IDENTIFICATION OF CO-REGULATED CANDIDATE GENES BY PROMOTER ANALYSIS

Elizabeth H.B. Hellen

A thesis submitted in partial fulfilment of the requirements of the University of Brighton and the University of Sussex for the degree of Doctor of Philosophy

April 2010

Abstract

Genes which are co-expressed in tissues or processes are often regulated by the same transcription factors. In this thesis, the analysis of transcription factors predicted to regulate a gene is conducted to attempt to discover co-regulated, possibly co-expressed, genes. The initial problem tackled in the thesis is to determine an accurate method of predicting transcription factor binding sites (TFBS), and therefore regulatory transcription factors. Three TFBS prediction methods are developed; one consensus method combining pairs of prediction algorithms, one consensus method combining 6 algorithms through a Naïve Bayes classifier and a phylogenetic footprinting method combining data from multiple organisms using a Naïve Bayes classifier. The methods are comparable to or an improvement on current TFBS prediction methods. The second problem, with which the thesis is concerned, is the utilisation of these methods to predict genes which may be co-expressed in certain systems. Two systems will be analysed; (a) the IFN-γ related immune response to *Mycobacterium tuberculosis*, and (b) the host processes regulated by the Epstein Barr transcription factor, Zta. Several sets of candidate genes are predicted in each of these systems to take forward for experimental verification. In particular, three genes from regions linked to the IFN-γ immune response to *M. tuberculosis* have been shown to be involved in similar processes and to contain a putative cis-regulatory module.

Contents

List of Figures

List of Tables

Acknowledgements

First and foremost I would like to acknowledge the help and support that Dr Susan Jones and Professor Melanie Newport, my supervisors, have given me.

I would also like to acknowledge the work of my collaborators and colleagues - Kirsty Flower for her experimental confirmation of the EBV predictions and for her advice on laboratory and EBV matters, Dr Alison Sinclair for her EBV expertise, David Damerrell for the use of his PSRSS modules and general advice on GO term statistics, Dr Lionel Ripley for his advice on ROC curves and Dr Chris Finan and Dr Ruth Spriggs for their invaluable assistance and support throughout the research. I would also like to thank my thesis panel for their advice: Dr Andrew Martin, Professor Diana Lawrence-Watt and Dr Sarah Newbury.

On a personal note, I would like to thank all of my family and friends who have supported me thoughout the Phd.

Declaration

I declare that the research contained in this thesis, unless otherwise formally indicated within the text, is the original work of the author. The thesis has not been previously submitted to these or any other university for a degree, and does not incorporate any material already submitted for a degree.

Signed

Dated

Abbreviations & Definitions

AUC	Area under the curve
BP	Biological Process
CRM	Cis-regulatory module
EBV	Epstein-Barr virus
EMSA	Electrophoretic mobility shift assay
FN	False negative
FP	False positive
GO	Gene ontology
IFN-γ	Interferon gamma
NB	Naïve bayes
PBMC	Peripheral blood mononuclear cells
PCO	Principal co-ordinate analysis
PFM	Position frequency matrix
PMNB	Pattern matching naïve bayes methods
PNB	Positive naïve bayes
PPMNB	Phylogenetic pattern matching naïve bayes method
PTAN	Positive tree augmented naïve bayes
PWM	Position weight matrix
qPCR	Quanatiative real-time polymerase chain reaction
ROC	Reciever operating characteristic
SiRNA	Small interfering RNA
SVM	Support vector machine
TAN	Tree augmented naïve bayes
TB	Tuberculosis disease

TF	Transcription factor
TFBS	Transcription factor binding site
TN	True negative
TP	True positive
TSS	Transcription start site
ZRE	Zta response element

1. Introduction

The analysis of regulatory motifs in a gene's promoter region can provide information on gene expression and function. This thesis centres on the hypothesis that two genes are more likely to be co-regulated if they share a common binding site for a specific transcription factor. Two systems will be analysed; (a) the IFN-γ related immune response to *Mycobacterium tuberculosis*, and (b) the host processes regulated by the Epstein Barr transcription factor, Zta.

1.1. Genetic Regulation

1.1.1. Transcription Factors

Transcription factors (TFs) are a diverse category of proteins that regulate the transcription of DNA, through direct binding to the nucleotide sequence, using a DNA binding domain (Mitchell & Tjian, 1989; Ptashne & Gann, 1997). There are numerous other proteins involved in the regulation of transcription, such as co-activators, chromatin remodelers and histone acetylases. However, these proteins do not contain a DNA binding domain and are therefore not classified as TFs for the purpose of this thesis (Brivanlou & Darnell, Jr., 2002).

TFs can be classified into two main categories in eukaryotic organisms. The first category consists of the basal transcription factors; TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH. The basal transcription factors form a large complex with RNA polymerase II and are involved in the regulation of all eukaryotic genes (Lee & Young, 2000). The mechanisms behind the initiation of transcription by basal transcription factors are well characterised (Gross & Oelgeschläger, 2006). The second category of TFs is more numerous and diverse than that of the basal transcription factors. TFs in this second category are able to bind to DNA in a sequence specific manner and effect the temporal and spatial regulation of a gene. The large numbers of these specific transcription factors, combined with the large number of possible binding sites means that our knowledge of their regulatory function is far from complete. It is these TFs, their binding sites and the genes that they regulate that are the focus of this thesis.

Specific transcription factors, which will be referred to simply as TFs for the remainder of the thesis, can be classified in two main ways; firstly, by function (Brivanlou & Darnell, Jr., 2002), and secondly, by structure (Stegmaier *et al*., 2004). Functional classification of TFs allows for the classification of TFs by the function of the genes that they regulate. The HNF TFs, for example, regulate genes involved in liver function processes, whereas SP1 is a ubiquitous TF, regulating genes involved in many varied processes.

Structural classification categorises TFs by the tertiary structure of their binding domains into one of five superclasses, each of which has numerous subclasses (Stegmaier *et al*., 2004). The five superclasses (Matys *et al*., 2006) are:

1. Basic domains, which includes TFs such as AP-1 (activator protein 1), a leucine zipper factor, and MyoD (Myogenic differentiation 1) (Ma *et al*., 1994), a helix-loop-helix factor.

Image not available due to copyright restrictions

Figure 1.1. MyoD a Helix-loop-helix transcription factor, in the basic domain superclass, bound to a DNA molecule (PDB code – 1mdy) (Ma *et al*., 1994).

2. Zinc co-ordinating DNA-binding domains, including the GATA family of TFs (Bates *et al*., 2008) which contain diverse Cys4 zinc fingers, and SP1, which contains a Cys2His2 zinc finger domain.

Image not available due to copyright restrictions

Figure 1.2. Two GATA transcription factors binding to a DNA molecule (PDB code -3dfv) (Bates *et al*., 2008).

3. Helix-turn-Helix factors, including the homeo domain containing POU domain factors, such as Oct (Octamer transcription factor), and the tryptophan clusters subclass which contains Myb (Ogata, 1994) and numerous interferon regulatory factors.

Image not available due to copyright restrictions

Figure 1.3. A Myb transcription factor bound to a DNA molecule (PDB code – 1mse) (Ogata, 1994).

4. β-Scaffold Factors with minor groove contacts, including NFκB (Nuclear Factor-kappa B) (Muller, 1995) in the Rel homology region class and the STAT (Signal transducer and activator of transcription) family of TFs.

Figure 1.4. Two NFκB p50 transcription factors bound as a homodimer to a DNA molecule (PDB code – 1svc) (Muller, 1995).

5. Other transcription factors, such as Copper fist proteins (Turner, 1998) and the AP2/EREBP-related factors.

Figure 1.5. The solution structure of a Yeast Copper Fist transcription factor (PDB code – 1co4) (Turner, 1998).

The amino acid sequence and the structure of the DNA binding domain in a TF affects the sequence of the transcription factor binding site (TFBS) that can be bound. However, not all nucleotides within a TFBS may be in direct contact with the TF, hence a degree of variability is often allowed in a binding site before the strength of the interaction becomes too weak for the TF to bind (Berg & von Hippel, 1987).

1.1.2. Other regulatory mechanisms

The position and sequence of TFBSs are extremely important in the regulation of gene expression. However the existence of a putative TFBS does not necessarily mean that the gene will be regulated by this TF at all times and in all tissues; many other issues can affect this.

The first issue to consider when determining if a TFBS is functional is the localisation of the gene product. If the gene is a housekeeping gene it will be switched on in most tissues, examples being glyceraldehydes-3-phosphate dehydrogenase (GAPDH) and succinate dehydrogenase (SDHA) (Silver *et al.*, 2006). However, if the gene product has a specific function it may only be switched on in certain tissues or at certain times. Some TFs, such as Sp1, are found ubiquitously and are involved in many diverse signalling pathways (Lin *et al*., 2009; Prasanna Kumar *et al.*, 2008; Sugawara *et al.*, 2007). Many other TFs are much more specific, occurring only in certain tissues. The Hepatocyte nuclear factors-3 alpha, -3 beta and -3 gamma (HNF-3α, HNF-3β, HNF-3γ) are good examples of this type of transcription factor. HNF-3α, HNF-3β and HNF-3γ are found in the liver and regulate a number of hepatocyte-specific genes (Clevidence *et al*., 1993). Binding sites for HNF factors found in genes relating to processes in the brain or the heart are much less likely to be functional binding sites than those found in genes relating to the liver.

The second mechanism to consider when differentiating between functional and non-functional TFBSs relates to the shape of the DNA. In the cell, DNA is not a linear molecule but is coiled around histone molecules (Kornberg, 1974). Histone molecules are octamers consisting of 2 copies of H2A, H2B, H3 and H4. The DNA wrapped around this molecule forms a nucleosome, approximately 147bp wrapped in 1.67 left-handed superhelical turns (Luger *et al.*, 1997). This DNA coiling affects the accessibility of some potential TFBSs and renders them non-functional. This effect has been shown to prevent the initiation of transcription (Lorch *et al.*, 1987).

1.2. Predicting function from Co-regulation

One of the working hypotheses of functional genomics is that genes with similar mRNA expression profiles are likely to be regulated via the same mechanisms. Microarray experiments are based on this hypothesis and allow mRNA expression data to be extrapolated to regulatory networks (Altman & Raychaudhuri, 2001; Brazma *et al.*, 1998). Clustering experiments have given indirect evidence for this hypothesis. When genes have been clustered, according to their mRNA expression profiles, the clusters often share a common TF (Brazma *et al.*, 1998; Tavazoie *et al.*, 1999; Wolfsberg *et al.*, 1999). Following these experiments, a wide scale analysis in *Saccharomyces cerevisiae* was carried out by Allocco *et al.* (Allocco *et al.*, 2004). In this study, data from a genome wide binding analysis was used in combination with mRNA expression data to quantify the likelihood of genes sharing a common TFBS, given a certain level of similarity in expression profiles. It was shown that genes with expression profiles with a correlation greater than 0.84 had a greater than 50% chance of sharing a common TFBS. At lower correlation levels a much lower fraction of the gene pairs shared a TFBS. However, when information about function, through the use of gene ontology (GO) terms, was included, it was shown that gene pairs with correlations between 0.5 and 0.8 were more likely to share a TFBS if they had a high level of GO term similarity (Figure 1.6).

Image not available due to copyright restrictions

Figure 1.6. Co-regulation vs. co-expression and functional similarity as defined by level of GO term match. A, B and C were created using the biological process, molecular function and cellular component ontologies, respectively. For each of the three ontologies, the fraction of gene pairs sharing a common transcription factor binder is shown as a function of both correlation in expression profile and the level of GO term match (Allocco *et al*., 2004).

Marco *et al*. (Marco *et al*., 2009) have carried out an analysis similar to that carried out by Allocco *et al*. (Allocco *et al*., 2004); however this analysis has been carried out in *Drosophila melanogaster*. Experimentally verified TFBSs were extracted from the REDFly 2.0 database (Halfon *et al*., 2008), plus some additional TFBS data from a study by Li *et al.* (Li *et al*., 2008). Expression profiles were extracted from microarray analyses (Spellman & Rubin, 2002) and co-expression levels were measured using Pearson's correlation coefficient.

Genes sharing at least one TFBS had a significantly higher correlation coefficient than genes with no common TFBSs ($p<0.01$). The average correlation coefficient was again higher for genes sharing two TFBSs; however the difference was not significant (Marco *et al.*, 2009). In the *S. cerevisiae* analysis 100% of genes with expression correlation ≥0.9 shared at least one TFBS in common (Allocco *et al.*, 2004). In the *D. melanogaster* analysis only 24% of genes with ≥0.9 expression correlation were found to share a TFBS. This implies that while some highly co-expressed genes may be found by comparing TFBSs, unfortunately there may also be a high rate of false negatives. It would be expected that when applying this analysis to a higher level organism, such as *Homo sapiens*, this effect would be even greater.

Recently, methods have been developed to combine transcriptional information with microarray experiment results. Microarray experiments are often clustered to find co-regulated genes. However the genes in these clusters may be co-expressed but not co-regulated. By combining the two sets of data together there is a much higher chance of finding genes which are both co-expressed and co-regulated (Clements *et al.*, 2007). The results obtained by Allocco *et al.* show a correlation between the level of the GO match and the gene expression correlation coefficient (Allocco *et al.*, 2004). Analysing the level of GO term at which the two genes match is not necessarily the most informative method of assessing the similarity between the annotations of the two genes. The problem arises due to the structure of the gene ontology graph, where leaf nodes can have any number of generations of parent nodes, causing leaf nodes to have varying levels of specificity. This means that nodes on a certain level may have different degrees of specificity and are therefore not immediately comparable. A number of statistical methods have been developed to address these problems (Rivals et al., 2007; Pesquita *et al.*, 2008; Tao *et al.*, 2007).

One such method is the relative specific similarity (RSS) method which compares the similarity between two GO terms of the same category i.e. biological process, molecular function or cellular component (Wu *et al.,* 2006). RSS uses three components to assess the similarity between two GO terms (term_i and term_j): α, β and γ. The α component measures the similarity between the paths from term_i to the root node e.g. biological process (GO:008150), and term_j to the same root node (equation 1.1), in effect it determines the most similar shared term (MRCA).

$$\alpha = \max_{\substack{\text{path}_m \in \text{Paths}(\text{term}_i), \\ \text{path}_n \in \text{Paths}(\text{term}_j)}} \left\{ \begin{matrix} \text{The number of common terms} \\ \text{between path}_m \text{ and path}_n \end{matrix} \right\} - 1 \qquad \textbf{(1.1)}$$

The β component measures the generality of term_i and term_j in the GO (equation 1.2). The generality of a term is defined as the minimum distance between it and the leaf terms decending from it.

$$\beta = \max\{\underset{u\in U}{\min}\{dist(term_i)\}, \underset{u\in U}{\min}\{dist(term_j)\}\} \tag{1.2}$$

The γ component measures the local distance between $term_i$ and $term_j$ in relation to the most similar shared term (MRCA)(equation 1.3).

$$\gamma = dist(MRCA, term_i) + dist(MRCA, term_j) \tag{1.3}$$

The RSS score is shown in equation 1.4, where MaxDepthGO is the maximum depth between root and leaf nodes in the graph.

$$RSS(term_i, term_j) = \frac{MaxDepthGo}{MaxDepthGo + \gamma} \cdot \frac{\alpha}{\alpha + \beta} \tag{1.4}$$

The statistic used to measure the similarity of GO terms in this thesis is the path specific relative similarity score (PSRSS), a statistic derived from RSS and developed by David Damerrell (personal communication).

The difference between the PSRSS and the RSS is in the MaxDepthGo component. In RSS the MaxDepthGo term is the same for all terms in the GO graph. However, it is entirely possible that the terms of interest have maximum paths which are much shorter than this value. In PSRSS this component is term specific, giving a different MaxDepthGo depending on the terms of interest. This prevents bias against terms which, while specific concepts, have short paths from the root node, although does bias the statistic to terms which have not been fully mapped to the GO.

1.3. Determining Transcription factor binding sites

1.3.1. Experimental methods

Experimental methods for determining TFBSs and their functionality include: electrophoretic mobility shift assays (EMSA), chromatin immunoprecipitation (ChIP) and ChIP-chip assays. Each of these methods has advantages and disadvantages.

Electrophoretic mobility shift assays (EMSAs) (Fried & Crothers, 1981; Garner & Revzin, 1981) are the traditional approach for assessing protein-DNA interactions. EMSAs separate a protein bound DNA molecule from the unbound DNA molecules and 'shift' them on the gel according to size and shape, the bound molecules moving different amounts to the unbound molecules. It allows for the determination of whether a particular protein has the ability to bind a specific DNA sequence *in vitro*, however it is unable to tell us whether this interaction is functional *in vivo* (Elnitski *et al.*, 2006).

Chromatin immunoprecipitation (ChIP) assays are an alternative method which allow for binding reactions to be analysed *in vivo* in real time. ChIP assays cause cross-linking to occur between proteins and their DNA recognition sites through the application of formaldehyde, the DNA is then fragmented and precipitated by a transcription factor-specific antibody. After precipitation the cross-links are dissolved and the DNA sequence is elucidated through a PCR amplification reaction. This method is therefore particularly useful when the specific DNA binding sequence is not known. A variation on the ChIP assay, ChIP-chip (Ren *et al.*, 2000), is a high throughput version of the reaction, the DNA binding sites recovered in a ChIP assay are hybridised to a microarray chip. The DNA from the ChIP assay is labelled with Cy5 and an equal amount of total input DNA is labelled with Cy3, the microarray chip can then be visualised in the usual way.

Experimental techniques, while the most comprehensive and reliable methods for discovering functional TFBSs, are time consuming and expensive. In many cases these costs would be decreased by a pre-processing step using *in silico* methods to make predictions about the location of TFBSs. A number of computational prediction methods currently exist and will now be discussed.

1.3.2. In Silico prediction of Transcription Factor Binding Sites

1.3.2.1. TFBS databases

Information known about TFBSs is stored in two principle databases TRANSFAC (Matys *et al*., 2006) and JASPAR (Bryne *et al*., 2008; Sandelin *et al*., 2004a; Vlieghe *et al*., 2006). Information is stored as position weight matrices (PWM) which describe the probability of each nucleotide occurring at each position in the TFBS.

TRANSFAC exists in two forms; the public version, containing a reduced set of TFBSs, and the professional version, containing the whole of the TRANSFAC database. TRANSFAC stores data in three primary tables; Gene, Site and Factor. These tables contain information on regulated genes, TFBSs and TFs respectively (Matys *et al.,* 2006). The site table contains both genomic binding sites and artificial binding sites from oligonucleotide selection assays or IUPAC consensus sequences. The information in these tables is manually extracted from the literature by a curator and is usually based on experimental evidence. Meta data is inferred via comparison and classification of this data (Matys et al., 2006). In addition to this primary data, TRANSFAC contains secondary data, the most important of this being information on TFBS matrices. Information from the Site table is collated for each TF and nucleotide distribution matrices are produced and stored in the Matrix table.

JASPAR, unlike TRANSFAC, is an open access database with all data free to the public. JASPAR CORE, the central database of JASPAR contains a set of curated, non-redundant models from multi-cellular eukaryotic organisms (Sandelin *et al.,* 2004a). Other JASPAR databases also exist, e.g. JASPAR PHYLOFACTS, a database of evolutionary conserved motifs in 5’ promoter regions of multi-cellular eukaryotes (Bryne *et al.,* 2008). The purpose of JASPAR is to act as a repository for TFBS matrix models, not for individual TFBSs (Bryne *et al*., 2008). This is the main difference between the purposes of the two databases.

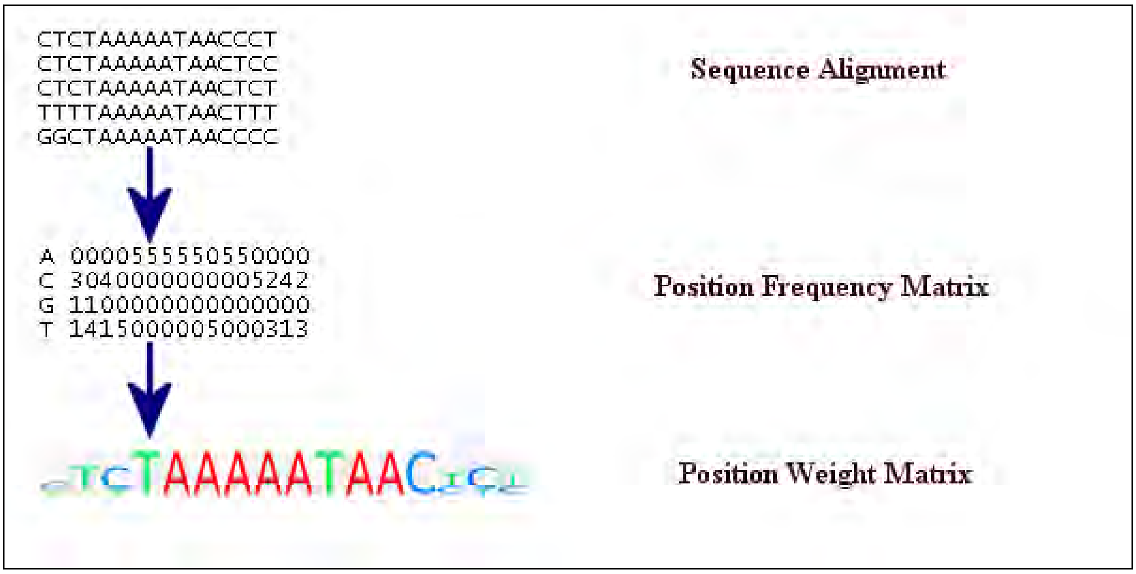
1.3.2.2. Prediction algorithms for TFBSs

With the large volume of DNA sequences deposited in public databases, algorithms have been developed to make predictions on the location of TFBSs in genomic DNA. Such prediction algorithms can be categorised into three groups. The first, and largest group, includes methods which match DNA motifs to those stored in a database of known TFBSs (Cartharius *et al.*, 2005; Kel *et al.*, 2003; Lenhard & Wasserman, 2002; Sandve *et al.,* 2007). Many of these algorithms match DNA motifs against TRANSFAC (Matys *et al.*, 2006). The second group uses phylogenetic footprinting, which involves the comparison of orthologous sequences, to detect conserved motifs (Carmack *et al.*, 2007; Friberg, 2007; Sandelin *et al*., 2004b; Tsai *et al*., 2007). The third group uses *ab initio* methods to search the genomic DNA for over expressed motifs in groups of genes known to have similar functions (Corà *et al.*, 2005; Dai *et al.*, 2007).

1.3.2.3. Methods using Position Weight Matrices for TFBS prediction

Most TFBS prediction algorithms which use TFBS databases use PWMs, also known as position specific scoring matrices (PSSM), to find probable binding sites. PWMs are created by aligning families of experimentally determined TFBSs and calculating the frequency of each nucleotide base, at each position in the alignment, assuming independence. This creates a position frequency matrix (PFM) which can then be converted into a PWM through the generation of log-likelihood scores for each nucleotide and position (Figure 1.7) (Crooks *et al.*, 2004; Wasserman & Sandelin, 2004). The algorithms used by TRANSFAC to create the PFMs have not been published. However, a study by Fu and Weng (Fu & Weng, 2005), showed TRANSFAC PWMs to have a superior ability to detect functional TFBSs than matrices created using the other algorithms tested; MotifSampler, MEME, AlignACE and Possum. The exception to this being matrices created using GLAM (Frith *et al*., 2004) which had a comparable performance to the TRANSFAC matrices.

Although hundreds of PWMs are available through TRANSFAC and JASPAR, many of these PWMs are not large enough nor specific enough to enable reliable predictions of TFBSs without large numbers of false positive predictions occurring by chance (Rahmann *et al*., 2003). Hence, algorithms have been developed which use additional features to reduce the number of false positive TFBS predictions. Features which have been used include phylogenetic data (Lenhard *et al*., 2003), nucleosome positioning (Westholm *et al*., 2008), clustering of binding sites (Zeng *et al*., 2008) and the distance from both the transcription start site (Makita *et al.*, 2005) and from other TFBSs (Larsson *et al.*, 2007; Singh *et al.*, 2007). Successful methods have been developed using each of these methods. However, none have completely solved the problem.



```
CTCTAAAAATAACCCT
CTCTAAAAATAACTCC
CTCTAAAAATAACTCT
TTTTAAAAATAACTTT
GGCTAAAAATAACCCC
```

```
A  0000555550550000
C  3040000000005242
G  1100000000000000
T  1415000005000313
```

Figure 1.7. Diagram describing the creation of a position weight matrix (PWM) for MEF2 (myogenic enhancer factor 2) from transcription factor binding site (TFBS) sequence alignments. A position frequency matrix (PFM) is created from a sequence alignment of the TFBS for the relevant transcription factor. The PWM is created from the PFM and can be represented numerically or graphically. Alignments and the PWM representation were retrieved from TRANSFAC.

1.3.2.4. Phylogenetic Footprinting for TFBS prediction

Phylogenetic footprinting is one way to improve the ratio of functional to non-functional TFBSs predicted using PWM based methods. It is used to good effect in the ConSite algorithm (Lenhard *et al.*, 2003; Sandelin *et al.*, 2004b). Phylogentic footprinting involves the retrieval of orthologous sequences, the alignment of the sequences, and the subsequent use of a PWM searching algorithm on the alignment. By comparing homologous sequences, and only retrieving TFBSs which score above a certain threshold value in both the sequences in the alignment, TFBSs can be filtered for those which have been evolutionarily conserved. These TFBSs are more likely to be functional binding sites; gene expression is known to be well conserved in vertebrates, and non-functional sequences are known to have a much higher mutation rate (Ganley & Kobayashi, 2007; Sorek & Ast, 2003). This is particularly the case if the orthologous sequences are carefully chosen. The work of Lenhard *et al*. shows that mouse-human comparisons are a particularly suitable evolutionary distance to consider (Lenhard *et al.*, 2003). Most phylogenetic methods for TFBS discovery in the human genome focus on the comparison of two species (Lenhard *et al.*, 2003), although methods which use multiple genomes have been developed for *Drosophila* or prokaryotic motif prediction (Maeder *et al.*, 2007; Neph & Tompa, 2006; Sandelin *et al.*, 2004b). The use of only one comparison genome makes the analysis prone to errors due to lack of orthologous sequences, or to alignment errors caused by the difficulty in aligning badly conserved sequences such as promoter regions (Satija *et al.*, 2008).

1.3.2.5. Other Computational methods

Ab-initio methods of TFBS prediction look for statistical over-representation of motifs in a sequence, or a set of sequences. The occurrence of a particular motif more frequently than would be expected for a random motif, implies that the sequence is being conserved. The conservation of the motif implies that it is a biologically relevant sequence. These methods are particularly useful when searching for motifs in genes which are assumed to be co-regulated. These methods are less useful for sets of sequences from genes not expected to be co-regulated. These sequences may not contain over-represented TFBS motifs. AlignAce (Hughes *et al.*, 2000; Roth *et al.*,

1998) and MEME (Zheng *et al*., 2003) are examples of *ab initio* algorithms and involve the analysis of DNA sequences for short motifs which occur more often than expected by chance. One of the main advantages of this type of method is to be able to predict unknown motifs. This is not possible using the PWM based methods as only well characterised binding sites can be searched for.

1.4. Finding functionally related genes in biological systems

This thesis focuses on the prediction of functional TFBSs to identify functionally related genes in two biological systems. The first system is the IFN-γ immune response to the *Mycobacterium bovis* Bacillus Calmette-Guérin vaccine (hereafter referred to as BCG) vaccination given to infants in The Gambia (Section 1.5 – 1.9). For this system the starting point was set of 532 genes from 3 different chromosomal regions identified in a genetic linkage study, which include genes linked to the the IFN-γ immune response to BCG. The chromosomal regions identified in the linkage study are large and the aim in this system was to use methods which compare TFBSs to identify co-regulated genes linked to the IFN-γ immune response. The second system is the interaction between the Epstein Barr Virus and its human host, through the Zta protein (Section 1.10 – 1.12). In this system TFBS prediction methods are used to identify candidate genes with Zta binding sites and further functional analyses are conducted to identify potential functional links between these genes.

1.5. System 1: IFN-γ immune response to BCG

1.5.1. Genetic Linkage Analysis

Following the success of linkage analysis in the identification of genes responsible for Mendelian disorders (Abou-Sleiman *et al*., 2004; Altshuler *et al.,* 2008; Pericak-Vance, 2001), the methodology was used to find genes involved in complex traits, such as the IFN-γ immune response to BCG. Linkage analysis is based on the assumption that, during meiosis, independent assortment has occurred. The Law of Independent Assortment states that alleles of different genes assort independently of one another during gamete formation. An exception to this occurs when genes are positioned closely together on a chromosome, reducing the likelihood of a crossover event occurring between them. Genes which are physically close are genetically linked and alleles are more likely to be inherited together. The distance between genes can then be calculated using this knowledge (Ott, 1991; Pericak-Vance, 2001). The more closely together the genes are positioned on the chromosome, the more frequently alleles occur together in families. Due to this linkage phenomenon, genetic markers can be used to determine where in the genome the genes related to a certain disease or condition lie. Markers are genotyped at intervals in the region of interest and inheritance of marker alleles in relation to the trait of interest is studied. Cosegregation of a particular allele of a marker and the trait within families suggests the two are linked (Pericak-Vance, 2001). However, in order to find the exact location of the gene the markers must be positioned very closely together and large numbers of families must be studied. Often a preliminary search with markers a greater distance apart will be carried out to determine the region in which the gene(s) of interest lie before continuing to a much more specific search in those regions to determine the actual gene(s). The aim of the TFBS prediction methods applied to this system is to take this coarse linkage data and identify candidate genes *in-silico* and hence avoid the need for further linkage analysis.

