, is a value between zero and thirteen. 112 ($2 \times 14 \times 2 \times 2$) combinations are possible for 'glm.model5' and 56 ($2 \times 14 \times 2$) for the models 'glm.model6.1' and 'glm.model6.2', respectively. The fact that there is only a limited number of possible combinations may be a problem for our next genome-wide analysis. In large genomes like human, there will be many random sequences that match the p53 consensus binding motif. We need to select an optimal cut-off value (above which a sequence is considered as a p53 binding site) such that we can minimize the number of such false positives. Using our 'glm' models, the set of sites classified as p53 binding sites cannot be reduced any further when a particular cut-off point is reached. It is questionable whether that cut-off value will be strict enough to differentiate true p53 binding sites from random ones. The 'logistf' models therefore seem to be the better models for the genome-wide prediction of p53 binding sites.

Having a closer look at the best performing model 'logistf.model2', the corresponding logistic regression equation with the estimated coefficients is defined as

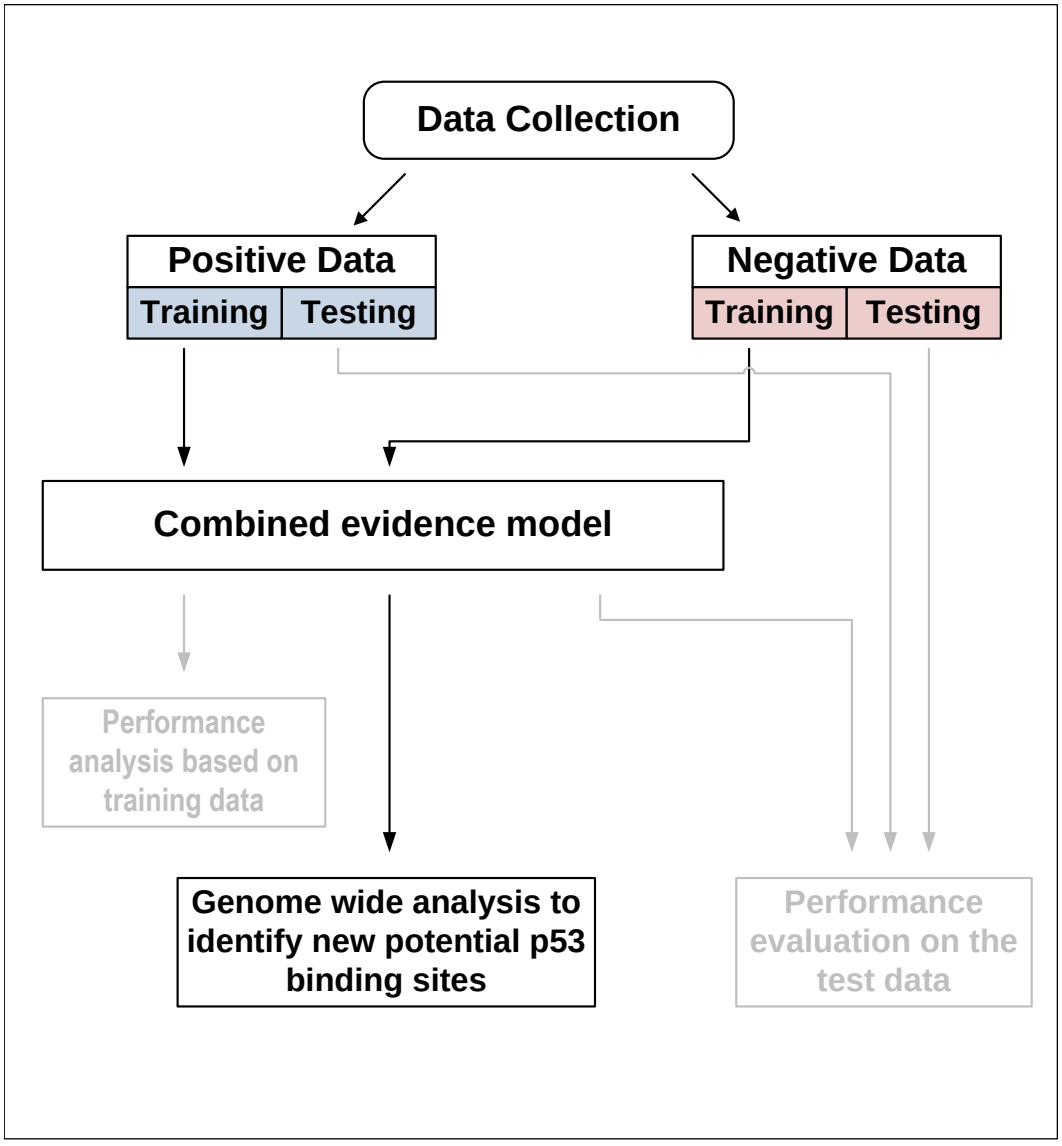
$$\begin{aligned}\mathsf{logit}[P(y=1)] &= 3.9932 + 0.8391[\mathsf{pair\_score.cont}] \\ &\quad -1.5315[\mathsf{spacer.cont}] + 0.1039[\mathsf{spacer.cont}^2] \\ &\quad +3.6782[\mathsf{in\_NhlfH3K4me1}] + 5.6133[\mathsf{in\_NhekH3K4me1}] \\ &\quad -7.0282[\mathsf{in\_NhekH3K4me3}] + 5.9039[\mathsf{in\_HuvecH3K4me3}].\end{aligned}$$

A negative coefficient for the *spacer.cont* seems to make sense, because the majority of our positive training sites had been reported to have no spacers between the two half-sites. According to the above equation, the existence of spacers will be penalized. Monomethylation of histone H3 lysine 4 is known to be associated with enhancers (Heintzman et al., 2009). Trimethylation, however, has been reported to be strongly associated with transcription start sites and not with p53 in any meaningful way. For this reason, the

coefficients for *in_NhlfH3K4me1* and *in_NhekH3K4me1* are positive, whilst *in_NhekH3K4me3* has a negative coefficient. The positive coefficient for *in_HuvecH3K4me3* is hardly to explain. It can be meaningful to have two different cell lines in our logistic regression model due to their different life histories. There will be some sites which are not accessible in a particular cell line. In that case, the methylation data in that cell line will only be reliable to some extent.

Chapter 4

Genome-wide prediction



Our combined evidence model showed a very good performance in accurately predicting human p53 binding sites using the testing data set. In order to test whether our model was capable of dealing with large and more complex data sets, we performed a genome-wide analysis by using human whole genome data. This chapter will demonstrate how well the model can deal with this challenging strategy and especially how well it can distinguish potential p53 binding sites from genome background. In this chapter, we will analyze the predicted binding sites in detail by examining their functional characteristics and compare them to predictions from other studies.

4.1 Introduction

We were interested in predicting p53 binding sites through genome-wide performance using our prediction model 'logistf.model2'. According to Smeenk et al. (2008), the number of p53 binding sites in the human genome is estimated to be between 300 and 3000. Hoh et al. (2002) identified about 300 p53 targets by using the p53MH algorithm which is based on discrete discriminant analyis. Cawley et al. (2004) performed ChIP-on-chip experiments to map the binding sites for p53 in vivo on human chromosomes 21 and 22. A total number of 48 high confidence p53 binding sites were observed along chromosomes 21 and 22 which would correspond approximately to 1600 binding sites when extrapolating to the whole genome.

Functionally important regions of the genome have been determined and predicted by experimental and computational approaches. Two general approaches have been undertaken to experimentally identify and validate p53 binding sites and potential p53 target genes (Cui et al., 2011). The first approach used by Riley et al. (2008) focuses on a specific target gene and checks whether the gene meets certain criteria that it needs to fulfill to be a potential p53 responsive gene. Riley et al. (2008) (1) tested the presence of a p53 RE in the DNA near or within the gene of interest, (2) searched for evidence that the gene was up- or down-regulated in response to wild-type p53, analyzed and validated the corresponding p53 RE (3) by using a luciferase reporter assay and (4) chromatin immuoprecipitation with a p53-specific antibody. This way, 129 genes and 160 p53 REs were identified which met at least three of the four criteria. The second experimental approach is genome-scale ChIP analysis. Wei et al. (2006) mapped p53 targets in the human genome by using chromatin immunoprecipitation with the paired-end ditag (ChIP-PET) and found 542 targets with high confidence of being involved in p53 interaction. Smeenk et al. (2008) identified

and characterized 1546 p53 binding sites using a genome-wide ChIP method which was performed in combination with DNA microarrays (ChIP-on-chip).

Modulation of the p53 transcriptional activity is mainly achieved by direct DNA-binding of p53 to their REs generally located within a few thousand base pairs of the target gene's transcriptional start site (TSS) (Horvath et al., 2007; Laptenko and Prives, 2006). Several studies using ChIP experiments have indicated that p53 binding sites also exist in intergenic (regions outside genes and promoters) and intragenic (regions within a gene) regions (Hearnes et al., 2005; Riley et al., 2008; Wei et al., 2006). Kaneshiro et al. (2007) reported that more than 80% of their detected p53 binding sites in the human ENCODE regions were intergenic (52%) or intragenic (29%) binding sites with 2% located in exons, 3% in first introns and 24% in other introns. More than half (60%) of the 1546 p53 binding sites identified by Smeenk et al. (2008) were mapped to intragenic or intergenic regions. Although many of the binding sites were located far away from the proximal promoter region of their target genes, Smeenk et al. (2008) showed that those sites could function as transcriptional enhancers. The long distance interaction between enhancer and target gene is mediated by DNA looping which brings distal transcription factor bound DNA binding sites close to the transcription start site to strongly affect transcription (Riley et al., 2008).

4.2 Methods

4.2.1 Applying our prediction model to the whole human genome data

Human genome data were downloaded from the Ensembl ftp site (Ensembl release 35, November 2005). For each possible pair of decameric sites in the human genome separated by a spacer of 0-13 bp, we estimated the probability of being a p53 binding site based on its observed values of the predictor variables.

Due to the large genome size, memory usage became a huge problem when running FIMO. Using Perl, we therefore had to split the genomic sequences into smaller sub-sequences of length 1 mega base pairs (Mb) with an overlapping area of 50 bp before we could run FIMO to compute the values for the *pair_score.cont* predictor. We scanned every sub-sequence with FIMO to determine the match scores of all possible decameric sequences to the two half-site motifs of the TRANSFAC M01651 matrix. The P value threshold was set at 1 to obtain the score for all possible decamers of the human genome. The score of a full site, as well as the values for the *in_NhlfH3K4me1*, *in_NhekH3K4me1*, *in_NhekH3K4me3* and the *in_HuvecH3K4me3* predictors were determined the same way as described in Chapter 3.

Our prediction model was finally used to estimate the probabilites of all possible pairs of decameric half-sites based on the observed predictor values.

4.2.2 Differentiating p53 binding sites from random sites

Given the set of all possible pairs of decameric sites in the human genome with their estimated probabilities, we were required to choose a cut-off value for the predicted outcome probability above which a site was classified as a p53 binding site. As mentioned in Section 4.1, there are between 300 and 3000 p53 binding sites estimated in the human genome. Based on this information we chose two optimal cut-off values that resulted in approximately 3000 and 300 sites, respectively.

4.3 Results

Two very stringent probability cut-off values were determined which gave the minimum and maximum estimated number of p53 binding sites in the human genome. The first cut-off value of 0.999999999998377 classified 2999 sites as potential p53 binding sites which might partially overlap with each other. The stricter cut-off value of 0.999999999999964 yielded 305 p53 binding sites.

4.3.1 Overlap with genome-wide ChIP data for p53

We compared our predictions with published ChIP-seq (Smeenk et al., 2008) and ChIP-PET (Wei et al., 2006) data. Of the 1545 binding targets identified by Smeenk et al. (2008), 300 (69) showed overlaps with any of our 2999 (305) predicted sites. A better result could be achieved with the ChIP-PET data from Wei et al. (2006). 129 (51) out of the 327 PET-3+ clusters (clusters with three or more overlapping DNA fragments) overlapped with any of our 2999 (305) predictions. In order to avoid "testing on training data", we distinguished the ChIP-PET data from training data and data that were not used in the training set. Among the 163 PET-3+ clusters that were not included in the training data set, 67 (25) were predicted by the combined evidence model. In particular, the three targets with high confidence of p53 interaction within the PET-11+ clusters, including the highest ranked target in the PET-18+ cluster, were all predicted by our model with the less stringent cut-off value for 2999 predictions (Table 4.1). The 100% overlap for the PET-11, PET-12 and PET-18 clusters, however, needs to be considered with care. The result would be unstable and would be challenged if more targets were present in the corresponding PET clusters.

Table 4.1 Overlapping results between our 2999/305 predictions and Wei's p53 targets that were not included in the training data set. A large amount of the high confidence targets that were identified by Wei et al. (2006) overlapped with our predicted sites. For the PET-10+ clusters (clusters with ten or more overlapping DNA fragments), for instance, the overlap to our 2999 predictions was approximately 0.80 (4 out of 5 were predicted).

Number of PETs per PET cluster	Number of PET clusters	Number of PET clusters not used for training	Combined evidence model-Comprehensive set of 2999 sites	Combined evidence model-Stringent set of 305 sites
3	158	83	27 (32.53%)	8 (9.64%)
4	63	31	14 (45.16%)	4 (12.90%)
5	37	18	6 (33.33%)	2 (11.11%)
6	28	13	8 (61.54%)	5 (38.46%)
7	13	6	5 (83.33%)	3 (50.00%)
8	10	5	2 (40.00%)	1 (20.00%)
9	7	2	1 (50.00%)	1 (50.00%)
10	5	2	1 (50.00%)	0 (0%)
11	1	1	1 (100.00%)	0 (0%)
12	2	1	1 (100.00%)	1 (100.00%)
13	1	0	0 (-)	0 (-)
16	1	0	0 (-)	0 (-)
18	1	1	1 (100.00%)	0 (0%)

4.3.2 Characteristics of the predicted p53 binding sites

We identified the localization of our predictions in the human genome relative to Ensembl genes the same way as described in Chapter 2. The binding sites were grouped into the six categories: intragenic (all introns and exons except the first exon and intron), TSS flanking (first intron, first exon and 5 kb upstream of TSS), 5 kb downstream (5 kb downstream of last exon), 5-25 kb downstream, 5-25 kb upstream and intergenic regions. A distinction between protein coding region and untranslated region (UTR) was not made. The protein coding region was considered as an intragenic region, whereas the UTR belonged to the TSS flanking region.

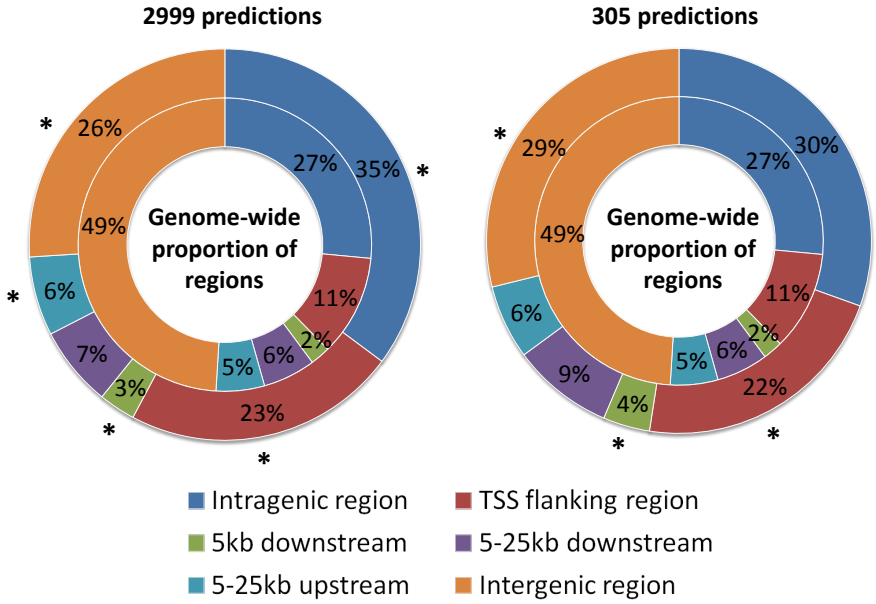


Figure 4.1 Distribution of the predicted p53 binding sites by 'logistf.model2' in intragenic, TSS flanking, 5 kb downstream, 5-25 kb downstream, 5-25 kb upstream and intergenic regions relative to Ensembl genes (outer ring) in comparison to the genome-wide proportions of the six regions of interest (inner ring). Significantly enriched or under-represented regions (G-test, P<0.05) are marked with an asterisk (*). Over-representation was observed among the 2999 binding sites in intragenic, TSS flanking, 5 kb downstream and 5-25 kb upstream regions. Under-represented binding sites were found in intergenic regions. Binding sites of the 305 predictions were statistically enriched in TSS flanking and 5 kb downstream regions and under-represented in intergenic regions.

The predicted p53 binding sites lay throughout the human genome. More than half of the 2999 (305) binding sites were found within or near a gene (Figure 4.1). Out of 2999 (305), 1055 (93) binding sites were in intragenic regions and 678 (67) sites were mapped to TSS flanking regions. 91 (12) of our predictions were located within a distance of 5 kb downstream of a gene, 199 (26) within 5-25 kb downstream, 196 (19) within 5-25 kb upstream regions and 780 (88) in intergenic regions.

A G-test analysis showed that the difference in proportions across the six regions between observed and expected sites was highly significant for the sets of 2999 (G=767.97, df=5, $P\approx 0$) and 305 (G=64.95, df=5, P=1.15$\times 10^{-12}$) predictions. Individual G-tests applied for testing each of the six regions identified several significantly enriched and under-represented

regions for the predictions. The p53 binding sites among the 2999 predictions were significantly over-represented in intragenic (G=96.32, df=1, $P\approx 0$), TSS flanking (G=328.14, df=1, $P\approx 0$), 5 kb downstream (G=14.15, df=1, P=1.68$\times 10^{-4}$) and 5-25 kb upstream (G=13.63, df=1, P=2.22$\times 10^{-4}$) regions and under-represented in intergenic (G=663.39, df=1, $P\approx 0$) regions. For the 305 binding sites, significant enrichment was observed for the TSS flanking (G=30.08, df=1, P=4.14$\times 10^{-8}$) and 5 kb downstream regions (G=4.56, df=1, P=0.03). Under-representation was reported for the intergenic regions (G=50.61, df=1, P=1.12$\times 10^{-12}$).

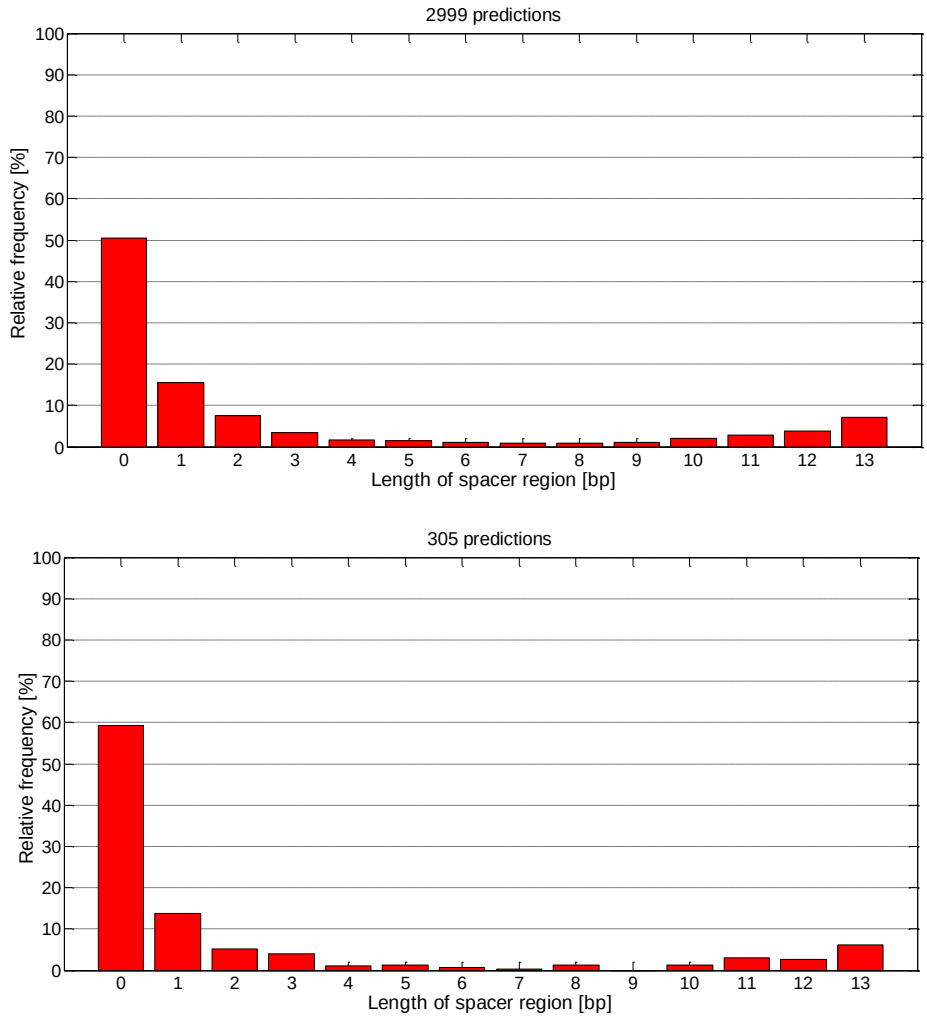


Figure 4.2 Spacer length distribution of the predicted p53 binding sites by the combined evidence 'logistf.model2' model based on logistic regression.

Investigating the length of spacers between the half-sites of the combined evidence predictions, we observed that more than half of the predicted sites had no spacer (Figure 4.2).