An alternative method of determining the genes involved in complex traits is by the use of genetic association studies. Association studies have been shown to be able to detect smaller genetic effects than linkage analysis studies. These studies involve the genotyping of single nucleotide polymorphisms (SNPs) and the statistical analysis of whether a certain SNP is causal for a disease (Jakobsdottir *et al*., 2009). However, until recently it has not been feasible to carry out study of the size required by association studies; thousands of individuals are required for testing by approximately 0.5 million markers (Barrett, J. *et al*., 2007; Collins, 2009). It was due to these extensive requirements and cost, for a genetic association study, that a linkage study was carried out.

1.5.2. Tuberculosis : an overview

Tuberculosis (TB) is one of the major causes of death in humans, particularly in Asian and African populations. It is estimated that 2 billion people are infected with TB; this is equal to one third of the world's population (Stop TB Partnership and World Health Organization, 2006). In 2006, 9.2 million new cases and 1.7 million deaths occurred across the world (World Health Organisation, 2009). In most developing countries the BCG vaccine is used as an integral part of the TB prevention program. The BCG vaccine, however, has variable efficacy. In different population BCG efficacy varies from 0-80% (Fine, 1995). TB cannot be entirely prevented through the use of this vaccine so treatments must be administered to individuals who develop disease. Traditionally TB is treatable, however due to inconsistent or partial treatment drug resistant TB strains have developed. Multidrug-resistant TB (MDR-TB) fails to respond to standard first line drugs and as such is difficult and expensive to treat (Heym *et al*., 1994; Jacobs, 1994). Extensively drug-resistant TB (XDR-TB) is caused by TB strains where second-line drug resistance has developed as well as first line drug resistance. XDR-TB is virtually untreatable as the situation stands (Matteelli *et al*., 2007).

The TB epidemic has not reached its peak in many countries, particularly in Africa. sub-Saharan African countries have the highest burden of TB morbidity and mortality (Corbett, 2003; Dolin *et al*., 1994), a large proportion of which is due to co-infection with HIV. Many social factors increase the incidence of TB in these regions such as; population growth, over crowding, war, smoking, alcoholism and natural disasters such as drought and famine (Daniel *et al*., 1994; Lönnroth *et al*., 2009). Resistant strains of TB have appeared in developing countries such as those in sub-Saharan Africa due to the lack of availability of first line drugs, and the difficulty in administering 6 month long treatments for TB.

1.5.3. Immune response to TB

Tuberculosis (TB) is caused by infection with *Mycobacterium tuberculosis*. *M. tuberculosis* is an intracellular parasite that grows within mononuclear phagocytes and as such requires the co-ordinated activity of a number of different immunological defence systems (Flynn & Chan, 2001). There is evidence that the host response plays an important role in determining the clinical symptoms and ultimate outcome for individuals infected with *M.tuberculosis* (Young, 1993).

1.5.4. Innate immunity

The human response to *M.tuberculosis* is provided primarily by the innate immunity of the host. The macrophage constitutes the primary defense against invasion. For the majority of microbes, the acidic and hydrolytically active environment of the phagosome is deadly (Dietrich & Doherty, 2009; Rhode *et al.*, 2007). For microbes where this response is not sufficient the macrophage undergoes a number of steps to change it into a fully activate macrophage. On entering the host, *M tuberculosis* is recognised by the macrophage through the activation of at least two pattern recognition receptor families: the Toll-like receptors (TLR) and the nucleotides oligomerization domain (NOC)-like receptors (Balaram *et al.*, 2009; Jo, 2008). The TLR-2, TLR-1 heterodimer activates NF-κB after recognition of a triacylated lipoprotein derived from *M. tuberculosis*. Activation of NF-κB leads to the production of inflammatory cytokines and direct antimicrobial activity (Brightbill *et al.*, 1999; Liu *et al.*, 2006; Thoma-Uszynski *et al.*, 2001). NOD2 recognises a peptidoglycan, muramyl dipeptide (MDP), which can be found on *M. tuberculosis* (Girardin *et al.*, 2003; Yang, 2007). The activation of NOD2 leads to the activation of a NF-κB mediated response, but also causes an activation of the inflamasome (Delbridge & O'Riordan, 2007). The activated macrophage has increased microbicidal activity through the expression of inducible nitric oxide sythase (iNOS) (MacMicking *et al.*, 1997; Nathan, 2002). This has been shown to be a highly effective mechanism in murine models of tuberculosis infection (Chan *et al.*, 1995; Chan *et al.*, 1992; MacMicking, 1997).

M. tuberculosis is able to survive within a macrophage by inhibiting phagolysosome biogenesis. It is thought that the induction of autophagy by the macrophage overcomes this maturation block and allows the *M. tuberculosis* to be eliminated (Vergne *et al.*, 2006). Stimulation of mouse macrophages with IFN-γ has been shown to induce autophagy (Gutierrez *et al.*, 2004). Serum starvation and rapamycin induced autophagy-dependent activity has been shown to be an active defense against *M. tuberculosis* in human macrophages (Gutierrez *et al.*, 2004).

1.5.5. Evidence for genetic factors

Susceptibility to *M. tuberculosis* was thought to involve genetic factors long before this hypothesis could be tested. In the nineteenth century, when European TB incidence had reached epidemic levels, it was observed that incidence of TB appeared to run in families. It has been shown that only 10% of individuals who contract the infection develop the disease (Murray *et al*., 1990). This suggests that susceptibility or resistance to mycobacterial infections varies between individuals. Convincing evidence for the genetic influence in the level of resistance to TB comes from an accidental study in Germany in 1926 (Birkhaug, 2005). Two hundred and fifty one children were immunised with a BCG vaccine which had been contaminated by the virulent Kiel strain of *M. tuberculosis*. Of the children infected with *M. tuberculosis*, 47 (20%) developed no clinical symptoms, 127 (50%) developed radiologically evident disease and 77 (30%) died. That not all children developed the same level of disease, if they developed the disease at all, suggests that there is variability in susceptibility to the disease between individuals and that this difference may be genetic. Further evidence for genetic factors, in the susceptibility to TB, come from twin studies (Kallman & Reisner, 1943). The prophet survey of the 1950s concluded that there was a 2.5 fold higher concordance rate for TB among monozygotic twins compared to dizygotic twins, clearly showing a large genetic component to development of the disease (Comstock, 1978). Further studies into this survey have questioned the strength of this relationship due to an imbalance of variables within subgroups (van der Eijk *et al*., 2007), however there is still undeniable evidence for a genetic element in the human susceptibility to TB.

1.5.6. Known genetic factors for susceptibility to TB

The identification of TB susceptibility genes has been attempted by many groups, on the grounds that the proteins they encode could be targets for new drugs and vaccines. Genes indentified in murine models have proved to be relevant in humans – a good example is NRAMP1 (also known as SLC11A1) (Awomoyi *et al*., 2002; Bellamy *et al*., 1998). Studies in human families with rare mendelian disorders have identified a role for genes within the IFN-γ/Interleukin-12 pathway (Casanova & Abel, 2002). However, studies in outbred populations have generally been disappointing (Newport, 2009). This led to the development of the neonatal BCG model in which to study the genetic regulation of important immune responses relevant in TB.

1.5.7. Simplifying the Tuberculosis problem

The response to the Bacillus Calmette-Guérin (BCG) vaccine by naïve individuals can be used as a model for infection by *M. tuberculosis* (Newport *et al*., 2005). Infant vaccination is an excellent model in which to study the role of host genetic variation in the regulation of immune responses: a vaccine represents a controlled antigenic challenge in which immunologically naive infants receive the same infecting dose of a well characterised antigen(s), and responses are measured at the same interval post-vaccination in all infants. Variation due to non-genetic causes is minimised as far as possible in a human study. A study was carried out with infant cohorts in the Gambia by *et al*. (Newport *et al*., 2004). BCG was given at birth, according to the Expanded Programme on Immunisation, to twin pairs. All study twins received BCG vaccination with the same dose of the same BCG strain within 24 hours of birth. IFN-γ responses to mycobacterial antigens including purified protein derivative (PPD), antigen 85, killed *Mycobacterium tuberculosis* (KMTB) and short term culture filtrate (STCF) were measured at 5 months of age (Newport *et al*., 2004). A genome-wide linkage scan was conducted in 155 DZ twin families using microsatellite markers spaced at 10 centimorgan intervals across the genome (Newport *et al*., see appendix).

Dizygous twins are genetically equivalent to siblings and non-parametric allele sharing methods were applied to identify chromosomal regions linked to IFN-γ production. Data were analysed using both variance components methods and more robust regression-based methods including the Haseman-Elston method (Haseman & Elston, 1972), which regresses the trait difference squared directly on the identical by descent (IBD) allele sharing, and the Visscher-Hopper method, which regresses a weighted combination of the trait difference squared and the mean-corrected trait sum squared on the IBD sharing (Visscher & Hopper, 2001).

The linkage analysis has identified three chromosomal regions, linked to IFN-γ response to BCG. The results show that these regions include 659 genes distributed across chromosomes 8, 10 and 11. A region of chromosome 8 has also been identified as a major TB susceptibility locus in Moroccan subjects (Baghdadi *et al*., 2006), the other regions do not appear to have been identified, although other chromosome regions such as Xq and 15q have been identified in other studies (Bellamy *et al*., 2000).

Traditionally, further genetic mapping techniques would be employed to narrow down the regions and the candidate genes systematically investigated. However, this approach is laborious, expensive and requires additional clinical cohorts. Instead, TFs and TFBSs in common between genes in each of the gene regions shall be searched for.

The similarity of TFBSs in common between the genes will give an indication of genes which may be functionally related.

1.6. System 2: Epstein Barr Zta binding in the human host

1.6.1. Epstein Barr Overview

Epstein-Barr virus (EBV) is a member of the herpes family and is also referred to as human herpesvirus 4 (HHV-4) (Epstein *et al*., 1964). Approximately 95% of the human population is infected with EBV and most infection is asymptomatic (World Health Organisation, 2008). Infection usually occurs in adolescence through contact with infected saliva or through the airborne virus; in adolescents and young adults (15-25) initial infection causes infectious mononucleosis, also known as glandular fever, in approximately 50% of cases. Infectious mononucleosis is usually a benign and self-limiting disease. However, EBV has also been associated with a number of malignant tumours which is a much more serious complication of EBV infection (Maeda *et al*., 2009).

Burkitt's lymphoma is a tumour of B cell origin, occurring primarily in children between 3 and 13 years of age, with a peak incidence at 6 or 7 years. The tumour occurs endemically in parts of Africa and in Papua New Guinea, particularly in areas in which malaria in also endemic (Bornkamm *et al.*, 2009; Wright *et al.*, 2009). Areas in which malaria has been eradicated have a much lower incidence of Burkitt's lymphoma, as do children who carry the sickle cell anaemia allele, which is known to give protection against malaria. Although it seems likely that malaria is a co-factor of EBV in the development of Burkitt's lymphoma, the mechanism is yet to be completely elucidated. It is thought, however, that the malarial infections may have reduced the patients' resistance to EBV (Ferry, 2006). Analysis of Burkitt's lymphoma tumour cells have shown multiple copies of the EBV genome to be present, and infectious EBV particles can be recovered from tumour cell lines. All patients with African Burkitt's lymphoma have antibodies to EBV antigens, which are present at much higher levels than infected subjects without Burkitt's lymphoma (Besson *et al.,* 2009; Bhaduri-McIntosh *et al*., 2007). EBV has also been shown to transform human B-lymphocytes *in vitro* and to cause tumours in subhuman primates (Bornkamn, 2009). Although evidence points to a role of EBV in Burkitt's Lymphoma, the exact role of the virus is still unknown (Bornkamn, 2009; Brady *et al*., 2008).

Nasopharyngeal carcinoma is a tumour of the nasopharynx, or throat, which is also associated with EBV. The tumour occurs most commonly in southern China, however it can also be found in localised areas in Africa, Malaysia, Alaska and Iceland. Environmental and genetic factors have been implicated in the susceptibility of subjects to the development of nasopharyngeal carcinomas and with the highly specific geographical locations in which the disease can be found. As in Burkitt's lymphoma, multiple copies of the EBV genome can be found in the malignant cells of undifferentiated nasopharyngeal carcinomas and infectious particles can be recovered from cell lines. The antibodies to EBV antigens are also found to be much higher in EBV infected individuals with nasopharyngeal carcinomas than those without (Ayadi *et al*., 2009; Lin, 2009; Mo *et al*., 2009).

1.6.2. Zta and ZRE3

EBV has a biphasic infection cycle, consisting of both a lytic and a latent phase. During the latent phase, EBV is in a dormant state, only a restricted number of genes are expressed. These include Epstein-Barr viral nuclear antigens 1, 2, 3A, 3B and 3C (EBNA-1, EBNA-2, EBNA-3A, EBNA-3B, EBNA-3C), Epstein-Barr nuclear antigen leader protein (EBNA LP), Latent membrane protein 1, 2A and 2B (LMP-1, LMP2A, LMP2B) and Epstein-Barr viral small RNA 1 and 2 (EBER1, EBER2) (Kieff, 1996). In the lytic phase the virus reactivates and starts to replicate itself, a part of the life cycle essential for propagation and transmission of the virus (Bornkamm & Hammerschmidt, 2001; Rodriguez *et al.*, 2001). The switch to the lytic phase is though to be initiated by the EBV immediate-early transcription factor Zta (ZEBRA, BZLF1, Z, EB1), encoded by the *BZLF1* gene (Chevallier-Greco *et al.*, 1986; Countryman & Miller, 1985; Takada *et al.*, 1986). *BZLF1* is silent during latency, but is expressed early in the lytic cycle and has been shown to disrupt latency in EBV (Sinclair, 2003). Zta binds to Zta response elements (ZREs) in order to activate the EBV lytic gene promoters, starting a cascade of over 50 genes (Petosa *et al*., 2006).

Zta belongs to the bZip TF family. bZip is a family of TFs found across all taxonomic groups from viruses to mammals. Other examples of TFs in this family are c-fos, MafG and AP-1. Zta belongs to the AP-1 sub family and has been shown to have a similar binding site to AP-1 (El-Guindy *et al*., 2006; Farrell *et al*., 1989). The Zta protein is composed of 245 amino acid residues and can be divided into five functional domains. The five domains are: a transactivation domain (aa 1-167), a regulatory domain (aa 168-177), a basic DNA binding domain (aa 178-194), a coiled-coil dimerisation domain (aa 195-227) and an accessory activation domain (aa 228-245) (Chang *et al*., 1990; Chi & Carey, 1993; Countryman & Miller, 1985; Farrell *et al*., 1989).

A number of ZREs have been identified and reported in the literature (Lehman *et al*., 1998), all with a binding sequence of 7 nucleotides, conserved for that particular ZRE. This study concentrates on ZRE3, which consists of a 7 base conserved sequence; TCGCGAA. ZRE3 is shown to preferentially bind Zta in a methylated state (El Guindy *et al*., 2006).

Figure 1.8. The Zta transcription factor in complex with a DNA binding site (PDB file 2C9I) (Petosa *et al*., 2006).

1.6.3. Known ZRE3 interactions

Zta has been shown to bind to motifs in human promoter regions and interact with cellular factors, activating genes in a manner out of the control of the human host. TRK-related tyrosine kinase has been shown to be up-regulated in Zta transfected cells but not in control cells (Lu *et al*., 2000). The up-regulation of a kinase suggests that host signalling cascades could be initiated by Zta. Zta is able to regulate AP1 protein expression and to compete with Fos-Jun heterodimers for AP1 sites suggesting that Zta also has the potential to interfere with AP1-mediated cell proliferation and differentiation (Speck *et al.*, 1997). Zta has also been shown to induce cell cycle arrest in the G0/G1 phase of epithelial tumour cell lines through the activation of p53, p21 and p27 (Cayrol & Flemington, 1996). Another effect of the Zta protein is an inhibition of the IFN-γ signalling pathway, altering host immune responses and suggesting a mechanism through which EBV may avoid host responses during initial infection (Morrison *et al.*, 2001).

1.7. Thesis Aims

The three main aims of the work presented in this thesis are to:

(a) Benchmark current TFBS prediction tools and develop novel methods for improving their accuracy for predicting functional TFBSs.

(b) Apply the improved TFBS prediction method(s) to identify candidate genes linked to the the IFN-γ response to BCG, in biological data derived from a genetic linkage study.

(c) Apply TFBS prediction methods to the identification of genes with TFBSs for the Zta transcription factor which activates lytic gene promoters in EBV.

2. Benchmarking and Combining TFBS prediction algorithms

2.1. Introduction

The first step in the identification of co-regulated genes through shared TFs is the accurate prediction of TFBSs (Section 1.2). A large number of TFBS prediction tools exist, which use a variety of different methods, and the field is continually growing (Section 1.3.2). One of the simplest groups of TFBS prediction tools use PWMs, or other pattern matching techniques, to search a DNA sequence for motifs contained in databases such as TRANSFAC (Matys *et al.*, 2006) or JASPAR (Bryne *et al.*, 2008).

One problem that commonly occurs with pattern matching TFBS prediction algorithms is that in order to achieve a high level of sensitivity the level of specificity is low. This results in a large number of false positive predictions. It is a common approach in many fields to combine different prediction algorithms and produce a consensus method with a higher rate of precision than the individual algorithms. The consensus method assumes that a prediction made by more than one algorithm is likely to be more accurate than one made by a single algorithm. Such an approach has been used in the prediction of structural domains and secondary structure in proteins (Frousios *et al.*, 2009; Kumar & Carugo, 2008; Vilar *et al.*, 2009). The integration of multiple databases or methods has also been used for TFBS prediction. PromAn is a tool developed for promoter analysis which integrates a phylogenetic footprinting method with multiple database sources to search for TFBSs (Lardenois *et al.*, 2006). However, what is lacking in this field is the systematic bench marking of algorithms to evaluate their reliablity and potential applications. Many programs have been tested on extremely small and highly specific datasets, but when scaled up to large gene datasets actually perform very poorly. In addition many TFBS prediction algorithms have restrictions on the length or number of sequences that can be processed which limits their application to large gene datasets. The research described in this chapter addresses this issue.

2.1.1. Algorithms to be compared

Seven algorithms, in the pattern matching group of TFBS prediction tools, were benchmarked and then used to create a consensus TFBS prediction method (Table 2.1). These algorithms were: (I) Alibaba 2.1 (Grabe, 2000), (II) Match (Kel *et al*., 2003), (III) MatInspector (Quandt *et al*., 1995), (IV) Patch (Striepe & Goessling, 2009), (V) P-Match (Chekmenev *et al*., 2005), (VI) TFsearch (Akiyama, 1998) and (VII) the forkhead TFBS perl modules (Lenhard & Wasserman, 2002). The algorithms use pattern matching techniques, primarily PWMs, to search TRANSFAC, JASPAR or an in house database of a similar type.

(I) AliBaba 2.1

AliBaba 2.1 was created as an alternative to the programs that rely on predefined matrices (Grabe, 2000). With programs which rely on predefined matrices it is difficult to make generalized statements about how accurate the predictions are; some matrices may be very specific, while others may have a much weaker specificity. Therefore, AliBaba 2.1 constructs context specific matrices for each sequence to be analysed. AliBaba 2.1 is based on the theory of Berg and von Hippel (Berg & von Hippel, 1987) which calculates how much energy is needed to bind a TF to DNA. The less conservation there is between the DNA sequence and the consensus sequence, the more energy is needed to bind the TF (Grabe, 2000). In an unpublished analysis, details of which are available on the AliBaba 2.1 website (Grabe, 2000), Alibaba 2.1 had a higher sensitivity and a higher sensitivity/specificity ratio than MatInspector, a commercially available algorithm also included in this analysis.

(II) Match

Match is a PWM based search program that uses the information contained in the Transfac database (Matys *et al.*, 2006) to search for likely TFBSs (Kel *et al*., 2003). Two scores are used to assess similarity between a prediction and a known TFBS matrix, the Matrix Similarity Score (MSS) and the Core Similarity Score (CSS). These scores range from 0.0 to 1.0, where 1.0 denotes an exact match and 0.0 denotes an incorrect match. The scores are calculated in the same way, but MSS analyses the whole sequence, whereas CSS analyses only the core sequence. The core sequence is defined as the first 5 conserved positions in the matrix. A TFBS prediction is only included if it has both an MSS and a CSS score higher than the threshold value, either the default or a user defined value (Kel *et al.*, 2003).

Three different threshold values are recommended for Match (Kel *et al*., 2003). The first is designed to minimise false negative results. The threhold is set so that 90% of correct results in the data set are found. An error rate of 10% is tolerated to take into account that the data used to determine the threshold values might contain weak representatives. This set of threshold values is designated minFN. The second score minimises false positive results. This is achieved by setting the score at a value where no TFBSs were predicted in sequences which are either known or assumed to contain no functional TFBSs. This set of threshold values is designated minFP. The final score, and default setting, is set to minimise the sum of both errors. This set of threshold values is designated minSum (Kel *et al.*, 2003).

There are two versions of Match. The first version of Match searches the public version of TRANSFAC, the second version searches the professional version and can only be accessed though a subscription to the professional database (Matys *et al.*, 2006). TRANSFAC Professional is regularly updated; the current version (2009.2) contains 892 matrices alongside information about 12,443 transcription factors and 37,255 genes. Transfac Public (version 7) was last updated in 2005, and contains 398 matrices alongside information about 6133 transcription factors and 2397 genes. Therefore, for predictions which represent the current knowledge of TFBSs, the professional database is superior.

(III) MatInspector

MatInspector (Quandt *et al.*, 1995) is another PWM weight matrix based search program. MatInspector uses both a CSS and MSS score, similar to Match, but the MSS score is only calculated if the CSS score exceeds a given threshold value. This pre-selection step reduces the total number of matches and increases the performance of the algorithm (Quandt *et al.*, 1995). A final difference between MatInspector and Match is that MatInspector uses a family concept to group matrices. This process reduces the number of redundant matches by only showing a match to a family of matrices rather than including matches to all similar matrices. MatInspector uses 634 matrices constructed by using information from the public Transfac database. This is the largest number of matrices available free of charge to the academic community (Cartharius *et al.*, 2005). The free version of MatInspector is restrictive in that each search is limited to 5000 TFBS predictions, and a user is only permitted 20 searches a month. However, a licensed version removes these restrictions.

(IV) Patch

Patch (Striepe & Goessling, 2009) is a pattern based search program, which, like Match, uses PWMs from the TRANSFAC Professional database to predict likely transcription factor binding sites. Unfortunately, no further literature exists to provide more details on the method.

(V) P-Match

P-Match (Checkmenev et al., 2005), as its name suggests, is a combination of the pattern matching method used in Patch and the PWM searching method used in Match. This combination of methods provides a higher level of accuracy than either method separately. P-Match searches PWM and binding site alignments found in TRANSFAC professional version 6.0, which while not the current version of TRANSFAC is far more up to date than that used by many other programs, e.g. TFSearch.

(VI) TFSearch

TFSearch searches sequence fragments against the 2,285 matrices found in the TFFACTOR section of the public TRANSFAC database version 3.3 (Akiyama, 1998) which was released in January 1998. In the current version of TRANSFAC Public, version 7.0, there are 6,133 entries in the TFFACTOR table of the database. TFSearch, therefore, is only able to search for a subset of the known TFBSs in TRANSFAC.

(VII) TFBS perl modules

The TFBS perl modules (Lenhard & Wasserman, 2002) are a set of integrated, object oriented perl modules designed to enable TFBS detection and analysis. The modules allow for PWM creation or retrieval from either a local database or a database available through the internet. The algorithm allows for the detection of putative TFBSs in a user given sequence via these PWM. Unlike the other TFBS prediction methods which are all based on external servers (and therefore must limit lengths or numbers of sequences which can be processed) the TFBS perl modules can be run locally. Hence the algorithm is far more flexible. The TFBS perl modules can also be used for phylogentic footprinting methods of TFBS detection, however in this analysis they were used without this additional capability.

The majorty of the algorithms assessed rely on PWM to predict TFBS. However, PWMs are not uniform, some allow much more degeneracy in the TFBS motif than others, some may contain information confirmed using less accurate techniques than others and the frequency of occurrence of motifs matching the PWM across the genome can be widely divergent. The algorithms do not all search using the same set of PWMs. This may affect the performance of the algorithms, as the PWMs used in one algorithm may be more specific than others, causing fewer false positive results. It is possible that bias is introduced to the analysis because of these factors. However, because may of the algorithms have not been published, it is not possible to determine the PWM datasets used to train the algorithms. Therefore we have assumed, for the purposes of this analysis, that the PWM sets for each algorithm are equally reliable. It is, however, important to remember that this may not be the case.

In the current chapter, the 7 TFBS prediction algorithms (Table 2.1) were benchmarked on a dataset of 151 genes extracted from TRANSFAC Public version 7.0 (Matys *et al.*, 2006). The performance of pairs of TFBS prediction algorithms were then evaluated to assess if combinations of methods gave results that had higher rates of precision, sensitivity and specificity than the algorithms separately. The highest performing combined methods were then applied to data from a genetic linkage study.

		Matching System	***Database Searched***	***Scoring System***	***Reference***
AliBaba 2.1	**(I)**	PWM	TRANSFAC Public v.7.0	No score	(Grabe, 2000)
Match	**(II)**	PWM	TRANSFAC Public v.7.0	MSS & CSS	(Kel *et al.*, 2003)
MatInspector	**(III)**	PWM	Internal	MSS & CSS	(Quandt *et al.*, 1995)
Patch	**(IV)**	PBS	TRANSFAC Public v.7.0	SS	(Striepe & Goessling, 2009)
P-Match	**(V)**	PWM & PBS	TRANSFAC Public v.7.0	MSS & CSS	
TFSearch	**(VI)**	PWM	TRANSFAC Public 3.3	SS	(Akiyama, 1998)
TFBS modules	**(VII)**	PWM	TRANSFAC Public v.7.0	SS	(Lenhard & Wasserman, 2002)

Table 2.1. Details of the 7 algorithms assessed. Matching system refers to whether the algorithms uses a Position Weight Matrix (PWM) or some other Pattern Based Searching mechanism (PBS). Only MatInspector uses a database other than TRANSFAC (public version 7), using an internal database instead. The scoring systems used by the algorithms fall into three categories; Matrix similarity scores (MSS), Core similarity scores (CSS) and other types of similarity scoring systems (SS).

2.1.2. Application to genetic linkage data

Three chromosomal regions, containing a total of 532 genes, have been identified in a genetic linkage study. The study was designed to find genes linked to the IFN-γ response to mycobacterial antigens following BCG vaccination (Section 1.9) (Newport *et al.*, 2004). It has been shown that genes with similar functions are likely to be regulated by the same, or similar mechanisms, and are likely to share regulatory TFs (Section 1.2) (Allocco *et al.*, 2004). The 3 chromosomal regions contain genes that are linked to the IFN-γ immune response but also large numbers of genes that have no functional role in the immune response. The working hypothesis was that the genes linked to the IFN-γ immune response will have similar transcription factor binding profiles. Hence, in order to identify these IFN-gamma linked genes, TFBSs would be predicted for each gene. The TFs that bind to predicted TFBSs in at least one gene from each of the three chromosomal regions, using combined pairs of algorithms, would be identified. The genes these shared TFs bind would then be investigated further to assess their potential functional links to the IFN-gamma response.

2.2. Methods

2.2.1. Dataset of genes for benchmarking

Information retrieved from the TRANSFAC database was used as a gold standard to measure the performance of the prediction algorithms. TRANSFAC is useful as a gold standard because the data, peer-reviewed and experimentally validated, can be assumed to be correct with a high degree of confidence. There are however problems with the use of TRANSFAC as a gold standard. The nature of curated data is that it is slow to accumulate; the input of new TFBS motifs relies on the publication of data and for the publication to be found and inputted by the curators. This naturally leads to the existence of a large number of false negatives – TFBS which cannot be found in the data. A further issue with the use of TRANSFAC is that the majority of the algorithms to be analysed were originally trained and tested using TRANSFAC data. However, regardless of these flaws, TRANSFAC is the largest most reliable dataset available for TFBS data and therefore is used as the gold standard in this study.

The distance from the start site of all of the TFBSs in the public TRANSFAC version 7.0 (Matys *et al*., 2006) was analysed to determine the optimum length of sequence in which to predict the TFBSs. It was determined that the majority of TFBSs could be found within 700bp in either direction from the transcription start site (Figure 2.1). Although the histogram in Figure 2.1 is not a normal curve, showing higher frequencies for the upstream (negative) TFBSs, it was decided to use the same length of sequence both up- and downstream from the transcription start site to adjust for a any possible bias in the database caused by more experimental determination of TFBSs occurring in the upstream promoter region (Collins & Hu, 2007; Jin *et al.*, 2007; Oh *et al.*, 2009; Xia *et al.*, 2008).

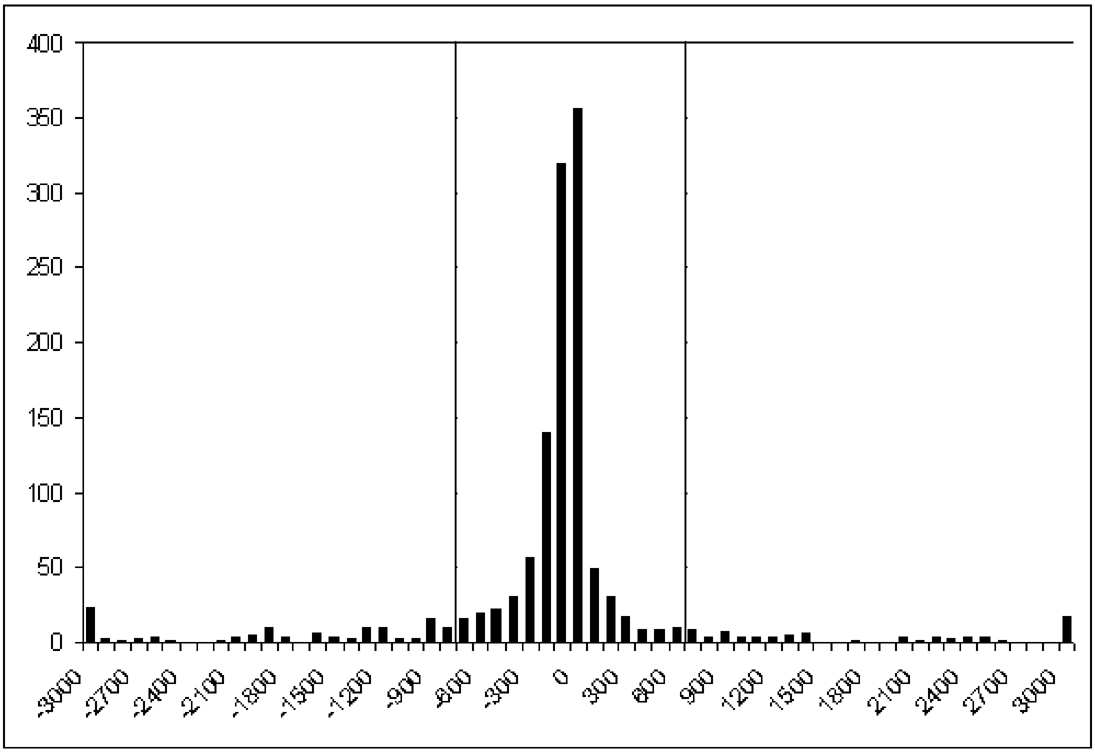


Figure 2.1. Histogram showing the distribution of start sites (basepairs) for the TFBSs found in Public Transfac version 39.

All TFBSs in public TRANSFAC, which occurred within 700bp up- or downstream of the TSS, were retrieved. TFBSs without start or end positions, information about which TF the motif binds to, or the sequence of the motif, were removed from the set. The 700bp sequences up- and downstream of the TSS, for each of the genes with TFBSs reported, were retrieved from Ensembl version 40 (Hubbard *et al.*, 2007) via Biomart (Smedley *et al.*, 2009; Durinck *et al.*, 2005). Any TFBSs that was not found in the reported position was removed from the dataset. This resulted in a dataset of 475 TFBSs located in 151 genes; designated Transfac_public_475.

2.2.2. Benchmarking the TFBS prediction algorithms

TFBSs were predicted in the 1400bp sequences for the 151 genes in the Transfac_public_475 dataset using each of the seven algorithms described in section 2.1.1: AliBaba 2.1, Match, MatInspector, Patch, P-Match, TFSearch and the TFBS perl modules.

The TFBS predictions made by each algorithm were compared against the location of known functional TFBSs in the Transfac_public_475 dataset. Each prediction was classified as true positive (TP), false positive (FP), true negative (TN) or false negative (FN). A TP was defined as a prediction which overlapped >80% of a Transfac_public_475 TFBS motif for the same TF and had an associated score above the given threshold value. A TN was defined as a prediction given by the algorithm which did not match a motif in the Transfac_public_475 set and which had an associated score below the given threshold. The assumption was made that the Transfac_public_475 dataset contained all TFBSs for the set of genes, therefore any TFBS predicted by the algorithm which were not classified as TP were assumed to be a FP. FNs are predictions matching functional TFBSs in the Transfac_public_475 dataset with scores below the threshold, not those TFBSs from Transfac_public_475 which were not predicted by the algorithm. The percentage of TFBSs from Transfac_public_475 which were predicted is measured as a separate statistic. Receiver Operating Characteristic (ROC) curves (Hanley & McNeil, 1983; Hanley & McNeil, 1982) were made for each algorithm with the exception of Alibaba 2.1 which did not have an associated scoring system. A ROC curve is constructed by plotting the sensitivity (equation 2.1) on the y axis and 1-specificity (equation 2.2) on the x axis.

Sensitivity = TP / (TP + FN) **(2.1)**

Specificity = TN / (TN + FP) **(2.2)**

The area under the ROC curve (AUC) statistic is used to measure the performance of the algorithm. It is based on the assumption of a binormal distribution of the data. A high AUC value indicates that the frequency curves consisting of either the positive or negative data have no, or only a small, overlap (Figure 2.2 A & B) and therefore by setting the correct threshold value both a high sensitivity and specificity is achieved. If the AUC value is low, between 0.5 and 0.7, the negative and positive results are not successfully split by the scoring system (Figure 2.2 C), and the threshold can be set to have either a high sensitivity level or a high specificity level, but not both (Hanley & McNeil, 1983; Hanley & McNeil, 1982).

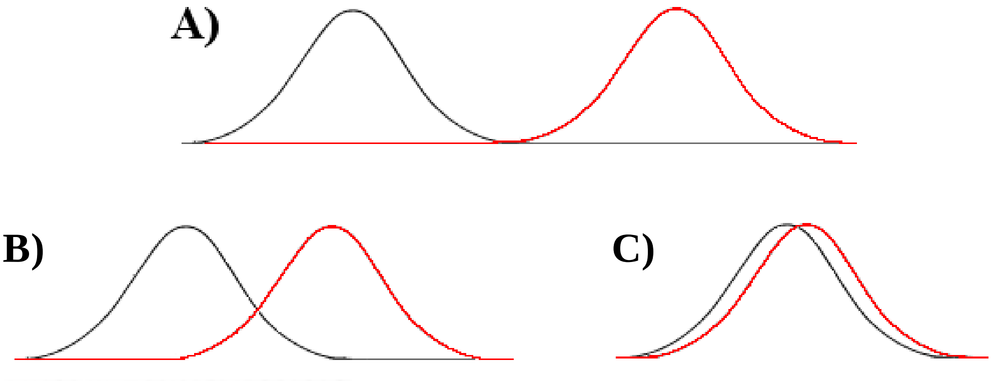


Figure 2.2. Frequency curves for positive data (red) and negative data (black) plotted against a hyperthetical scoring system (x axis). A) The curves underlying a ROC curve with AUC=1. B) The curves underlying a ROC curve with AUC ≈ 0.8. C) The curves underlying a ROC curve with AUC ≈ 0.6.

ROC curves were constructed using the statistical software package MedCalc (version 9.2.1.0) (Schoonjans *et al.*, 1995). The software uses the Hanley and McNeil method (Hanley & McNeil, 1983; Hanley & McNeil, 1982) for determining the significance of the difference between the AUC values for each curve. This method relies on the similarity between the ROC curve and the wilcoxin statistic to derive the standard error of the curve, from which the standard error of the difference between the AUC values can be calculated (equation 2.3) (Hanley & McNeil, 1982). However, the difference between the two curves must be calculated differently when comparing scoring systems for the same dataset than when using different datasets (equation 2.4) (Hanley & McNeil, 1982). The Z statistic is used to compare the differences between the AUC values of the two curves (equation 2.5) (Hanley & McNeil, 1983).

$$SE(Area_1 - Area_2) = \sqrt{SE^2(Area_1) + SE^2(Area_2)} \qquad \textbf{(2.3)}$$

$$SE(Area_1 - Area_2) = \sqrt{SE^2(Area_1) + SE^2(Area_2) - 2rSE(Area_1)SE(Area_2)} \quad \textbf{(2.4)}$$

$$z = \frac{A_1 - A_2}{SE(Area_1 - Area_2)} \qquad \textbf{(2.5)}$$

The optimum threshold value was calculated from the ROC curves by determining the score at which the distance between the curve and the point (0,1) was the smallest. This point represents the best possible classification of the data into positive and negative sets (Hanley & McNeil, 1983; Hanley & McNeil, 1982). The precision value (equation 2.6) at each threshold value was also calculated and a second threshold value was determined using this criteria.

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \qquad \textbf{(2.6)}$$

2.2.3. Pairwise consensus prediction method

The predictions from the algorithms were combined in a pairwise manner. For each pair of algorithms the TFBS predictions from each algorithm were collated and exact matches were removed to create a non-redundant set of predictions (Figure 2.3). This was done to prevent the results in the intersect from being counted twice in the analysis.

The pairwise prediction dataset was compared to the Transfac_public_475 dataset. Predictions which have a >80% overlap with a known TFBS in the Transfac_public_475 dataset, and which predicted the same TF, were assigned as functional predictions (TP or FN). Predictions which did not meet these criteria were designated non-functional predictions (FP or TN).

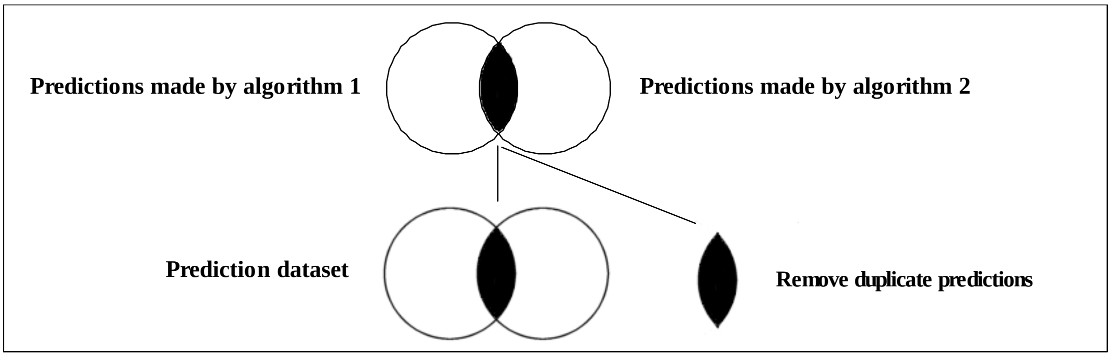

Predictions made by algorithm 1
Predictions made by algorithm 2
Prediction dataset
Remove duplicate predictions

Figure 2.3. Creation of the pairwise prediction data set. One set of predictions made by both algorithms 1 and 2 was removed to create a non-redundant dataset.

The pairwise prediction dataset was compared against the predictions made by each of the two algorithms (Figure 2.4A). Each prediction in the non-redundant set was assigned a score for each of the two algorithms, normalised to a range from 0 to 1. If the prediction from the pairwise dataset overlapped a prediction made by the algorithm by >80%, and predicted the same transcription factor, the score given by the algorithm was assigned to the prediction (Figure 2.4A). If more than one prediction made by the algorithm matched the prediction from the pairwise dataset, the highest score was used. This can occur if more than one matrix predicts the TFBS, or if overlapping TFBSs are predicted due to repeating nucleotide sequences. If there was no matching algorithm prediction the score was set to 0. After each prediction was assigned a score from each of the two algorithms, a final score was assigned by taking the mean of the two algorithm scores. This resulted in scores between 0 and 0.5 for all TFBSs predicted by only one of the algorithms and scores ≤1 for TFBSs predicted by both algorithms (Figure 2.4B).

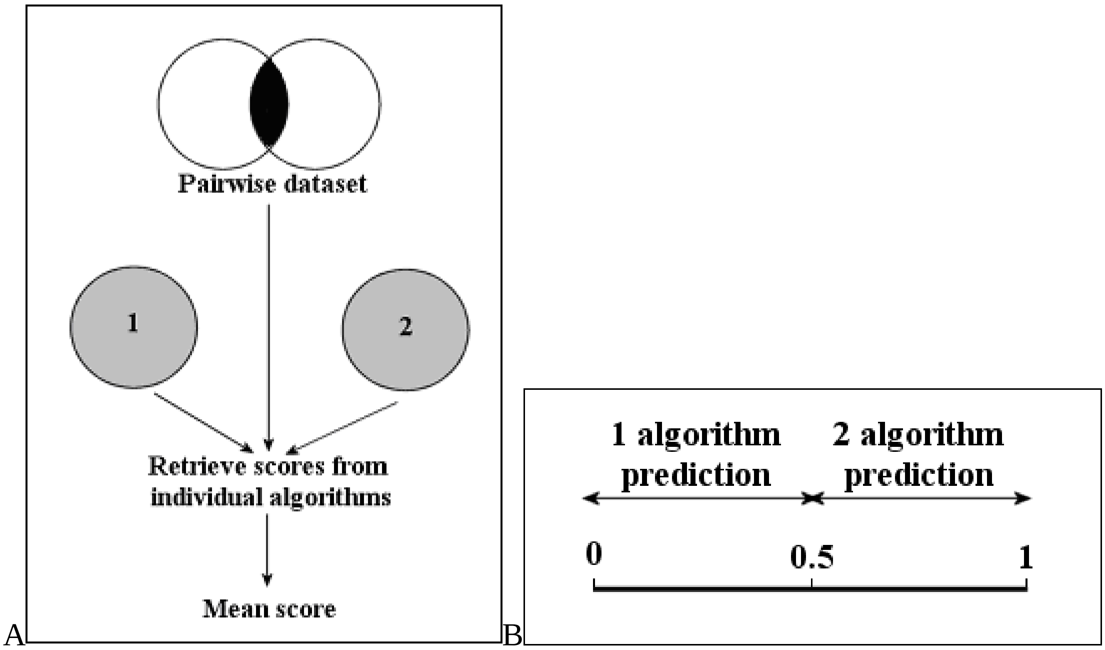


Figure 2.4. (A) Calculating the scores for each algorithm in the dataset. Each prediction in the pairwise dataset was compared to the original algorithm 1 and algorithm 2 data and if a match existed the score for the TFBS was retrieved. If no score existed a zero was retrieved. The final score for the predicted TFBS was the mean of these two scores. (B) The range of scores achieved for TFBSs predicted by either one or both of the algorithms. TFBS predictions made by only one algorithm were

assigned scores in the region 0 to 0.5, scores in the region 0.5 to 1 were only assigned to TFBS predictions made by both algorithms.

ROC curves were constructed for each pair of algorithms using the final scores for each prediction in the pairwise prediction set. The line y=x with an AUC of 0.5 is considered to represent a random classification. Scoring systems with an AUC of ≥0.7 are considered to be successful classifiers (Hanley & McNeil, 1983; Hanley & McNeil, 1982). The smallest distance between a ROC curve and the optimum classification point at (0,1) is a statistic that evaluates the compromise between sensitivity and specificity. The threshold value was calculated by determining the closest point to (1,0), the perfect classification, on the ROC curve, as in section 2.2.2 (Hanley & McNeil, 1983; Hanley & McNeil, 1982). Precision scores were also calculated and an alternate threshold value was determined by the score at which the precision was the highest.

2.2.4. Applying the algorithm to the IFN-γ linked genes

700bp up- and downstream of the TSS were retrieved for each of the 532 genes in the genetic linkage analysis regions from chromosomes 8, 10, 11 (Section 1.9). TFBS predictions were made for these sequences using the three consensus prediction algorithms with the highest performance. TFBS predictions with scores above the threshold values calculated in section 2.3.2 were defined as correct predictions.

The set of genes predicted to be regulated by TFs binding TFBSs in at least one gene from each chromosomal location were investigated further using the Gene Ontology (GO) (Ashburner *et al.*, 2000) and text-mining techniques. Biological process (BP) terms from GO were used to assess the processes and mechanisms that each gene was involved with. The set of BP GO terms for each gene was retrieved from the Ensembl annotations via BioMart. These are known to be slightly different to those retrieved through AmiGO, the Gene ontology browser. However AmiGO does not have functionality allowing for batch download. The set of BP GO annotations for each gene was compared to the set of annotations for every other gene in the set. Path specific relative similarity statistic (PSRSS) scores were used to determine the similarity between the sets of terms for each gene. The scores were calculated using a set of Java classes developed by David Damerell (personal communication). Pairs of genes with PSRSS score <0.4 were assumed to be very functionally similar.

Text mining was carried out using the web based tool PubGene (Jenssen, *et al.*, 2001). PubGene automatically creates graphs showing the relationship in the text between genes, proteins, or other biological concepts such as a disease, the edges represent a co-occurrence in an abstract, in the literature, between the two concepts represented by the nodes. Direct connections were searched for between each of the genes in the dataset. This search was then expanded to allow a connection between two genes to have a linked node, i.e. another gene that both genes were found to have a co-occurrence with in the literature.

2.3. Results

2.3.1. Benchmarking the TFBS prediction algorithms

The AUC values (Table 2.2b) for the ROC curves constructed using each algorithm varied from 0.54 for MatInspector, to 0.7 for Match and the TFBS perl modules. Precision rates were very low (<10%) for all algorithms, including those with AUC values greater than 0.7. The TFBS perl modules had the highest overall performance with the joint highest AUC value (Table 2.2), the highest precision value and the second highest percentage of known TFBSs found. The AUC value for Patch and MatInpsector was 0.54, indicating a performance barely better than random. The reason for the poor performance of Patch may be the much higher number of predicted TFBSs per sequence found using this algorithm than is found using other algorithms (Table 2.2). Equally, MatInspector is designed to have a high sensitivity level, predicting as many known TFBSs as possible (Quandt *et al*., 1995). This aim can be seen to be achieved by the high percentage of TFBSs found (Table 2.2c). However with high sensitivity levels comes a large number of FP predictions and thus low specificity levels, reducing the AUC.

		(a)	***(b)***	***(c)***	***(d)***	***(e)***	***(f)***
N^o	***Algorithm***	***Mean n^o TFBSs predicted per sequence***	***Area Under Curve (AUC)***	***% TFBSs found***	***Best Precision %***	***Best Threshold Score : mean AUC***	***Best Threshold Score : precision***
2	Match	100	0.70	27.8	1.3	0.98	0.80
7	TFBS modules	395	0.70	69.9	8.5	7.7	14.9
5	P-Match	20	0.60	14.3	4.8	0.99	1.00
6	TFSearch	107	0.55	34.3	1.9	89.0	98.0
4	Patch	1954	0.54	42.7	0.7	90.3	87.5
3	MatInspector	252	0.54	88.4	2.7	0.87	1.00
1	AliBaba 2.1	153	N/A	39.0	N/A	N/A	N/A

Table 2.2. Results for the analysis of the TFBS prediction algorithms. N/A = not applicable to this algorithm.

Precision values were low for all of the algorithms. This was due to the large number of FP predictions made by each algorithm. By combining the results from pairs of algorithms it was hoped that the precision of the predictions could be increased.

2.3.2. Pairwise consensus prediction method

In seven of the paired combinations the AUC values show an improvement over using either of the algorithms separately. Four pairs of algorithms achieved AUC values higher than 0.70. These pairs were: Match with MatInspector, the TFBS modules with Patch, TFSearch with Patch and TFsearch with the TFBS modules (Table 2.3). One of these pairs (TFSearch with Patch) achieved an AUC of >0.80.

	Match	*MatInspector*	*Patch*	*P-Match*	*TFBS*	*TFSearch*
Match	-	**0.71**	0.54	0.56	0.57	0.47
MatInspector	-	-	**0.63**	**0.65**	0.48	0.52
Patch	-	-	-	**0.61**	**0.79**	**0.81**
P-Match	-	-	-	-	0.66	0.57
TFBS	-	-	-	-	-	**0.79**
TFSearch	-	-	-	-	-	-

Table 2.3. Empirical AUC values for ROC curves created from the combination of two TFBS prediction algorithms. Values in bold represent AUCs which are higher than those achieved by the constituent algorithms separately.

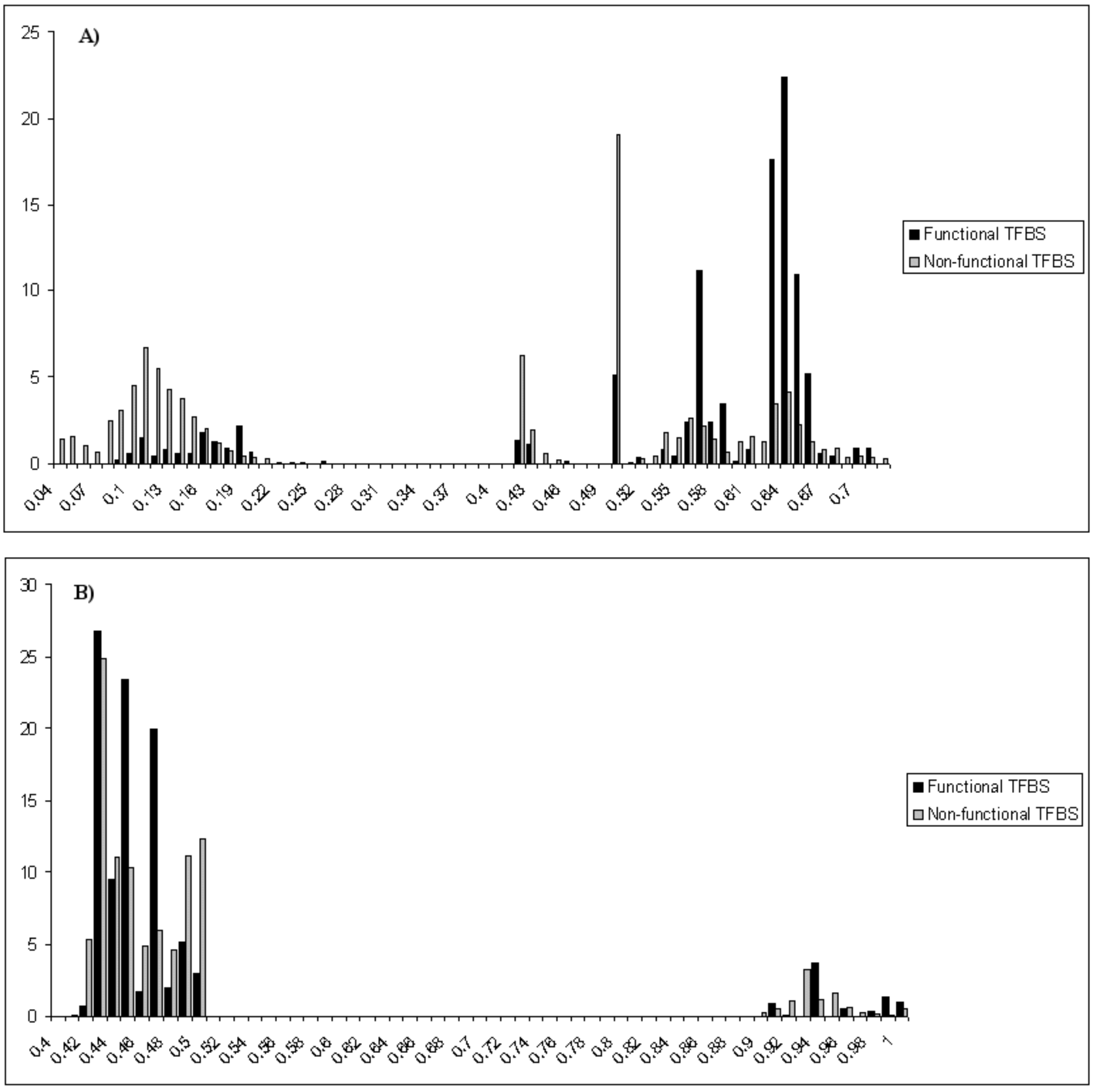


Figure 2.5. Graphs showing the percentage of functional TFBSs and non-functional TFBSs predicted at each score. A) The results from combining Patch and TFSearch, the combination with the highest AUC of 0.81. B) The results from combining TFSearch and Match, the combination with the lowest AUC of 0.47.

Pairs of algorithms with high AUC values, i.e. those with AUC values >0.70, all gave bimodal graphs when the percentage of functional and non-functional TFBSs predicted at each score threshold was plotted (Figure 2.5A). This confirmed that the combined algorithm was assigning high scores to functional TFBSs and low scores to non-functional TFBSs. Combinations with low AUC values did not show the same distinction between the scores at which functional and non-functional TFBSs were predicted (Figure 2.5B). The combination of Match and TFSearch (Figure 2.2B) predicted the majority of both functional and non-functional TFBSs at low scores; this is because, in the majority of cases, there was little overlap between the results predicted by the two algorithms.

The AUC statistic gave a good overall evaluation of the scoring systems for each pair of algorithms. The closest point to (0,1) on the x=y line designating a random classification had a distance of 0.71. Hence, the distance between the threshold point and (0,1) must be less than 0.71 if the distribution is to be considered better than random. On this basis all but one of the pairs of algorithms (Match with P-Match) had a better than random performance. The pairs of algorithms with the highest difference between the threshold value and the optimum value were: TFSearch with Patch at a distance of 0.32, TFsearch with TFBS at a distance of 0.32 and TFBS with Patch at a distance of 0.31 (Table 2.4).

	Match	*MatInspector*	*Patch*	*P-Match*	*TFBS*	*TFSearch*
Match	-	0.42 (0.85)	0.67 (0.50)	0.75 (0.98)	0.60 (0.14)	0.69 (0.94)
MatInspector	-	-	0.51 (0.50)	0.55 (0.48)	0.70 (0.14)	0.65 (0.45)
Patch	-	-	-	0.63 (0.50)	0.31 (0.57)	0.32 (0.88)
P-Match	-	-	-	-	0.50 (0.14)	0.59 (0.45)
TFBS	-	-	-	-	-	0.32 (0.57)
TFSearch	-	-	-	-	-	-

Table 2.4. Distance from (0,1) to the closest point on ROC curves created from a combined scoring system between pairs of TFBS prediction algorithms. Optimum threshold values are shown in parenthesis.

	Match	*MatInspector*	*Patch*	*P-Match*	*TFBS*	*TFSearch*
Match	-	0.96 (0.94)	0.94 (0.50)	3.47 (0.98)	2.44 (0.71)	0.90 (0.45)
MatInspector	-	-	1.18 (0.88)	2.44 (0.95)	1.82 (0.66)	1.22 (0.95)
Patch	-	-	-	2.21 (0.99)	2.36 (0.63)	4.73 (0.97)
P-Match	-	-	-	-	4.05 (0.67)	3.75 (0.98)
TFBS	-	-	-	-	-	2.40 (0.63)
TFSearch	-	-	-	-	-	-

Table 2.5. Precision percentage scores for each combination of paired algorithms. The threshold value at which this precision occurs is displayed in parenthesis.

Combining the results from two algorithms did not, in the majority of cases, improve the precision (Table 2.5). This was most likely to be due to an increase in both positive and negative results. It is interesting that the two threshold values, one based on the nearest point to (0,1) (Table 2.4), and one based on the highest precision value (Table 2.5), were rarely found to be the same for any pair of algorithms. This shows that depending on the type of information the user wishes to extract from the prediction algorithm it may be useful to vary the threshold value. For example, if the user wants predictions which are predominantly TPs, the precision based threshold would be the most useful. Conversely, if the user wants to find all possible TPs but still limit the number of false positives retrieved, the threshold based on the distance from (0,1) would be the most useful.

2.3.3. Applying the algorithm to the genes from the IFN-γ genetic linkage study

The three paired algorithms with the highest AUC values and the closest point to (0,1) were: TFSearch with TFBS (TFS_TFBS), TFSearch with Patch (TFS_P) and TFBS with Patch (TFBS_P). Each paired method was used to make predictions for the 532 genes in the datasets from the genetic linkage studies. TFS_P predicted an average of 162.4 TFBSs per gene when using a threshold value of >0.88. TFS_TFBS and TFBS_P both gave a smaller number of predictions (Table 2.6.) with an average of 7.9 and 14.2 TFBSs per gene respectively.

Method	*Total TFBSs predicted for each Chromosome*			*Mean TFBSs / gene*
	Chr 8	Chr 10	Chr 11	All
TFS_TFBS	1227	1688	1298	7.9
TFS_P	25914	36690	23809	162.4
TFBS_P	2064	3035	2429	14.2

Table 2.6. TFBS predictions, given by the three best pairs of algorithms, for the genes from the IFN-γ linked regions.

Method	*Total TFs predicted for each Chromosome*			*Total TFs*
	Chr 8	Chr 10	Chr 11	
TFS_TFBS	16	16	16	16
TFS_P	28	31	31	32
TFBS_P	10	11	11	11

Table 2.7. Numbers of transcription factors (TFs), predicted by each of the three best pairs of algorithms, in each chromosomal region in the dataset.

The TFS_TFBS pairing predicted TFBSs for 16 different transcription factors. All 16 transcription factors had a corresponding TFBS in at least one gene from each of the three regions (Table 2.7). The TFBS_P pairing predicted TFBSs for 11 different transcription factors, 10 of these were found in genes from all 3 regions, 1 (p53) was found only in genes from the regions in chromosomes 10 and 11 (Table 2.7). The TFS_P pairing found TFBSs for 27 different transcription factors in all three regions and TFBSs for 32 unique transcription factors in total (Table 2.7). Only TFs with predicted binding sites in each of the three regions were further analysed.

A large proportion of TFs, predicted to bind genes in all three regions, were only identified by one of the paired algorithms (Figure 2.6). For example TFS_P predicted 21 TFs not identified by TFBS_TFS or TFBS_P. These data indicate that more complex combinations of TFBS prediction algorithms may result in a method with even greater sensitivity.

The five TFs (Figure 2.6) predicted to bind to genes from each of the three chromosomal regions, by the three algorithm pairings, are: cAMP response element binding protein (CREB), TFs from the E2F family (with indistinguishable binding sites), hepatic nuclear factor 1 (HNF1), serum response factor (SRF) and upstream stimulatory factor (USF).

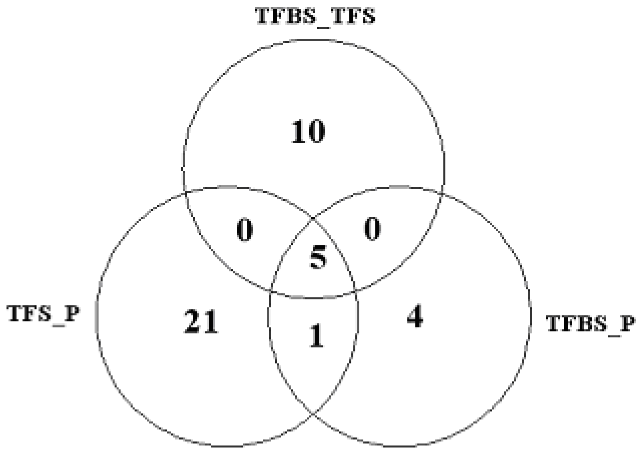


Figure 2.6. Venn diagram showing the number of transcription factors predicted to bind to at least one gene from each of the chromosomal regions (identified in the genetic linkage study) for the three pairs of algorithms. The three pairs of algorithms are those with the highest performance according to ROC curve statistics.