4.3.3 Functional annotation of the detected p53 binding sites

Using the Ensembl Perl API (Ensembl release 35, November 2005), the predicted p53 binding sites were assigned to their nearest genes as described in Chapter 2. The 2999 (305) predictions with their 3969 (545) unique genes were functionally analyzed using Gene Ontology (GO) categories. For given gene lists, we used DAVID to identify statistically enriched GO terms associated with the genes in the input lists.

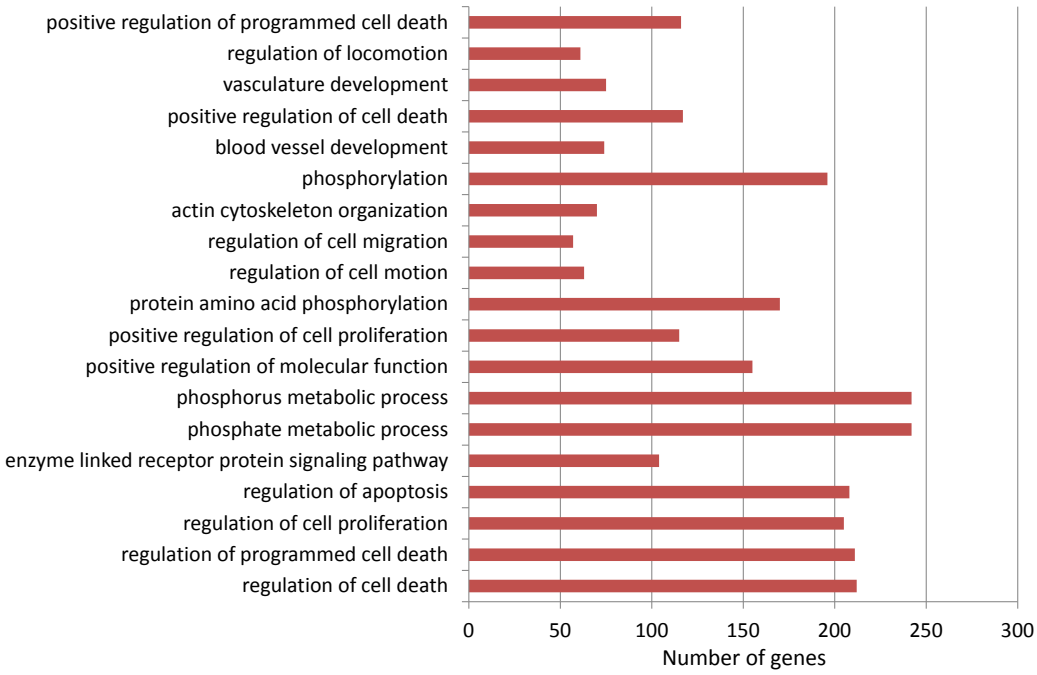


Figure 4.3 Biological process terms in the 'GO FAT' annotation category significant at $P<1\times10^{-5}$ for 2467 involved genes. The most statistically significant term is shown at the bottom.

Figure 4.3 shows the biological process GO terms for the 'GO FAT' annotation category which were strongly enriched in the gene list of the 2999 predictions. 2467 genes were associated with GO terms of biological process. We observed an enrichment of the biological GO terms 'regulation of apoptosis', 'regulation of cell death' and 'regulation of programmed cell death'. In addition, numerous metabolism related GO terms, such as 'phosphate metabolic process' and 'phosphorus metabolic process', were also found to be statistically enriched. Statistically significant GO molecular function terms associated with our gene list are presented in Figure 4.4. The number of genes involved in the molecular function category was 2383. The most significant group of molecular function was the 'protein kinase activity'.

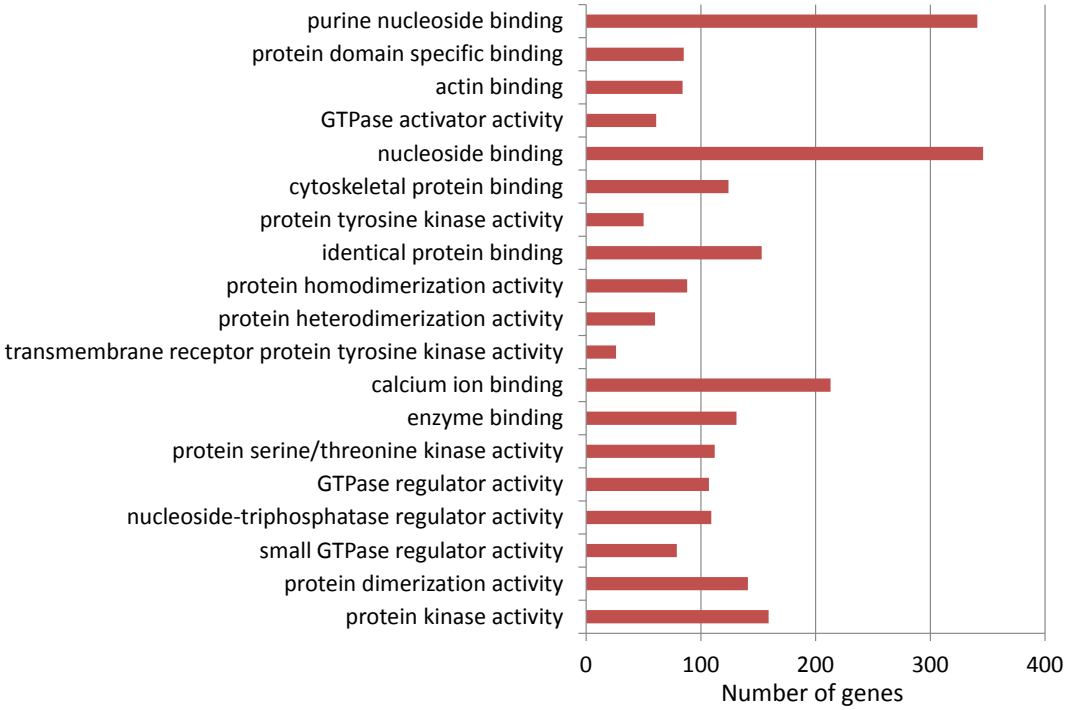


Figure 4.4 Molecular function terms in the 'GO FAT' annotation category significant at $P<1\times10^{-3}$ for 2383 involved genes.

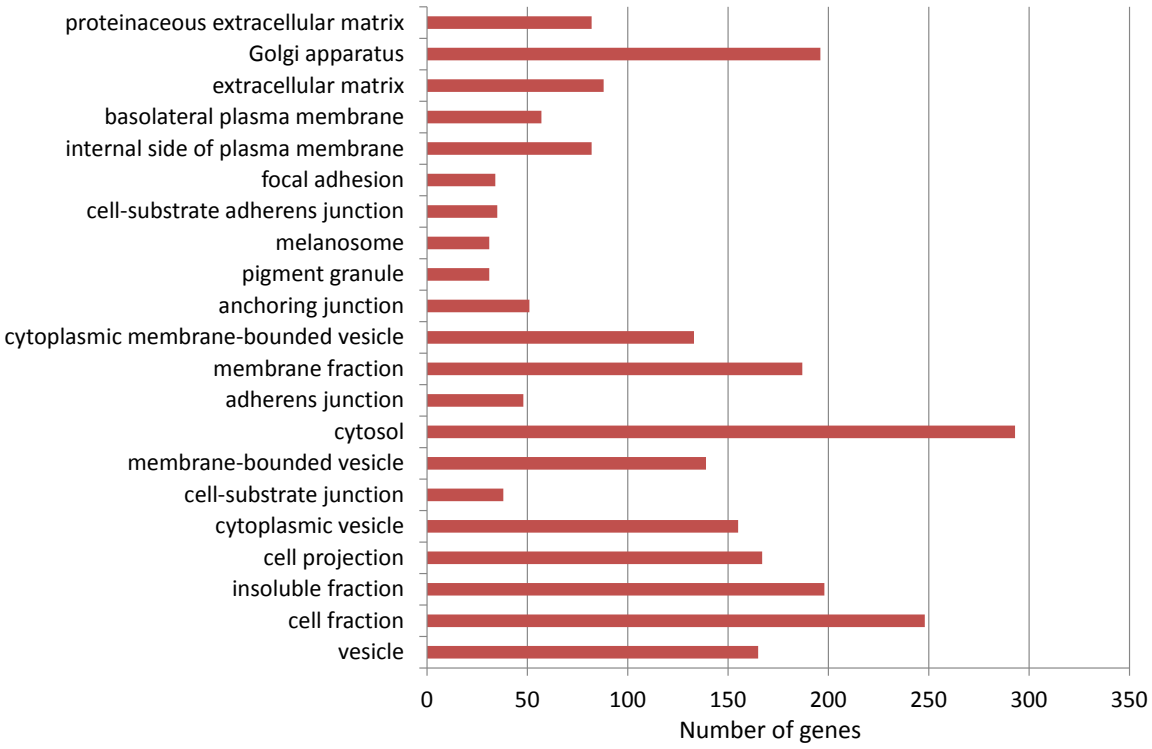


Figure 4.5 Cellular component terms in the 'GO FAT' annotation category significant at $P<1\times10^{-3}$ for 2324 involved genes.

These observations obtained from the statistically significant biological process and molecular function GO terms were in broad accordance with the results from our positive data set (Chapter 2). 2324 genes were associated with GO terms of cellular component. The most statistically significant cellular component term was 'vesicle' and the most numerous one was 'cytosol' (Figure 4.5).

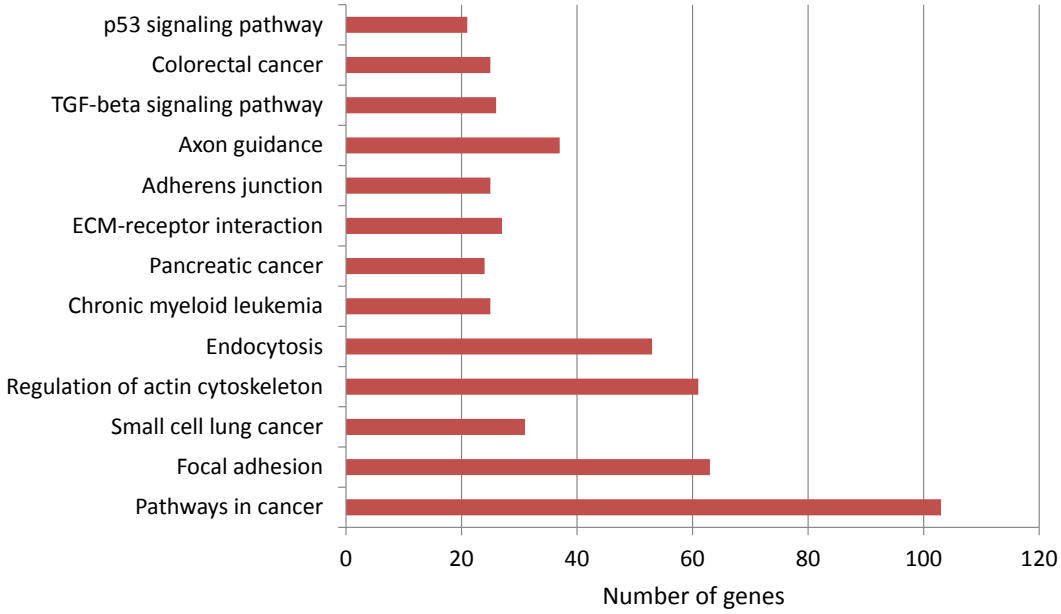


Figure 4.6 KEGG pathways significant at P<0.05 for 1037 involved genes.

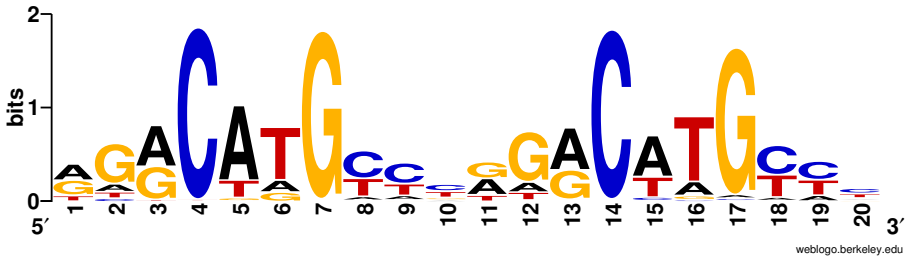


Figure 4.7 Sequence logo for the 2999 predicted p53 binding sites visualized using WebLogo.

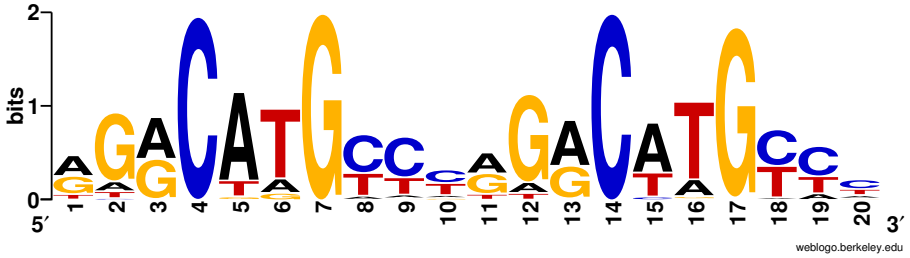


Figure 4.8 Sequence logo for the 305 predicted p53 binding sites visualized using WebLogo.

Performing an analysis for KEGG pathway participation of our 2999 predicted p53 binding

sites, many genes of them were significantly enriched in cancer-related pathways with the most significant KEGG pathway 'pathways in cancer' (Figure 4.6). The GO enrichment and KEGG pathway analyses of sets of genes of the 2999 and 305 (Subsection C.2.1) combined evidence predictions showed promising results.

Our combined evidence model seems to be reliable and able to predict high probability p53 binding sites. This finding can be confirmed by the two sequence logos presented in Figure 4.7 and Figure 4.8 which were generated with WebLogo (Crooks et al., 2004; Schneider and Stephens, 1990) (`http://weblogo.berkeley.edu/`). Both sequence logos resemble the consensus sequence for human p53 binding sites ([AG][AG][AG]C[AT][TA]G[TC][TC][TC] (El-Deiry et al., 1992)).

4.4 Discussion

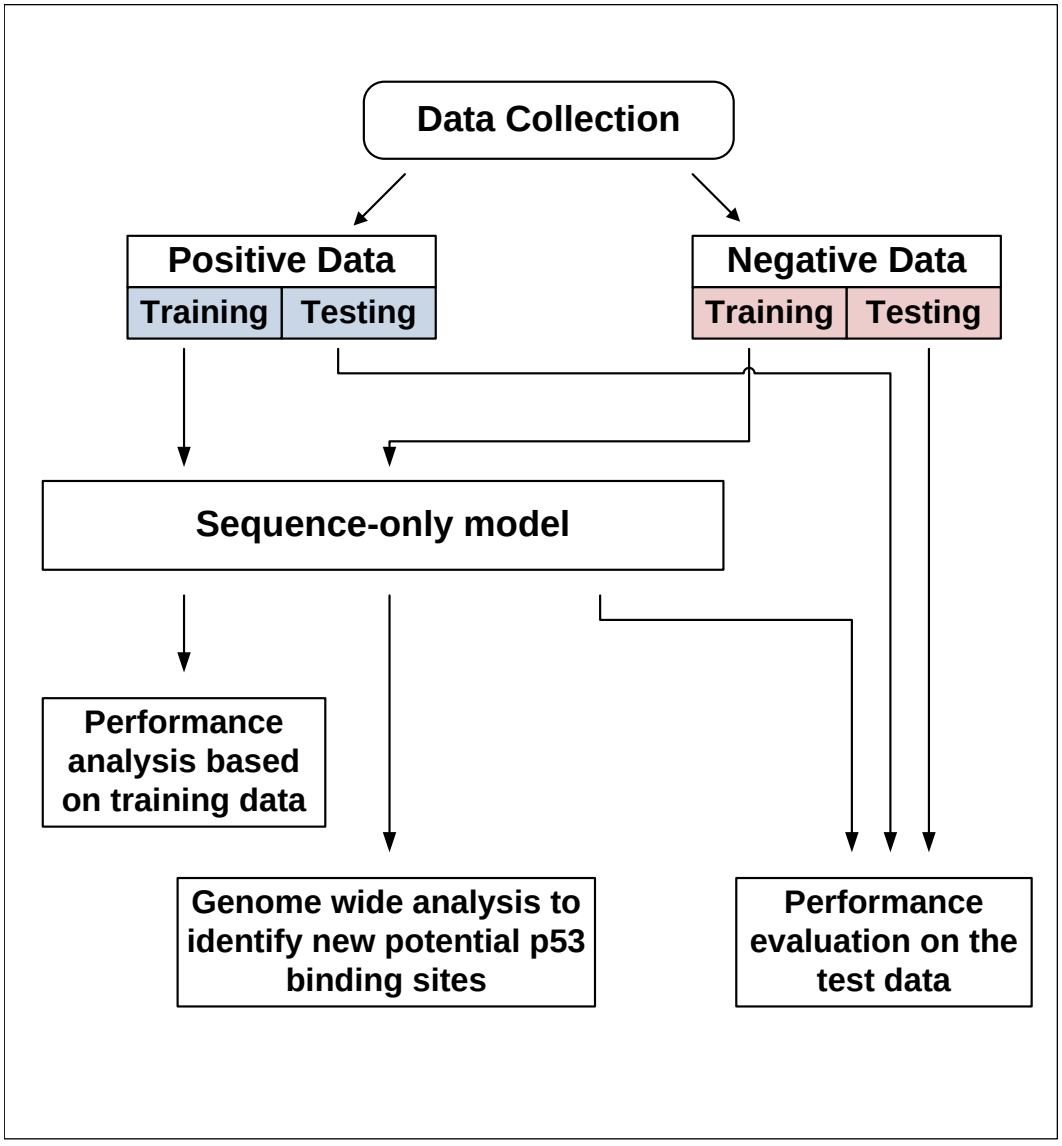
In order to make practical use of our model, we performed the combined evidence model to the entire human genome to predict p53 binding sites. Performance on a test data set (unseen data) that is limited to a few thousand sequences is not informative enough to fully ensure that the model has acceptable predictive power. A genome-wide application, however, is much more challenging and will help measure the performance of a model more accurately. A standard sequence-based discovery of functional transcription factor binding sites that performs well on bacteria and yeast sequence data, will have problems with sequences from higher eukaryotic systems, such as humans and mouse, due to their large genome size and would result in many false positive binding sites (MacIsaac and Fraenkel, 2006). A genome-wide analysis is therefore important to carry out and can also be used to identify novel binding sites.

We built a simple and easily interpretable model based on logistic regression which integrates sequence information and epigenetic information. Our prediction model was able to predict potential p53 binding sites from the whole human genome whose estimated probabilities were all above a stringent cut-off point. The characteristics of the predicted binding sites were mostly in accordance with previous studies of human p53 binding sites identified by ChIP experiments (Smeenk et al., 2008; Wei et al., 2006). Using our combined evidence model on the whole human genome, we were able to make a good prediction on human p53 binding sites. The GO enrichment and KEGG pathway analyses, as well as the resulting sequence logo showed promising results. Taking a detailed look at the two sequence logos in Figures 4.7 and 4.8, we observe that the central 'CWWG' nucleotides

known to be the most conserved positions within the half-sites of the human p53 binding sites (Riley et al., 2008) are clearly presented among the sets of our combined evidence predictions. Despite the promising findings, our computationally predicted binding sites do not necessarily represent actual p53 binding sites. Thus, our predictions require further experimental investigations to provide strong evidence that these are indeed binding sites bound by p53 in vivo. In addition, special considerations need to be taken into account which include the involvement of specific cellular factors known to interact with p53 by examining the flanking regions of the potential binding sites and chromatin state which have been shown to control access of transcription factors to their binding sites.

Chapter 5

Comparison with a sequence-only model



5.1 Introduction

The transcription function of p53 is mainly regulated through direct sequence-specific binding to the p53 response elements (REs) in DNA (Menendez et al., 2009; Riley et al., 2008). Most transcription factors have preferred sequence motifs on the DNA to which they bind. Many computational methods therefore predict transcription factor binding sites based on their regulatory motifs. There are different ways of representing such a motif. One simple way is the pattern-based model which uses a consensus sequence of preferred nucleotides. A motif M of length l is presented as

$$M = a_1 a_2 \dots a_l \; a_i \in \Sigma_{DNA} = \{A, C, G, T\} \; \forall i \in \{1, \dots, l\}.$$

Sinha and Tompa (2000) improved the simple pattern-based model by extending the alphabet Σ to symbols 'R' (purine), 'Y' (pyrimidine), 'S' (strong), 'W' (weak) and the wildcard character 'N' for spacers within the motif. A motif by Sinha and Tompa is, thus, a length-l sequence over $\Sigma_{DNA} = \{A, C, G, T, R, Y, S, W, N\}$. Some motif finding algorithms which use the pattern-based model are, for example, PROJECTION, PatternBranching, Voting, MITRA, MULTIPROFILER and cWINNOWER (Buhler and Tompa, 2002; Chin and Leung, 2005; Eskin and Pevzner, 2002; Keich and Pevzner, 2002; Liang, 2003; Rajasekaran, 2006).

The most common model of describing a motif is the matrix representation. The motif is represented by a position frequency matrix (PFM) or a position weight matrix (PWM), also known as a position-specific score matrix (PSSM), which contains log odds weights of every nucleotide a at each position in the motif of length l:

$$M = (a_{i,j})_{|\Sigma| x l} \; , 0 \leq i, j < l \; and \; a \in \Sigma.$$

The computational tools, such as MEME, GLAM, ProfileBranching, MITRA-PSSM and CONSENSUS are all based on probabilistic models to discover transcription factor binding sites (Bailey et al., 2009; Eskin, 2004; Frith et al., 2004; Hertz and Stormo, 1999; Price et al., 2003; Timothy and Elkan, 1995). The FIMO tool, for example, which is part of the MEME Suite software searches a biological sequence database for occurrences of motifs represented as probability matrices (in MEME format) provided by the user.