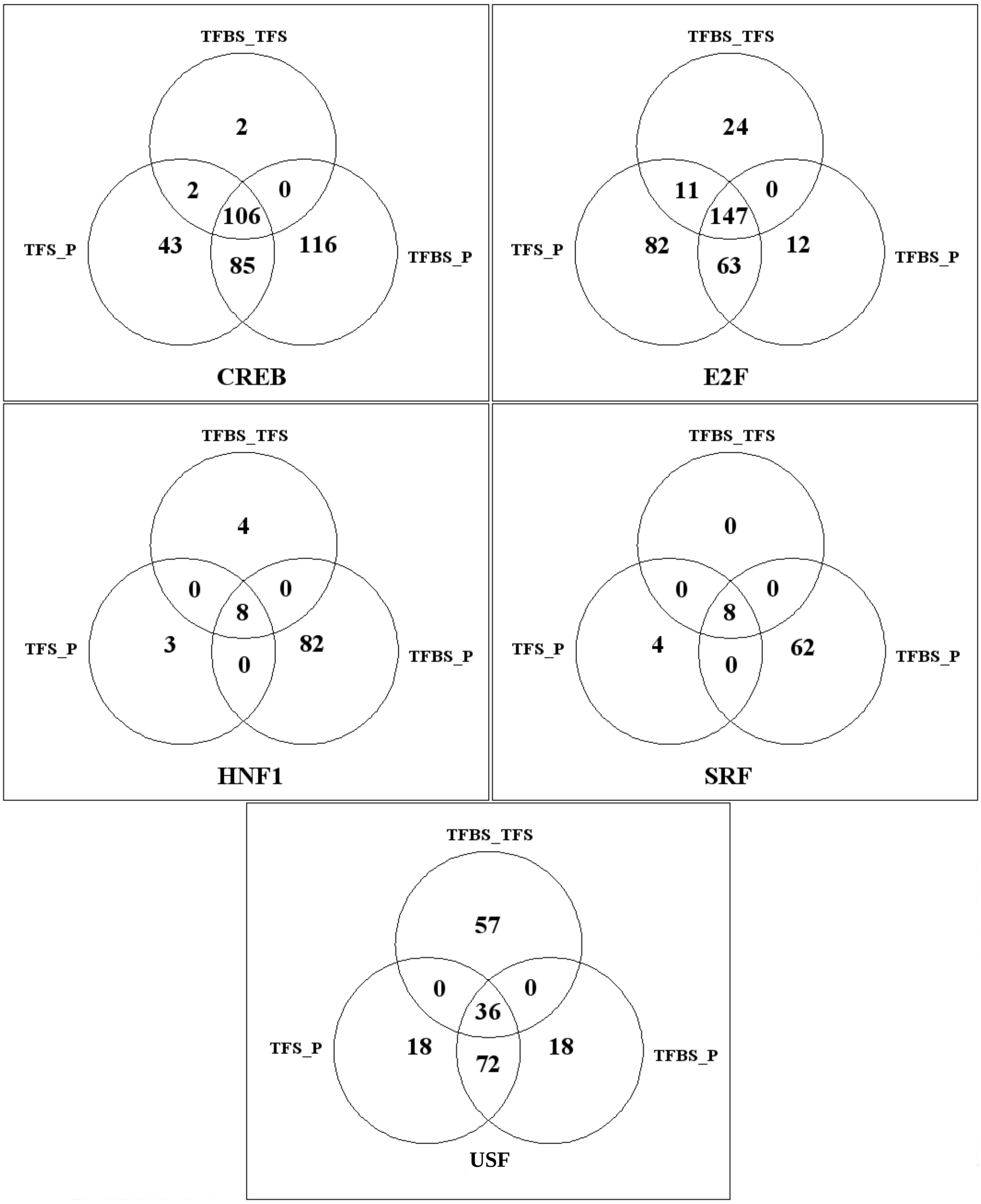


Figure 2.7. Venn diagrams showing the number of genes predicted by each of the paired methods to bind to a certain transcription factor. Venn diagrams were created for each transcription factor predicted, by all three methods, to bind to at least one gene in each of the three chromosomal regions.

Figure 2.7 shows the number of genes predicted to be regulated by the five TFs (common to all three chromosomal regions) for each of the 3 paired methods. The TFBS_TFS method showed a particularly large overlap with the other two paired algorithms; in most cases the majority of predictions made by TFBS_TFS were made by both of the other paired algorithms. In the results from three of the five transcription factors (CREB, E2F, USF) the TFS_Patch and TFBS_Patch algorithms showed a large overlap of predictions which are not predicted by the TFBS_TFS algorithms. In these cases the score from the Patch algorithm elevated the overall score above the threshold, whilst the TFBS_TFS predictions had a score lower than the threshold value.

Transcription Factor	*Total genes predicted*	*Genes with PSRSS<0.4*	*Immune Response*
CREB	106	73	39
E2F	147	96	64
HNF1	8	0	0
SRF	8	2	1
USF	36	21	10

Table 2.8. Total number of genes predicted to contain TFBSs, the number of these genes which have a PSRSS score <0.40 when paired with a gene from at least one of the other regions and the number of these genes to have a co-occurrence with the term ‘immune response’ detected by PubGene for each transcription factor,

151 genes were predicted by all three pairwise algorithms to have binding sites for at least one of the 5 TFs of interest. The functions of these genes were investigated. Genes with similar functions from different chromosomal regions were determined by a PSRSS score of <0.40. Of the genes with <0.40 PSRSS score, when paired with a gene

from one of the other chromosomes, those with text mined associations to the term 'immune response', were determined (Table 2.8). The genes which achieved these criteria were assumed to be the most likely candidate genes.

Graphs showing the connections between genes from different chromosome regions with PSRSS scores <0.40 could only be constructed for four of the TFs: CREB, E2F, SRF and USF (Figures 2.8, 2.10, 2.12). None of the genes found with HNF1 binding sites were functionally related to each other as determined by the PSRSS score. It was therefore assumed that HNF1 is not the regulatory TF in this system. The CREB and E2F graphs in particular show a large number of possible connections between genes from each of the chromosome regions, the USF graph shows five complete subgraphs containing three nodes, one from each of the regions, while the SRF graph contains only 2 genes and therefore cannot show connections between genes from all three regions. By looking for co-occurance in the literature of the genes with predicted connections, through Pubgene, the number of candidates was significantly reduced (Figures 2.9, 2.11 & 2.14).

Two complete 3 node subgraphs containing one node from each of the three chromosomal regions were found by looking for both a PSRSS score <0.40 and for co-occurrence of genes in the literature for CREB TFBS containing genes (Figures 2.8, 2.9). In both cases these subgraphs contain *RIPK2* (receptor-interacting serine/threonine kinase 2) and *NOX4* (NADPH oxidase 4), with the addition of *BLNK* (B-cell linker protein) in one case and *SMC3* (structural maintenance of chromosomes protein 3) in the other. The most promising of these is the *RIPK2*, *NOX4*, *BLNK* group as these genes are connected by a CREB TFBS, a PSRSS score <0.4 and co-occurrence in the literature found using PubGene. Both *RIPK2* and *BLNK* are also annotated with the GO term ‘immune response’. *RIPK2* is of particular interest given its role in nuclear factor kappa B activation which in turn can lead to transcription of the IFN-γ gene (Hawiger, 2001).

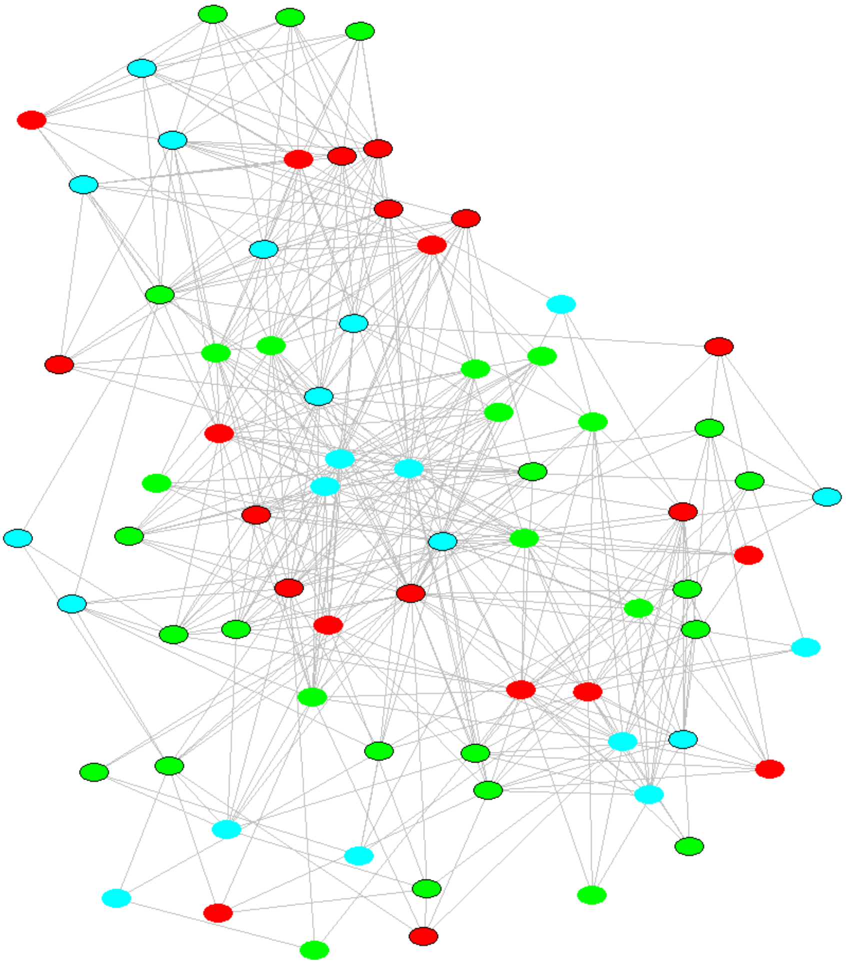


Figure 2.8. Graph showing the <0.4 PSRSS relations between genes with a predicted CREB binding site, from the different chromosomal regions (red represents genes from chromosome 8, green from chromosome 10 and cyan from chromosome 11). Genes with a co-occurrence in the literature, found using PubGene (Jenssen, *et al.*, 2001), with the term ‘immune response’ have a black border (see also figure 2.9).

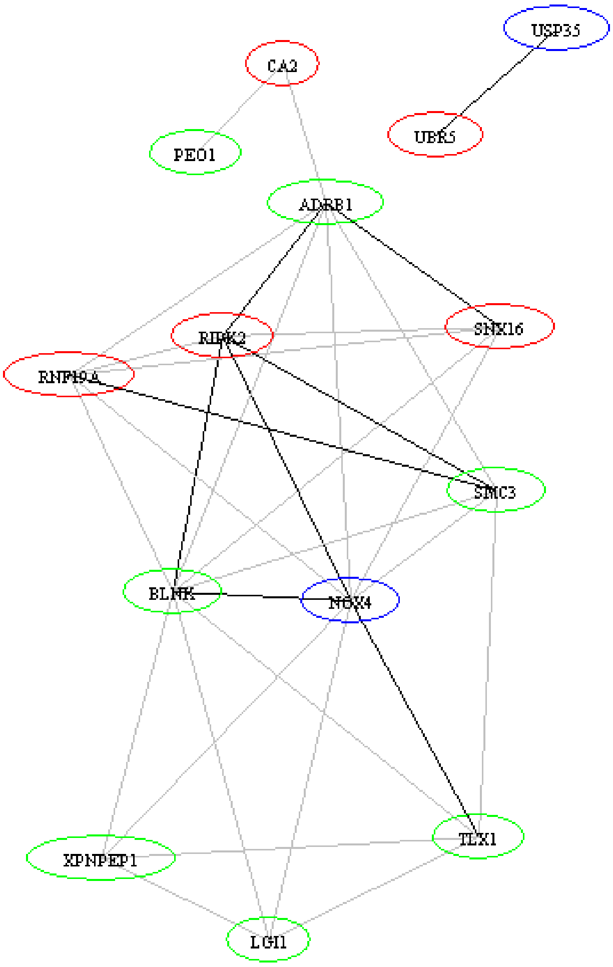


Figure 2.9. A graph showing the CREB binding site genes with a PSRSS score <0.4 with at least one other CREB binding site containing genes that have connections found through PubGene (Jenssen, *et al*., 2001). PubGene connections were allowed to either be direct connections between the genes, or have one gene connecting both genes. Black edges connect gene which have a PSRSS score <0.4 as well as a PubGene connection. Grey edges connect genes with a PubGene connection, but a PSRSS score ≥0.4. Red nodes represent genes from chromosome 8, green nodes from chromosome 10 and blue nodes from chromosome 11.

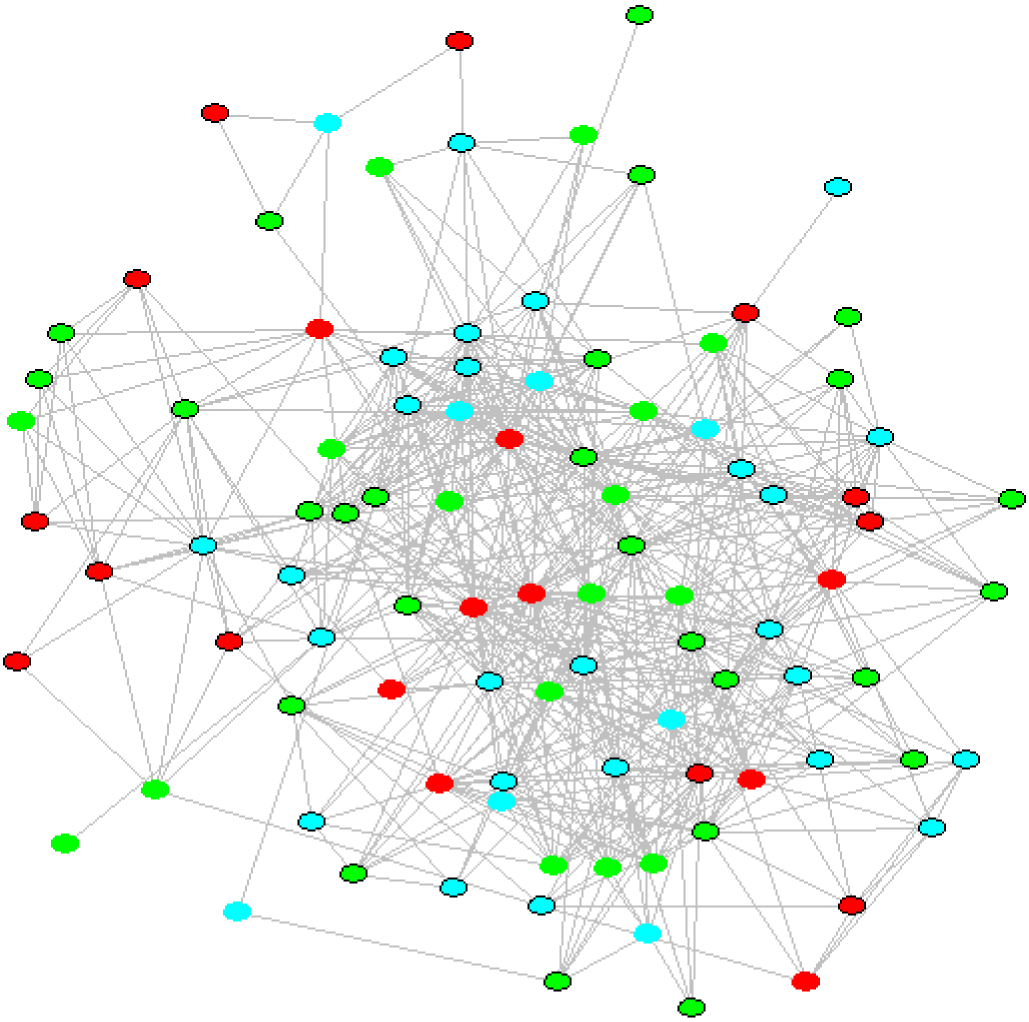


Figure 2.10. Graph showing the <0.4 PSRSS relations between genes with a predicted E2F binding site, from the different chromosomal regions. Genes with a co-occurrence in the literature, found using PubGene (Jenssen, *et al.*, 2001), with the term 'immune response' have a black border. Red nodes represent genes from chromosome 8, green nodes from chromosome 10 and cyan nodes from chromosome 11.

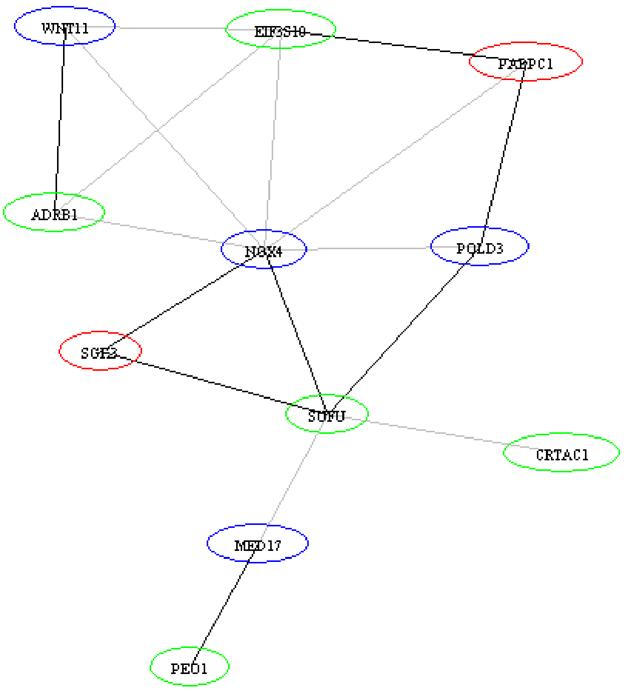


Figure 2.11. A graph showing the E2F binding site genes with a PSRSS score <0.4 with at least one other E2F binding site containing gene that have connections found through PubGene (Jenssen, *et al.*, 2001). PubGene connections are allowed to either be direct connections between the genes, or have one gene connecting both genes. Black edges connect gene which have a PSRSS score <0.4 as well as a PubGene connection. Grey edges connect genes with a PubGene connection, but a PSRSS score ≥0.4. Red nodes represent genes from chromosome 8, green nodes from chromosome 10 and blue nodes from chromosome 11.

Two complete 3 node subgraphs containing one node from each of the three chromosomal regions were found by looking for literature connections between the E2F binding site containing genes (Figure 2.8). Both subgraphs contain *NOX4* with one subgraph containing *SGK3* (serum/glucocorticoid-regulated kinase 3) and *SUFU* (homologue of drosophila suppressor/fused gene) as the remaining nodes and the other containing *PABPC1* (polyadenylate-binding protein, cytoplasmic, 1) and *EIF3S10* (eukaryotic translation initiation factor 3). *NOX4* was also found in the most likely combinations when comparing the genes with CREB binding sites. *SUFU* is the only node involved in the subgraphs which is annotated with the 'immune response' GO term. However this does not preclude the other genes from being involved in the 'immune response', merely that they have not been annotated as such. This fact coupled with the interactions being found using both PSRSS scores and co-occurrence in the literature makes the *NOX4*, *SGK3*, *SUFU* combination of genes the most likely when looking at literature connections of E2F binding site containing genes. *NOX4* has been shown to be functionally positively regulated by the E2F transcription factor E2F1 in vascular smooth muscle (Zhang *et al.*, 2008).

Only two of the genes with SRF binding sites, *CRH* (chromate resistance) and *MAML2* (homologue of Drosophila mastermind-like 2), were found to have a PSRSS score <0.4 between them. This prevents the formation of a complete three node subgraph containing a gene from each of the chromosomal regions. No co-occurrence was found in the literature between the genes *CRH* and *MAML2*. It would still be worth considering these genes as candidates, however, as *CRH* is annotated as an 'immune response' gene.

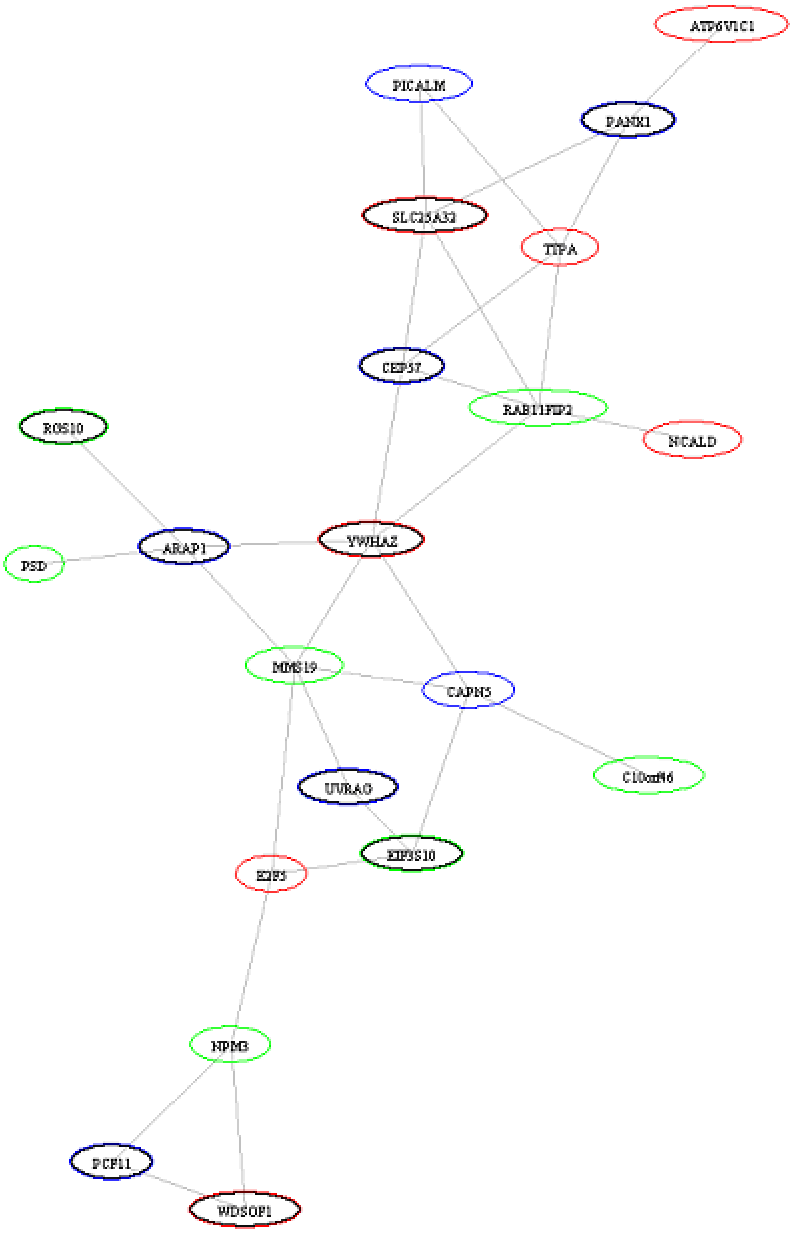


Figure 2.12. Graph showing the <0.40 PSRSS relations between genes with a predicted USF binding site, from the different chromosomal regions. Genes with a co-occurrence in the literature, found using PubGene, with the term ‘immune response’ contain an inner circle of black. Red nodes represent genes from chromosome 8, green nodes from chromosome 10 and blue nodes from chromosome 11.

The smaller number of genes and edges involved in the USF binding site containing gene network (Figure 2.12) allows for the closer identification of possible co-regulated genes before the PubGene filter is applied. Five 3 node complete subgraphs, with one gene from each chromosome, were identified. These were; i) *RAB11FIP2* (Rab11 family-interacting protein 2), *CEP57* (centrosomal protein 57kD) and *SLC25A32* (Solute carrier family 25 member 32) where both *CEP57* and *SLC25A32* were annotated with the ‘immune response’ GO term, ii) *YHWAZ* (tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, zeta polypeptide), *ARAP1* (ankyrin repeat and PH domain 1) and *MMS19* (MMS19 nucleotide excision repair homologue) where *YHWAZ* and *ARAP1* were annotated with the ‘immune response’ GO term, iii) *YHWAZ*, *MMS19* and *CAPN5* (calpain 5) where only *YHWAZ* was annotated with the ‘immune response’ GO term, iv) *NPM3* (nucleoplasmin 3), *PCF11* (cleavage and polyadenylation factor subunit) and *WDSOF1* (WD repeats and SOF1 domain containing) where both *PCF11* and *WDSOF1* were annotated with the ‘immune response’ GO term and v) *TTPA* (Alpha-tocopherol transfer protein), *CEP57*(centrosomal protein 57kD) and *RAB11FIP2* (Rab11 family-interacting protein 2) where CEP57 was annotated with the term ‘immune response’. The inclusion of at least one ‘immune response’ annotated term in each of these subgraphs warrants the inclusion of each of these genes as candidates.

When the connections found through pubgene were included in the network (Figure 2.13), two complete three node subgraphs containing genes from each of the three regions were found using the USF binding site containing genes. The subgraphs both contain *PCF11* and *MMS19*. The third nodes in the subgraphs are *WDSOF1* and *ATP6V1C1* (ATPase, H+ transporting, lysosomal 42kDa, V1 subunit C1). Only one of the edges is representative of a link through both PSRSS score and through the literature, this edge linked the two ‘immune response’ GO term annotated genes; *WDSOF1* and *PCF11*. Therefore the most likely of the subgraphs is that containing *PCF1*, *MMS19* and *WDSOF1*.

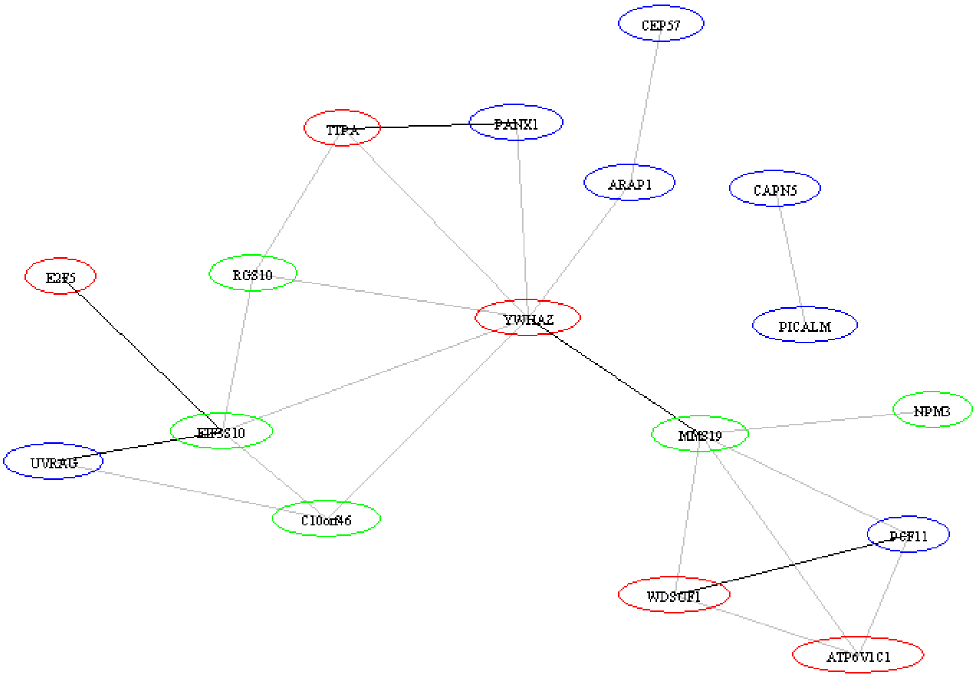


Figure 2.13. A graph showing the USF binding site genes with a PSRSS score <0.40 with at least one other USF binding site containing gene that have connections found through PubGene. PubGene connections are allowed to either be direct connections between the genes, or have one gene connecting both genes. Black edges connect gene which have a PSRSS score <0.40 as well as a PubGene connection. Grey edges connect genes with a Pubgene connection, but a PSRSS score ≥0.40. Red nodes represent genes from chromosome 8, green nodes from chromosome 10 and blue nodes from chromosome 11.

2.4. Conclusions

The TFBS prediction algorithms analysed showed varying levels of performance. Using the benchmark dataset only TFBS and Match performed to a satisfactory level giving AUC values ≥ 0.70. The great variations in performance observed in the current work highlights the importance of using a single standard dataset to benchmark the in-silico TFBS prediction tools.

By combining certain pairs of TFBS prediction algorithms the performance as measured by AUC value was improved. Three pairs of algorithms, TFS_P, TFBS_TFS and TFBS_P, showed a significant increase in AUC over the highest AUC values for single TFBS prediction algorithms. The TFBS modules, one of the highest performing algorithms with an AUC of 0.70 was included in two of the three pairings. However, Match, the other high performing algorithm showed a significant decrease in AUC value when combined with almost all other algorithms. The improvement in AUC achieved by combining pairs of algorithms observed in the current study indicates that the performance may be further improved by combining more than two prediction algorithms. This question is investigated in chapter 3, where more than two prediction algorithms are combined using machine learning techniques.

Three of the best paired TFBS prediction methods revealed five transcription factors shared by at least one gene from each of the 3 chromosomal regions from the IFN-γ response genetic linkage study. These transcription factors were CREB, E2F, HNF-1, SRF and USF.

CREB is known to be directly involved in the activation of IFN-γ in T cells (Pasquinelli *et al*., 2009; Samten *et al*., 2008). CREB, enhanced via phosphorylation, binds to the proximal promoter of IFN-γ. Stimulation of peripheral blood mononuclear cells (PBMC) with *M. tuberculosis* has been shown to induce phosphorylation and binding of CREB to the IFN-γ gene promoter. *M. tuberculosis* infected patients with ineffective immunity show both diminished IFN-γ production and reduced amounts of CREB binding to the IFN-γ proximal promoter (Samten *et al.*, 2005). Although there is evidence for a direct link between IFN-γ and the CREB TF, if CREB is regulating the IFN-γ linked genes, it would imply that CREB is also involved in another indirect method of IFN-γ regulation. This may be via the MHC pathways. CREB has been shown to transcriptionally regulate MHC class II expression (Burd *et al*., 2004) which plays a critical role in mycobacterial immunity through antigen presentation to CD4+ T-cells leading to the production of cytokines such as IFN-γ.

The E2F family of TFs bind to similar TFBSs with the consensus sequence; TTTCGCGC, although it has been recently shown that many E2F transcription factors also bind to alternative sites (Rabinovich *et al*., 2008). The E2F family is involved in a large number of processes including cell cycle, apoptosis, nucleotide synthesis, DNA repair and DNA replication. Many of the of genes with E2F TFBSs were associated with the term ‘immune response’ in the literature and had known functional links to other genes with E2F TFBSs. There is little evidence in the literature of a link between E2F transcription factors and IFN-γ, although, it has been shown that *e2f1* deficient mice have an increase in numbers of CD4+ cells, a decrease in CD4+25+ cells and a significant increase of IFN-γ following antigenic stimulation when measure in-vitro (Salam *et al*., 2004).

HNF1 is a hepatocyte nuclear factor, expressed mainly in the liver. HNFs regulate a wide variety of targets involved in processes such as development and metabolic homeostasis (Pontoglio, 2000; Tronche & Yaniv, 1992). Due to its location and functional bias it seems unlikely that HNF1 is involved in the regulation of IFN-γ and therefore unlikely that the IFN-γ linked genes are regulated by HNF1. There is a singular lack of evidence in the literature for a link between, HNF1 and IFN-γ, tuberculosis or the immune response in general. The conclusion that HNF1 is not one of the TFs regulating IFN-γ linked genes is further defended by the lack of genes with similar functions in the set of those with HNF1 TFBSs, although this does not preclude the existence of genes with similar functions that are at present unknown.

Serum response factor, SRF, is involved in the transformation of extracellular signals into specific nuclear responses (Ling *et al*., 1998). The SRF DNA binding domain has been shown to act as a docking site for multiple TFs. Therefore, it is unlikely that by analysing SRF binding sites, functionally related genes could be predicted. The prediction of whether genes were co-regulated would depend on which other transcription factors were co-operating with SRF. Only two of the genes with predicted SRF TFBSs were linked by known functions. However, one of these had a co-occurrence with the ‘immune response’ term in the literature and so the two genes should still, therefore, be considered candidate genes.

USF is reported to be an integral part of the IFN-γ inducibility of the type IV CIITA promoter (O'Keefe *et al*., 2001). The type IV CIITA promoter contains three cis-acting elements the proximal IFN-γ activation sequence (GAS) element, the E box, and the proximal IFN regulatory factor (IRF) element. The E box binds USF-1 while Gas binds

STAT-1 and IRF binds to the IFN regulatory factor. The binding of STAT-1 to the promoter appears to be dependent on URF-1 binding to the promoter (Dong *et al*., 1999; O'Keefe *et al*., 2001). Several possible groups of genes from each of the three chromosomal regions exist with USF TFBSs, many of these groups contain at least one gene with co-occurrence to the ‘immune response’ term, although none of the grouping contain immune response related genes for all three regions.

151 unique candidate genes have been identified by analysing co-occurrence of TFBSs for the five selected transcription factors. This number has been reduced to 21 (Table 2.9) when the most likely groups of three genes are considered.

However, due to the limitations of the method, it is possible that additional candidates have been missed. The main cause of this is the inability to predict TFBSs for certain transcription factors by some of the algorithms due to an incomplete set of PWM. The development of a search method that utilised information from a greater number of algorithms and was not restricted by the version of TFBS database used as a reference, would be an improvement over the methods described here. The first element of such a method is described in Chapter 3.

Candidate Gene	Chromosome	Transcription Factor(s)	Immune Response GO term
ARAP1	11	USF	Yes
ATP6V1C1	8	USF	
BLNK	10	CREB	Yes
CAPN5	11	USF	
CEP57	11	USF	Yes
CRH	8	SRF	Yes
EIF3S10	10	E2F	
MAML2	11	SRF	
MMS19	10	USF	
NOX4	11	CREB, E2F	
NPN3	10	USF	
PABPC1	8	E2F	
PCF11	11	USF	Yes
RIPK2	8	CREB	Yes
SGK3	8	E2F	
SLC25A32	8	USF	Yes
SMC3	10	CREB	
SUFU	10	E2F	
RAB11FIP2	8	USF	
WDSOF1	8	USF	Yes
YHWAZ	8	USF	Yes

Table 2.9. The candidate genes most likely to be linked to the IFN-γ immune response from the co-regulatory evidence in the current analysis.

3. Bayesian Analysis

3.1. Introduction

A number of the combinations of current TFBS prediction programs analysed in Chapter 2 show an improvement in accuracy compared to the performance of single algorithms. However, this is a simplistic method of combining the information from the two algorithms, and does not easily allow for scaling up to use information from additional prediction algorithms. By using machine learning techniques a variety of information can be integrated and applied to the TFBS prediction problem. In this problem a machine learning classifier would be used to classify the predicted TFBSs into one of two classes. The first class contains TFBSs which are functional binding sites *in vivo* while the second class contains TFBSs which are not functional *in vivo*. Particular advantages of machine learning include the ability to assign different weighting to different types of information, allowing some features to influence the classification decision more than others, and the use of real data to train the algorithm, computationally calculating weightings to attach to the features. In this analysis two different sets of features were used to describe the data for the classification problem: firstly, the TFBS prediction algorithms assessed in Chapter 2 and secondly, phylogenetic data utilised through the TFBS Perl modules (Lenhard & Wasserman, 2002) in a similar fashion to Consite (Sandelin *et al*., 2004b).

The TFBS Perl modules, when analysed as an individual algorithm, achieved the highest precision at 8.5% and the joint highest AUC at 0.70. In the analysis in Chapter 2 the TFBS modules were used as a PWM searching algorithm. However there are additional modules which allow the algorithm to include additional information in the form of phylogenetic footprinting. Phylogenetic footprinting algorithms are mainly used on sequences from only two species, one being the sequence and species of interest and the other being an orthologous sequence from another species. This allows for the identification of regions which have a higher level of conservation between the two organisms than would be expected. The method suggested here uses the addition of extra species to phylogenetic footprinting. This is a proposition which has been suggested before, but usually for lower level organisms and often for very closely related species. For example, phylogenetic footprinting methods have recently been used to identify microRNA cis-regulatory elements across 12 *Drosophila* species (Wang *et al*., 2008).

3.1.1. Classic Naïve Bayes

Bayes' theorem is a probabilistic theorem which describes conditional dependencies. In the classic Naïve Bayes classifier, Bayes' theorem (equation 3.1) is used to calculate the conditional probability of a data point or example belonging to a certain class (A), given a set of features which describe the data point (B) (Zhang, 2004).

$$P(A \mid B) = \frac{P(B \mid A)(A)}{P(B)} \tag{3.1}$$

The classic Naïve Bayes classifier is based on applying Bayes' theorem with strong, naïve, independent assumptions. It is a supervised learning algorithm which creates a model based on parameters estimated using maximum likelihood. Each feature in the classic Naïve Bayes classifier is assumed to independently contribute to the probability of a prediction, although this may not in reality be the case. Therefore, independent probability distributions are calculated, using a training dataset, for each feature describing the data. The training process uses both positive and negative examples to calculate these distributions and other parameters.

One of the most important other parameters in the Naïve Bayes classifier is the *a priori* value. This value gives the likelihood of any given example belonging to a class before any additional knowledge has been gained by the analysis of features by the classifier. It is often calculated by the number of examples belonging to each class in the data that the classifier will be trained on. If the training set contained equal numbers of examples of each of two classes, this would set the *a priori* value at 0.5 as without further information all that can be predicted about the data is that a given example has the same likelihood of belonging to either class. Further conditional probabilities of a data point belonging to a class are calculated in the training phase. The classifier gives a probability of a datapoint belonging to a certain class by combining these probabilities. For each attribute associated with the data point, a distribution is created showing the probabilities of the data point belonging to a certain class, given a certain value of the attribute. For each datapoint the probabilities of it belonging to a certain class given each attribute value are combined with the prior probability to give a final posterior probability of the data point belonging to the class.

Bayesian classifiers have been shown to perform at a much higher level than would normally be assumed and are able to estimate parameters for a model with a small amount of training data compared to many other classifiers (Zhang, 2004).

3.1.2. Positive Naïve Bayes

Positive Naïve Bayes (PNB) is a variation of the classic Naïve Bayes algorithm which allows the data to be trained with positive examples but without negative examples. This is particularly interesting when dealing with biological problems as negative data is often missing. If there is no negative data the usual strategies are either to generate negative data or use real data which is unknown but likely to contain mostly negative examples (Erill & O'Neill, 2009). Parameters used to train the classifier are calculated differently in PNB compared to the classic Naïve Bayes classifier. This is due to the difference in classification datasets (Calvo *et al.*, 2007a). As in classical Naïve Bayes, the probability distribution of feature occurrence given that the datapoint is positive (D|1) is calculated from the training dataset. However, the lack of negative training data means that the probability distribution of feature occurrence given that the datapoint is negative (D|0), must be estimated. In PNB equation 1 is used to estimate the probability of (D|0) (Denis et al, 2003).

$$\Pr(D \mid 0) = \frac{\Pr(D) - \Pr(D \mid 1)\Pr(1)}{1 - \Pr(1)} \qquad \textbf{(3.2)}$$

The PNB algorithm (Denis *et al.*, 2003) was first proposed as a tool for classifying text documents, represented as bags of words. The algorithm has since been adapted for use in other situations and has been used productively in a number of different situations, including biological problems (Calvo *et al.*, 2007a). One problem where the PNB classifier has been successfully used is in the classification of dominant and recessive human disease genes (Calvo *et al.*, 2007b). Another problem involves the prediction of true or false, donor or acceptor splice sites as defined by the database ACCDON (Calvo *et al.*, 2007a).

3.1.3. Positive Tree Augmented Naïve Bayes

The tree augmented Naïve Bayes classifier (TAN) is a variation on the Naïve Bayes classifier (see figure 3.2). Unlike the Naïve Bayes classifier in which prior probabilities for each attribute are only dependent on the class, additional relationships and dependencies between features are allowed (Figure 3.1). If all possible combinations of relationships between features were allowed this could quickly increase the computational power and time needed to perform the classification. Friedman et al., suggest limiting the parents of each attribute to the class and, at most, one other attribute. It has been shown that this approximation of the dependencies between features is optimal and that the TAN classifier can be learned quickly without the need for large amounts of computer power (Friedman *et al.*, 1997). An example of the dependencies in the TAN classifier can be seen in Figure 3.1. In the Naïve Bayes classifier (NB) the probability distribution associated with each attribute (A1-A4) are calculated using only the class feature (C). In the TAN classifier the probability distribution of each attribute is dependent on both the class and at most, one other attribute. For example, A2 is dependent on both the class and A1; however A1 is only dependent on the class. Figure 3.1 shows a very simplistic view of the TAN classifier, in reality one feature may have many dependents while other features may have none.

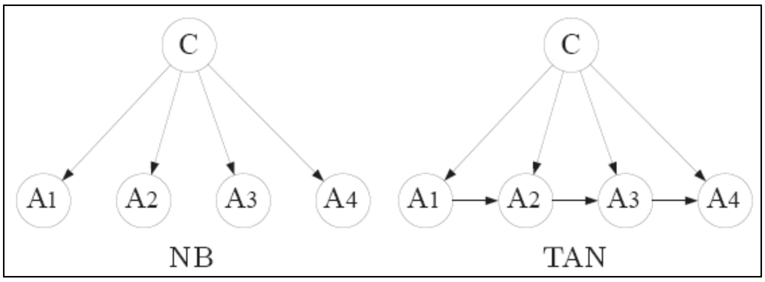


Figure 3.1. Simplistic representation of the differences in the structure of the Naïve Bayes classifier and the tree augmented Naïve Bayes classifier. C is the class variable, A's are features, and arrows indicate dependence among variables.

The optimum relationships between attributes in the TAN classifier are chosen by estimating conditional frequencies using the training data. One disadvantage of this method is that a lack of instances representing certain conditions can often occur, causing less reliable estimates of conditional probabilities. This is less likely to occur when training a Naïve Bayes classifier where the conditional probabilities are only reliant on the class.

The Positive Tree Augmented Naïve (PTAN) Bayes classifier (Calvo *et al.*, 2007a) is a partially supervised classifier based on the TAN Bayes classifier (Friedman *et al.*, 1997) in much the same way as the PNB classifier is based on the classic Naïve Bayes classifier (Figure 3.2). The PTAN algorithm is an adaptation of the TAN algorithm which allows for the use of only positive examples in the training of the classifier. This involves changes in the parameter estimation step to adapt for the lack of negative variables. The PTAN algorithm was assessed using datasets which were both biological and non-biological. The biological example involved the prediction of donor or acceptor splice sites from the ACCDON dataset (Castelo & Guigó, 2004). This dataset contains a set of known canonical donor splice sites and a set of known canonical acceptor splice sites extracted from the RefSeq genes (Pruitt *et al.*, 2009). Using the ACCDON-based data PTAN underperformed compared to PNB when assessed by F measure. However, when the classifiers were trained and tested using synthetic data, PTAN significantly out performed PNB (Calvo *et al.*, 2007a).

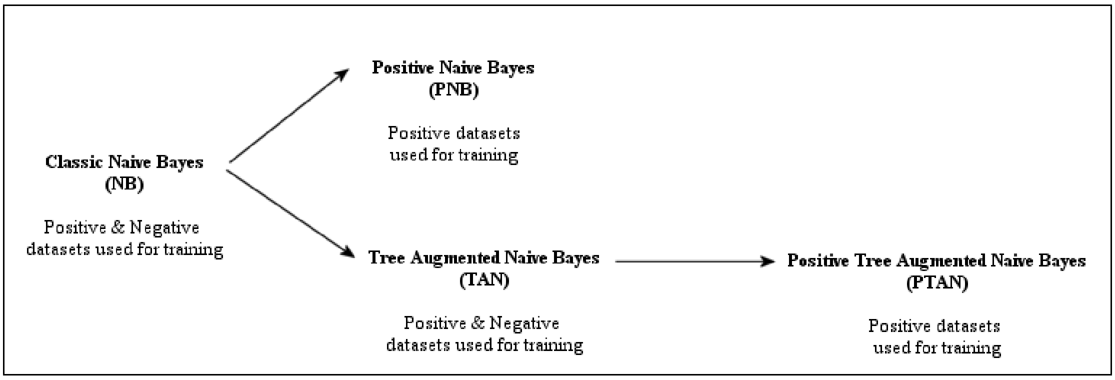


Figure 3.2. The relationships between the various Naïve Bayes classifiers used in the study. Both Positive Naïve Bayes (PNB) and Tree Augmented Naïve Bayes (TAN) are types of Naïve Bayes classifier. Positive Tree Augmented Naïve Bayes (PTAN) is a type of TAN classifier.

3.1.4. Feature selection

For many supervised machine learning techniques, to optimally train a set of data, the features which describe the data must be carefully chosen. Failure to do this can lead to incorrect weights being assigned to features and to overtraining of the classifier, reducing performance when an unseen dataset is introduced. Feature selection is the technique of selecting a subset of the most relevant features for use in machine learning algorithms. Naïve Bayes classifiers, in particular, have been shown to perform optimally after feature selection (Ratanamahatana & Gunopulos, 2002; Zhang, 2004). If a feature set chosen for a classifier is particularly large, it is helpful to reduce the number of features needed to classify the data. This often improves both classification, reducing the chance of over fitting, and the efficiency of the classifier (Hall & Smith, 1998).

There are two main methods of reducing the number of features used in machine learning. The first method is by combining existing features into a smaller number of abstract features. Principal component analysis is one method of this type which is often used. The second method is feature selection. Ideally the feature selection should involve an exhaustive search of all possible combinations of features. If large numbers of features are available this is not practical, therefore feature selection algorithms are used to search for a local maximum. There are a large number of feature selection algorithms which have been developed. Four of these methods are used in this study: Chi squared, Support Vector Machines (SVM), Information gain and Gain ratio.

(I) Chi squared

Chi squared (χ^2) is one method used for feature selection (Liu & Setiono, 1995). Chi squared is used to test the independence of two events. When this test is applied to feature selection the two events are: the value associated with the feature and the class associated with the example. Features are selected based on a low probability of independence between the occurrence of the term and occurrence of the class (Doraisamy *et al*., 2008; Manning *et al*., 2008).

(II) Support Vector Machines (SVM)

Support vector machines attempt to find a linear decision boundary where objects with one classification fall one side and objects with the other classification fall on the other side. Although the decision boundary is linear, the space through which it runs can be multidimensional. SVMs, used as a feature selection method, determine how closely a boundary, calculated by using any given feature, represents the known classification of the data (Doraisamy *et al.*, 2008; Guyon *et al*., 2002).

(III) Information Gain

Information gain, or the Kullback-Leibler divergence (Kullback, 1987) is a measure of the difference between two probability distributions P and Q. The algorithm is used to measure the expected difference in the number of bits required to code examples from P using a code based on P and using a code based on Q. P represents the actual data whereas Q represents a theory or model. When used for feature selection Q represents the class as determined by the feature being assessed and P represents the actual class (Kullback, 1987).

(IV) Gain Ratio

The gain ratio (Doraisamy *et al.*, 2008) is a statistic which is derived from the information gain statistic. Gain ratio is the information gain divided by the intrinsic information measure. The intrinsic information measure shows how well the data is split by a certain feature. It has been suggested that information gain, when applied to a dataset with a large number of distinct variables, may learn the data too well and over fit the data. Information gain ratio (Gain Ratio) biases the decision against features with a large number of distinct values through the use of the intrinsic information measure and therefore solves this problem (Doraisamy *et al*., 2008).

3.2. Pattern Matching Algorithm Naïve Bayes Methods

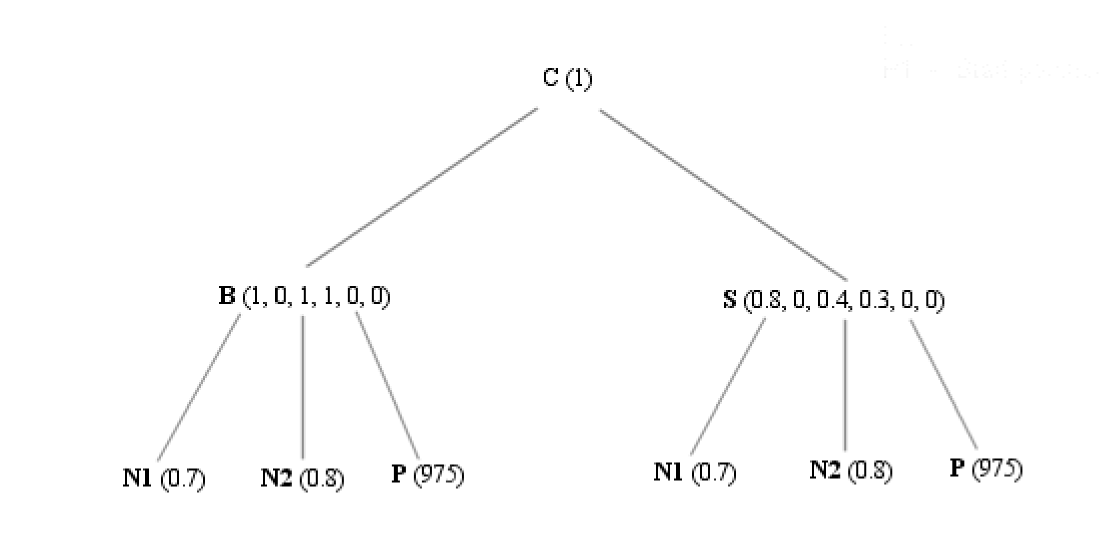
3.2.1. Feature sets

A set of data which can be used to test the Pattern Matching Algorithm Naïve Bayesian classifier (PMANB) exists from the work in Chapter 2. Only six of the algorithms used to predict TFBSs in the dataset are used as input to the PMANB classifier. AliBaba 2.1 is left out due to the lack of a score associated with the predictions. The set of total predictions made by the six TFBS prediction algorithms is checked for redundancy, removing all TFBSs which have the same start, end and binding factor. This is the dataset used in the training and testing of the classifier and is denoted PublicTFBS_206384. The name represents the fact that the data comes from the public version of TRANSFAC and that it contains 206384 TFBS predictions. Subsequent datasets were named using these conventions.

Each predicted TFBS in the dataset was described by two different feature sets. Feature set ‘B’ was a binary array where each element in the array represented one prediction algorithm (Figure 3.3). The final element in the array coded whether the TFBS prediction was found in the TRANSFAC database, using a 1 to represent a positive match and a 0 to represent no match to a TFBS from the database.

The second feature set, ‘S’, was prepared in a similar way. In this array the elements contained the score given by the algorithm. If no matching prediction existed then the element contained a 0 (Figure 3.3). Using the scoring systems of the algorithms potentially allowed the classifier to bias the features depending on how successful the scoring system of each algorithm was.

Extra features were added to feature sets B and S. Firstly, nucleosome positioning information was included. The probability of a nucleosome occurring at each position of the sequence was calculated using the Recon algorithm (Levitsky, 2004). The data from Recon was used in two ways, firstly, by using the score at the first position of the TFBS prediction, secondly by taking the mean of the nucleosome prediction scores at each position of the TFBS. The distance and direction from the transcription start site (TSS) start site was also included as feature. This was given by the start position of the TFBS prediction.



Feature	Description	Possible Values
C	Class	Functional or Non Functional
B	Binary Feature Set	0 or 1
S	Score Feature Set	0.0 - 1.0
N1	Nucleosome probability at start site	0.0 - 1.0
N2	Mean nucleosome probability	0.0 - 1.0
P	Positon on sequence	1 - 1400

Figure 3.3. Representation of the feature sets used in the Naïve Bayes classifier. Each example has a binary class representation 'C'. Either feature set 'B' or 'S' can then be used to describe the main features. The 6 elements in the 'B and 'S' arrays represent one TFBS prediction algorithm each. Feature set B contains binary representations of whether the example was found by an algorithm or not, in this example, algorithms 1, 3 and 4 predicted the TFBS, algorithms 2, 5 and 6 did not. Feature set 'S' contains the score returned by each algorithm for the example, the example shows a high scoring match for algorithm 1, low scoring matches for algorithms 3 and 4 and no match for algorithms 2, 5 and 6. As additional features to the 'B' or 'S' feature set, N1, N2 or P can be added. These features are added to the array singly.

3.2.2. Naïve Bayes Classifier

The Naïve Bayes classification algorithm was implemented using the Weka suite of machine learning programs (Witten *et al.*, 1999). The Weka suite has been successfully implemented in other studies using both Naïve Bayes and other machine learning techniques (Chen *et al.*, 2009; Ivanciuc, 2008; Pirooznia *et al.*, 2008). It contains a large selection of supervised and unsupervised classification algorithms and feature selection methods and is a reliable and fast method of implementing the classic Naïve Bayes classifier.

The classic Naïve Bayesian classifier was trained and tested using 5 fold cross validation. The classic Naïve Bayes classification was, firstly, carried out using 921 positive examples. These were the predicted TFBSs from the PublicTFBS_206384 dataset which matched a known TFBS in TRANSFAC. The negative examples used in the classifier were the 205463 remaining predicted TFBSs from the PublicTFBS_206384 dataset which did not match any known TFBS. When performing the cross validation with these positive and negative examples the positive *a priori* value for the Naïve Bayes classifier was very low. The Weka suite calculates the *a priori* value based on the proportion of positive and negative examples in the training dataset. In this case there was a much larger amount of negative data than positive data, skewing the *a priori* value and causing the classifier to predict a similar ratio of positive and negative examples in the testing data.

The classifier was then trained and tested a second time. The number of negative examples used in the classifier was reduced. This gave a smaller dataset consisting of equal numbers of positive and negative examples, 921 examples of each. The negative data was sampled using a random number generator to pick the examples (http://www.random.org/). This process was performed three separate times to create three alternative negative datasets. The new datasets were designated

PublicTFBS_1842a, PublicTFBS_1842b and PublicTFBS_1842c. The mean result of the classifier using each of the three datasets was used in the analysis.

The features initially used to train the classifier consisted of only the prediction algorithm results, either in the binary or the scoring based forms. The extra features, nucleosome probability at the start position (N1), average nucleosome probability (N2) and position (P) were added to the array singly to determine whether they improved the classification.

Sensitivity, specificity and accuracy scores were calculated at the 0.5 probability threshold. All TFBSs with Naïve Bayes predictions of 0.5 or greater were assumed to be positive predictions; those with less than 0.5 were assumed to be negative predictions. ROC curves were created, and the AUC was calculated, for each classification undertaken. The AUC was measured using a ROC curve which had been fitted using the maximum likelihood fit of a binormal model. Where a small number of points are plotted on a ROC curve, or the points fall very closely together, empirical ROC curves often severely under estimate the AUC (Park *et al.*, 2004). The curves were calculated using JROCFIT and JLABROC4 as found on the John Hopkins Web based calculator for ROC curves (www.jrocfit.org). Overall performance was measured by a combination of these statistics.

3.3. Phylogenetic Pattern Matching Naïve Bayes Methods (PPMNB)

3.3.1. Dataset

The PublicTFBS_206384 dataset used for the PMANB classification contained enough examples to successfully train and test the classifier when using either the ‘B’ or ‘S’ feature sets. The use of phylogenetic data is able to predict functional important TFBSs through the identification of evolutionarily conserved motif. However, the inclusion of this data can also prevent a prediction being made for a gene; this occurs if there are no known homologs from the species in question. This can reduce the available dataset considerably; hence to implement the phylogenetic pattern matching Naïve Bayes method (PPMNB) the size of the dataset was increased.

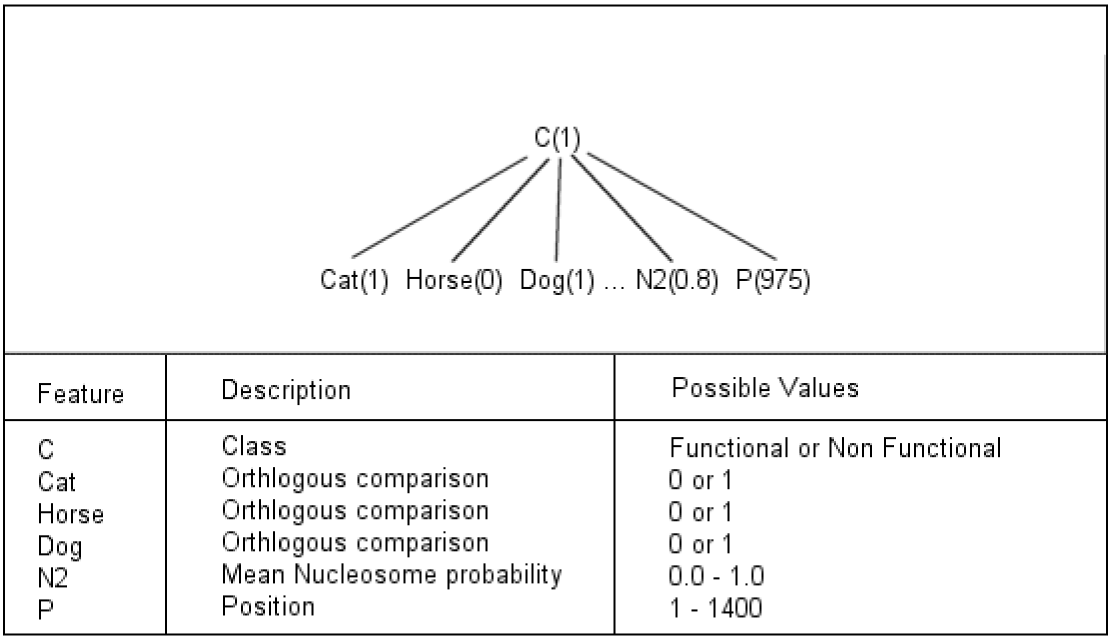
By using TRANSFAC Professional the number of known TFBSs and genes containing TFBSs is increased. All TFBSs with associated start sites, end sites and sequences were retrieved from TRANSFAC Professional and the position of the TFBSs in TRANSFAC Professional were analysed to determine the most useful length of sequence to use for the dataset (Section 3.6.1, Figure 3.9).

Sequences of length 9.5Kb, 5.4Kb upstream from the transcription start site and 4.1 Kb downstream from the transcription start site were used. This distance was one standard deviation in length from the mean position of the TFBSs found in TRANSFAC Professional (Section 3.6.1., Figure 3.9). ConSite is a phylogenetic TFBS prediction program based on the TFBS Perl modules, it allows the user to compare two orthologous sequences. It is a useful tool to use as a comparison to the PPMNB because the TFBSs are predicted using the same algorithm. The publicly available web version of ConSite limits the length of sequences to 10Kb which makes this a useful length of sequence to allow for comparison between the PPMNB classification and predictions made by ConSite in the standard way. The 9.5Kb sequences containing a TFBS were retrieved from Ensemble 49 (Hubbard *et al*., 2007). The genes were searched for the TFBSs to ensure that the binding site could be found in the correct position, with reported start and end positions and with a reported binding factor. This resulted in 362 known TFBSs which could be found in 114 genes.