Further representations which are commonly used are regular grammars and tree data structures (Rigoutsos and Floratos, 1998; Sagot, 1998).

5.2 Methods

The sequence-only model scans sequences with a probabilistic matrix to score and predict p53 binding sites.

5.2.1 Scoring DNA sequences

We used FIMO (Grant et al., 2011) from the MEME Suite motif finding tools (Bailey et al., 2009) and the TRANSFAC p53 binding site motif (matrix accession M01651) to score DNA sequences. The two individual *decamer1_score.cont* and *decamer2_score.cont* scores for determining the total *pair_score.cont* score were calculated as described earlier in Chapter 3. The higher the total score, the more likely it was that a given site was a p53 binding site.

5.2.2 Process of training and testing

In the training step, every site in the training data was scored using FIMO. An optimal threshold score was specified that defined a p53 binding site based on the training set which gave the best performance results in terms of specificity and sensitivity.

In the testing step, we measured the performance of the sequence-only model on the basis of the testing sites using the threshold from the training process. If the total score of a testing site was greater than the threshold score, the testing site was predicted as a p53 binding site (Figure 5.1).

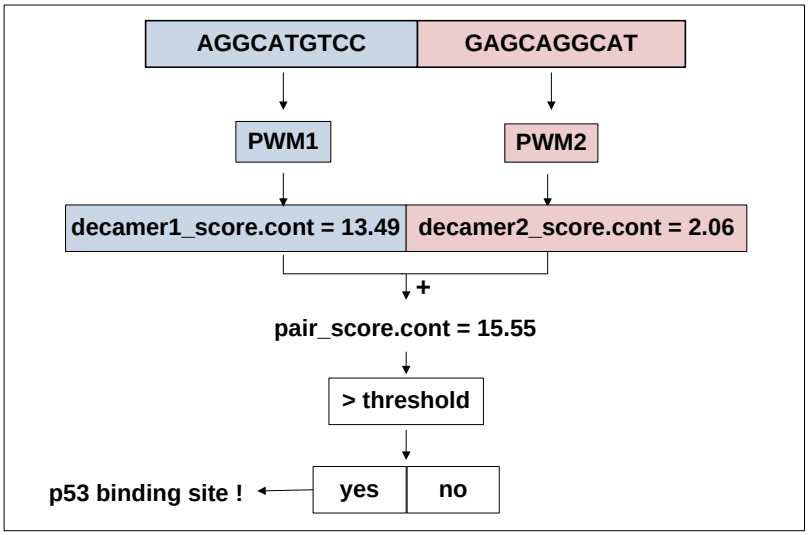


Figure 5.1 Predicting p53 binding sites using the sequence-only model for an example input 20-mer.

5.2.3 Applying the sequence-only model to the whole human genome data

To perform a genome-wide analyis, the sequences in the whole human genome were scanned. We identified sequences which were most likely to be p53 binding sites on the basis of the sequence-only model by considering and scoring every possible pair of decameric half-sites which were either directly adjacent or separated by a spacer region of length 1-13 bp using FIMO (see also Chapter 4). Given the set of all possible pairs of decamers from the human genome with their total scores, we selected two threshold scores that gave the minimum (300) and maximum (3000) number of estimated binding sites, respectively.

5.3 Results

5.3.1 Prediction accuracy using training and testing data sets

Figure 5.2 shows the two ROC curves of the sequence-only and combined evidence models along with a diagonal reference line based on the training sites. Since the two ROC curves show intersections, it is difficult to say which is better. A zoom in on the grey square rectangle in the plot on the left hand side, however, suggests that the combined evidence model is the better model in terms of its ability to correctly classify and predict observations. Indeed, the sequence-only model had an AUC of 0.9999573, which was slightly less than that of combined evidence model (AUC=0.9999974).

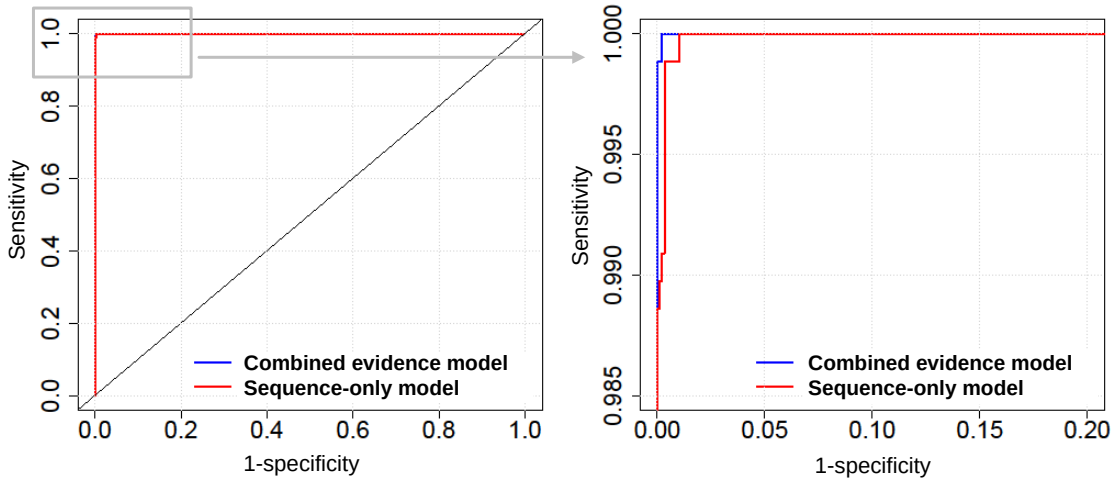


Figure 5.2 ROC curves of the sequence-only and the combined evidence models for the training data set. The second graph on the right hand side is an enlarged version of the highlighted region in the first graph.

The training data set was used to find the optimal threshold value for the scores (of the motif matches) of the binding sites. All sites with a score above this value would then be classified as p53 binding sites. One possible strategy for choosing the best suitable threshold is to determine the minimized difference threshold (*MDT*) which is the point on the curve shown in the first plot of Figure 5.3 where sensitivity and specificity are equal (see also Figure 3.2 in Chapter 3). We determined a value of *MDT*=-1.455 and both the sensitivity and specificity were equal to 0.997 at that point. Using the *MDT* as a score threshold, we obtained the following confusion matrix for the training data set.

	observation	
prediction	1	0
1	876	3
0	3	876

Another useful value is the maximized sum threshold (*MST*) representing the cut-off value that maximizes the sum of sensitivity and specificity (see Figure 3.2 of Chapter 3). The value of the score threshold *MST* was reported to be -3.873 with a sensitivity of 0.999 and a specificity of 0.997 (Figure 5.3). The corresponding confusion matrix for the sequence-only model and the training set was

	observation	
prediction	1	0
1	878	3
0	1	876 .

With the MST we could get a slightly better performance result in terms of sensitivity and specificity. We therefore chose the MST as the score cut-off.

Using a threshold of -3.873 the sequence-only model was evaluated on the testing set.

	observation	
prediction	1	0
1	875	4
0	3	874

The values of 0.997 and 0.995 for the sensitivity and specificity, respectively were less than those for the training set, but slightly better than those for the combined-evidence model on the basis of the testing set (Figure 5.1).

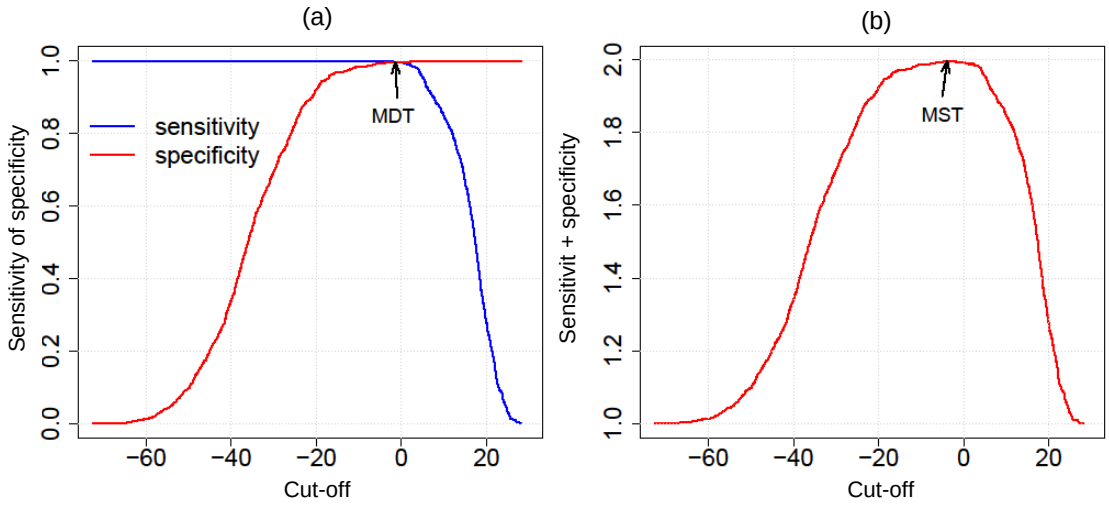


Figure 5.3 (a) The sensitivity and specificity for the sequence-only model using the training data. Both curves cross at *MDT*=-1.455 corresponding to a sensitivity and specificity of 0.997. (b) The sum of sensitivity and specificity for different cut-off points. The point indicated by the arrow on the plot represents the place where the sum of sensitivity and specificity is maximized (*MST*). *MST* has a score threshold of -3.870 corresponding to a sensitivity of 0.999 and a specificity of 0.997.

Table 5.1 Comparison of performance of the sequence-only and combined evidence models. We used a score threshold of -3.873 and a probability threshold of 0.610 for the sequence-only and combined evidence models, respectively.

Method	Accuracy	Precision	Sensitivity	Specificity
Training set				
Sequence-only model	0.998	0.997	0.999	0.997
Combined evidence model	0.999	1	0.999	1
Testing set				
Sequence-only model	0.996	0.995	0.997	0.995
Combined evidence model	0.994	0.993	0.994	0.993

5.3.2 Genome-wide prediction

For the genome-wide analysis, we redetermined the score threshold. Two stringent score cut-off values were determined which gave the minimum and maximum estimated number of p53 binding sites in the human genome as in Chapter 4 for the genome-wide predictions by our combined evidence model. The first cut-off value we selected for the motif match

score was 21.690 corresponding to 2998 binding sites. The second cut-off value of 24.758 classified 305 sites as potential p53 binding sites.

Overlap with genome-wide ChIP data for p53

The 2998 (305) predicted sites obtained from the sequence-only model were compared with published ChIP data (Smeenk et al., 2008; Wei et al., 2006). Of the 1545 p53 binding targets identified by Smeenk et al. (2008), 195 (29) overlapped with any of the 2998 (305) predicted sites.

Table 5.2 Overlapping results between the sequence-only/combined evidence predictions and Wei's p53 targets that were not included in the training data set. Fractions of non-training PET clusters predicted by the combined evidence and sequence-only models that were statistically significantly different from each other are marked with an asterisk (*) for the comprehensive sets and with a sharp (#) for the stringent sets.

Number of PETs per PET cluster	Number of PET clusters	Number of PET clusters not used for training	Combined evidence model - Comprehensive set of 2999 sites	Combined evidence model - Stringent set of 305 sites	Sequence-only model - Comprehensive set of 2998 sites	Sequence-only model - Stringent set of 305 sites
>2	327	163	67 (41.10%)	25 (15.34%)	62 (38.04%)	19 (11.66%)
>3	169	80	40 (50.00%)	17 (21.25%)	47 (58.75%)	17 (21.25%)
>4	106	49	26 (53.06%)	13 (26.53%)	30 (61.22%)	13 (26.53%)
>5	69	31	20 (64.52%)	11 (35.48%)	22 (70.97%)	10 (32.26%)
>6	41	18	12 (66.67%)	6 (33.33%)	13 (72.22%)	9 (50.00%)
>7 #	28	12	7 (58.33%)	3 (25.00%)	9 (75.00%)	7 (58.33%)
>8	18	7	5 (71.43%)	2 (28.57%)	4 (57.14%)	3 (42.86%)
>9	11	5	4 (80.00%)	1 (20.00%)	3 (60.00%)	3 (60.00%)
>10 *	6	3	3 (100.00%)	1 (33.33%)	1 (33.33%)	1 (33.33%)
>11	5	2	2 (100.00%)	1 (50.00%)	1 (50.00%)	1 (50.00%)
>12 *	3	1	1 (100.00%)	0 (0%)	0 (0%)	0 (0%)

Among Wei's PET-3+ clusters, 140 (43) sites out of 327 overlapped with our 2998 (305) sequence-only predictions. Of the 163 PET-3+ clusters that were not included in the training data set, 62 (19) were predicted by the sequence-only model (Table 5.2). In comparison,

more overlapping binding sites were observed for our predictions based on combined evidence (see Chapter 4). The number of Smeenk's targets which overlapped with our 2999 (305) predictions was 300 (69). The number of Wei's PET-3+ clusters overlapping our 2999 combined evidence predictions was 129 (51) for the entire data set and 67 (25) for the data set excluding PET clusters that were used for training.

To test whether the fraction of PET-3+, PET-4+, ..., PET-13+ clusters covered by a prediction that were not included in the training data set was the same for the combined evidence and sequence only predictions, we applied individual exact binomial tests using R. Significantly different fractions of non-training PET clusters that were identified by combined evidence and sequence-only predictions were observed for PET-11+ (exact binomial test, P=0.03704) and PET-13+ clusters (exact binomial test, P=2.2$\times 10^{-16}$) when using the less stringent cut-off value (comprehensive set of predictions). For the PET-11+ and PET-13+ clusters, the combined evidence model showed a better performance in predicting p53 targets compared to the sequence-only model. For the stringent set, the fraction of PET-8+ clusters predicted by the combined evidence model significantly differed from the fraction of PET-8+ clusters predicted by the sequence-only model (exact binomial test, P=0.03542). Significantly more PET-8+ clusters were predicted by the sequence-only model with the more stringent cut-off value.

Characteristics of the sequence-only predictions

When analyzing the locations of the seqeuence-only predictions relative to Ensembl genes, a large amount of the predictions were found to be located in intergenic regions (Figure 5.4). In detail, out of 2998 (305), 1254 (142) predicted sites were mapped to intergenic, 872 (71) to intragenic and 378 (30) to TSS flanking regions. 99 (11) sites were located in 5 kb downstream of a gene, 225 (30) and 170 (21) within 5-25 kb downstream and 5-25 kb upstream regions, respectively. We compared the observed counts across the six genomic regions with the counts expected by chance by using G-tests. The G-test analyses showed a very strongly significant difference between the observed and expected data for the 2998 predictions (G=76.84, df=5, P=3.89$\times 10^{-15}$) and a significant difference for the 305 predictions (G=13.78, df=5, P=0.02). Individual G-tests reported significant enrichment of the 2998 predicted binding sites in intragenic (G=4.52, df=1, P=0.03), TSS flanking (G=5.61, df=1, P=0.02) , 5 kb downstream (G=19.73, df=1, P=8.93$\times 10^{-6}$) and 5-25 kb downstream (G=9.20, df=1, P<0.01) regions. Significant under-representation was observed in intergenic regions (G=60.04, df=1, P=9.33$\times 10^{-15}$).

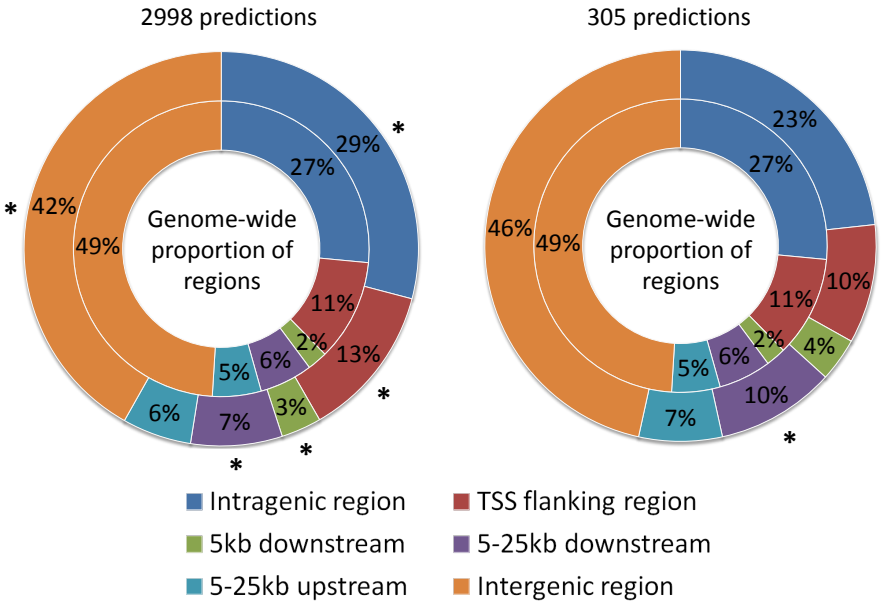


Figure 5.4 Distribution of the sequence-only predictions in intragenic, TSS flanking, 5 kb downstream, 5-25 kb downstream, 5-25 kb upstream and intergenic regions relative to Ensembl genes (outer ring) in comparison to the genome-wide proportions of the six regions of interest (inner ring). Significantly enriched or under-represented regions are marked with an asterisk (*).

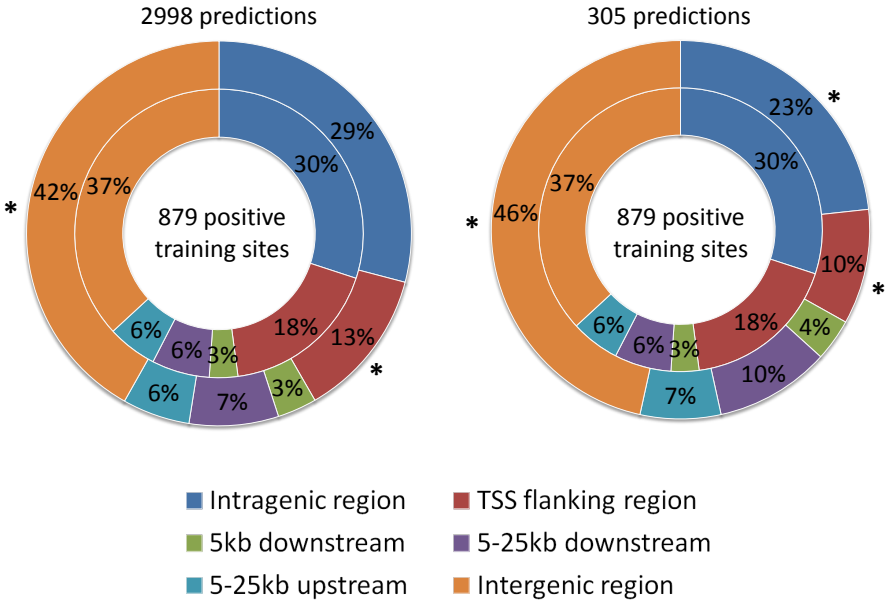


Figure 5.5 G-test of independence to compare the genomic location distributions between the positive training and sequence-only prediction sets. The genomic location distribution of the binding sites was different for the positive training and the sequence-only prediction data at the 5% significance level. Significantly different regions between the two data sets are marked with an asterisk (*).

For the 305 predictions, binding sites were only significantly enriched in 5-25 kb downstream regions (G=6.74, df=1, P=0.01).

We performed a G-test of independence to test whether the genomic location distribution of the binding sites was significantly different for the positive training set of 879 p53 binding sites and the sequence-only prediction set (Figure 5.5). The G-test showed statistically significant differences between the two data sets (2998 sites: G=18.19, df=5, P<0.01; 305 sites: G=23.04, df=5, P<0.001). Individual G-tests reported significant differences in TSS flanking (G=14.96, df=1, P<0.001) and intergenic regions (G=7.00, df=1, P<0.01) for the positive training sites and 2998 sequence-only predictions and in intragenic (G=5.22, df=1, P=0.02), TSS flanking (G=11.87, df=1, P<0.001) and intergenic regions (G=8.81, df=1, P<0.01) for the positive training and 305 prediction sets.

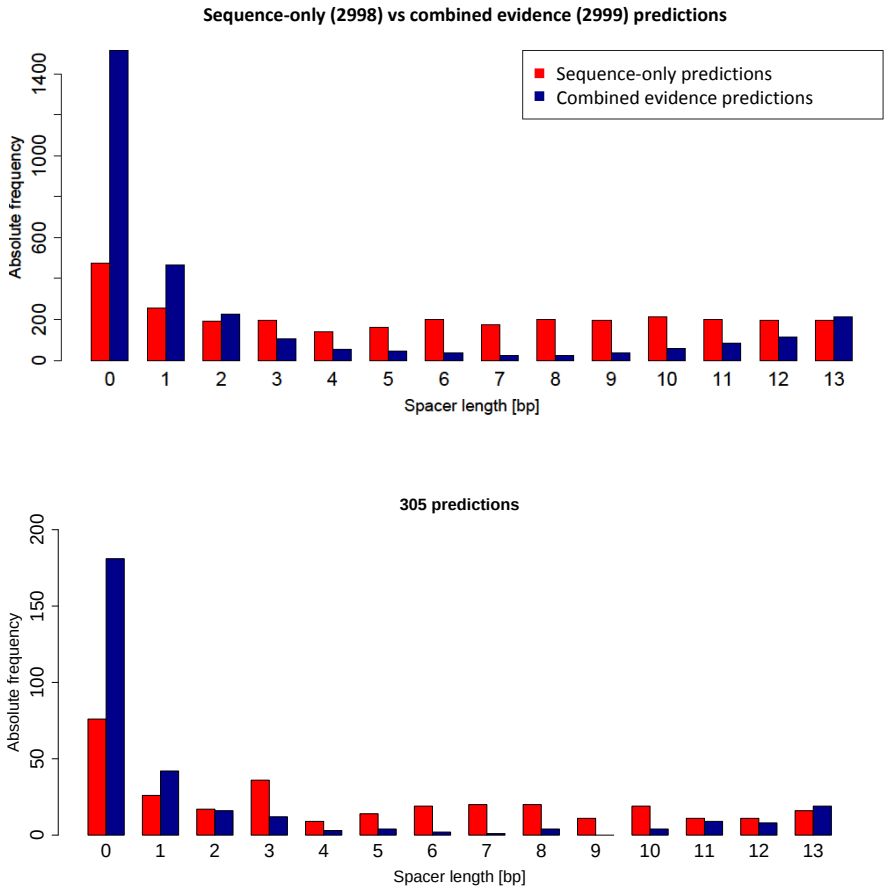


Figure 5.6 Spacer length distribution of the predictions based on the sequence-only model in comparison to the combined evidence predictions.

The binding sites predicted by only using sequence information showed a different distribution of spacer lengths compared to the combined evidence predictions. The sequence-

only predictions were not predominantly spacerless. Furthermore, the absolute counts in the different spacer length categories were relatively similar for the set of 2998 predictions (Figure 5.6).

Functional annotation of the sequence-only predictions

4618 (356) nearby genes were identified for the 2998 (305) binding sites predicted by the sequence-only model. To analyze GO and KEGG pathway enrichment, we used the functional annotation tool provided by DAVID. 3287 genes were involved in the GO enrichment analysis, of which 2256 were associated with GO terms of biological process, 2195 with GO terms of molecular function and 2122 genes were assicated with GO cellular component terms.

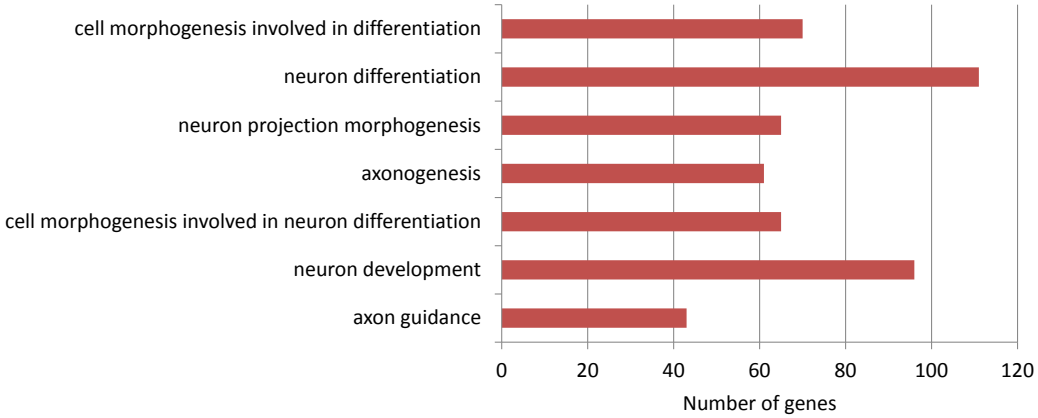


Figure 5.7 Significantly enriched ($P<1\times10^{-5}$) biological process terms in the 'GO FAT' for the 2998 sequence-only predictions.

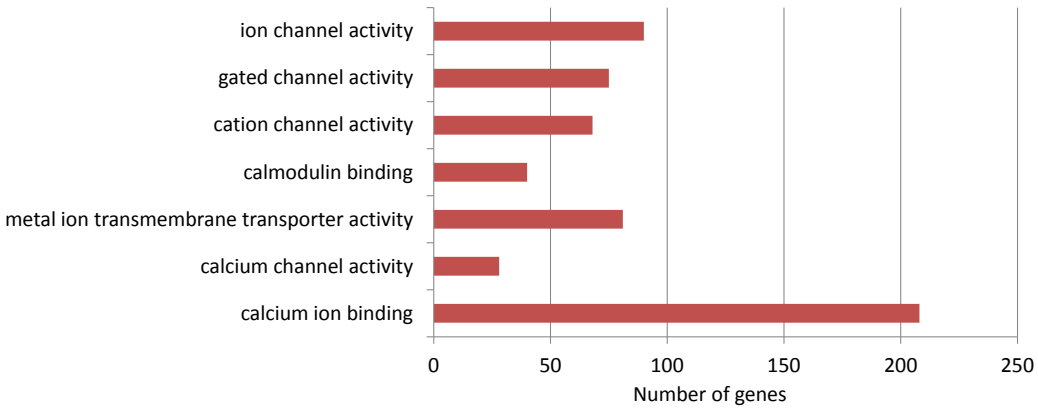


Figure 5.8 Significantly enriched ($P<1\times10^{-3}$) molecular function terms in the 'GO FAT' for the 2998 sequence-only predictions.

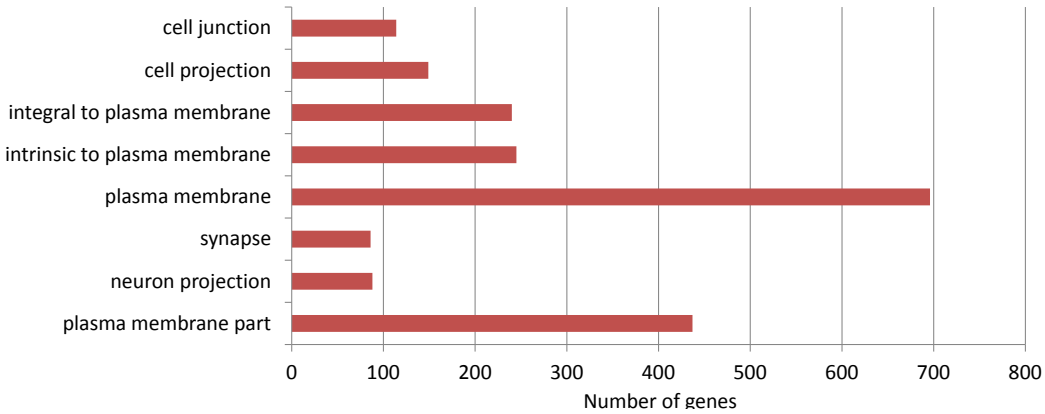


Figure 5.9 Significantly enriched (P<1×10^{-3}) cellular component terms in the 'GO FAT' for the 2998 sequence-only predictions.

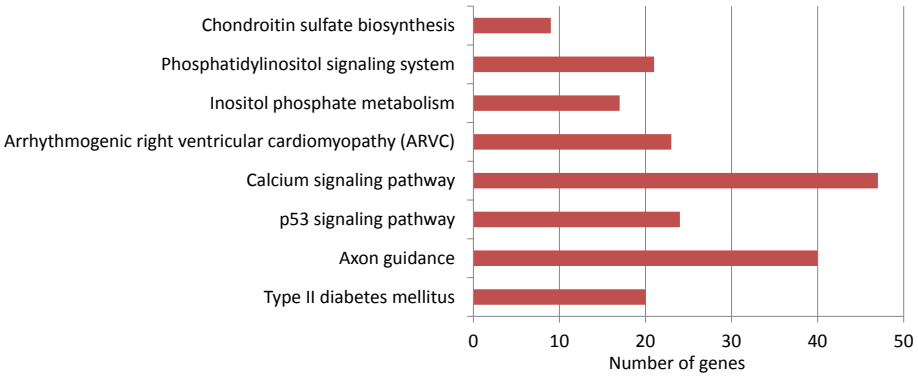


Figure 5.10 Significantly enriched (P<0.05) KEGG pathways for the 2998 sequence-only predictions.

Figure 5.7, Figure 5.8 and Figure 5.9 present highly statistically significant GO terms found in the biological process (P<1×10^{-5}), the molecular function (P<1×10^{-3}) and the cellular component (P<1×10^{-3}) 'GO FAT' annotation categories which were associated with our gene list of the 2998 predictions. Enriched GO terms for the 305 predictions are listed in Appendix C). The most statistically significant GO term of biological process was 'axon guidance' and the most numerous one was represented by 'neuron differentiation'. Unlike the positive p53 binding sites and the combined evidence predictions, we observed that here, the common biological process terms related to the p53 pathway, such as 'regulation of apoptosis', 'regulation of cell death' and 'regulation of programmed cell death', were not found among the top enriched GO terms. The 'calcium ion binding' was the most statististically significant and the most numerous GO molecular function term associated with the genes of the 2998 predictions. Interestingly, we did not observe the common GO

terms related to kinase activity among the top ranked terms. The most statistically significant and the most numerous GO cellular component terms were related to 'plasma membrane'. The KEGG pathway analysis revealed 'Type II diabetes mellitus' as the most statistically significant KEGG pathway and 'calcium signaling pathway' as the most numerous pathway.

Figure 5.11 and Figure 5.12 show the sequence logos for the predicted p53 binding sites, both generated from the sequence-only predictions. We can well recognize the core sequence 'CATG' pattern, which is not surprising, because the model we used to make our prediction was based on sequence information only.

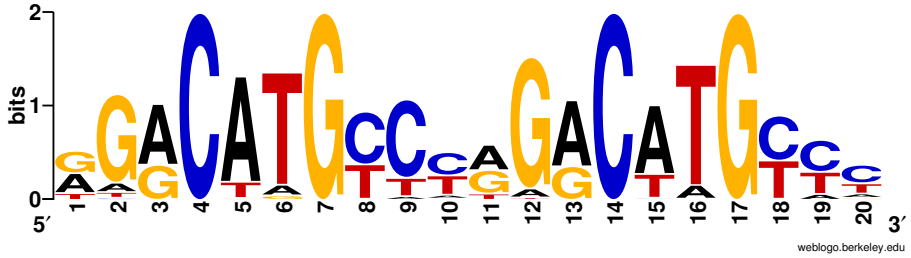


Figure 5.11 Sequence logo for the 2998 predicted p53 binding sites by the sequence-only model, visualized using WebLogo.

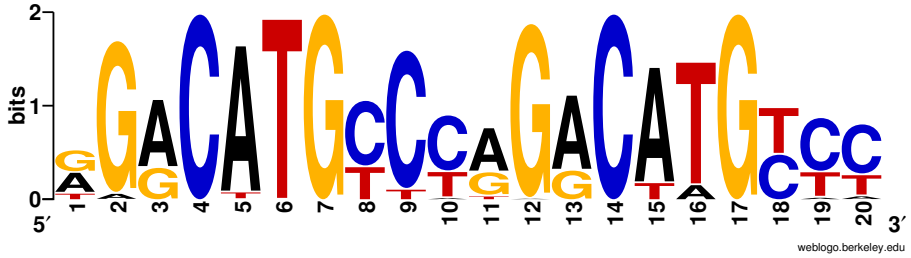


Figure 5.12 Sequence logo for the 305 predicted p53 binding sites by the sequence-only model, visualized using WebLogo.

5.4 Discussion

We observed a very good performance of the sequence-only model on the basis of the testing set. This was not surprising, but expected, because in Chapter 3, the *pair_score.cont* variable had already been reported to be an extremely good predictor for the discovery of p53 binding sites. The majority of the positive training sites were characterized by positive scores, while most of the negative training sites had negative scores (see Table 3.6).

To fully ensure that the sequence-only model had acceptable predictive power as the combined evidence model had, we used the simple model to examine the entire human genome. On complex systems such as humans, there is a large amount of uninformative background DNA which makes the motif discovery very difficult and challenging. We therefore expected a large number of false-positives with the sequence-only model. Indeed, the sequence-only model seemed to perform less successfully on the whole genome than on the training and testing data. Although the simple model was able to find some of the p53 binding sites which had been identified by Wei et al. (2006) and Smeenk et al. (2008), no overlapping was observed for the high-confidence Wei's targets in the PET-12+ clusters. Furthermore, the distribution of the spacer lengths in the sequence-only predicted sites was not what we expected to see. A majority of the predicted sites had spacers between their two half-sites which was not consistent with previous studies suggesting that most functional p53 binding sites had no spacer between the half-sites. The gene enrichment analyses using GO also failed to provide any clear confirmation that the predictions were indeed functional p53 binding sites.

Obviously, unlike the combined evidence model, the sequence-only model was less successful in the genome-wide prediction of potential p53 binding sites. The comparison analysis between the two models showed that using epigenetic information, such as Lys4 methylation of H3, improved the performance of sequence-only prediction. Epigenetic information is therefore an important and useful factor for the prediction of p53 binding sites.

Chapter 6

Discussion

With the remarkable progress in the development and improvement of ChIP techniques (ChIP-on-chip and ChIP-seq), significant findings have been made in the field of epigenetic research. Various ChIP experiments have provided strong evidence showing the important role of chromatin states, such as histone modifications (Barski et al., 2007; Guccione et al., 2006; Heintzman et al., 2009; Wang et al., 2008) and nucleosome positioning (He et al., 2010), in the regulation of gene expression. Post-translational modifications of histone tails, such as acetylation, methylation, phosphorylation and ubiquitination, can regulate transcription by causing structural and functional changes in chromatin to modulate the accessibility of DNA to regulatory proteins (Pawlak and Deckert, 2007).

In recent years, several strategies have been developed to identify transcriptional sequence elements on the basis of chromatin signatures (Cuellar-Partida et al., 2012; Heintzman et al., 2007; Shen et al., 2012; Won et al., 2010), but only a few of them were evaluated on p53 binding sites (Ernst et al., 2010). Promoter regions of active genes have been shown to be generally marked by histone acetylation and methylation state of the lysine 4 (Lys4) residue of histone H3 in mammalian systems (Barski et al., 2007; Bernstein et al., 2005; Kim et al., 2005; Roh et al., 2005, 2006; Wang et al., 2008). Heintzman et al. (2007) reported high levels of trimethylated H3K4 and depletion of H3K4 monomethylation in active promoters and found that enhancers were marked by monomethylated H3K4, but not trimethylation of H3K4. In contrast, dimethylation of H3K4 has been shown to be associated with enhancers by Bernstein et al. (2005) and Barski et al. (2007), whereas Barski et al. (2007) reported that enhancers were marked by all three methylation states (mono-, di- and trimethylation) of H3K4.

6.1 Summary of contributions

In this thesis, I present a computational model which integrates sequence information and epigenetic information to predict p53 binding sites in the human genome on the basis of multiple logistic regression. Using p53 binding sites from high-resolution ChIP data, such as ChIP-PET (Wei et al., 2006) and ChIP-on-chip (Smeenk et al., 2008), the model was trained to learn and identify the features of the binding sites which could explain the binding specificity of the p53 protein. Starting with an initial complex model based on logistic regression and multiple features, including mono-, di- and trimethylation states of H3K4 (H3K4me1, H3K4me2 and H3K4me3) from multiple cell lines, the overlap information with known enhancers and the PWM score of a binding site as well as the spacer length between the two half-sites within a binding site, the variables which seemed to have the largest effect on the prediction of p53 binding sites, individually or in combinations, were selected by the model selection procedure. Our final prediction model, which does not include the two initial feature variables H3K4me2 state and the overlap information with enhancers, produced very high degrees of sensitivity and specificity when testing on the test data, demonstrating a great level of prediction accuracy (Chapter 3). Furthermore, our model provided evidence showing the importance of mono- and trimethylations of H3K4 for the DNA binding of p53.

In contrast to many studies which test their computational models on data sets whose sizes are limited to thousands of sequences, we used our model to scan the whole human genome for potential binding sites specific for p53. The results obtained from the genome-wide analysis confirmed the ability of our model to predict p53 binding sites. The characteristics of the detected binding sites were consistent to a great extent with previous studies (Smeenk et al., 2008; Wei et al., 2006). The comparison analysis with a simple model, which only used the PWM score of a binding site for the identification, demonstrated that the epigenetic information we used in our logistic regression based model helped improve the prediction of p53 binding sites. The good performance of the sequence-only model on the testing set could not be confirmed by the genome-wide analysis (Chapter 5).

The importance of histone modification information for finding regions of likely transcription factor binding sites has also been stressed by Ernst et al. (2010). Their method, which combines the so-called general binding preference (GBP) score resulted from a logistic regression based model with 29 features for the transcription factor binding prediction with a PWM based search, is presented in detail in Chapter 1. Of the 29 features, the one

combining 20 histone modification levels has been shown to be the most informative feature. The GBP score demonstrated a powerful approach for the prediction of transcription factor binding sites. Only one (Wei et al., 2006) of the 14 tested data sets with experimentally derived binding sites of various transcription factors showed a bad performance. The GBP failed to correctly predict p53 binding sites (AUC = 0.57). However, in combination with the PWM score a better performance result could be achieved (AUC = 0.80). Comparing their AUC value of 0.80 with that of our prediction model when using the testing set (AUC = 0.99), our model obtained a better performance accuracy on the basis of AUC.

6.2 Future directions

Based on the obtained results, our model offers a promising start in gaining insights into how the tumour suppressor p53 protein recognizes and binds to specific DNA sequences. It is important to note that our computationally predicted p53 binding sites described in Chapter 4 do not necessarily represent actual p53 binding sites. Further experimental investigations are required to test the binding specificity of the predicted binding sites. The first step should be to test by ChIP for the sequences in question.

DNA binding is the starting point to study gene expression. To know how genes are expressed and regulated, multiple factors have to be explored, including several cis-acting elements, such as the regulatory DNA sequence motifs or response elements, and trans-acting factors, such as activators, cofactors, transcription factors, protein-protein interactions, post-translational modifications and epigenetic factors (chromatin environment). To incorporate epigenetic information into our prediction, we used the histone mono- and trimethylation patterns of H3K4. In addition to histone methylations, H3 lysine 27 acetylation (H3K27ac) is becoming an increasingly important factor for identifying enhancers. In recent years, several studies have determined the genomic locations of the histone modification signal H3K27ac in mammalian systems (Creyghton et al., 2010; Heintzman et al., 2009; Rada-Iglesias et al., 2011; Shen et al., 2012). H3K27ac has been shown to be a useful mark for enhancer regions, especially for distinguishing between active and inactive enhancers which are only marked by H3K4me1 (Creyghton et al., 2010). We expect that more patterns of histone modification, which are associated with enhancer regions, such as H3K27ac, will help further improve our prediction model. In particular for the genome-wide analysis, better and more accurate performance is expected to be obtained with our model if H3K27ac signals are incorporated. Since the p53 protein is known to interact with

various transcription factors and mediate the recruitment of additional factors required for p53-dependent transcriptional activation, it will be interesting to examine the flanking regions of p53 to find the binding sites of the p53-related proteins. Smeenk et al. (2008) identified the response elements of eight different transcription factors (Krüppel-like factors (KLF), Sp1/Sp3, the group of basic helix-loop-helix (bHLH) proteins, AP1, AP2, MZF1, CP2 and ETS2) which were significantly enriched in the regions surrounding their predicted p53 binding sites. Many of these transcription factors are known to affect p53 stability and activity. A further improvement may be achieved by using comparative genomics which represents a useful way of distinguishing functional binding sites (true positives) from false binding sites (false positives) across a whole genome. The resulting p53 binding sites predicted by a model, which takes into account all the factors known to influence p53 function, can then be analyzed in more detail. An interesting task would be to discover restricted sets of p53-related genes which contribute to a specific p53-induced response, such as apoptosis, DNA repair, cell-cycle arrest and senescence, which in turn may be stress- and cell-specific.

A high-performing model for predicting p53 binding sites can then be generalized to any transcription factor with the correct PWM for the specific sequence motif. With the improved (but not perfect) model which integrates all the information and data necessary for a good prediction of transcription factor binding sites, we will yield a better understanding of the process by which a transcription factor recognizes its DNA binding sites and gain new insights into the complex regulatory mechanisms mediated by the interaction of various transcription factors with their DNA binding sequences.

Appendix A

Cluster analysis of decameric half-sites based on sequence similarity

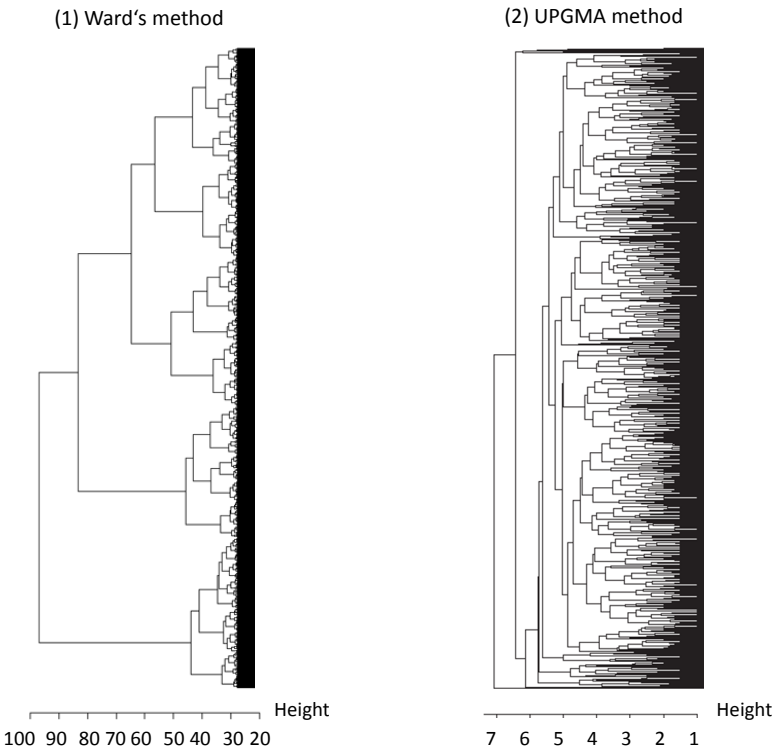


Figure A.1 Dendrograms obtained by cluster analyses of the 1688 unique decamers of the 1757 p53 binding sites using the Hamming distance with (1) Ward's method and (2) UPGMA. The dendrogram lists all decamers and reports at what level of similarity any two clusters were joined. The horizontal axis (x-axis) shows the fusion level, i.e. the similarity measure at which clusters were merged. The vertical axis (y-axis) specifies the decameric half-site samples and shows how the different clusters are formed.