3.3.2. Feature Set

The TFBS algorithm (Lenhard & Wasserman, 2002) was applied to the 114 9.5 Kb human sequences using the search criteria of a 50 nucleotide search window and a score threshold of 80%. The TFBS algorithm is essentially the ConSite algorithm used without a homologous sequence for comparison. This created the total set of possible TFBS predictions which could be made for each sequence when the ConSite algorithm was applied. This dataset was designated ProfessionalTFBS_20467. Each predicted TFBS was categorised as a positive i.e. a match against a TRANSFAC TFBS, or a negative, a TFBS prediction which did not match a known TFBS. The same criteria as that used with the PublicTFBS_206384 dataset was used to determine a match (Section 3.2.1).

Orthologs for each of the genes in the dataset were retrieved from each of the 38 eukaryotic full genomes available in Ensemble 50 via Biomart. TFBS predictions were made for each ortholog, paired with its corresponding human sequence, using the orca alignment method (unpublished) recommended by ConSite, and the ConSite phylogenetic footprinting algorithm (Lenhard & Wasserman, 2002; Sandelin *et al.*, 2004b). The search parameters used were the presence of 80% sequence conservation in a 50 nucleotide search window and a score threshold of 80%. The conservation level of 80% was chosen to select for regions which were more conserved than average for the majority of species. For example, mouse-human promoters have been shown to have a mean similarity level of 77% (Sorek & Ast, 2003).



Feature	Description	Possible Values
C	Class	Functional or Non Functional
Cat	Orthlogous comparison	0 or 1
Horse	Orthlogous comparison	0 or 1
Dog	Orthlogous comparison	0 or 1
N2	Mean Nucleosome probability	0.0 - 1.0
P	Position	1 - 1400

Figure 3.4. Representation of the phylogenetic feature set used with the Naïve Bayes classifier. 38 possible phylogenetic features are used, but are represented in the diagram as the three features, Cat, Horse and Dog. These features are represented in a binary format with 1 representing a conserved TFBS in the species. Additional features are ‘N2’ representing the likelihood of a nucleosome occurring at the same position as the TFBS prediction and ‘P’, the start site of the TFBS.

For each TFBS prediction an array consisting of 40 elements was created (Figure 3.4). Each element in the array represents a feature. The first 38 features represent a possible phylogenetic match to a TFBS predicted using ConSite with each of the 38 orthologs. The remaining two features represent the distance, given by the position on the sequence, and the predicted probability of nucleosome occupancy. Each element contains a value for the particular feature, a binary value for the phylogenetic features, the position on the sequence for the distance and the score retrieved from Recon, averaged across the TFBS motif, for the nucleosome occupancy probability.

3.3.3. Feature selection

Four feature selection methods were chosen from Weka (Witten *et al*., 1999); SVM, Chi Squared, Info gain and Gain Ratio. The features were ranked by each method using 5 fold cross validation on the test dataset. The features to be used in the classifiers were selected incrementally starting with the feature ranked highest by each classifier. After the addition of each feature the performance of the classifier was assessed. While performance increased, features continued to be added. Once performance stopped increasing, the optimum selection of features was assumed to have been reached.

3.3.4. Naïve Bayes Classifier

Three types of classifiers were used; Naïve Bayes (NB) from Weka (Witten *et al*., 1999), Positive Naïve Bayes (PNB) and Positive Tree Augmented Naïve Bayes (PTAN) (Calvo *et al.*, 2007a). The Naïve Bayes classifier was first used to determine the optimum set and number of features, through 5 fold cross validation. All three classifiers were then used to classify the data based on the optimum set of features. This was to determine whether classification based only on positive examples improved performance.

The classic Naïve Bayes classifier was trained and tested using 5 fold cross validation using the Weka implementation. The first dataset that the classifier used, ProfessionalTFBS_20467, contained all examples, both positive and negative, predicted by the TFBS algorithm; as with the PMANB dataset, PublicTFBS_206384, this gave a very low positive *a priori* value due to a much larger negative data set than positive data set. The classifier was secondly trained using a smaller dataset. The second dataset contained a random selection of the negative examples of the same size as the positive examples. This meant that the *a priori* value was set to 0.5 as the algorithm calculated that there was an equal chance of a data point being classified as positive or negative. The creation of the second dataset and the testing of the classifier with the dataset were repeated three times, as in the PMANB classifier. The three resulting datasets were designated ProfessionalTFBS_230a, ProfessionalTFBS_230b and ProfessionalTFBS_230c. Neither PNB nor PTAN needed negative classification data however, meaning that the *a priori* value was user defined. The *a priori* value of 0.5 was used as this was the value used by the Naïve Bayes classifier.

Performance was measured and compared by the sensitivity, specificity and accuracy values obtained at a 0.5 threshold value. ROC curves were also created and the AUC value made an additional measure of performance (Hanley & McNeil, 1982; McNeil & Hanley, 1984). All three Bayesian methods were compared to predictions made using the ConSite algorithm using mouse and human paired orthologs.

3.4. Applying the methods to the genes from the IFN-γ genetic linkage study

Once the optimum PMANB and PPMNB TFBS classifiers had been ascertained, they were used to predict TFBSs in the set of genes from the IFN-γ genetic linkage study (Section 1.9 & Section 2.2.4). Sequences for each of the genes in the linked regions were retrieved from Ensembl 49. The sequences were of the same length as those of the dataset used to test and train the prediction model. For the PPMNB classifier this was 5.4Kb upstream from the transcription start site and 4.1Kb downstream from the transcription start site. For the PMANB classifier this was 0.7 KB both up and downstream from the transcription start site. The dataset and feature information was obtained in the same way as the testing and training data for each classifier. TFBS predictions which were assigned a probability value greater than 0.5 by the classifier were assumed to be correct.

3.5.Pattern Matching Algorithms (PMANB) Results

3.5.1. Datasets

The full dataset, PublicTFBS_206384, consisted of 921 examples in the positive set and 205463 examples in the negative set. When using the 'B' feature set, with the PublicTFBS_206384 dataset, there were no positive predictions made by the classifier. This was due to the positive *a priori* value being automatically set as very low. The 'S' feature set predicted a small number of positive predictions, but most of the predictions made were still negative. This resulted in a high specificity score of 99.0% and an accuracy score of 98.6%. The sensitivity score was very low at 7.9%. These statistics show that the classifier was not accurately predicting the data; the results were being skewed by the large amount of negative data.

Smaller datasets, PublicTFBS_1842a, PublicTFBS_1842b or PublicTFBS_1842c, were created by taking all positive examples and subset of random negative examples from PublicTFBS_206384. This gave positive and negative datasets of the same size. The sensitivity was much higher when using these datasets with either the 'B' or 'S' feature sets (Table 3.1). Specificity and accuracy values were reduced, when using the PublicTFBS_1842 datasets with either feature set, however this reduction is very small compared to the increase in sensitivity.

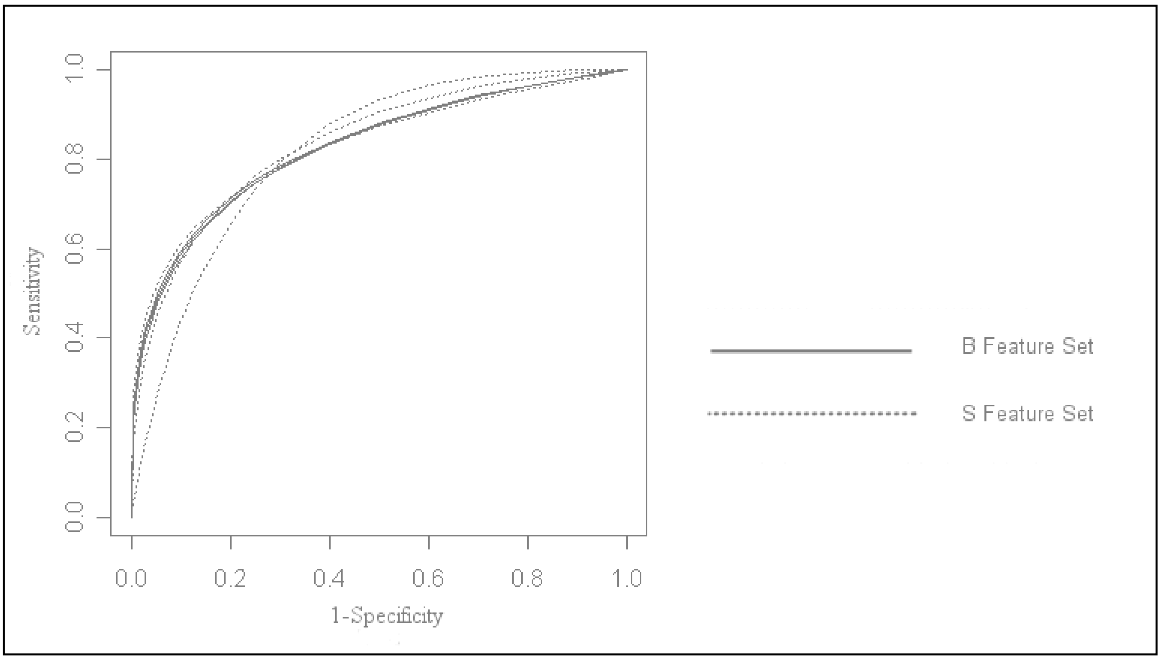


Figure 3.5. Fitted ROC curve showing the classification of predicted TFBSs by Naïve Bayes using either the ‘B’ or ‘S’ features sets. The three solid lines represent the results given when the ‘B’ feature set was used to test and train the classifier on each of the three datasets: PublicTFBS_1842a, PublicTFBS_1842b and PublicTFBS_1842c. The three dotted lines represents the results obtained when the ‘S’ feature set was used to test and train the classifier on each of the three datasets: PublicTFBS_1842a, PublicTFBS_1842b and PublicTFBS_1842c.

The three ROC curves created using the ‘B’ feature set (one curve for each of the small datasets) and the three ROC curves created using the ‘S’ feature set all showed an AUC of over 0.82 (Figure 3.5). The repetitions using the PublicTFBS_1842a, PublicTFBS_1842b and PublicTFBS_1842c datasets gave ROC curves of the same shape and size in each case when the ‘B’ feature set was used to train and test the classifier. There was a much greater variation in both the shape and AUC of the ROC curves when the ‘S’ feature set was used for training and testing purposes (Figure 3.5).

Neither feature set was shown to consistently outperform the other feature set as a means of representing the data. There was a slight decrease of less than 1% in mean sensitivity when using the ‘S’ feature set rather than the ‘B’ features set (Table 3.1). In contrast to this, specificity and accuracy scores were increased when using the ‘S’ feature set rather than the ‘B’ features set (Table 3.1). Both feature sets gave similar AUC values in the ROC curves but the repetitions using the ‘B’ feature set gave a much more consistent ROC curve shape and AUC than the repetitions carried out using the ‘S’ feature set.

3.5.2. Feature selection

3.5.2.1. Nucleosome probability at start position (N1)

The effect of the addition of nucleosome positioning data was shown to be inconsistent in the PublicTFBS_1842a, PublicTFBS_1842b and PublicTFBS_1842c datasets. In two of the PublicTFBS_1842 datasets, when using the ‘B’ features set, the addition of the N1 feature did not improve the classification as measured by AUC, in the third dataset an improvement was seen (Figure 3.6). When using the ‘S’ feature set the N1 feature increased the AUC for two of the PublicTFBS_1842 datasets and reduced the AUC for the third PublicTFBS_1842 dataset. In all cases the average sensitivity, specificity and accuracy were reduced by the addition of the nucleosome data (Table 3.1). It can therefore be deduced that the N1 feature is not a useful addition to the classifier.

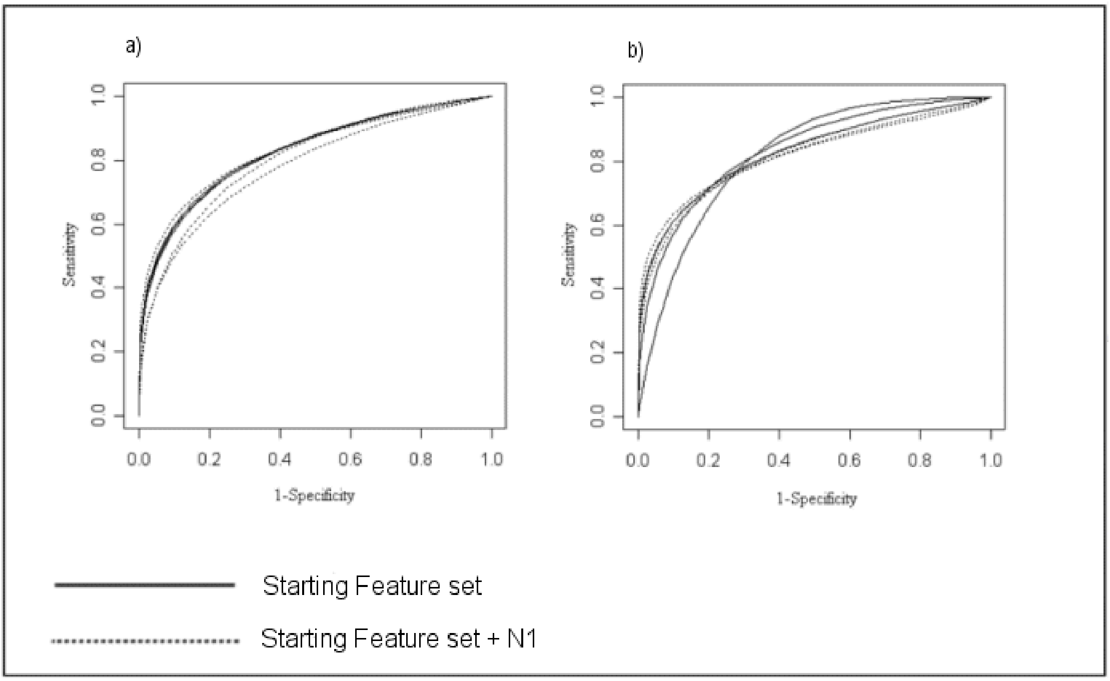


Figure 3.6. a) shows the classification of TFBSs using the ‘B’ feature set, b) shows the ROC curves when using the ‘S’ feature set. The solid lines show the classification if only the ‘B’ or ‘S’ feature sets are used with each of the PublicTFBS_1842 datasets. The dotted lines show the classification if the nucleosome prediction at the start site (N1) is included as a feature with each of the PublicTFBS_1842 datasets.

		Sensitivity (%)	*Specificity (%)*	*Accuracy (%)*
B features	6 features	78.8	74.3	76.5
	+ N1	75.5	73.5	74.4
	+ N2	78.9	77.7	78.3
	+ P	77.9	83.3	80.6
S features	6 features	78.3	77.8	78.1
	+ N1	77.6	77.6	77.6
	+ N2	78.1	79.3	78.7
	+ P	77.7	83.4	80.5

Table 3.1. Table containing the mean sensitivity, specificity and accuracy values obtained when using the Naïve Bayes classifier. These values were recorded by taking the mean across the three datasets: PublicTFBS_1842a, PublicTFBS_1842b and PublicTFBS_1842c, when using different feature sets. In each case the base S or B feature set is used, with 0 or 1 additional feature.

3.5.2.2. Average nucleosome probability (N2)

Using the N2 feature (mean value of the nucleosome scores over all positions of the TFBS) showed improved performance, as measured by ROC curve, compared to using the 'B' or 'S' feature sets without either nucleosome feature (Figure 3.7). When using either of the 'B' or the 'S' feature sets with the additional N2 feature, the AUC of the ROC curves was larger in all three of the PublicTFBS_1842 datasets, than when using the 'B' or 'S' feature sets alone.

The change in the shape of the ROC curve with the addition of the N2 feature is interesting. The ROC curves based on the 'S' feature set and the ROC curves based on the 'B' feature set do not improve in the same way as each other. The 'B' feature set with N2 (Figure 3.7a) has improved sensitivity at low specificity values; this implies that fewer FN results will be predicted by the classifier. The 'S' feature set with N2, however, has improved sensitivity at high specificity values; this will result in fewer FP results. The three curves created using the ‘S’ + N2 feature set can also be seen to have very similar shapes to each other, showing that the addition of the N2 feature has reduced the variability in the classification of the data from that seen in the ‘S’ feature set.

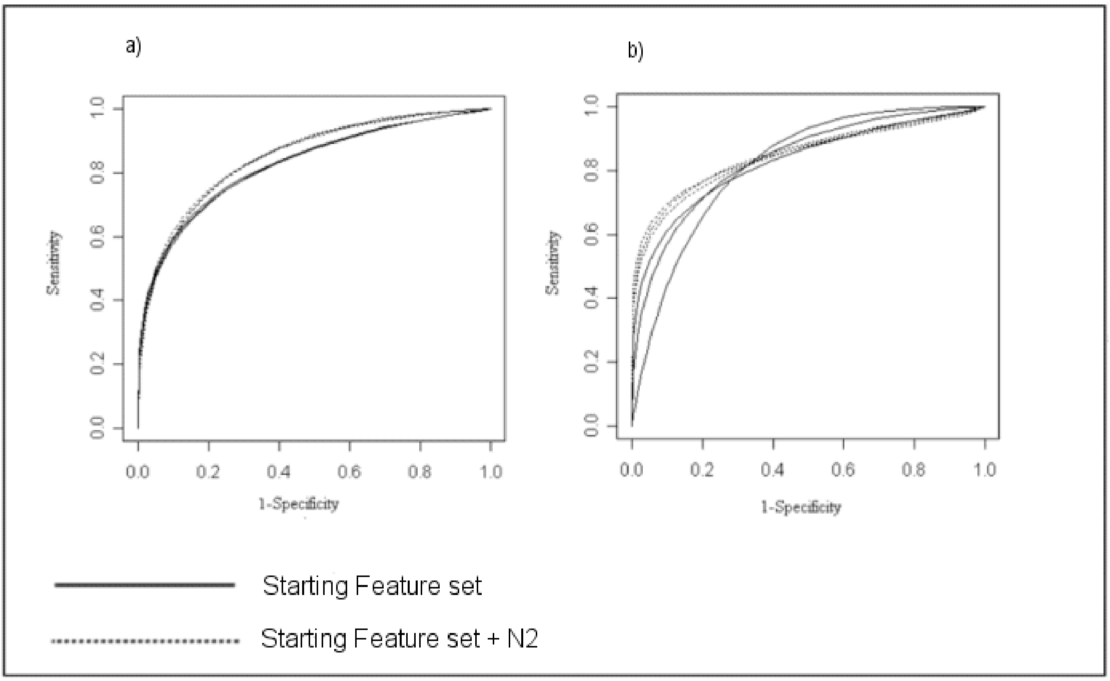


Figure 3.7. a) shows the classification of TFBSs using the ‘B’ feature set, b) shows the ROC curves when using the ‘S’ feature set. The solid lines show the classification if only the ‘B’ or ‘S’ feature sets are used with each of the PublicTFBS_1842 datasets. The dotted lines show the classification if the mean nucleosome prediction values (N2) are included with each of the PublicTFBS_1842 datasets.

When using the ‘B’ features set plus N2, the sensitivity, specificity and accuracy scores were all improved when compared to the ‘B’ feature set alone (Table 3.1). When using the ‘S’ feature set plus N2, there was no real change to the mean sensitivity, a drop of 0.20%, compared to the ‘S’ feature set alone. Improvement could be seen, however, in the specificity and accuracy values obtained using the ‘S’ + N2 feature set compared to the ‘S’ feature set (Table 3.1).

3.5.2.3. Position on sequence (P)

The addition of ‘P’ (TFBS start site), to either feature set ‘B’ or ‘S’, increased the AUC of the ROC curves (Figure 3.8). All ROC curves created using the ‘B’ + ‘P’ feature set had AUC scores of above 0.88. The AUC scores for the ROC curves based on the ‘S’ +’P’ feature set were 0.86 and above. The average sensitivity scores, using either feature set, were slightly decreased, a reduction of less than 1.00%, with the addition of ‘P’. However, the specificity and accuracy scores were substantially increased.

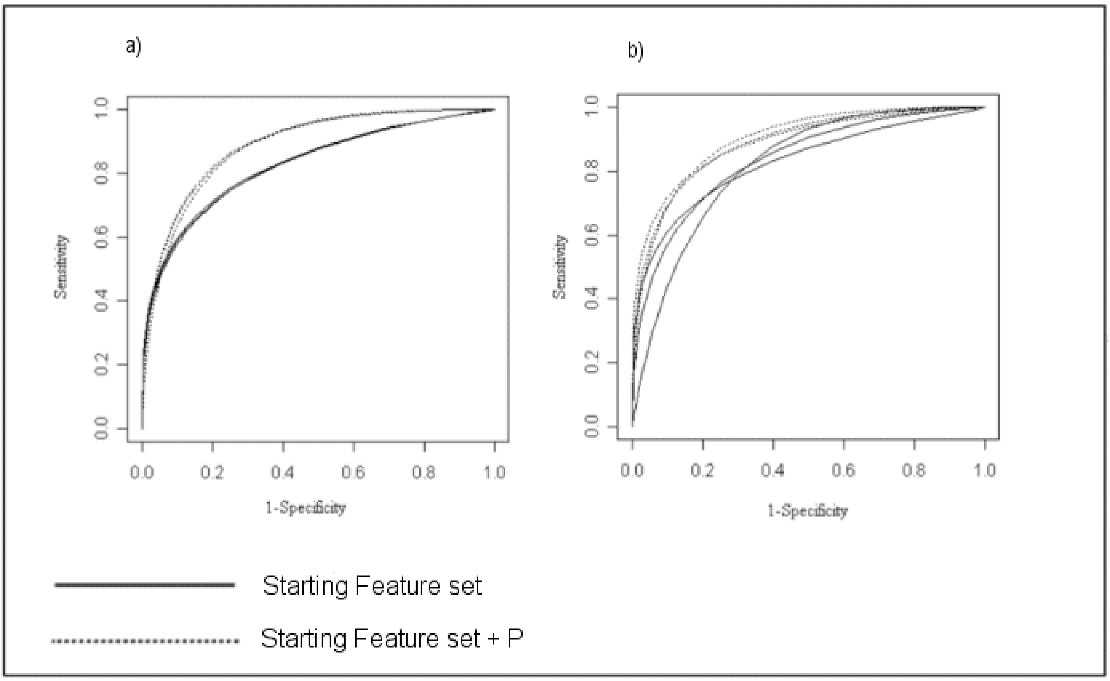


Figure 3.8. (a) shows the classification of TFBSs using the ‘B’ feature set, (b) shows the ROC curves when using the ‘S’ feature set. The solid lines show the classification if only the ‘B’ or ‘S’ feature sets are used with each of the three PublicTFBS_1842 datasets. The dotted lines show the classification of each of the three PublicTFBS_1842 datasets if ‘P’ is included as a feature.

The final PMANB classifier consisted of the 6 TFBS prediction algorithm features and the position of the TFBS relative to the TSS.

3.6. Phylogenetic Naïve Bayes Results (PPMNB)

3.6.1. Dataset

Phylogenetic conservation across species was used as another method of generating features describing the functionality of TFBSs. These features were used to train a Naïve Bayes classifier. The classifier trained using these features was designated PPMNB.

The histogram showing the start positions of the TFBSs in TRANSFAC peaks at 250bp upstream of the transcription start site (Figure 3.9). The curve is skewed towards the upstream positions. This is consistent with the literature which states that a peak in TFBS occurrence can be found within 300bp upstream of the TSS (Koudritsky & Domany, 2008).

Sequences of length one standard deviation from the mean position of the binding sites were found to be 5.4Kb upstream from the transcription start site and 4.1Kb downstream from the transcription start site. This gave sequences of 9.5Kb in total. This length of sequence contained over 50% of the TFBSs, to increase this percentage the length of sequence would need to be substantially increased.

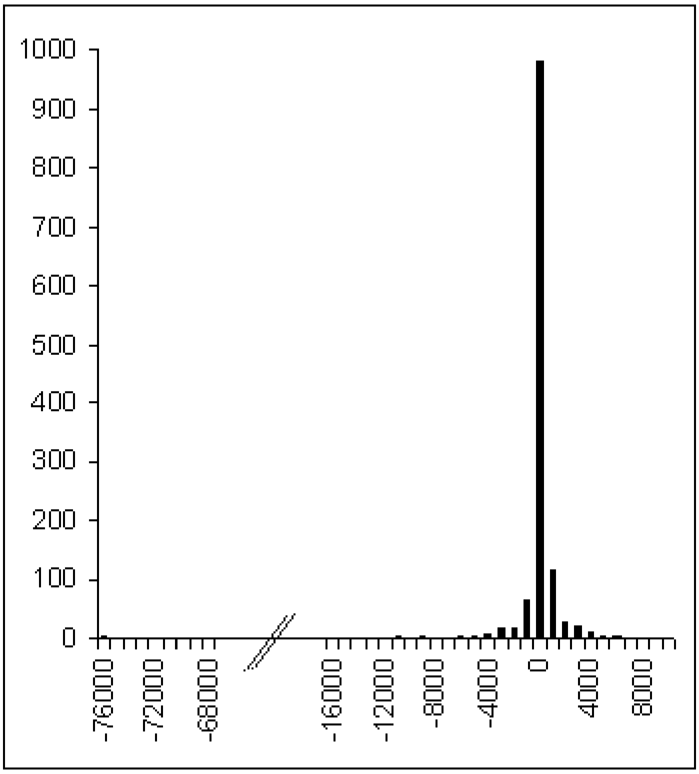


Figure 3.9. Histogram showing the frequency of TFBSs found different distances from the TSS in TRANSFAC Professional. Frequency is measured in number of TFBSs found at the position, start site is measured in nucleotides upstream (-) or downstream (+) from the transcription start site (0).

3.6.2. Feature Selection for PPMNB

All four feature selection methods used, SVM, Chi Squared, InfoGain and GainRatio, found distance from the TSS, armadillo (*Dasypus novemcinctus*) orthologs and cat (*Felis catus*) orthologs to be the highest scoring features. This convergence of results indicates that the features chosen are likely to represent the real optimum rather than a local optimum in the search space. Chi squared and Info gain methods both selected Guinea Pig (*Cavia porcellus*) orthologs as the fourth feature. The SVM algorithm selected horse (*Equus caballus*) orthologs as the fourth feature and GainRatio selected Stickleback (*Gasterosteus aculeatus*) orthologs. All three features selected as the fourth possible feature were assessed to determine which gave the most improved performance.

Three of the feature selection algorithms, Chi Squared, Info Gain and Gain Ratio, gave very similar rankings for each of the organisms (Figure 3.10). The SVM classifier gave very different rankings to these other methods. The SVM classifier uses a supervised machine learning method of feature selection compared to the other methods which use simpler statistical functions for feature selection and it is most likely to be this difference which is showing in the rankings.

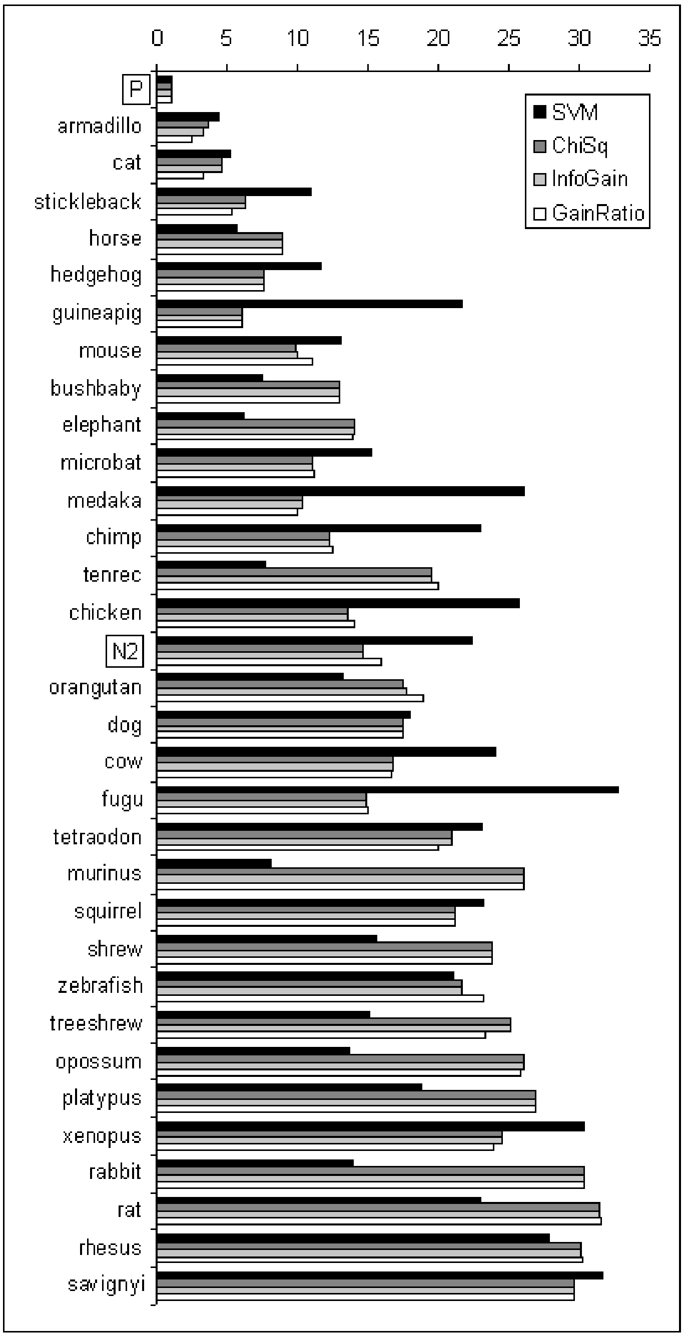


Figure 3.10. Histogram showing the rank given by each of the four feature selection methods. The features are ordered by average rank. The features are named according to the naming scheme used in the Biomart entry point to Ensembl which accounts for the mixture of common and Latin names for the organisms. Features which are not orthologous comparisons are shown in boxes. These features are 'P', the position of the TFBS on the sequence and 'N2' the mean probability of nucleosome occupancy across all positions in the TFBS.