Clustering is a commonly used method which finds structure in a data by grouping similar objects into classes or categories. Dissimilar objects are put into different clusters (Fielding, 2007).

We performed clustering on the 1688 unique decameric half-sites found in our 1757 positive binding sites to find meaningful sequence patterns which might be biologically important. In addition, we were interested in finding out whether the pairings of half-sites in the p53 binding sites were not arbitary. We used the Hamming distance and hierarchical clustering to cluster the decamers into subgroups of sites which were similar in sequence using the 'hclust' function in R (R version 2.14.1).

The Hamming distance (Hamming, 1950) is a distance function suitable for categorical data. For binary data, the Hamming distance is defined as the number of different bits between two binary vectors (Tan et al., 2006). It is similar to the simple matching coefficient which can be calculated by dividing the Hamming distance (number of different bits) by the number of bits. The Hamming distance is also often used for analyzing similarity of biological sequences (Federico and Pisanti, 2009). The Hamming distance between two sequences of equal length is the number of positions that differ.

Definition 9. (Hamming distance) *Given two sequences s_1 and s_2 of equal length l, the Hamming distance $d_H : \Sigma^l \times \Sigma^l \mapsto \mathbb{N}_0$ between those two sequences is the number of positions for which the corresponding characters are not equal (Gogol-Döring and Reinert, 2009):*

$$d_H(s_1, s_2) := |\{i \in \{1, ..., l\} \mid s_1[i] \neq s_2[i]\}|.$$

A.1 Clustering using Hamming distance and Ward's method

The results of the cluster analysis on the 1688 unique decamers using the Hamming distance measure and the Ward's (Ward, 1963) and UPGMA (Sokal and Michener, 1958) methods are shown by a dendrogram (Figure A.1). Here, we will only present the results obtained by the cluster analysis with the Ward's method, because the UPGMA method generated highly unbalanced groups which gave non-significant results. In theory, the Ward's method assumes a Euclidean space. The Hamming distance, however, is a distance measure for non-Euclidean spaces. Nevertheless, a number of studies have carried out a hierarchial cluster analysis using Ward's method with Hamming distance as the distance measure (Poage et al., 2010). We transformed our distance matrix to a Euclidean matrix by using the 'lingoes' function from the 'ade4' package in R as suggested in the R help mailing list archive (`https://stat.ethz.ch/pipermail/r-help/2008-September/173843.html`).

When drawing a line vertically across the dendrogram to divide the data into an appropriate set of reasonably distinct clusters, the first intersection passes through two horizontal lines, suggesting the presence of two major clusters (k=2). The next grouping would result in three major clusters (k=3). Further divisions are possible, but would result in clusters which are less unique. Here, we present a cluster solution with two distinct clusters (k=2) which produced significant results.

Table A.1 Ward cluster group composition 2 distinct clusters (k=2).

Cluster	Number of decamers
1	1290
2	398

A.1.1 Cluster solution with two different clusters (k=2)

In our cluster analysis, the decameric sites were placed in two clusters. Cluster 1 was the larger cluster with 1290 and cluster 2 the smaller one with 398 unique members (Table A.1). Among the 1757 positive binding sites, the majority were composed of half-sites from cluster 1 (Table A.2).

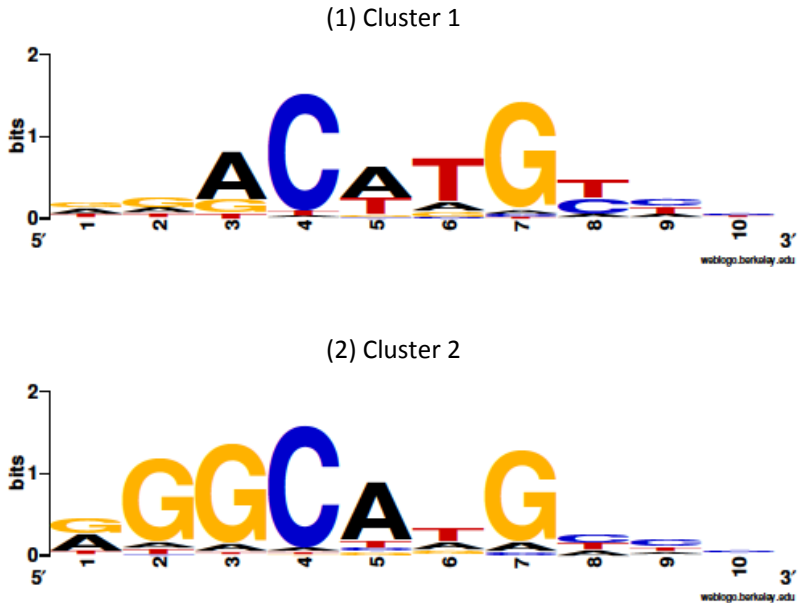


Figure A.2 Sequence logo for (1) cluster 1 containing 1290 unique decamers and (2) cluster 2 containing 398 unique decamers.

Table A.2 Observed (O) and expected (E) counts of cluster pairings within the p53 binding sites formed by Ward's method with k=2.

Pairing	O	E	O/E	$O\ln(O/E)$
cl11	1036.00	1004.50	1.03	31.99
cl12	282.00	324.00	0.87	-39.15
cl21	303.00	324.00	0.94	-20.30
cl22	136.00	104.50	1.30	35.83
Total	1757.00	1757.00		$\ln L$=8.36

To test whether the observed pairings of clusters in the data set departed from null (random) expectations, we used a G-test which was based on the ratio of observed to expected counts. For the expected counts of cluster pairings, we determined the expected counts of each cluster group on the basis of our 1757 positive sites (intrinsic hypothesis).

Definition 10. (Expected frequency of a cluster) *Given a data set of n p53 binding sites, the total number of decameric half-sites in the data set is equal to* $2n$*. If there are k distinct clusters, the expected frequency of a specific cluster x is*

$$\tilde{f_x} = \frac{n_x}{2n}, x \in \{1, ..., k\},$$

where n_x *is the number of cluster x occurred in the data set.*

Definition 11. (Expected count of a cluster pair) *The following calculation is based on the product rule of probability. Let n be the number of binding sites (cluster pairs) in a data set. The expected count of a cluster pair is*

$$E_{ij} = \tilde{f_i}\tilde{f_j}n,$$

where $\tilde{f_i}$ *and* $\tilde{f_j}$ *are the expected frequencies of clusters i and j, respectively.*

The G-test for goodness of fit investigating whether the observed data differed from the expected values resulted in G=16.72 adjusted by William's correction. Comparing our result with a χ^2-distribution with df=3 degrees of freedom, we found that the observed value of G from our sample was statistically significant (P<0.001). Thus, the observed values for cluster pairings showed significant deviation from the expected ones. Applying individual G-tests for the four categories (*cl*11, *cl*12, *cl*21, *cl*22), p53 binding sites were found to be significantly

enriched in *cl*12 (G=6.91, df=1, P<0.01) and *cl*22 (G=9.26, df=1, P<0.01).

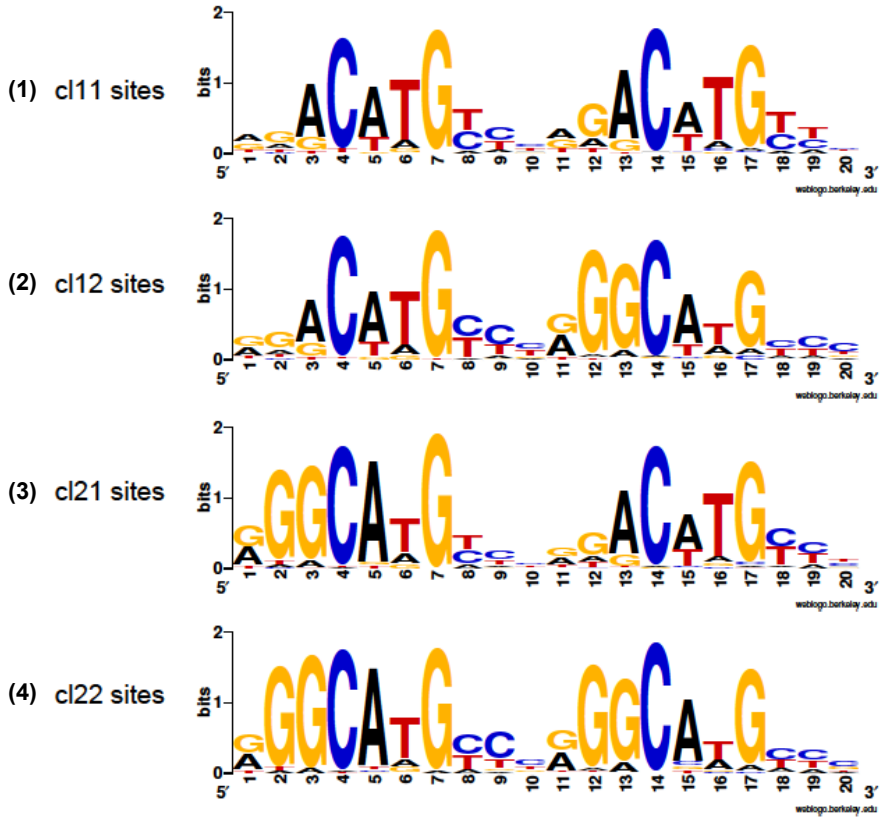


Figure A.3 Sequence logos for (1) 1036 p53 binding sites consisting of half-sites, both from cluster 1, (2) 282 p53 binding sites consisting of pairs of clusters 1 and 2, (3) 303 p53 binding sites consisting of pairs of clusters 2 and 1 and (4) 136 p53 binding sites consisting of half-sites, both from cluster 2.

GO enrichment analyses using DAVID revealed that the p53 binding sites representing *cl*22 sites were mainly related to cell death processes and apoptosis (Figure A.7). The most statistically significant GO term in the biological process FAT annotation category was 'induction of apoptosis' and the most numerous terms were 'regulation of cell death', 'regulation of programmed cell death' and 'regulation of apoptosis'.

In contrast, the p53 binding sites from the *cl*12 category seemed to be mainly involved in DNA damage response/checkpoint processes which do not lead to cell death (Figure A.5). Among the biological GO terms, we found that 'DNA repair', 'regulation of cell cycle', 'cell cycle checkpoint' and many other GO terms related to metabolism were strongly enriched.

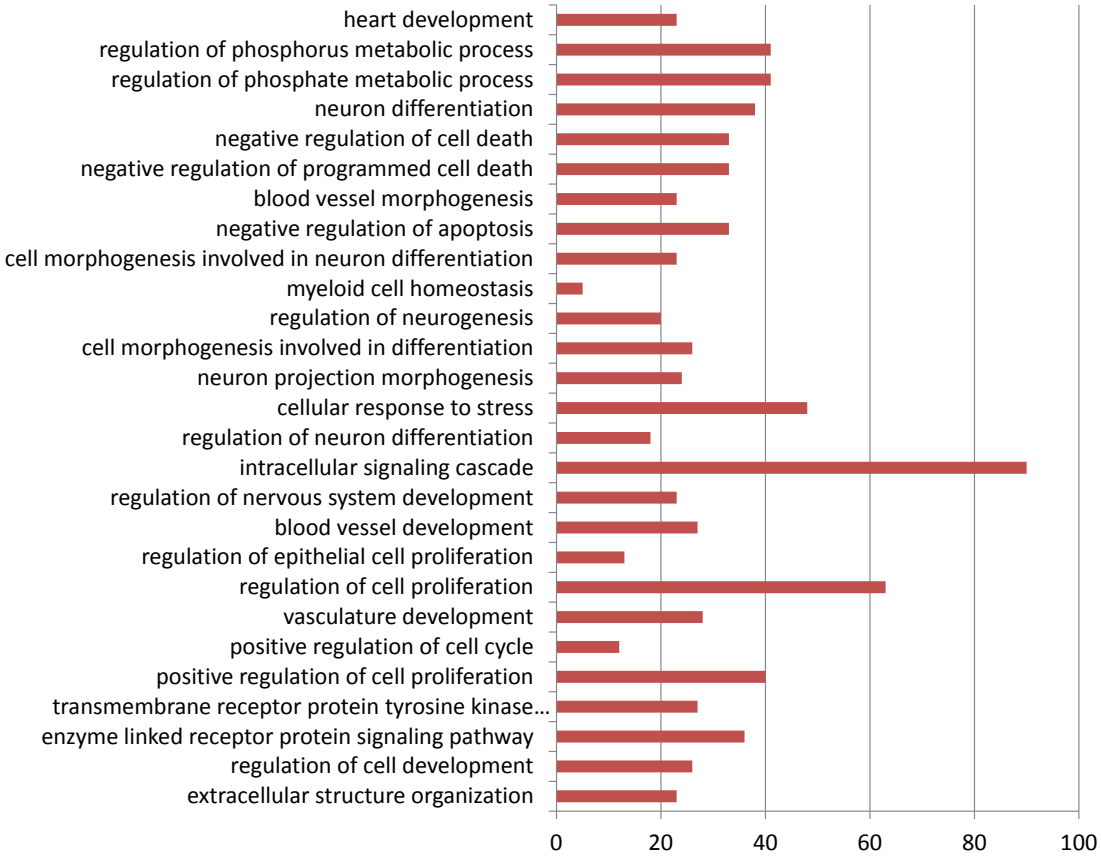


Figure A.4 Biological process terms in the 'GO FAT' annotation category found to be highly significantly enriched (P<0.001) in our list of 663 (out of 938) genes of the 1036 cl11 binding sites. The horizontal axis (x-axis) shows the gene counts (number of genes involved in the analyzed GO FAT annotation category).

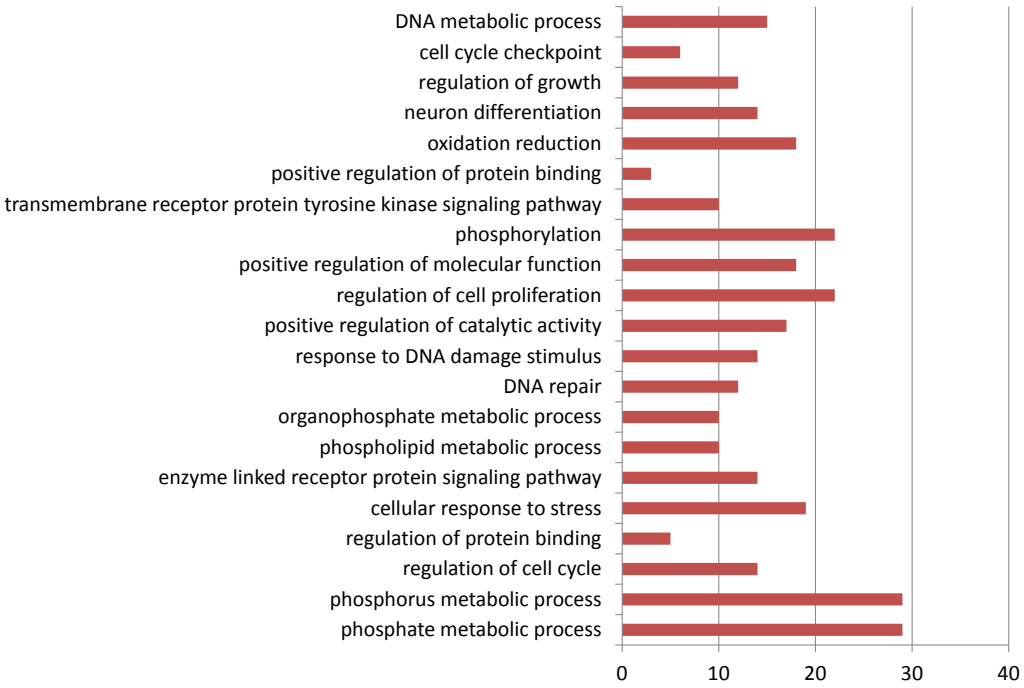


Figure A.5 Biological process terms in the 'GO FAT' annotation category found to be highly significantly enriched (P<0.01) in our list of 183 (out of 272) genes of the 282 cl12 binding sites. The horizontal axis (x-axis) shows the gene counts (number of genes involved in the analyzed GO FAT annotation category).

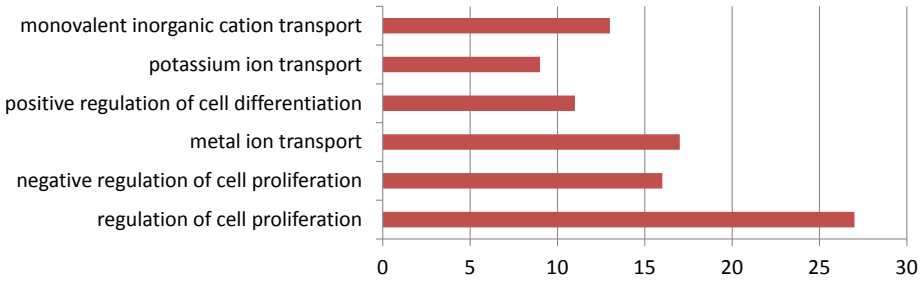


Figure A.6 Biological process terms in the 'GO FAT' annotation category found to be highly significantly enriched (P<0.01) in our list of 222 (out of 312) genes of the 303 cl21 binding sites. The horizontal axis (x-axis) shows the gene counts (number of genes involved in the analyzed GO FAT annotation category).

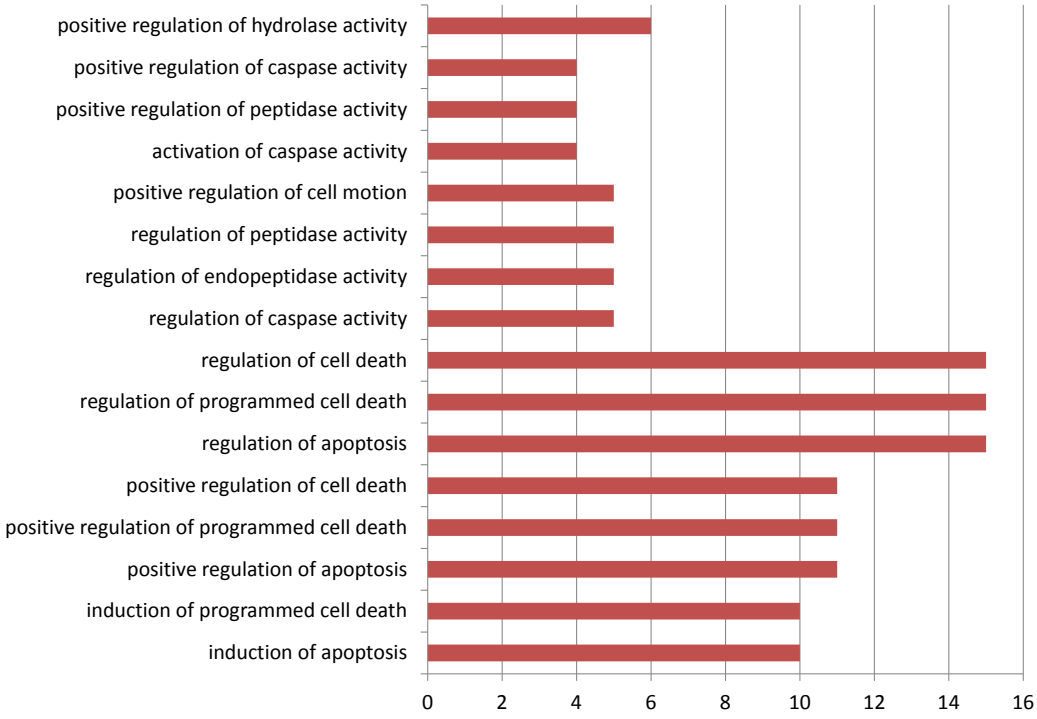


Figure A.7 Biological process terms in the 'GO FAT' annotation category found to be highly significantly enriched (P<0.01) in our list of 97 (out of 139) genes of the 136 cl22 binding sites. The horizontal axis (x-axis) shows the gene counts (number of genes involved in the analyzed GO FAT annotation category).

Appendix B

Genomic location analysis

The following calculations were made for the G-test. Here, we use the following abbreviations: f for 'observed count', E for 'expected probability' and e for 'expected count'.