3.6.3. Naïve Bayes

As with the PMANB classifier, the PPMNB classifier consistently classified all examples as negative when the ProfessionalTFBS_20467 dataset was used. The negative classification of the entire data set occurred with any and all feature combinations attempted. This was due to the very small positive *a priori* value that is calculated when a large imbalance in numbers of positive and negative training data exists. Due to the small number of positive test examples compared to the large number of negative test examples the specificity was close to 100% while the sensitivity was 0%.

The alternative dataset included equal numbers of positive and negative examples. There were three datasets: ProfessionalTFBS_230a, ProfessionalTFBS_230b and ProfessionalTFBS_230c. When these datasets were used for the training and testing of the classifier both positive and negative predictions were made. This was due to the reduced bias towards negative prediction.

Using only two features, ‘P’ (TFBS start site position) and the phylogenetic conservation of TFBS with armadillo orthologs, the classifier gave high AUC values >0.90. The mean sensitivity was 93.9%, although the mean specificity was lower at 71.3%. The addition of another feature, phylogenetic conservation with cat, improved both mean specificity and sensitivity scores (Figure 3.11). However the variation between ROC curves increased considerably with the addition of this extra feature (Figure 3.11). This variation made it difficult to determine whether the added feature improved the performance of the classifier. Ideally more datasets would be used to find a consensus for the performance of the classifier. Although it is difficult to see if there was an increase in classification performance using the ROC curves, the sensitivity and specificity showed an improvement in performance so the third feature, phylogenetic conservation with cat was retained. The fourth features were included in the classifier to determine if this would further improve the classification.

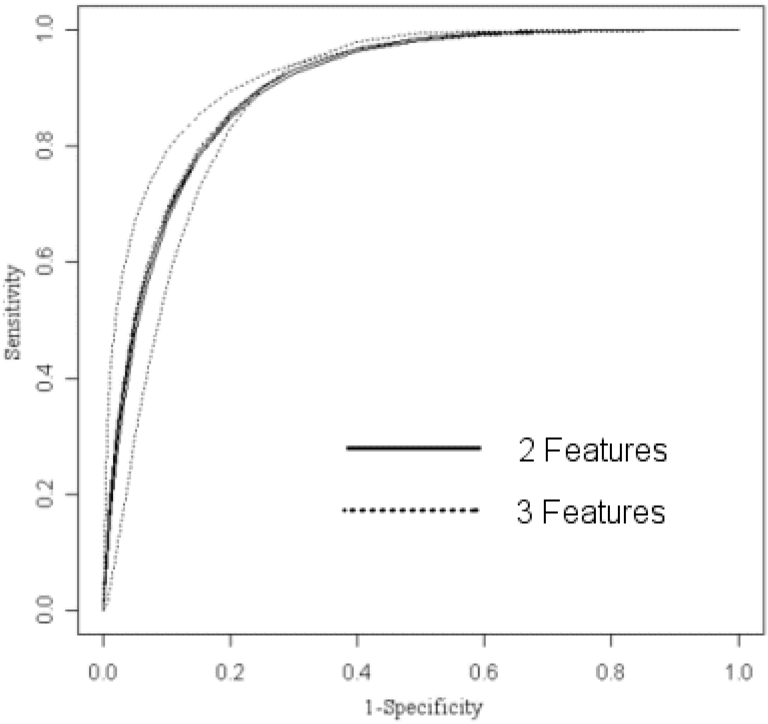


Figure 3.11. ROC curve showing the performance of the classifier using distance and armadillo orthologs as features (solid) and using distance, armadillo orthologs and cat orthologs as features (dotted). Each feature set is trained and tested using each of the three ProfessionalTFBS_230 datasets and the resulting ROC curves are plotted separately.

Three possible fourth features were selected by the feature selection algorithms. The three options were conservation with horse orthologs, guinea pig orthologs or stickleback orthologs. All features selected by these algorithms as possible fourth features were used in the classifier to determine which gave the optimum performance.

The conservation with hedgehog orthologs feature was the fourth feature when calculated by mean rankings, however it was not the fourth feature for any given feature selection algorithm so was not included in this analysis. Figure 3.14 shows the mean sensitivity, specificity and accuracy observed when using each of the possible fourth features, along with the three features previously chosen, in the Naïve Bayes classifier. In each case conservation with horse orthologs gave higher scores than with guinea pig orthologs or stickleback orthologs. However, this was only a slight increase in performance, particularly when the standard deviation, shown as error bars on Figure 3.12, was taken into account.

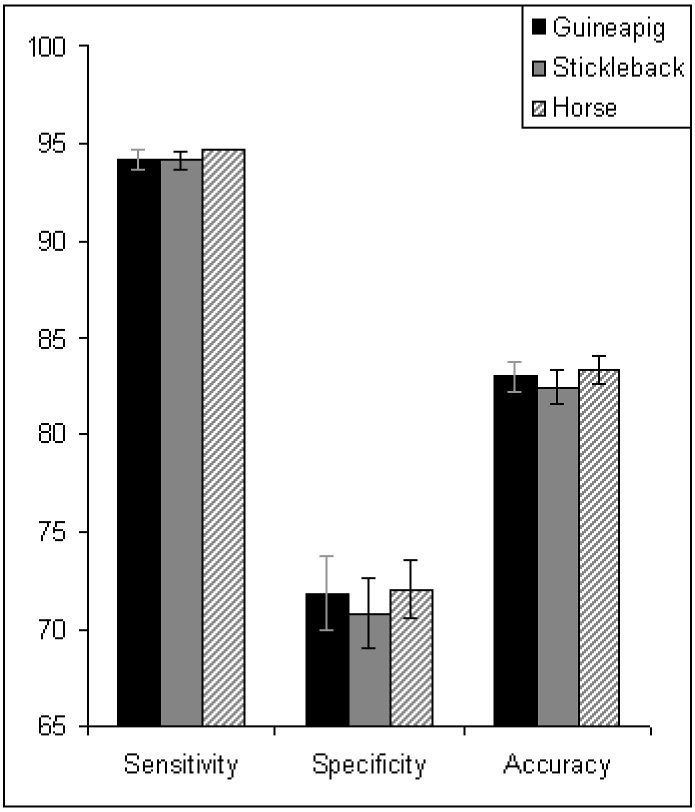


Figure 3.12. Mean sensitivity, specificity and accuracy scores for the classified dataset when using distance, armadillo and cat orthologs as features plus one of the fourth features identified by the feature selection algorithms. Error bars represent one standard deviation from the mean.

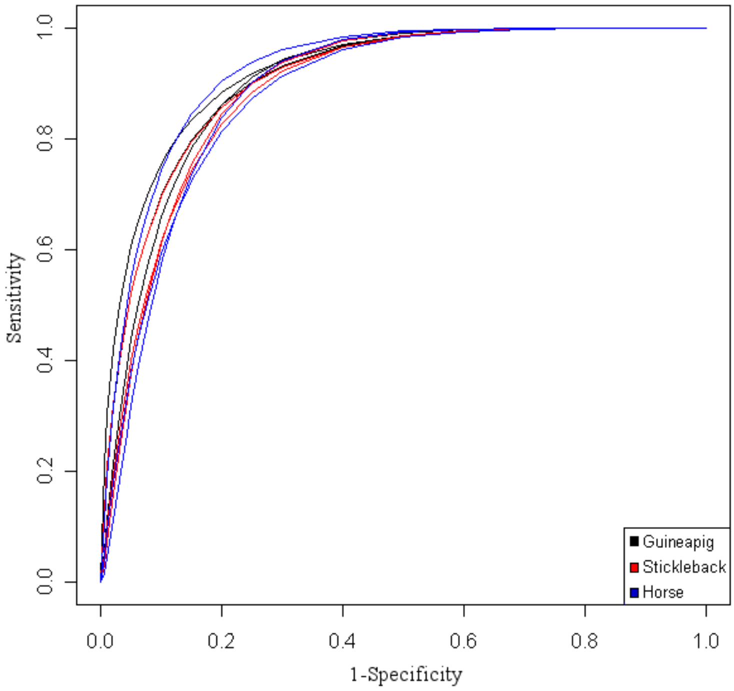


Figure 3.13. ROC curves showing the classification performance using distance and armadillo and cat ortholog features, plus one of the fourth features chosen by the feature selection algorithms. The fourth features analysed are: guinea pig orthologs, stickleback orthologs and horse orthologs. The ROC curves created by using each of the ProfessionalTFBS_230a datasets are plotted individually.

The ROC curves describing the performance of the classifier when using each of the different possible fourth features did not show a consensus hierarchy of possibilities for the fourth feature. One of the ‘conservation with horse orthologs’ datasets had the highest AUC: however, another of the ‘conservation with horse orthologs’ datasets had the lowest AUC. The average AUC for the ROC curves was the highest when using ‘conservation with guinea pig orthologs’ as the fourth feature. Although the ‘conservation with horse orthologs’ gave the highest sensitivity and specificity values, the difference between these results and the results obtained using the other possible features was very small and no greater than the error limits for the other two orthologs. Therefore, ‘conservation with guinea pig orthologs’ was taken as the fourth feature for the classifier.

The addition of ‘conservation with stickleback orthologs’ as the fifth feature did not alter the classification of the data. However, the addition of the ‘conservation with horse orthologs’ feature improved the mean sensitivity and specificity scores for the classifier (Figure 3.14). The size of the standard deviations of the sensitivity and specificity results showed that the difference in means between using four or five features was not a significant result. The AUC values were not improved using either fifth feature.

As the addition of an extra feature added to the training and running time of the classifier, it was decided to stop adding features at this point. The optimum feature set was taken to be; ‘P’, ‘conservation with armadillo orthologs’, ‘conservation with cat orthologs’ and ‘conservation with guinea pig orthologs’.

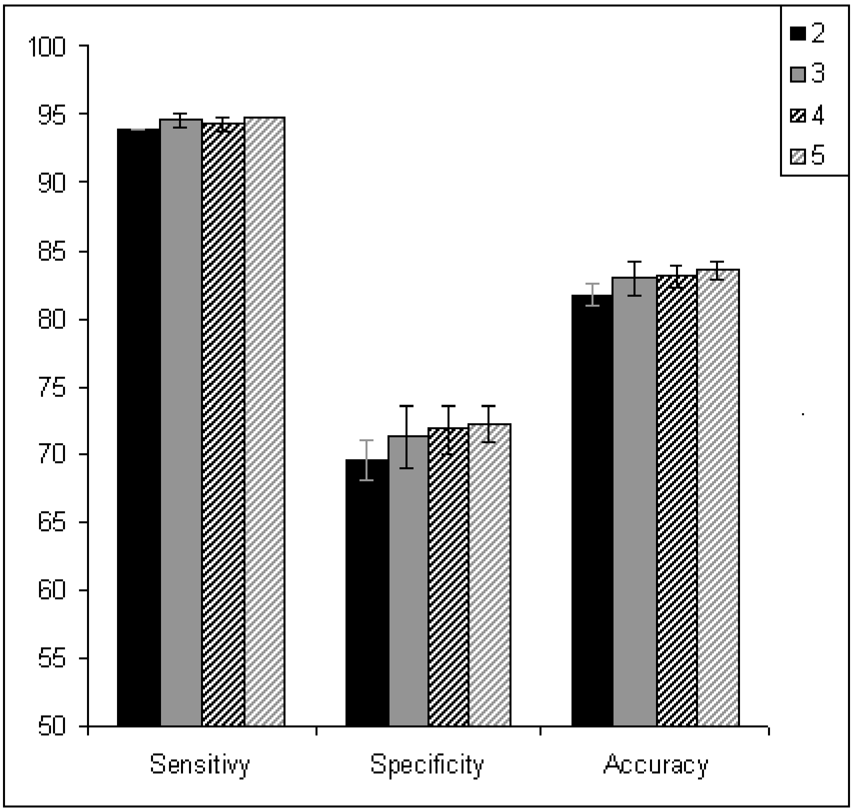


Figure 3.14. Mean Sensitivity, Specificity and Accuracy scores achieved by the Naïve Bayes classifier when using 2, 3, 4 or 5 features. Error bars show 1 standard deviation from the mean. 1st feature = distance, 2nd feature = Armadillo orthologs, 3rd feature = Cat orthologs, 4th feature = guinea pig orthologs, 5th feature = Horse orthologs.

3.6.4. PPMNB Naïve Bayes method comparison

There was no significant difference found between using the positive Naïve Bayes methods, PNB and PTAN, and the classic Naïve Bayes method in the PPMNB classifier (Figure 3.15). This was the case when assessment was made based on sensitivity and specificity values at 0.5 or using AUC. Significant differences between the methods were calculated using z statistics as described in Hanley and McNeil's work for comparing related ROC curves (Hanley & McNeil, 1983; McNeil & Hanley, 1984). The three Bayesian methods had significantly higher AUC values than the AUC for predictions made using ConSite with mouse orthologs.

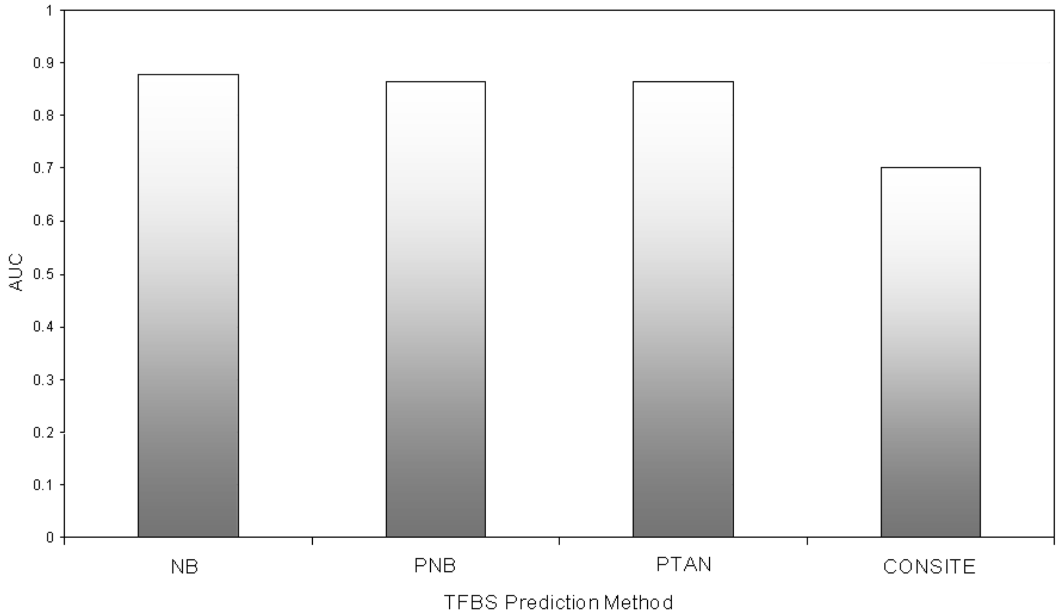


Figure 3.15. Empirical AUC values (derived from ROC curves) for the 3 best performing naive bayes methods using ConSite scores with ortholgues from 3 species (Armadillo, Cat and Guinea Pig) and distance from the TSS as features. The AUC value for ConSite (using mouse orthologues only) is shown for comparison. AUC values are shown as a bar chart rather than a ROC curve due to the similarity between NB, PNB and PTAN curves.

3.7. Applying the methods to the genes from the IFN-γ genetic linkage study

Both the PMANB and the PPMNB classifiers were used to predict TFBSs for the genes in the gene set from the IFN-γ linkage study. The PMANB method predicted at least one transcription factor binding site for 523 of the 532 genes. The PPMNB method predicted only 35 genes with TFBSs. The discrepancy between the numbers of genes predicted by each method was due to the limitations imposed by using phylogenetic features as described in section 3.3.1.

75 different transcription factors had TFBS predictions made by the PMANB classifier in genes from all three regions of interest (chromosomes 8, 10 & 11). 53 transcription factors were predicted by the PPMNB classifier to bind to genes in each of the three regions. In each case there were too many combinations to successfully search for co-regulated genes without recourse to computational clustering methods.

Although the PPMNB classifier predicted a smaller range of transcription factors than did the PMANB classifier, a small number of transcription factors were predicted using the PPMNB classifier that were not predicted using the PMANB classifier e.g. Pax4 and Pax6 (Figure 3.16). However, there were far more transcription factors which are predicted to bind by the PMANB classifier which were not predicted by the PPMNB classifier. Examples of these transcription factors were the EGR family of transcription factors, the GATA family of transcription factors and SP1.

Binding sites for the majority transcription factors which were predicted by the PPMNB classifier were found in high numbers of genes across the dataset. In some cases the transcription factor was predicted to bind every gene in the set. By contrast, the PMANB classifier predicted that each transcription factor bound to a much smaller percentage of the genes. This was most likely to be due to the different lengths of sequence across which the transcription factors were predicted. The algorithm based classifier searched a sequence of length 1400b while the phylogenetic classifier searched sequences of length 9.5Kb. It would be much more likely for a transcription factor binding site to occur by chance in the longer sequences used by the PPMNB classifier.

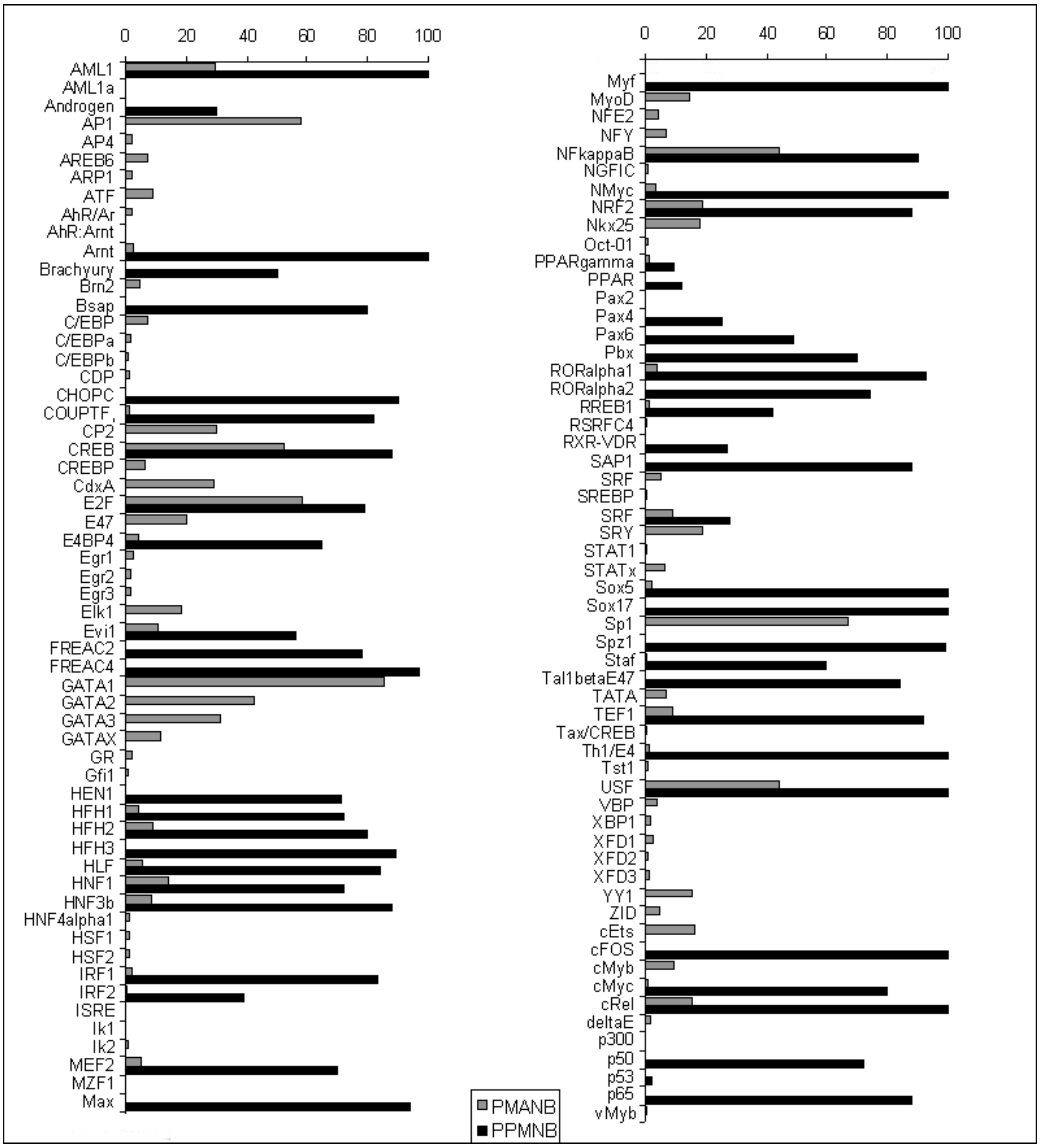


Figure 3.16. Chart showing the percentage of genes in each the relevant dataset with binding site occurrence for specific transcription factors. Transcription factors were predicted either through the Phylogenetic Pattern Matching Naïve Bayes classifier (PPMNB) or the Pattern Matching Algorithm Naïve Bayes classifier (PMANB).

3.8. Conclusions

3.8.1. Comparison of Naïve Bayes methods

PMANB and PPMNB, while both high performing classifiers, had different strengths in the classification of TFBS predictions. The PMANB classifier had a better performance as measured by AUC but the distance from the transcription start site at which TFBSs can be found was small. The classifier could not be trained with longer sequences because of the limits imposed by some of the algorithms used as features. PPMNB, therefore, had an advantage when searching for TFBSs further from the TSS, as it was trained and tested using much longer sequences.

The PMANB classifier had an advantage in the number of transcription factors it was able to predict. The variety in algorithms used to train the model allowed for a greater number of TFBSs and, therefore, regulatory transcription factors, while the use of Naïve Bayes to train the model overcame the problem of some algorithms being unable to predict the TF in questions. PPMNB, in its current version, was restricted to the TFBSs which are searched for by ConSite. This resulted in the identification of a much smaller number of possible regulatory TFBSs (Figure 3.16). However, as PPMNB was built using the extremely versatile TFBS Perl modules (Lenhard & Wasserman, 2002), it would be possible to increase the number of TFBSs searched for with the method. Any new TFBS matrices added to the PMANB classifier could have different orthologous conservation profiles than those currently used by the classifier. Therefore, if adding in additional TFBS matrices, it would be prudent to re-test the classifier to ensure the high level of performance was retained and that the current features chosen are still the optimum selection.

The comparison between PPMNB using the classic Naïve Bayes algorithm and PPMNB using either PNB or PTAN, the positive Naïve Bayes algorithms, showed no significant difference in performance. The positive Naïve Bayes classification methods had been included due to the lack of experimentally verified negative training examples for use in the Naïve Bayes classifier. The negative examples which were used in the classic Naïve Bayes classifier were defined by their lack of a positive label, not by experimental evidence showing that the prediction was not a functional TFBS. PNB and PTAN, the positive Naïve Bayes methods, did not, however, have a higher AUC than the classic naive Bayes classification. Although the problem appears to be an example where the positive naive Bayes methods would be useful, the PNB and PTAN classifiers did not have a higher performance than the classic Naïve Bayes classifier; neither did they perform any worse.

If there were a clear distinction between the characteristics which define the positive examples and the negative examples, the difference between the classic and positive classifiers would be expected to be small. In this case it would not matter whether the classifier works by assigning a high probability that an example was positive and a low probability that the example was negative, as in the classic Naïve Bayes classifier, or simply a high probability of the example being positive as in the positive Naïve Bayes classifier. Using either classifier the example would be predicted as a positive datapoint. It is only when the negative and positive classifications overlap in their feature space that the classifiers would perform differently. If an example was to fit the criteria for being a positive example and for being a negative example the classic Naïve Bayes classifier would assign the example a probability of approximately 0.5. If the probability was just less than 0.5 the example would be pushed into the negative predictions, if just over 0.5 it would be included in the positive predictions. The positive Naïve Bayes classifiers would categorise this example as positive because it fit the criteria for a positive example (table 3.2).

Example	*Class*	*Positive probability*	*Negative probability*	*NB result*	*PNB/PTAN result*
1	1	High	Low	1 (TP)	1 (TP)
2	1	High	High	0 (FN) or 1 (TP)	1 (TP)
3	0	Low	High	0 (TN)	0 (TN)
4	0	Low	Low	0 (TN) or 1 (FP)	0 (TN)

Table 3.2. Examples of the different classification of data by PNB or PTAN and classic NB.

When the example fits the positive criteria, giving a high positive probability, the PNB and PTAN classifiers will predict it as a functional binding site, however for the NB classifier it also depends on the fit of the negative data. When the example does not fit the positive criteria, giving a low positive probability, the PNB and PTAN classifier will predict it as a non-functional binding site, the prediction made by the classic Naïve Bayes classifier again depends on the fit of the negative data.

3.8.2. Application to the genes from the genetic linkage IFN-γ study

In the results from the simple combinations of programs, in Chapter 2, genes which shared predicted regulatory transcription factors could be identified. The methods of TFBS prediction outlined here predicted many more TFBSs, and therefore, regulatory transcription factors, for each gene than the analysis in chapter 2. For the prediction made by the phylogenetic classifier, it was not feasible to group the genes by single transcription factor. This was due to the fact that many transcription factors have binding sites occurring on all or most of the genes for which predictions were made.

The PMANB classifier found a number of genes associated with a larger number of transcription factors. This would still give a large number of genes which may be co-regulated by each transcription factor. Gene regulation is more likely to be reliant on groups of transcription factors than on single transcription factors. Therefore it is a more useful and accurate analysis to look at all of the transcription factors predicted to regulate the gene in concert rather than singly. This was investigated further in chapter 4.

4. Predicting function from regulation

4.1. Introduction

The TFBS predictions from the consensus method in Chapter 2 were used to find genes which may be co-regulated by a single transcription factor. Although a number of candidate genes were found, only a small number of transcription factors could be analysed using this method. This is because each single prediction method scores motifs against only those TFBS matrices in its database. The method of combining these algorithms reduced the number of possible TFs still further to only those TFs which multiple algorithms are able to predict. Conversely, the TFBS prediction machine learning methods developed in Chapter 3 predicted a large number of TFBS in a large number of genes. In the case of PMNB this is because the combining of algorithms is much less restrictive, still allowing the prediction of TFBS which are predicted by only one algorithm. The PPMNB method only uses one algorithm, with a large database of TFBS matrices and so is not restricted in the same way. The large number of TFBS matrices for which predictions can be made makes the space in which to search for co-regulated genes very large. Using these predictions it would be difficult to predict co-regulation by analysing the co-occurrence of single TFs. To address this issue two approaches can be taken; the inclusion of additional experimental data, and the identification of and search for groups of TFs that work co-operatively.

4.1.1. The use of experimental data to aid the prediction of co-regulated genes

The addition of experimental data would significantly aid the discovery of functionally similar genes in the IFN-related gene dataset. Many methods have been developed which use experimental evidence to improve the prediction of functionally similar groups of genes. For example, the combination of in-silico prediction techniques with large-scale genetic screening has been used as a tool for functional gene discovery in *Drosophila* (Aerts *et al.*, 2009) and a combination of genome-wide expression data, TFBS predictions and ChIP-Chip data were used to identify a PU.1 regulatory network in macrophages (Weigelt *et al.*, 2009).

Several publically available databases exist which contain data from microarray experiments. Examples of these are the NCBI database, Gene Expression Omnibus (GEO) (Barrett, T. *et al.*, 2007), and the EBI database, ArrayExpress (Parkinson *et al.*, 2009). However, we were unable to obtain a complete set of data suitable for this analysis – ideally this would be an analysis of the expression of the genes in the IFN-γ linked gene regions in BCG vaccinated naïve individuals from the Gambia. Additional microarray experiments would need to be conducted if the methods were to be applied to the genes from the regions linked to the IFN-γ response. Therefore, within the limitations of this thesis, the addition of experimental data was not possible.

4.1.2. Cis-Regulatory modules

Transcription factors rarely act autonomously in the regulation of a gene. A much more common situation involves groups of TFs working co-operatively or antagonistically (Lee *et al.*, 2008; Santalucía *et al.*, 2001; Takahashi, *et al.*, 2002). Cis-regulatory modules (CRMs), also known as composite regulatory modules, are groups of two or more closely situated TFBSs which enable multiple TFs to regulate a gene. The analysis of genes with correlated expression, measured using experimental techniques such as microarrays, has allowed for the identification of many CRMs (Baldwin *et al.*, 2005; Kim & Jung, 2006). Recent analysis (Danko *et al.*, 2009) into degenerative heart disease identified a CRM consisting of a TATA-box followed by a CACC-box. This motif was found in 9 genes involved in myocardial contraction. A second CRM was found in 20 genes involved in translation. This module consisted of a pyrimidine-rich initiator, Elk1, Sp1, and a novel motif with a GCGC core (Danko *et al*., 2009).

Known CRMs can be found in TRANSFAC's sister database TRANSCompel (Kel-Margoulis *et al*., 2002) and in other databases such as PReMod (Ferretti *et al*., 2007) or CisRed (Robertson *et al*., 2006). Some of these CRMs are experimentally verified, such as those in TRANSCompel, others, such as those in CisRed are statistically inferred from either known or predicted TFBS. The annotation of CRMs is progressing at a slower rate than that of the transcriptome (Van Loo & Marynen, 2009), hence the need for the statistically inferred modules. However, even with the statistically inferred modules the number of known CRMs is still small. A search of TRANSCompel public found only one CRM known to be involved with the immune response (ZIP$REL_004). The module consisted of a C/EBPbeta site at -179 to -168 and a NFκB or NFκB p65 site at -91 to -82 (Xia *et al*., 1997). Both of these TFBS are able to be predicted using the PMNB classifier; however no genes were found with predictions at positions similar to the reported CRM. It does not appear likely from the annotations associated with other known CRMs found in TRANSCompel that our system is regulated by one of these known modules.