Genomic region	Size (bp)	E
Intragenic region	815470630	27
TSS flanking region	342582384	11
5kb downstream	67670618	2
5-25kb downstream	178570332	6
5-25kb upstream	164015357	5
Intergenic region	1508472566	49
Total	3076781887	100

B.1 1757 p53 binding sites

Region	f	E	e	(f/e)	$f\ln(f/e)$
Intragenic region	523.00	27.00	474.39	1.10	51.02
TSS flanking region	316.00	11.00	193.27	1.64	155.36
5kb downstream	54.00	2.00	35.14	1.54	23.20
5-25kb downstream	124.00	6.00	105.42	1.18	20.13
5-25kb upstream	120.00	5.00	87.85	1.37	37.42
Intergenic region	620.00	49.00	860.93	0.72	-203.54
Total	1757.00	100.00	1757.00		$\ln L$=83.59

Region	f	E	e	(f/e)	$f\ln(f/e)$
Intragenic region	523.00	27.00	474.39	1.10	51.02
Other regions	1234.00	73.00	1282.61	0.96	-47.68
Total	1757.00	100.00	1757.00		$\ln L$=3.34

Region	f	E	e	(f/e)	$f\ln(f/e)$
TSS flanking region	316.00	11.00	193.27	1.64	155.36
Other regions	1441.00	89.00	1563.73	0.92	-117.78
Total	1757.00	100.00	1757.00		$\ln L$=37.58

Region	f	E	e	(f/e)	$f\ln(f/e)$
5kb downstream	54.00	2.00	35.14	1.54	23.20
Other regions	1703.00	98.00	1721.86	0.99	-18.76
Total	1757.00	100.00	1757.00		$\ln L$=4.44

Region	f	E	e	(f/e)	$f\ln(f/e)$
5-25kb downstream	124.00	6.00	105.42	1.18	20.13
Other regions	1633.00	94.00	1651.58	0.99	-18.48
Total	1757.00	100.00	1757.00		$\ln L$=1.65

Region	f	E	e	(f/e)	$f\ln(f/e)$
5-25kb upstream	120.00	5.00	87.85	1.37	37.42
Other regions	1637.00	95.00	1669.15	0.98	-31.84
Total	1757.00	100.00	1757.00		$\ln L$=5.58

Region	f	E	e	(f/e)	$f\ln(f/e)$
Intergenic region	620.00	49.00	860.93	0.72	-203.54
Other regions	1137.00	51.00	896.07	1.27	270.75
Total	1757.00	100.00	1757.00		$\ln L$=67.21

B.2 Combined evidence predictions

B.2.1 2999 combined evidence predictions

Region	f	E	e	(f/e)	$f\ln(f/e)$
Intragenic region	1055.00	27.00	809.73	1.30	279.15
TSS flanking region	678.00	11.00	329.89	2.06	488.42
5kb downstream	91.00	2.00	59.98	1.52	37.93
5-25kb downstream	199.00	6.00	179.94	1.11	20.04
5-25kb upstream	196.00	5.00	149.95	1.31	52.49
Intergenic region	780.00	49.00	1469.51	0.53	-494.04
Total	2999.00	100.00	2999.00		$\ln L$=383.99

Region	f	E	e	(f/e)	$f\ln(f/e)$
Intragenic region	1055.00	27.00	809.73	1.30	279.15
Other regions	1944.00	73.00	2189.27	0.89	-230.99
Total	2999.00	100.00	2999.00		$\ln L$=48.16
TSS flanking region	678.00	11.00	329.89	2.06	488.42
Other regions	2321.00	89.00	2669.11	0.87	-324.35
Total	2999.00	100.00	2999.00		$\ln L$=164.07
5kb downstream	91.00	2.00	59.98	1.52	37.93
Other regions	2908.00	98.00	2939.02	0.99	-30.86
Total	2999.00	100.00	2999.00		$\ln L$=7.08
5-25kb downstream	199.00	6.00	179.94	1.11	20.04
Other regions	2800.00	94.00	2819.06	0.99	-19.00
Total	2999.00	100.00	2999.00		$\ln L$=1.04
5-25kb upstream	196.00	5.00	149.95	1.31	52.49
Other regions	2803.00	95.00	2849.05	0.98	-45.68
Total	2999.00	100.00	2999.00		$\ln L$=6.82
Intergenic region	780.00	49.00	1469.51	0.53	-494.04
Other regions	2219.00	51.00	1529.49	1.45	825.74
Total	2999.00	100.00	2999.00		$\ln L$=331.69

B.2.2 305 combined evidence predictions

Region	f	E	e	(f/e)	$f\ln(f/e)$
Intragenic region	93.00	27.00	82.35	1.13	11.31
TSS flanking region	67.00	11.00	33.55	2.00	46.34
5kb downstream	12.00	2.00	6.10	1.97	8.12
5-25kb downstream	26.00	6.00	18.30	1.42	9.13
5-25kb upstream	19.00	5.00	15.25	1.25	4.18
Intergenic region	88.00	49.00	149.45	0.59	-46.61
Total	305.00	100.00	305.00		$\ln L$=32.47

Region	f	E	e	(f/e)	$f\ln(f/e)$
Intragenic region	93.00	27.00	82.35	1.13	11.31
Other regions	212.00	73.00	222.65	0.95	-10.39
Total	305.00	100.00	305.00		$\ln L$=0.92
TSS flanking region	67.00	11.00	33.55	2.00	46.34
Other regions	238.00	89.00	271.45	0.88	-31.30
Total	305.00	100.00	305.00		$\ln L$=15.04
5kb downstream	12.00	2.00	6.10	1.97	8.12
Other regions	293.00	98.00	298.90	0.98	-5.84
Total	305.00	100.00	305.00		$\ln L$=2.28
5-25kb downstream	26.00	6.00	18.30	1.42	9.13
Other regions	279.00	94.00	286.70	0.97	-7.60
Total	305.00	100.00	305.00		$\ln L$=1.54
5-25kb upstream	19.00	5.00	15.25	1.25	4.18
Other regions	286.00	95.00	289.75	0.99	-3.73
Total	305.00	100.00	305.00		$\ln L$=0.45
Intergenic region	88.00	49.00	148.96	0.59	-46.32
Other regions	216.00	51.00	155.04	1.39	71.62
Total	304.00	100.00	304.00		$\ln L$=25.31

B.3 Sequence-only predictions

B.3.1 2998 sequence-only predictions

Region	f	E	e	(f/e)	$f\ln(f/e)$
Intragenic region	872.00	27.00	809.73	1.08	64.61
TSS flanking region	378.00	11.00	329.89	1.15	51.46
5kb downstream	99.00	2.00	59.98	1.65	49.61
5-25kb downstream	225.00	6.00	179.94	1.25	50.28
5-25kb upstream	170.00	5.00	149.95	1.13	21.33
Intergenic region	1254.00	49.00	1469.51	0.85	-198.87
Total	2998.00	100.00	2999.00		$\ln L$=38.42

Region	f	E	e	(f/e)	$f\ln(f/e)$
Intragenic region	872.00	27.00	809.73	1.08	64.61
Other regions	2126.00	73.00	2189.27	0.97	-62.35
Total	2998.00	100.00	2999.00		$\ln L$=2.26
TSS flanking region	378.00	11.00	329.89	1.15	51.46
Other regions	2620.00	89.00	2669.11	0.98	-48.66
Total	2998.00	100.00	2999.00		$\ln L$=2.80
5kb downstream	99.00	2.00	59.98	1.65	49.61
Other regions	2899.00	98.00	2939.02	0.99	-39.75
Total	2998.00	100.00	2999.00		$\ln L$=9.86
5-25kb downstream	225.00	6.00	179.94	1.25	50.28
Other regions	2773.00	94.00	2819.06	0.98	-45.68
Total	2998.00	100.00	2999.00		$\ln L$=4.60
5-25kb upstream	170.00	5.00	149.95	1.13	21.33
Other regions	2828.00	95.00	2849.05	0.99	-20.97
Total	2998.00	100.00	2999.00		$\ln L$=0.36
Intergenic region	1254.00	49.00	1469.51	0.85	-198.87
Other regions	1744.00	51.00	1529.49	1.14	228.89
Total	2998.00	100.00	2999.00		$\ln L$=30.02

B.3.2 305 sequence-only predictions

Region	f	E	e	(f/e)	$f\ln(f/e)$
Intragenic region	71.00	27.00	82.35	0.86	-10.53
TSS flanking region	30.00	11.00	33.55	0.89	-3.36
5kb downstream	11.00	2.00	6.10	1.80	6.49
5-25kb downstream	30.00	6.00	18.30	1.64	14.83
5-25kb upstream	21.00	5.00	15.25	1.38	6.72
Intergenic region	142.00	49.00	149.45	0.95	-7.26
Total	305.00	100.00	305.00		$\ln L$=6.89

Region	f	E	e	(f/e)	$f\ln(f/e)$
Intragenic region	71.00	27.00	82.35	0.86	-10.53
Other regions	234.00	73.00	222.65	1.05	11.63
Total	305.00	100.00	305.00		$\ln L$=1.11
TSS flanking region	30.00	11.00	33.55	0.89	-3.36
Other regions	275.00	89.00	271.45	1.01	3.57
Total	305.00	100.00	305.00		$\ln L$=0.22
5kb downstream	11.00	2.00	6.10	1.80	6.49
Other regions	294.00	98.00	298.90	0.98	-4.86
Total	305.00	100.00	305.00		$\ln L$=1.63
5-25kb downstream	30.00	6.00	18.30	1.64	14.83
Other regions	275.00	94.00	286.70	0.96	-11.46
Total	305.00	100.00	305.00		$\ln L$=3.37
5-25kb upstream	21.00	5.00	15.25	1.38	6.72
Other regions	284.00	95.00	289.75	0.98	-5.69
Total	305.00	100.00	305.00		$\ln L$=1.03
Intergenic region	142.00	49.00	149.45	0.95	-7.26
Other regions	163.00	51.00	155.55	1.05	7.63
Total	305.00	100.00	305.00		$\ln L$=0.36

Appendix C

Gene Ontology and KEGG pathway enrichment analyses

C.1 1757 positive p53 binding sites

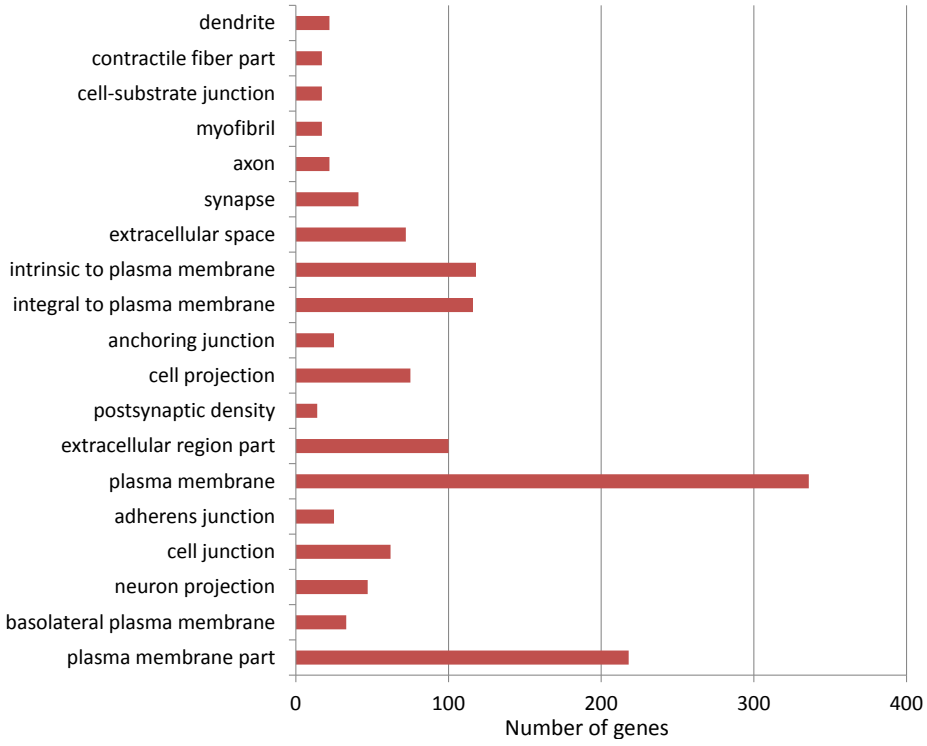


Figure C.1 Statistically enriched (P<0.014) cellular component terms in the 'GO FAT' annotation category. The most statistically significant GO term is displayed at the bottom.

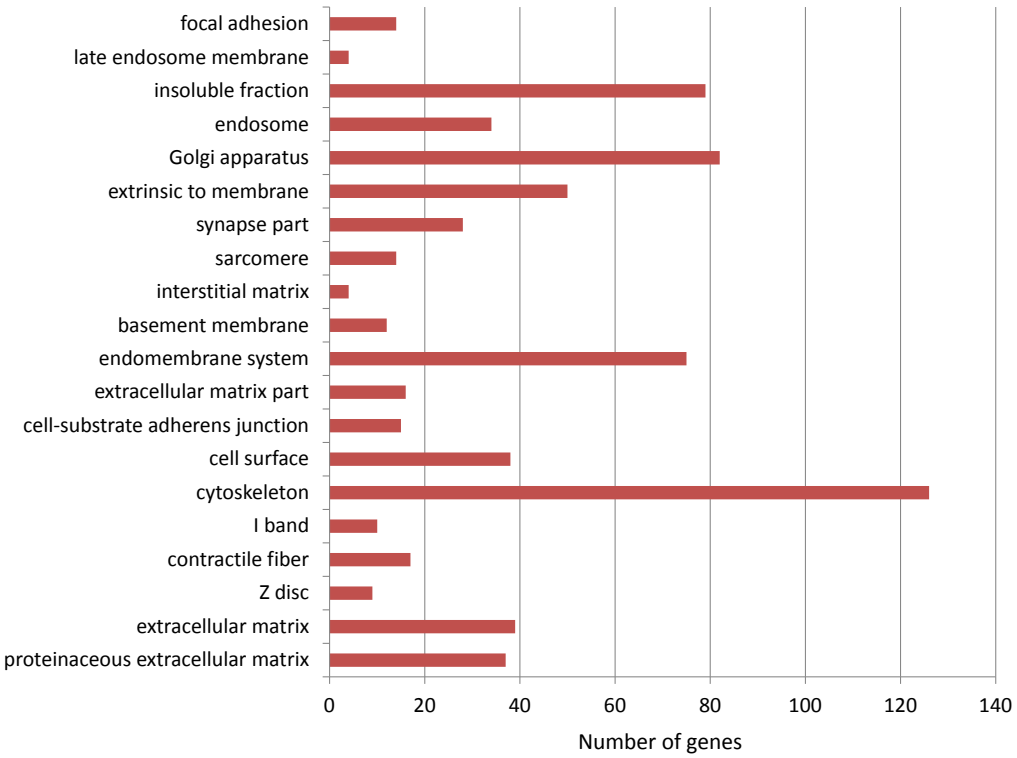


Figure C.2 Statistically enriched ($0.014<P<0.05$) cellular component terms in the 'GO FAT' annotation category. The most statistically significant GO term is displayed at the bottom.

Table C.1 and Table C.2 present statistically significant ($P<0.05$) cellular component GO terms which were associated with our 1757 positive binding sites. 987 genes were involved in the cellular component 'GO FAT' annotation category. The most statistically significant and the most numerous GO terms in the cellular component ontology were 'plasma membrane part' and 'plasma membrane', respectively.

C.2 Combined evidence predictions

C.2.1 305 predicted p53 binding sites by genome-wide analysis

545 nearby genes were identified for the 305 human p53 binding sites predicted by our combined evidence model. The total number of genes involved in the GO enrichment analyses by DAVID was 483, of which 339, 329 and 313 were associated with particular GO terms of biological process, molecular function and cellular component, respectively. 158 genes were involved in the KEGG pathway analysis.

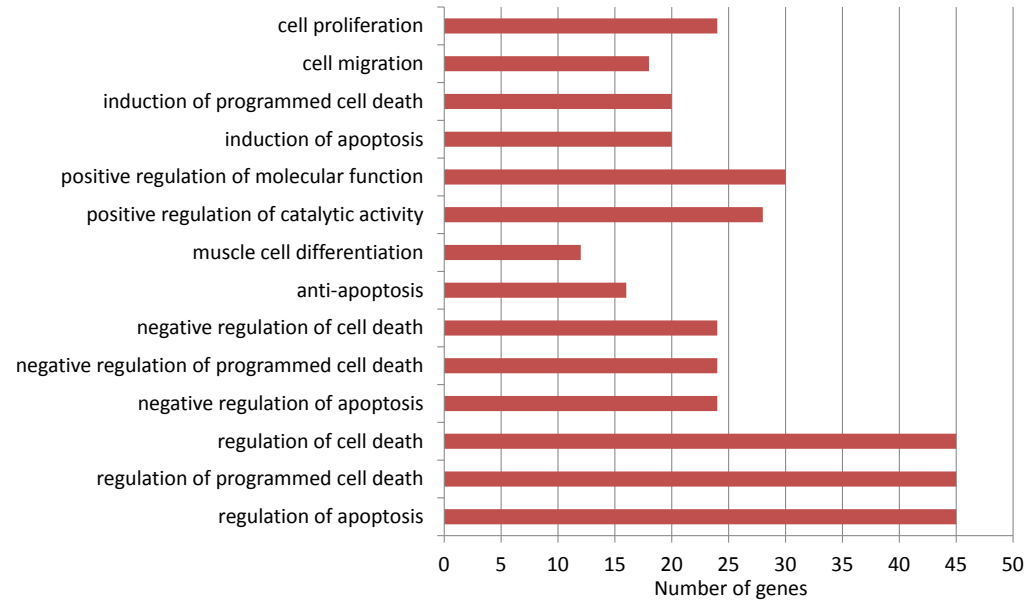


Figure C.3 339 genes of the 305 combined evidence predictions were involved in the biological process 'GO FAT' annotation category. Strongly enriched (P<1×10^{-3}) GO terms of biological process are presented with the most statistically significant term 'regulation of apoptosis' at the bottom.

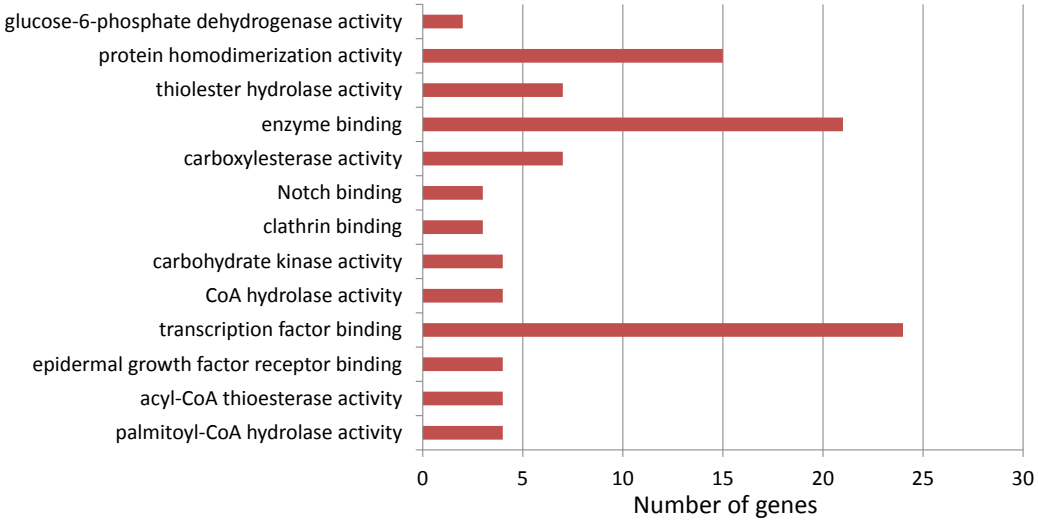


Figure C.4 Functional enrichment analysis associated with GO molecular function terms from the 'GO FAT' annotation category involving 329 genes. The statistically significant (P<0.05) GO terms are listed in order from most significant to less significant with the most statistically significant term 'palmitoyl-CoA hydrolase activity' at the bottom.

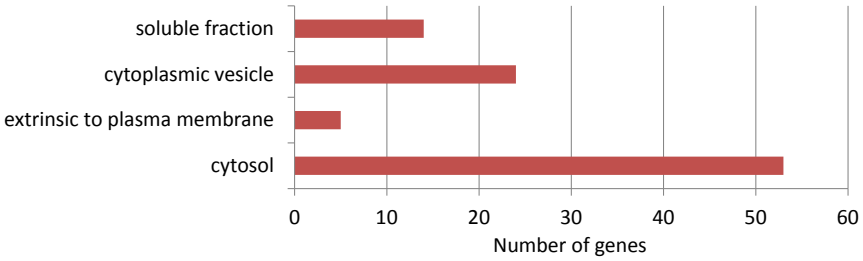


Figure C.5 313 genes of the 305 combined evidence predictions were involved in the cellular component 'GO FAT' annotation category. Strongly enriched (P<0.05) GO terms of cellular component are presented with the most statistically significant term 'cytosol' at the bottom.

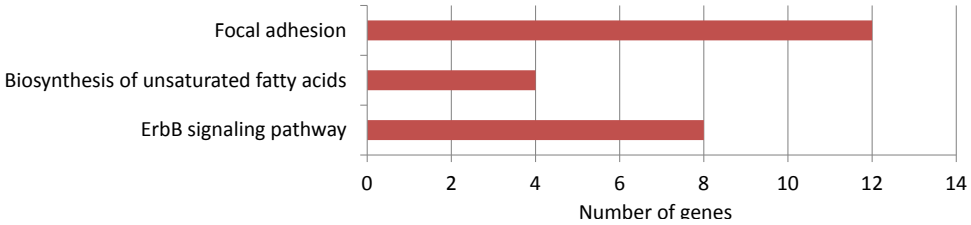


Figure C.6 158 genes of the 305 combined evidence predictions were associated with a particular KEGG pathway. Statistically significant (P<0.05) pathways are presented with 'ErbB signaling pathway' as the most significant KEGG pathway associated with our gene list.

C.3 Sequence-only predictions

C.3.1 305 predicted p53 binding sites by genome-wide analysis

423 nearby genes were identified for the 305 human p53 binding sites predicted by the simple sequence-only model. The total number of genes involved in the GO enrichment analyses by DAVID was 356, of which 329, 225 and 217 were associated with particular GO terms of biological process, molecular function and cellular component, respectively. 98 genes were involved in the KEGG pathway analysis.

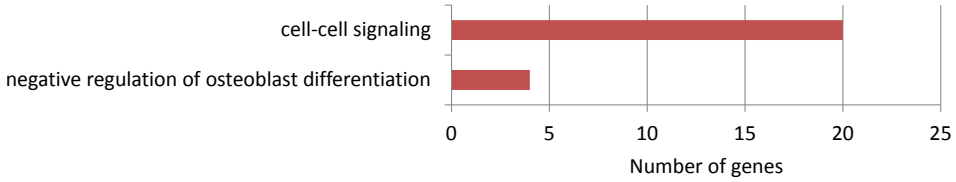


Figure C.7 329 genes of the 305 sequence-only predictions were involved in the biological process 'GO FAT' annotation category. Strongly enriched (P<0.01) GO terms of biological process are presented with the most statistically significant term 'negative regulation of osteoblast differentiation' at the bottom.

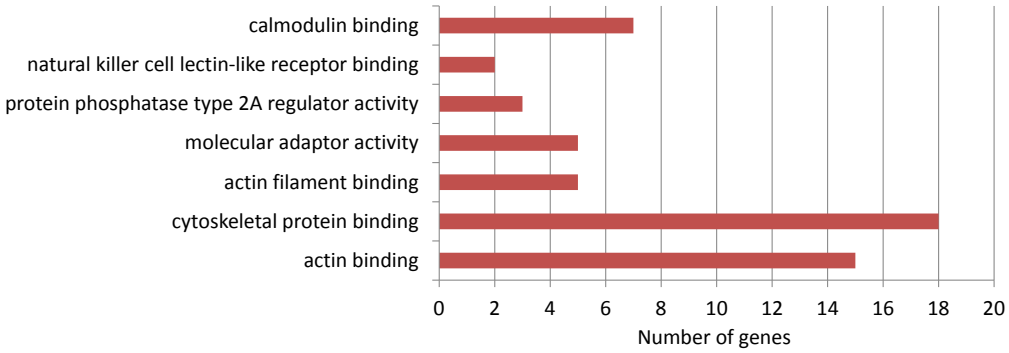


Figure C.8 Functional enrichment analysis associated with GO molecular function terms from the 'GO FAT' annotation category involving 225 genes. The statistically significant (P<0.05) GO terms are listed in order from most significant to less significant with the most statistically significant term 'actin binding' at the bottom.

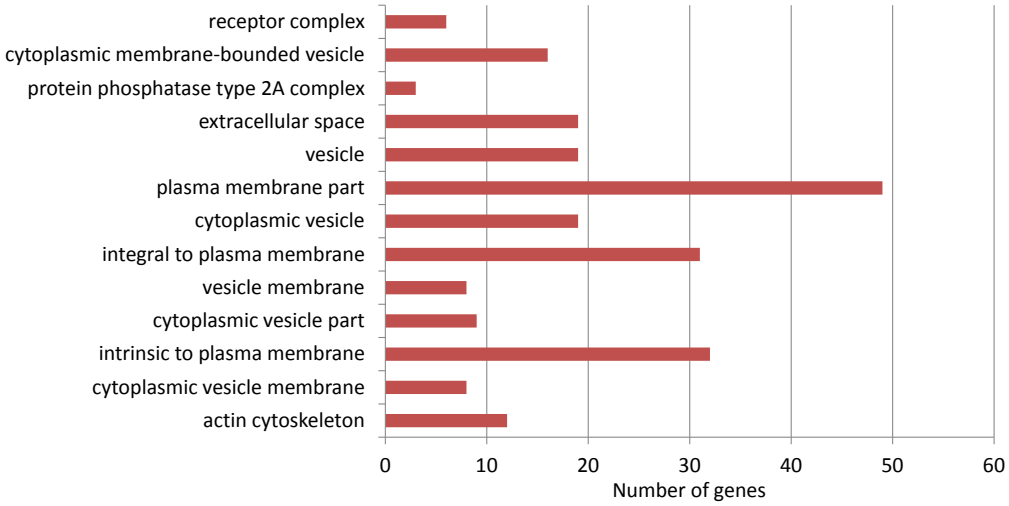


Figure C.9 217 genes of the 305 sequence-only predictions were involved in the cellular component 'GO FAT' annotation category. Strongly enriched (P<0.05) GO terms of cellular component are presented with the most statistically significant term 'actin cytoskeleton' at the bottom.

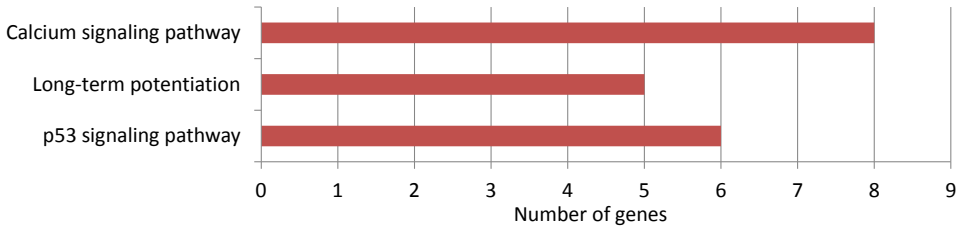


Figure C.10 98 genes of the 305 sequence-only predictions were associated with a particular KEGG pathway. Statistically significant (P<0.05) pathways are presented with 'p53 signaling pathway' as the most significant KEGG pathway associated with our gene list.

Bibliography

Agresti, A. (2002). *Categorical data analysis*. Wiley-Interscience.

Agresti, A. and B. Finlay (2007). *Statistical Methods for the Social Sciences II*. Prentice Hall.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *The Second International Symposium on Information Theory*, pp. 267–281.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman (1990). Basic local alignment search tool. *Journal of Molecular Biology 215*, 403–410.

An, W., J. Kim, and R. G. Roeder (2004). Ordered cooperative functions of PRMT1, p300, and CARM1 in transcriptional activation by p53. *Cell 117*, 735–748.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock (2000). Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nature Genetics 1*, 25–29.

Bailey, T. L., M. Bodén, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, and W. S. Noble (2009). Meme suite: tools for motif discovery and searching. *Nucleic Acids Research 37*, W202–W208.

Bailey, T. L. and C. Elkan (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proceedings of the International Conference on Intelligent Systems for Molecular Biology*, Volume 2, pp. 28–36.

Baker, S. J., E. R. Fearon, J. M. Nigro, S. R. Hamilton, A. C. P. J. M. Jessup, P. vanTuinen, D. H. Ledbetter, D. F. Barker, Y. Nakamura, R. White, and B. Vogelstein (1989). Chromosome 17 deletions and p53 gene mutations in colorectal carcinomas. *Science 244*, 217–221.

Baptiste, N., P. Friedlander, X. Chen, and C. Prives (2002). The proline-rich domain of p53 is required for cooperation with anti-neoplastic agents to promote apoptosis of tumor cells. *Oncogene 21*, 9–21.

Barski, A., S. Cuddapah, K. Cui, T. Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao (2007). High-resolution profiling of histone methylations in the human genome. *Cell 129*, 823–837.

Bayle, J. H., B. Elenbaas, and A. J. Levine (1995). The carboxyl-terminal domain of the p53 protein regulates sequence-specific DNA binding through its nonspecific nucleic acid-binding activity. *Proceedings of the National Academy of Sciences of the United States of America 92*, 5729–5733.

Beckerman, R. and C. Prives (2010). Transcriptional regulation by p53. *Carcinogenesis 2*, a000935.

Bergamaschi, D., Y. Samuels, A. Sullivan, M. Zvelebil, H. Breyssens, A. Bisso, G. D. Sal, N. Syed, P. Smith, M. Gasco, T. Crook, and X. Lu (2006). iASPP preferentially binds p53 proline-rich region and modulates apoptotic function of codon 72-polymorphic p53. *Nature Genetics 38*, 1133–1141.

Bernstein, B. E., M. Kamal, K. Lindblad-Toh, S. Bekiranov, D. K. Bailey, D. J. Huebert, S. McMahon, E. K. Karlsson, E. J. 3rd Kulbokas, T. R. Gingeras, S. L. Schreiber, and E. S. Lander (2005). Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell 120*, 169–181.

Blat, Y. and N. Kleckner (1999). Cohesins bind to preferential sites along yeast chromosome III, with differential regulation along arms versus the centric region. *Cell 98*, 249–259.

Bode, A. M. and Z. Dong (2004). Post-translational modification of p53 in tumorigenesis. *Nature Reviews Cancer 4*, 793–805.

Böttger, V., A. Böttger, C. Garcia-Echeverria, Y. F. Ramos, A. J. van der Eb, A. G. Jochemsen, and D. P. Lane (1999). Comparative study of the p53-mdm2 and p53-MDMX interfaces. *Oncogene 18*, 189–199.

Brain, R. and J. R. Jenkins (1994). Human p53 directs DNA strand reassociation and is photolabelled by 8-azido ATP. *Oncogene 9*, 1775–1780.

Buhler, J. and M. Tompa (2002). Finding motifs using random projections. *Journal of computational biology : a journal of computational molecular cell biology 9*, 225–242.

Burley, S. K. and R. G. Roeder (1996). Biochemistry and structural biology of transcription factor IID (TFIID). *Annual Review of Biochemistry 65*, 769–799.

Campbell, M. K. and S. O. Farrell (2007). Transcription of the genetic code: the biosynthesis of RNA. In *Biochemistry*, pp. 287–330. Cengage Learning.

Carstensen, B., M. Plummer, E. Laara, and M. Hills (2012). *Epi: A Package for Statistical Analysis in Epidemiology*.

Cawley, S., S. Bekiranov, H. H. Ng, P. Kapranov, E. A. Sekinger, D. Kampa, A. Piccolboni, V. Sementchenko, J. Cheng, A. J. Williams, R. Wheeler, B. Wong, J. Drenkow, M. Yamanaka, S. Patel, S. Brubaker, H. Tammana, G. Helt, K. Struhl, and T. R. Gingeras (2004). Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell 116*, 499–509.

Chan, H. M. and N. B. L. Thangue (2001). p300/CBP proteins: HATs for transcriptional bridges and scaffolds. *Journal of Cell Science 114*, 2363–2373.

Chehab, N. H., A. Malikzay, E. S. Stavridi, and T. D. Halazonetis (1999). Phosphorylation of Ser-20 mediates stabilization of human p53 in response to DNA damage. *Proceedings of the National Academy of Sciences of the United States of America 96*, 13777–13782.

Chin, F. Y. L. and H. C. M. Leung (2005). Voting algorithms for discovering long motifs. *Proceedings of the Third Asia-Pacific Bioinformatics Conference (APBC 2005), Singapore*, 261–271.

Choong, M. L., H. Yang, M. A. Lee, and D. P. Lane (2009). Specific activation of the p53 pathway by low dose actinomycin D: a new route to p53 based cyclotherapy. *Cell Cycle 8*, 2810–2818.

Chuikov, S., J. K. Kurash, J. R. Wilson, B. Xiao, N. Justin, G. S. Ivanov, K. McKinney, P. Tempst, C. Prives, S. J. Gamblin, N. A. Barlev, and D. Reinberg (2004). Regulation of p53 activity through lysine methylation. *Nature 432*, 353–360.

Commowick, O. and G. Malandain (2007). Efficient selection of the most similar image in a database for critical structures segmentation. In *Proceedings of the 10th Int. Conf. on Medical Image Computing and Computer-Assisted Intervention - MICCAI 2007, Part II*, Volume 4792, pp. 203–210.

Cook, J. L., R. N. Ré, J. F. Giardina, F. E. Fontenot, D. Y. Cheng, and J. Alam (1995). Distance constraints and stereospecific alignment requirements characteristic of p53 DNA-binding consensus sequence homologies. *Oncogene 11*, 723–733.

Craig, A. L., J. P. Blaydes, L. R. Burch, A. M. Thompson, and T. R. Hupp (1999). Dephosphorylation of p53 at Ser20 after cellular exposure to low levels of non-ionizing radiation. *Oncogene 18*, 6305–6312.

Creyghton, M. P., A. W. Cheng, G. G. Welstead, T. Kooistra, B. W. Carey, E. J. Steine, J. Hanna, M. A. Lodato, G. M. Frampton, P. A. Sharp, L. A. Boyer, R. A. Young, and R. Jaenisch (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences of the United States of America 107*, 21931–21936.

Crooks, G. E., G. Hon, J. M. Chandonia, and S. E. Brenner (2004). WebLogo: A sequence logo generator. *Genome Research 14*, 1188–1190.

Cuellar-Partida, G., F. A. Buske, R. C. McLeay, T. Whitington, W. S. Noble, and T. L. Bailey (2012). Epigenetic priors for identifying active transcription factor binding sites. *Bioinformatics 28*, 56–62.

Cui, F., M. Sirotin, and V. Zhurkin (2011). Impact of Alu repeats on the evolution of human p53 binding sites. *Biology Direct 6*, 2.

Deleo, A. B., G. Jay, E. Appella, G. C. Dubois, L. W. Law, and L. J. Old (1979). Detection of a transformation-related antigen in chemically induced sarcomas and other transformed cells of the mouse. In *Proceedings of the National Academy of Sciences of the United States of America*, Volume 76, pp. 2420–2424.

Dornan, D., H. Shimizu, L. Burch, A. J. Smith, and T. R. Hupp (2003). The proline repeat domain of p53 binds directly to the transcriptional coactivator p300 and allosterically controls DNA-dependent acetylation of p53. *Molecular and Cellular Biology 23*, 8846–8861.

Down, T. A. and T. J. Hubbard (2005). NestedMICA: sensitive inference of over-represented motifs in nucleic acid sequence. *Nucleic Acids Research 33*, 1445–1453.

Dumaz, N. and D. W. Meek (1999). Serine15 phosphorylation stimulates p53 transactivation but does not directly influence interaction with HDM2. *The EMBO journal 18*, 7002–7010.

Dumaz, N., D. M. Milne, L. J. Jardine, and D. W. Meek (2001). Critical roles for the serine 20, but not the serine 15, phosphorylation site and for the polyproline domain in regulating p53 turnover. *The Biochemical Journal 359*, 459–464.

Dumaz, N., D. M. Milne, and D. W. Meek (1999). Protein kinase CK1 is a p53-threonine 18 kinase which requires prior phosphorylation of serine 15. *FEBS Letters 463*, 312–316.

El-Deiry, W., S. Kern, J. Pietenpol, K. Kinzler, and B. Vogelstein (1992). Definition of a consensus binding site for p53. *Nature Genetics 1*, 45–49.

Eliason, S. R. (1993). *Maximum Likelihood Estimation: Logic and Practice*. SAGE.

Ernst, J., H. L. Plasterer, I. Simon, and Z. Bar-Joseph (2010). Integrating multiple evidence sources to predict transcription factor binding in the human genome. *Genome Research 20*, 526–536.

Eskin, E. (2004). From profiles to patterns and back again: a branch and bound algorithm for finding near optimal motif profiles. In *RECOMB '04: Proceedings of the eighth annual international conference on Resaerch in computational molecular biology*, pp. 115–124.

Eskin, E. and P. A. Pevzner (2002). Finding composite regulatory patterns in DNA sequences. *Bioinformatics 18*, 354–363.

Espinosa, J. M. and B. M. Emerson (2001). Transcriptional regulation by p53 through intrinsic dna/chromatin binding and site-directed cofactor recruitment. *Molecular Cell 8*, 57–69.

Euskirchen, G. M., J. S. Rozowsky, C. L. Wei, W. H. Lee, Z. D. Zhang, S. Hartman, O. Emanuelsson, V. Stolc, S. Weissman, M. B. Gerstein, Y. Ruan, and M. Snyder (2007). Mapping of transcription factor binding regions in mammalian cells by ChIP: comparison of array- and sequencing-based technologies. *Genome Research 17*, 898–909.

Federico, M. and N. Pisanti (2009). Suffix tree characterization of maximal motifs in biological sequences. *Theoretical Computer Science 410*, 4391–4401.

Fielding, A. (2007). Cluster analysis. In *Cluster And Classification Techniques for the Biosciences*, pp. 46–77. Cambridge University Press.

Fingerman, I. M. and S. D. Briggs (2004). p53-mediated transcriptional activation: from test tube to cell. *Cell 117*, 690–691.

Finlan, L. and T. R. Hupp (2004). The N-terminal interferon-binding domain (IBiD) homology domain of p300 binds to peptides with homology to the p53 transactivation domain. *The Journal of Biological Chemistry 279*, 49395–49405.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika 80*, 27–38.

Flores, E. R., K. Y. Tsai, D. Crowley, S. Sengupta, A. Yang, F. McKeon, and T. Jacks (2002). p63 and p73 are required for p53-dependent apoptosis in response to DNA damage. *Nature 416*, 560–564.

Frith, M. C., U. Hansen, J. L. Spouge, and Z. Weng (2004). Finding functional sequence elements by multiple local alignment. *Nucleic Acids Research 32*, 189–200.

Fu, W., P. Ray, and E. P. Xing (2009). DISCOVER: a feature-based discriminative method for motif search in complex genomes. *Bioinformatics 25*, i321–i329.

Fujita, P. A., B. Rhead, A. S. Zweig, A. S. Hinrichs, D. Karolchik, M. S. Cline, M. Goldman, G. P. Barber, H. Clawson, A. Coelho, M. Diekhans, T. R. Dreszer, B. M. Giardine, R. A. Harte, J. Hillman-Jackson, F. Hsu, V. Kirkup, R. M. Kuhn, K. Learned, C. H. Li, L. R. Meyer, A. Pohl, B. J. Raney, K. R. Rosenbloom, K. E. Smith, D. Haussler, and W. J. Kent (2011). The UCSC Genome Browser database: update 2011. *Nucleic Acids Research 39*, D876–D882.

Fullwood, M. J., C. L. Wei, E. T. Liu, and Y. Ruan (2009). Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome Research 19*, 521–532.

Gogol-Döring, A. and K. Reinert (2009). Pattern matching. In *Biological Sequence Analysis Using the SeqAn C++ Library*, pp. 135–162. Cambridge University Press.

Grant, C. E., T. L. Bailey, and W. S. Noble (2011). Fimo: Scanning for occurrences of a given motif. *Bioinformatics 27*, 1017–1018.

Gu, W., J. Luo, C. L. Brooks, A. Y. Nikolaev, and M. Li (2004). Dynamics of the p53 acetylation pathway. In *Novartis Foundation Symposium*, pp. 197–205.

Gu, W., X. L. Shi, and R. G. Roeder (1997). Synergistic activation of transcription by cbp and p53. *Nature 387*, 819–823.

Guccione, E., F. Martinato, G. Finocchiaro, L. Luzi, L. Tizzoni, V. D. Olio, G. Zardo, C. Nervi, L. Bernard, and B. Amati (2006). Myc-binding-site recognition in the human genome is determined by chromatin context. *Nature Cell Biology 8*, 764–770.

Guimond, A., J. Meunier, and J.-P. Thirion (2000). Average brain models: A convergence study. *Computer Vision and Image Understanding 77*, 192–210.

Hamming, R. W. (1950). Error Detecting and Error Correcting Codes. *Bell System Techincal Journal 29*, 147–160.

Han, K. A. and M. F. Kulesz-Martin (1992). Altered expression of wild-type p53 tumor suppressor gene during murine epithelial cell transformation. *Cancer Research 52*, 749–753.

Harris, S. L. and A. J. Levine (2005). The p53 pathway: positive and negative feedback loops. *Oncogene 24*, 2899–2908.

He, H. H., C. A. Meyer, H. Shin, S. T. Bailey, G. Wei, Q. Wang, Y. Zhang, K. Xu, M. Ni, M. Lupien, P. Mieczkowski, J. D. Lieb, K. Zhao, M. Brown, and X. S. Liu (2010). Nucleosome dynamics define transcriptional enhancers. *Nature Genetics 42*, 343–347.

Hearnes, J. M., D. J. Mays, K. L. Schavolt, L. Tang, X. Jiang, and J. A. Pietenpol (2005). Chromatin immunoprecipitation-based screen to identify functional genomic binding sites for sequence-specific transactivators. *Molecular and Cellular Biology 25*, 10148–10158.

Heintzman, N. D., G. C. Hon, R. D. Hawkins, P. Kheradpour, A. Stark, L. F. Harp, Z. Ye, L. Lee, R. K. Stuart, C. W. Ching, K. A. Ching, J. E. Antosiewicz-Bourget, H. Liu, X. Zhang, R. D. Green, V. V. Lobanenkov, R. Stewart, J. A. Thomson, G. E. Crawford, M. Kellis, and B. Ren (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature 459*, 108–112.

Heintzman, N. D., R. K. Stuart, G. Hon, Y. Fu, C. W. Ching, R. D. Hawkins, L. O. Barrera, S. V. Calar, C. Qu, K. A. Ching, W. Wang, Z. Weng, R. D. Green, G. E. Crawford, and B. Ren (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics 39*, 311–318.

Heinze, G. and M. Ploner (2003). Fixing the nonconvergence bug in logistic regression with SPLUS and SAS. *Computer Methods and Programs in Biomedicine 71*, 181–187.

Heinze, G. and M. Schemper (2002). A solution to the problem of separation in logistic regression. *Statistics in medicine 21*, 2409–2419.

Hertz, G. Z. and G. D. Stormo (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics 15*, 563–577.

Ho, J. W., E. Bishop, P. V. Karchenko, N. Négre, K. P. White, and P. J. Park (2011). ChIP-chip versus ChIP-seq: lessons for experimental design and data analysis. *BMC Genomics 12*, 134.

Hoh, J., S. Jin, T. Parrado, J. Edington, A. J. Levine, and J. Ott (2002). The p53MH algorithm and its application in detecting p53-responsive genes. *Proceedings of the National Academy of Sciences of the United States of America 99*, 8467–8472.

Horvath, M. M., X. Wang, M. A. Resnick, and D. A. Bell (2007). Divergent evolution of human p53 binding sites: Cell cycle versus apoptosis. *PLoS Genetics 3*, e127.

Huang, D. W., B. T. Sherman, and R. A. Lempicki (2009a). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research 1*, 1–13.

Huang, D. W., B. T. Sherman, and R. A. Lempicki (2009b). Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature Protocols 1*, 44–57.

Huang, J. and S. Li (2005). Mining p53 binding sites using profile hidden markov model. In *Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'05)*, Volume 1, pp. 146–151.

Hudson, M. E. and M. Snyder (2006). High-throughput methods of regulatory element discovery. *BioTechniques 41*, 673–681.

Jenkins, L. M., S. R. Durell, S. J. Mazur, and E. Appella (2012). p53 N-terminal phosphorylation: a defining layer of complex regulation. *Carcinogenesis*, Epub ahead of print.

Jiang, B., M. Q. Zhang, and X. Zhang (2007). OSCAR: one-class SVM for accurate recognition of cis-elements. *Bioinformatics 23*, 2823–2828.

Jiménez-Valverde, A. and J. M. Lobo (2007). Threshold criteria for conversion of probability of species presence to either-or presence-absence. *Acta Oecologica 31*, 361–369.

Kadane, J. B. and N. A. Lazar (2004). Methods and criteria for model selection. *Journal of the American Statistical Association 99*, 279–290.

Kanehisa, M., S. Goto, Y. Sato, M. Furumichi, and M. Tanabe (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research 40*, D109–D114.

Kaneshiro, K., S. Tsutsumi, S. Tsuji, K. Shirahige, and H. Aburatani (2007). An integrated map of p53-binding sites and histone modification in the human ENCODE regions. *Genomics 89*, 178–188.

Keich, U. and P. A. Pevzner (2002). Finding motifs in the twilight zone. *Bioinformatics 18*, 1374–1381.

Kent, W. J., C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler (2002). The human genome browser at UCSC. *Genome Research 12*, 996–1006.

Kim, T. H., L. O. Barrera, M. Zheng, C. Qu, M. A. Singer, T. A. Richmond, Y. Wu, R. D. Green, and B. Ren (2005). A high-resolution map of active promoters in the human genome. *Nature 436*, 876–880.

Kress, M., E. May, R. Cassingena, and P. May (1979). Simian virus 40-transformed cells express new species of proteins precipitable by anti-simian virus 40 tumor serum. *Journal of Virology 31*, 472–483.

Kuang, R., E. Ie, K. Wang, K. Wang, M. Siddiqi, Y. Freund, and C. Leslie (2005). Profile-based string kernels for remote homology detection and motif extraction. *Journal of Bioinformatics and Computational Biology (JBCB) 3*, 527–550.

Lackner, D. H. and J. Bähler (2008). Translational Control of Gene Expression: From Transcripts to Transcriptomes. In *International Review of Cell and Molecular Biology*, Volume 271, pp. 199–251. Elsevier.

Lambert, P. F., F. Kashanchi, M. F. Radonovich, R. Shiekhattar, and J. N. Brady (1998). Phosphorylation of p53 serine 15 increases interaction with CBP. *The Journal of Biological Chemistry 1*, 33048–33053.

Lane, D. P. and L. V. Crawford (1979). T antigen is bound to a host protein in sv40-transformed cells. *Nature 278*, 261–263.

Laptenko, O. and C. Prives (2006). Transcriptional regulation by p53: one protein, many possibilities. *Cell death and differentiation 13*, 951–961.

Larose, D. T. (2006). *Data Mining Methods And Models*. Wiley-Interscience.

Latchman, D. S. (2008a). DNA Sequences, transcription factors and chromatin structure. In *Eukaryotic transcription factors*, pp. 1–28. Academic Press.

Latchman, D. S. (2008b). Methods for studying transcription factors. In *Eukaryotic transcription factors*, pp. 29–67. Academic Press.

Latchman, D. S. (2008c). RNA polymerases and the basal transcriptional complex. In *Eukaryotic transcription factors*, pp. 68–95. Academic Press.

Lavin, M. F. and N. Gueven (2006). The complexity of p53 stabilization and activation. *Cell Death and Differentiation 13*, 941–950.

Lee, I. and E. M. Marcotte (2009). Effects of functional bias on supervised learning of a gene network model. In *Computational Systems Biology*, Volume 541, pp. 1–13. Humana Press.

Lee, T. I., S. E. Johnstone, and R. A. Young (2006). Chromatin immunoprecipitation and microarray-based analysis of protein location. *Nature Protocols 1*, 729–748.

Levine, A. J., W. Hu, and Z. Feng (2006). The P53 pathway: what questions remain to be explored? *Cell Death and Differentiation 13*, 1027–1036.

Levkovitz, L., N. Yosef, M. C. Gershengorn, E. Ruppin, R. Sharan, and Y. Oron (2010). A novel HMM-based method for detecting enriched transcription factor binding sites reveals RUNX3 as a potential target in pancreatic cancer biology. *PLoS One 5*, e14423.

Lewin, B. (2004). Activating transcription. In *Genes VIII*, pp. 631–655. Pearson Prentice Hall.

Liang, S. (2003). cwinnower algorithm for finding fuzzy dna motifs. In *Proceedings of the IEEE Computer Society Conference on Bioinformatics*, CSB '03, pp. 260–265.

Lifton, R. P., M. L. Goldberg, R. W. Karp, and D. S. Hogness (1978). The organization of the histone genes in drosophila melanogaster: functional and evolutionary implications. *Cold Spring Harbor symposia on quantitative biology 42*, 1047–1051.

Linzer, D. I. H. and A. J. Levine (1979). Characterization of a 54k dalton cellular sv40 tumor antigen present in sv40-transformed cells and uninfected embryonal carcinoma cells. *Cell 17*, 43–52.

Liu, L., D. M. Scolnick, R. C. Trievel, H. B. Zhang, R. Marmorstein, T. D. Halazonetis, and S. L. Berger (1999). p53 sites acetylated in vitro by PCAF and p300 are acetylated in vivo in response to DNA damage. *Molecular and Cellular Biology 19*, 1202–1209.

Liu, X., D. L. Brutlag, and J. S. Liu (2001). Bioprospector: discovering conserved dna motifs in upstream regulatory regions of co-expressed genes. In *Pacific Symposium on Biocomputing*, Volume 6, pp. 127–138.

Liu, X. S., D. L. Brutlag, and J. S. Liu (2002). An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nature Biotechnology 20*, 835–839.

Luo, J., M. Li, Y. Tang, M. Laszkowska, R. G. Roeder, and W. Gu (2004). Acetylation of p53 augments its site-specific DNA binding both in vitro and in vivo. *Proceedings of the National Academy of Sciences of the United States of America 101*, 353–360.

Lupien, M., J. Eeckhoute, C. A. Meyer, Q. Wang, Y. Zhang, W. Li, J. S. Carroll, X. S. Liu, and M. Brown (2008). FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. *Cell 132*, 958–970.

Lyakhov, I. G., A. Krishnamachari, and T. D. Schneider (2008). Discovery of novel tumor suppressor p53 response elements using information theory. *Nucleic Acids Research 36*, 3828–3833.

MacIsaac, K. D. and E. Fraenkel (2006). Practical strategies for discovering regulatory DNA sequence motifs. *PLoS Computational Biology 2*, e36.

Martin, C. and Y. Zhang (2005). The diverse functions of histone lysine methylation. *Nature Reviews Molecular Cell Biology 6*, 838–849.

Matys, V., E. Fricke, R. Geffers, E. Gössling, 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. Münch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, and E. Wingender (2003). TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Research 31*, 374–878.

Meek, D. W. (1999). Mechanisms of switching on p53: a role for covalent modification? *Oncogene 18*, 7666–7675.

Meek, D. W. and C. W. Anderson (2009). Posttranslational modification of p53: cooperative integrators of function. *Cold Spring Harbor Perspectives in Biology 1*, a000950.

Melero, J. A., D. T. Stitt, W. F. Mangel, and R. B. Carroll (1979). Identification of new polypeptide species (48-55k) immunoprecipitable by antiserum to purified large t antigen and present in sv40-infected and -transformed cells. *Virology 93*, 466–480.

Mendes, N. D., A. C. Casimiro, P. M. Santos, I. Sá-Correia, A. L. Oliveira, and A. T. Freitas (2006). MUSA: a parameter free algorithm for the identification of biologically significant motifs. *Bioinformatics 22*, 2996–3002.

Menendez, D., A. Inga, and M. A. Resnick (2009). The expanding universe of p53 targets. *Nature Reviews Cancer 9*, 724–737.

Millau, J. M., N. Bastien, and R. Drouin (2009). P53 transcriptional activities: A general overview and some thoughts. *Mutation Research/Reviews in Mutation Research 681*, 118–133.

Miyakoda, M., K. Suzuki, S. Kodama, and M. Watanabe (2002). Activation of ATM and phosphorylation of p53 by heat shock. *Oncogene 21*, 1090–1096.

Müller-Tiemann, B. F., T. D. Halazonetis, and J. J. Elting (1998). Identification of an additional negative regulatory region for p53 sequence-specific DNA binding. *Proceedings of the National Academy of Sciences of the United States of America 95*, 6079–6084.

Nowak, S. J. and V. G. Corces (2004). Phosphorylation of histone H3: a balancing act between chromosome condensation and transcriptional activation. *Trends in Genetics 20*, 214–220.

Oakes, D. (1999). Direct calculation of the information matrix via the EM algorithm. *J. R. Statistical Society 61*, 479–482.

Patel, S., R. George, F. Autore, F. Fraternali, J. E. Ladbury, and P. V. Nikolova (2008). Molecular interactions of ASPP1 and ASPP2 with the p53 protein family and the apoptotic promoters PUMA and Bax. *Nucleic Acids Research 36*, 5139–5151.

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

Pawlak, S. and J. Deckert (2007). Histone modifications under environmental stress. *Biological Letters 44*, 56–73.

Pierce, J. R. (1980). *An introduction to information theory. Symbols, signals & noise*. Dover Publications.

Poage, G. M., B. C. Christensen, E. A. Houseman, M. D. McClean, J. K. Wiencke, M. R. Posner, J. R. Clark, H. H. Nelson, C. J. Marsit, and K. T. Kelsey (2010). Genetic and epigenetic somatic alterations in head and neck squamous cell carcinomas are globally coordinated but not locally targeted. *PLoS One 5*, e9651.

Price, A., S. Ramabhadran, and P. A. Pevzner (2003). Finding subtle motifs by branching from sample strings. *Bioinformatics 19*, ii149–ii155.

Qian, H., T. Wang, L. Naumovski, C. D. Lopez, and R. K. Brachmann (2002). Groups of p53 target genes involved in specific p53 downstream effects cluster into different classes of DNA binding sites. *Oncogene 21*, 7901–7911.

Rada-Iglesias, A., R. Bajpai, T. Swigut, S. A. Brugmann, R. A. Flynn, and J. W. J (2011). A unique chromatin signature uncovers early developmental enhancers in humans. *Nature 470*, 279–283.

Rajasekaran, S. (2006). Algorithms for motif search. In *Handbook of Computational Molecular Biology*, pp. 37–1–37–21. Chapman&Hall/CRC.

Ren, B., F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T. L. Volkert, C. J. Wilson, S. P. Bell, and R. A. Young (2000). Genome-wide location and function of DNA binding proteins. *Science 290*, 2306–2309.

Rigoutsos, I. and A. Floratos (1998). Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. *Bioinformatics 14*, 55–67.

Riley, T., E. Sontag, P. Chen, and A. Levine (2008). Transcriptional control of human p53-regulated genes. *Nature Reviews Molecular Cell Biology 9*, 402–412.

Riley, T., X. Yu, E. Sontag, and A. Levine (2009). The p53HMM algorithm: using profile hidden markov models to detect p53-responsive genes. *BMC Bioinformatics 10*, 111.

Rodriguez, M. S., J. M. P. Desterro, S. Lain, D. P. Lane, and R. T. Hay (2000). Multiple C-terminal lysine residues target p53 for ubiquitin-proteasome-mediated degradation. *Molecular and Cellular Biology 20*, 8458–8467.

Roh, T., S. Cuddapah, and K. Zhao (2005). Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. *Genes & Development 19*, 542–552.

Roh, T. Y., S. Cuddapah, K. Cui, and K. Zhao (2006). The genomic landscape of histone modifications in human T cells. *Proceedings of the National Academy of Sciences of the United States of America 19*, 15782–15787.

Roth, F. P., J. D. Hughes, P. W. Estep, and G. M. Church (1998). Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature Biotechnology 16*, 939–945.

Sagot, M. (1998). Spelling approximate repeated or common motifs using a suffix tree. In *LATIN'98: Theoretical Informatics*, Volume 1380, pp. 374–390. Springer Berlin / Heidelberg.

Sakaguchi, K., S. Saito, Y. Higashimoto, S. Roy, C. W. Anderson, and E. Appella (2000). Damage-mediated phosphorylation of human p53 threonine 18 through a cascade mediated by a casein 1-like kinase. Effect on Mdm2 binding. *The Journal of Biological Chemistry 275*, 9278–9283.

Sakamuro, D., P. Sabbatini, E. White, and G. C. Prendergast (1997). The polyproline region of p53 is required to activate apoptosis but not growth arrest. *Oncogene 15*, 887–898.

Samuels-Lev, Y., D. J. O'Connor, D. Bergamaschi, G. Trigiante, J. K. Hsieh, S. Zhong, I. Campargue, L. Naumovski, T. Crook, and X. Lu (2001). ASPP proteins specifically stimulate the apoptotic function of p53. *Molecular Cell 8*, 781–794.

Sandelin, A., W. Alkema, P. Engström, W. W. Wasserman, and B. Lenhard (2004). JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research 32*, D91–D94.

Schneider, T. D. (1997). Information content of individual genetic sequences. *Journal of Theoretical Biology 189*, 427–441.

Schneider, T. D. and R. M. Stephens (1990). Sequence Logos: A New Way to Display Consensus Sequences. *Nucleic Acids Research 18*, 6097–6100.

Schon, O., A. Friedler, M. Bycroft, S. M. Freund, and A. R. Fersht (2002). Molecular mechanism of the interaction between MDM2 and p53. *Journal of Molecular Biology 323*, 491–501.

Shannon, C. E. (1948). A Mathematical theory of communication. *Bell System Technical Journal 27*, 379–423.

Shen, Y., F. Yue, D. F. McCleary, Z. Ye, L. Edsall, S. Kuan, U. Wagner, J. Dixon, L. Lee, V. V. Lobanenkov, and B. Ren (2012). A map of the cis-regulatory sequences in the mouse genome. *Nature*, Epub ahead of print.

Shieh, S. Y., M. Ikeda, Y. Taya, and C. Prives (1997). DNA damage-induced phosphorylation of p53 alleviates inhibition by MDM2. *Cell 91*, 49395–49405.

Sigal, A. and V. Rotter (2000). Oncogenic mutations of the p53 tumor suppressor: The demons of the guardian of the genome. *Cancer Research 60*, 6788–6793.

Sing, T., O. Sander, N. Beerenwinkel, and T. Lengauer (2005). Rocr: visualizing classifier performance in r. *Bioinformatics 21*, 3940–3941.

Sinha, A. U., M. Phatak, R. Bhatnagar, and A. G. Jegga (2007). Identifying functional binding motifs of tumor protein p53 using support vector machines. In *Sixth International Conference on Machine Learning and Applications (ICMLA 2007)*, pp. 506–511.

Sinha, S. and M. Tompa (2000). A statistical method for finding transcription factor binding sites. *Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology 8*, 344–354.

Smale, S. T. and J. T. Kadonaga (2003). The rna polymerase ii core promoter. *Annual Review of Biochemistry 72*, 449–479.

Smeenk, L., S. J. van Heeringen, M. Koeppel, M. A. van Driel, S. J. J. Bartels, R. C. Akkers, S. Denissov, H. G. Stunnenberg, and M. Lohrum (2008). Characterization of genome-wide p53-binding sites upon stress response. *Nucleic Acids Research 36*, 3639–3654.

Smith, A. D., P. Sumazin, and M. Q. Zhang (2005). Identifying tissue-selective transcription factor binding sites in vertebrate promoters. *Proceedings of the National Academy of Sciences of the United States of America 102*, 1560–1565.

Smith, A. E., R. Smith, and E. Paucha (1979). Characterization of different tumor antigens present in cells transformed by simian virus 40. *Cell 18*, 335–346.

Sokal, R. R. and C. D. Michener (1958). A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin 28*, 1409–1438.

Sokal, R. R. and F. J. Rohlf (1995). *Biometry: the principles and practice of statistics in biological research*. W. H. Freeman and Company.

Stommel, J. M. and G. M. Wahl (2004). Accelerated mdm2 auto-degradation induced by dna-damage kinases is required for p53 activation. *The EMBO journal 23*, 1547–1556.

Sun, X. X., M. S. Dai, and H. Lu (2007). 5-fluorouracil activation of p53 involves an MDM2-ribosomal protein interaction. *The Journal of Biological Chemistry 282*, 8052–8059.

Tan, P.-N., M. Steinbach, and V. Kumar (2006). Data. In *Introduction to Data Mining*, pp. 19–96. Pearson Addison Wesley.

Thijs, G., M. Lescot, K. Marchal, S. Rombauts, B. D. Moor, P. Rouzé, and Y. Moreau (2001). A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics 17*, 1113–1122.

Timothy, L. B. and C. Elkan (1995). The value of prior knowledge in discovering motifs with meme. In *ISMB'95*, pp. 21–29.

Tompa, M., N. Li, T. L. Bailey, G. M. Church, B. D. Moor, E. Eskin, A. V. Favorov, M. C. Frith, Y. Fu, W. J. Kent, V. J. Makeev, A. A. Mironov, W. S. Noble, G. Pavesi, G. Pesole, M. Régnier, N. Simonis, S. Sinha, G. Thijs, J. van Helden, M. Vandenbogaert, Z. Weng, C. Workman, C. Ye, and Z. Zhu (2005). Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnology 23*, 137–144.

Unger, T., T. Juven-Gershon, E. Moallem, M. Berger, R. V. Sionov, G. Lozano, M. Oren, and Y. Haupt (1999). Critical role for Ser20 of human p53 in the negative regulation of p53 by Mdm2. *The EMBO journal 18*, 1805–1814.

Valouev, A., D. S. Johnson, A. Sundquist, C. Medina, E. Anton, S. Batzoglou, R. M. Myers, and A. Sidow (2008). Genome-wide analysis of transcription factor binding sites based on ChIP-seq data. *Nature Methods 5*, 829–834.

Venot, C., M. Maratrat, C. Dureuil, E. Conseiller, L. Bracco, and L. Debussche (1998). The requirement for the p53 proline-rich functional domain for mediation of apoptosis is correlated with specific PIG3 gene transactivation and with transcriptional repression. *The EMBO journal 17*, 4668–4679.

Vittinghoff, E., D. V. Glidden, S. C. Shiboski, and C. E. McCulloch (2011). Logistic regression. In *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*, pp. 139–260. Springer.

Vousden, K. H. (2002). Activation of the p53 tumor suppressor protein. *Biochimica et Biophysica Acta 1602*, 47–59.

Vukojevic, V., T. Yakovleva, and G. Bakalkin (2010). Modes of p53 Interactions with DNA in the Chromatin Context. In *p53*, pp. 127–141. Springer.

Walker, K. K. and A. J. Levine (1996). Identification of a novel p53 functional domain that is necessary for efficient growth suppression. *Proceedings of the National Academy of Sciences of the United States of America 93*, 15335–15340.

Wang, Y., M. Reed, P. Wang, J. E. Stenger, G. Mayr, M. E. Anderson, J. F. Schwedes, and P. Tegtmeyer (1993). p53 domains: identification and characterization of two autonomous DNA-binding regions. *Genes & Development 7*, 2575–2586.

Wang, Y., J. F. Schwedes, D. Parks, K. Mann, and P. Tegtmeyer (1995). Interaction of p53 with its consensus DNA-binding site. *Molecular and Cellular Biology 15*, 2157–2165.

Wang, Z., C. Zang, J. A. Rosenfeld, D. E. Schones, A. Barski, S. Cuddapah, K. Cui, T.-Y. Roh, W. Peng, M. Q. Zhang, and K. Zhao (2008). Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature Genetics 40*, 897–903.

Ward, J. H. J. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association 58*, 236–244.

Warnock, L. J. and S. A. Raines (2004). Restoration of wild-type conformation to full-length and truncated p53 proteins: specific effects of ATP and ADP. *Cancer Biology & Therapy 3*, 634–637.

Wei, C.-L., Q. Wu, V. B. Vega, K. P. Chiu, P. Ng, T. Zhang, A. Shahab, H. C. Yong, Y. Fu, Z. Weng, J. Liu, X. D. Zhao, J.-L. Chew, Y. L. Lee, V. A. Kuznetsov, W.-K. Sung, L. D. Miller, B. Lim, E. T. Liu, Q. Yu, H.-H. Ng, and Y. Ruan (2006). A global map of p53 transcription-factor binding sites in the human genome. *Cell 124*, 207–219.

Won, K. J., B. Ren, and W. Wang (2010). Genome-wide prediction of transcription factor binding sites using an integrated model. *Genome Biology 11*, R7.

Xie, X., J. Lu, E. J. Kulbokas, T. R. Golub, V. Mootha, K. Lindblad-Toh, E. S. Lander, and M. Kellis (2005). Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature 434*, 338–345.

Zhang, Y. and D. Reinberg (2001). Transcription regulation by histone methylation: interplay between different covalent modifications of the core histone tails. *Genes & Development 15*, 2343–2360.

Zhu, J., J. Jiang, W. Zhou, K. Zhu, and X. Chen (1999). Differential regulation of cellular target genes by p53 devoid of the PXXP motifs with impaired apoptotic activity. *Oncogene 18*, 2149–2155.

Zweig, M. H. and G. Campbell (1993). Receiver-operating characteristic (roc) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry 39*, 561–577.