4.1.3. CRM prediction algorithms

A number of CRM prediction algorithms have been recently developed. The majority of these algorithms search for statistically over represented groups of TFs. The search for CRMs is usually carried out by comparing groups of TFBS in sets of co-regulated genes (Zeng *et al*., 2008). Ab-inito methods of CRM discovery, involving the search for statistically over-represented motifs within groups of co-expressed genes, have been successful in a number of studies. The results from one such study are described in Section 4.1.2 (Danko *et al*., 2009) and many other CRMs have been discovered using such methods (Choi, D. *et al*., 2008; Niida *et al*., 2009; Tuteja *et al*., 2008). Methods which search for groups of over-represented motifs rely on the existence of a set of genes which exhibit similar expression profiles. These methods are not applicable to the set of genes from the regions linked to the IFN-γ immune response because we are trying to find a subset of co-regulated genes within a data set of genes that are unlikely to have similar expression profiles.

The ab-initio methods described above are not suitable for the identification of co-expressed IFN-γ linked genes in the set of genes in the IFN-γ linked regions. These methods aim to find CRMs from genes which are known to be co-expressed, while this analysis aims to identify a subset of co-regulated genes within the IFNγ-linked dataset. Another challenge is that our set of TFBS is predicted. With the use of predicted TFBS it must be assumed that some TFBSs are false positive results and that other TFBSs are not predicted by the method. This is problematic when searching for similar clusters of TFBS. In this chapter, we analyse the presence of multiple TFs, but not their positions. Three approaches have been taken in this chapter (a) a Principal Coordinate Analysis (PCO), (b) a Naïve Bayes classifier (described in Section 3.1.1) and (c) a simple graph analysis. These approaches will be used to look for groups of shared TFs that correspond to genes with similar functions.

4.1.4. Principal Coordinate Analysis

PCO is a multidimensional scaling method (Gower, 1966) which allows for data points to be clustered using a smaller set of features than they were originally described with. The new features which are used to describe the data are entirely new, unlike the features in Principal component analysis which are a combination of existing features. The new features in PCO are created by representing the data in multidimensional Euclidean space and determining the smallest set of axes to use to accurately describe this data. The data is described as a distance matrix; this is the representation in Euclidean space. PCO determines the number of axes on which to plot the data from the distance matrix in order to minimise a loss function referred to as stress (Borg & Groenen, 2005). When stress is at a minimum level this implies that the data is correctly described using the current number of axes. The axes represent the new features with which the data is described. The level of stress declines as the number of axis increases, however the aim of the PCO analysis is to reduce the data to a smaller number of features, represented by the axis. Therefore a compromise must be reached.

A screeplot can be used to describe the level of stress occurring in the model. Screeplots give a measure of the variance described by each axis given by the PCO. The level of variance is measured using the eigenvalues of the matrix. Eigenvalues are properties of a matrix, used in matrix arithmetic. When a matrix acts upon a vector it affects both the magnitude and direction of the vector. However when a matrix acts upon certain vectors, eigenvectors, the magnitude is altered, but not the direction. The factor by which the magnitude is altered is the eigenvalue. The optimum number of axes used to describe the data is given by the ‘elbow’ of the screeplot (Steyvers, 2002; Zhu & Ghodsi, 2005). The elbow can be seen between the 3rd and 4th axes on the example screeplot shown in Figure 4.1. The optimum number of axes in this case would be 3.

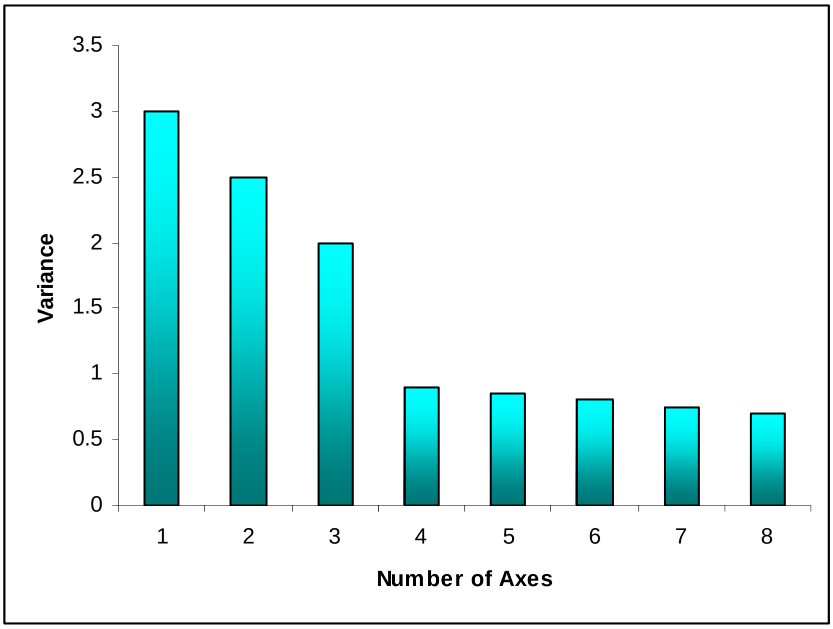


Figure 4.1. Example screeplot showing an elbow in the data between the 3rd and 4th axes.

PCO has not been used to cluster genes according to TF occurrence before. However the method has been used in numerous biological problems to allow for the clustering of data by a smaller number of features than the original representation contained. One use of PCO has been the clustering of genomes and assessment of variance. The data in these studies was originally described using polymorphisms as markers (Sobral *et al*., 2009; Lam *et al*., 2009; Perumal *et al.,* 2008; Su *et al*., 2009; Zhang *et al*., 2009). A second use has been in the categorisation of protein structures and folding states (Palyanov *et al*., 2007; Manson *et al*., 2009; Yang *et al*., 2006)

4.2. Methods

4.2.1. Dataset

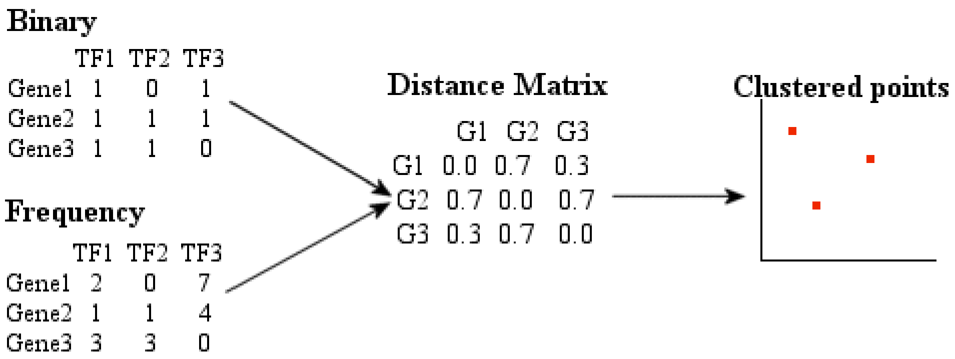
Biological process GO (BP GO) annotations (Ashburner *et al.*, 2000) were retrieved using BioMart (Smedley *et al.*, 2009) for the 532 genes in the regions genetically linked to an IFN-γ immune response to TB. The BP GO annotations were available for 392 genes in this dataset. These genes were used as the dataset for the co-regulation analysis, designated as the IFNγ-regions dataset. A path specific relative similarity statistic (PSRSS) score based on the BP GO annotations was calculated between each possible pair of genes in the IFNγ-regions dataset using a set of Java modules developed by D. Damerell (personal communication). The PSRSS score denotes the level of functional similarity between pairs of genes; the smaller the score the more similar the functions of the genes.

TFBS predictions were made for each of the 392 genes in the IFN-γ-regions dataset using the PMNB method described in sections 3.2.2 and 3.5.2. The PMNB method, rather than the PPMNB method, was used to predict TFBS because binding sites could be predicted for the greatest range of transcription factors.

4.2.2. Principle co-ordinate analysis

A matrix was constructed containing the transcription factors predicted to bind to each gene sequence in the IFNγregions dataset. The rows in the matrix represented genes (392 genes) and the columns represented transcription factors (84 TFs). The matrices represented the TFBS predictions in two ways: a binary representation and a frequency representation (Figure 4.2). The binary matrix represented only whether at least one

TFBS for each TF was predicted (1) or not predicted (0) for each gene. The frequency matrix contained the number of TFBS predicted for each TF in each gene.



Binary

	TF1	TF2	TF3
Gene1	1	0	1
Gene2	1	1	1
Gene3	1	1	0

Frequency

	TF1	TF2	TF3
Gene1	2	0	7
Gene2	1	1	4
Gene3	3	3	0

Distance Matrix

	G1	G2	G3
G1	0.0	0.7	0.3
G2	0.7	0.0	0.7
G3	0.3	0.7	0.0

Figure 4.2. Figure describing the processes of clustering genes by frequency or binary TF matrix. Matrices are first created describing the number or occurrence of each type of TF in each gene. The matrices are converted to a distance matrix to which a clustering algorithm (cmdscale) was applied.

Both types of matrices (binary and frequency), were converted into distance matrices using the R (dist) function (http://cran.r-project.org/). The dist function is used to calculate the dissimilarities between the rows of a numeric data matrix using a specified the distance. In this analysis we have used the Euclidean distance (Equation 4.1).

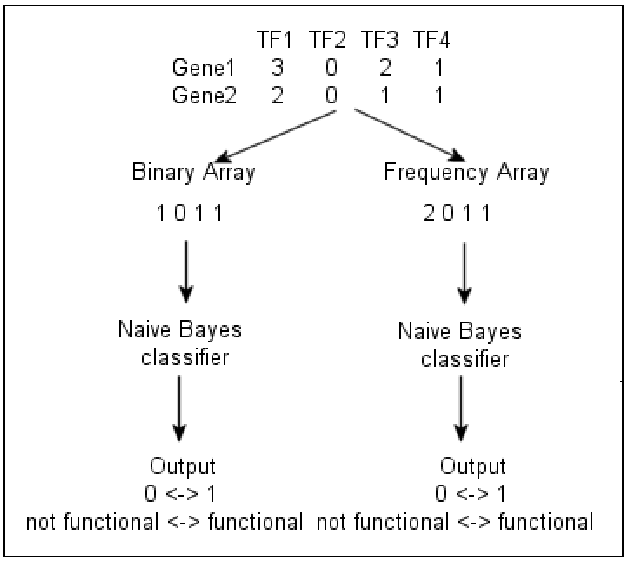
$$\sqrt{}((x_1 - x_2)^2 + (y_1 - y_2)^2)) \qquad \textbf{(4.1)}$$

PCO was used to cluster the genes based on the distance matrices (Gower, 1966). The implementation of PCO was conducted using the R (cmdscale) algorithm. Screeplots were calculated for both the binary and frequency matrices to determine the number of axis needed to represent the data. The optimum number of axes was identified using the elbow technique (section 4.1.4). Euclidean distances were calculated between the points plotted using the co-ordinates given by the PCO analysis, for each gene pair. These distances were compared to the PSRSS scores. The correlation between these two scores was measured by Pearson's correlation coefficient.

The PCO analysis carried out on the frequency matrix identified a problem with the scaling of the TF features (Section 4.3.1). Due to a large range of frequencies associated with some TFs, and a small range of frequencies associated with other TFs, the TFs were analysed according to different scales. PCO analysis relies on the features being described using the same scale. This problem was solved by removing the TFs which had a range of frequencies greater than 1 times the standard deviation from the mean. This is a stringent criterion for removal of values. However, the criterion was used because the ranges of frequencies observed were substantially skewed towards 1 (Figure 4.7). Although this is a simplistic method of solving the problem it has been used successfully in other studies which search for patterns in TFBS occurrence to reduce the impact of frequently occurring TFBS (Hatanaka *et al*., 2008).

4.2.3. Naïve Bayes

A Naive Bayes classifier has also been used to classify gene pairs as functionally related, based on the TFs found to bind each of the genes. In this classifier the data points for classification were pairs of genes and the features were TFs. The classes into which the datapoints were classified were: functionally related, as measured by a PSRSS score below a given threshold (denoted by 1), and not functionally related, measure by a PSRSS score above a given threshold (denoted by 0). The Naive Bayes classifier was implemented using Weka interface as described in section 3.2.2 (Witten *et al*., 1999).



	TF1	TF2	TF3	TF4
Gene1	3	0	2	1
Gene2	2	0	1	1

Figure 4.3. Diagram showing the inputs and outputs for the naïve Bayes classifier. The raw input contains a frequency matrix of genes and transcription factors (TF) binding to predicted sites within the gene or promoter sequence. The naïve Bayes input consists of an array for each gene pair, either a binary encoding or a frequency encoding. The output of the classifier is a probability score representing a gene pair which is very functionally similar at 1 and a gene pair is which is very functionally different at 0.

The features for the classifier were encoded in two ways, but each used individual TFs as features (Figure 4.3). For each gene pair, the specific TFs for each TFBS predicted by PMNB were compared. The first coding, (a), was a binary representation. An array was created for each pair of genes where each element in the array represented one of the 84 TFs predicted using the PMNB method. If both genes contained predicted binding sites for the TF in question the array position was set to 1, otherwise the position was set to 0. In the second coding, (b), the feature was encoded with the number of TFBS both genes contained, e.g. if one gene contained 12 TFBS for a transcription factor and the other gene contained 8, the array element for that position would contain 8, as both genes contain at least 8 TFBS for that transcription factor. The class (functionally related or not functionally related) was defined by the PSRSS threshold value. A varying threshold of PSRSS scores was used to determine the classifier's performance at different levels of similarity.

The proportion of functional and non-functional pairs in the dataset is altered as the PSRSS threshold score changes. When the threshold is set at a low level most pairs of genes are classified as not-functionally similar. As the threshold level is raised, the number of gene pairs classified as functionally similar increases. The Weka implementation of the naïve Bayes classifier calculates the *a priori* values from the number of data points in the training set classified as positive (functional) or negative (non-functional). When using a 5 fold cross validation method, as described in sections 3.2.2 and 3.3.4, 80% of the data is used for training and 20% for testing. If the entire dataset was used with each analysis, when the threshold level was changed, the *a priori* values would also change. Some bias in the *a priori* value may be useful, for example, with the PSRSS score set at 0.9 it can be assumed that most gene pairs should be classified as functionally similar. However, it was shown in Sections 3.5.1 and 3.6.1 that the classifier was unable to overcome a large bias in *a priori* value, classifying every data point according to the bias.

A smaller subset of data was created for each threshold, artificially setting the *a priori* value to 0.5. Datasets were created by taking a randomly sampled set of data points so that the set of functional and non-functional data points were the same size, and of the largest size possible. The datasets created this way were used to train and test the classifier using 5 fold cross validation. This process of creating datasets with *a priori* values of 0.5 produced datasets of different sizes for each threshold value. This meant that the datasets for the threshold values in the middle of the scale were much larger than those at the extremes i.e. thresholds of 0.1 or 0.9. These datasets had a larger number of training examples and as a consequence were more likely to accurately train the classifier than the smaller datasets used for the extreme threshold values. For the task of classifying gene pairs as functionally or non-functionally related, we were most interested in threshold values in the middle of the range (0.40 - 0.60). Very low thresholds would only allow for the identification of very similar gene pairs and very high thresholds would only allow for the exclusion of very different gene pairs. Therefore it was decided to tolerate this bias in classification accuracy.

Chi squared feature selection (Liu and Setiono, 1995) was used to rank the features for the Naïve Bayes classifier. Feature selection was carried out for each of the coding type (binary (a) and frequency (b)), and for each of the PSRSS thresholds. After the classifier had been trained using all TF features, an analysis of the optimum features needed to train the classifier was carried out. The optimum number of features needed was identified by training the classifier on a dataset with features removed incrementally. Features were removed from the dataset in the reverse order selected by the Chi squared feature selection rankings.

4.2.4. Graph analysis

The use of graph theory to examine gene expression networks is a common one (Brustolini *et al.*, 2009; Janky *et al.*, 2009; Yanashima *et al*, 2009). It allows for the visualisation of interactions between genes and transcription factors and therefore the discovery of previously unknown relationships between biological entities. Often in these analyses both gene and TF will be represented as nodes (Janky *et al.*, 2009), however to simplify matters only the gene are represented as nodes in this analysis and the possible co-regulation of two genes via at least one TF is represented as an edge.

The TFBS predictions made in Chapter 2 were used to create graphs representing possible co-regulated genes (Section 2.3.3). A similar approach was used to determine possible co-regulated genes from the PMNB predictions made in Chapter 3. The analysis carried out in Section 2.3.3 created edges between genes which both contained a TFBS for a certain TF and which had a PSRSS ≤0.40. In this analysis the criteria is widened to allow edges between genes which share any TF. Initially for two genes to have an edge connecting them, both genes must contain at least one TFBS for the same TF. Subsequent analyses were carried out where the number of TFs in common between a pair of genes was increased. The largest number of TFs, for which both genes must have a predicted TFBS, used in the analysis was 6 because no genes were found with a larger number of TFs in common.

For each of the graphs created, the edges were annotated with the relevant PSRSS score. The number of known functionally related pairs (≤0.40) and the number of known non-functionally related pairs (>0.40) was analysed for each graph. The rate, at which the size of the graph decreased, as the number of TFs in common increased, was analysed. The size of the graph was calculated in two ways; the number of nodes and the number of edges. This was compared between the graphs where edges were determined by TFs in common and the graphs where the edges were determined both by TFs in common and by PSRSS score. This was to assess whether graphs with more stringent conditions produced more edges with a PSRSS score ≤0.4, therefore determining if the smaller graphs contained more accurate results.

Candidate genes retrieved from the graphs were analysed for possible CRM motifs. A search was carried out for TFBS with similar distances from each other and from the TSS in genes linked in the network graphs.

4.3. Results

4.3.1. Principal co-ordinate analysis

The screeplot calculated for the binary matrix showed a gradual decline in variance (measuring the level of stress). An elbow could be seen after the 5th axis (Figure 4.4). The elbow was not very pronounced, but it was assumed that this was the optimum number of axes. The elbow is small; this may imply that further axes are needed to describe the data. A large number of axes needed to describe the data can imply that either the original features do not combine easily to describe the data, or that the data is noisy (Zhu & Ghodsi, 2006). Is it likely that the data was noisy due to its predicted nature; therefore we shall assume that this was the main reason for the small elbow.

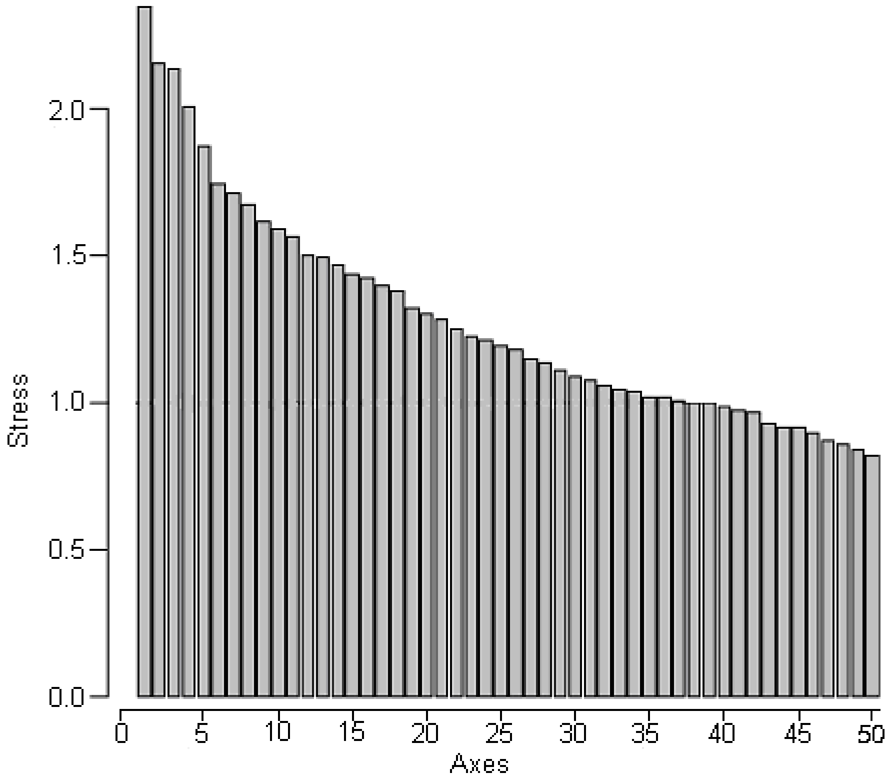


Figure 4.4. Screeplot showing the variances of the Eigenvectors in the binary PCO clustering for each axis. An elbow in the plot can be seen between axes 5 and 6.

The PSRSS score and the Euclidean distance, measured using the PCO coordinates, for each pair of genes was compared by Pearson's correlation coefficient. No correlation was found. Several sets of guidelines have been suggested for interpreting the size of a correlation given by the Pearson's correlation coefficient (Cohen, 1977). A general guideline, used for this analysis, is shown in table 4.1.

Correlation	*Negative*	*Positive*
Small	-0.3 to -0.1	0.1 to 0.3
Medium	-0.5 to -0.3	0.3 to 0.5
Large	-1.0 to -0.5	0.5 to 1.0

Table 4.1. Guidelines for interpreting the Pearson's Correlation Coefficient (Cohen, 1977).

When using the binary PCO clustering with 5 axes, the correlation was <0.1. According to the guidelines described in table 4.1, this shows that there was no correlation between the PSRSS score and the distances between genes as measure on the PCO axes.

When PCO was used with the frequency matrix the screeplot showed an elbow after the first axis. A strong bias towards only one axis is often a sign that the features used originally to describe the data, transcription factors in this case, are not all scaled equivalently. The bias towards the first axis in this analysis was caused by a small number of transcription factors which had a very high variation in the number of TFBS present in each gene. This meant that the data was clustered using only a small number of the transcription factors predicted.

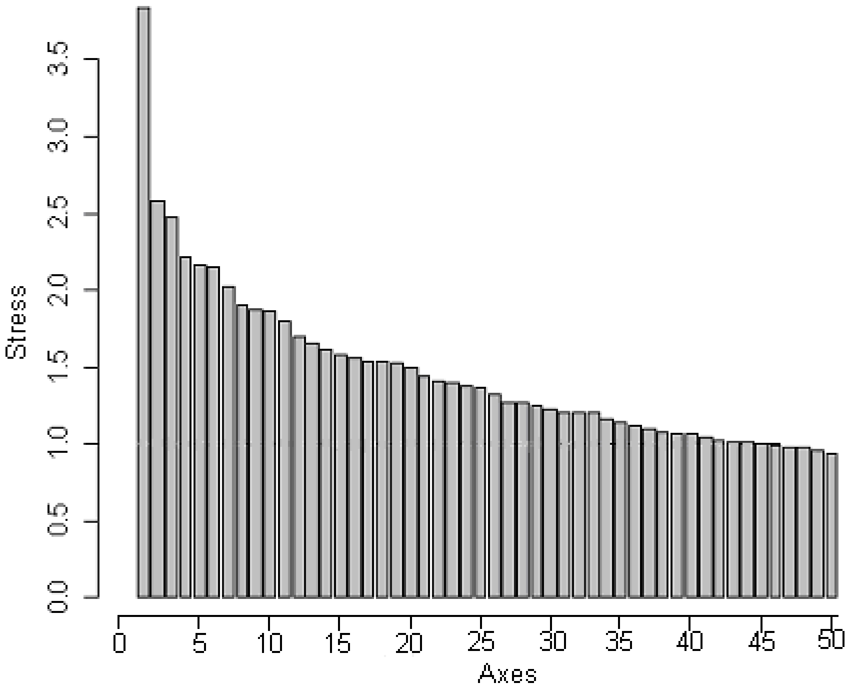


Figure 4.5. Screeplot showing the variances of the Eigenvectors in the frequency cmdscale clustering for each component. Axis 1 has a much higher variance than other axes.

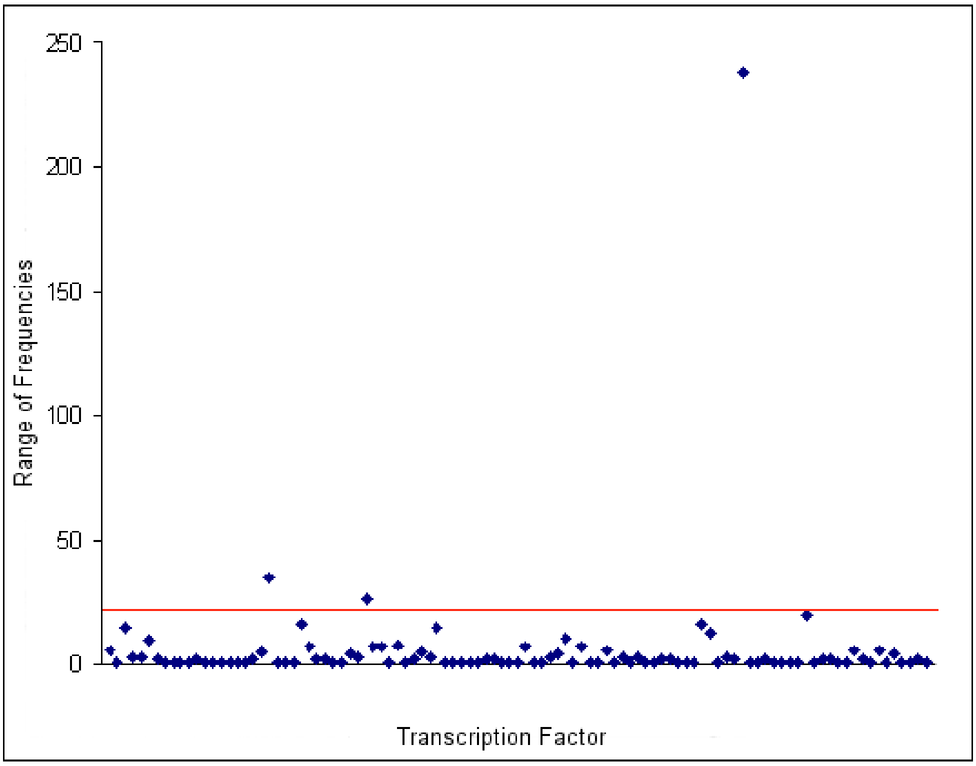


Figure 4.6. Chart showing the ranges of frequency of TFBS prediction shown for each TF. The threshold above which TFs were removed from the analysis (mean+standard deviation) is shown by the red line.

Three TFs were found above the threshold value of 1 times the standard deviation from the mean; Sp1, GATA1 and CREB (Figure 4.6). The most prominent example was Sp1 for which some genes were found to have 238 TFBS while other genes were found to have none. This created a different scale for the frequency of Sp1 TFBS than for other TFs. An example of this can be seen in the case of two genes, one with 2 TFBS for a certain TF and one with 4 TFBS for the same TF. If the TF was Sp1, this would be a negligible difference in TFBS frequencies between the two genes, however if this TF was one where 4 was the largest number of TFBS found in any gene, the difference in frequencies between the two genes would be large. In order to measure all TFs using the same scale, the 3 TFs with a larger range of frequencies were removed from the analysis.

The screeplot for the frequency matrix without Sp1, GATA1 and CREB TFs (Figure 4.8) still showed an elbow after the first axis. However, the amount of stress present was much lower than in the screeplot for the frequency matrix containing all TFs (Figure 4.6).

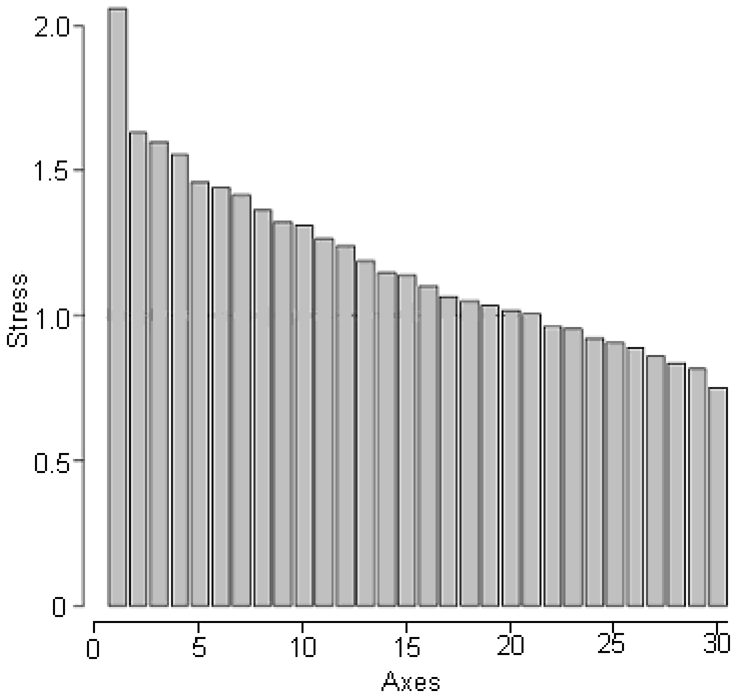


Figure 4.7. Screeplot showing the variances of the Eigenvectors in the frequency PCO clustering (without the Sp1, GATA1 and CREB transcription factors) for each axis. The most obvious elbow in the plot can be seen between axes 1 and 2.

The correlation coefficients for the comparison of PCO co-ordinates and PSRSS score, using the frequency clustering, are <0.1, this signifies that there was no correlation. The coefficients are smaller than those achieved using the binary matrix, however this is not a significant result (p>0.05).

In summary, the correlation coefficients, for either the binary matrix or the frequency matrix, (Table 4.2) between PSRSS score and distance between genes calculated by PCO, do not show a correlation.

4.3.2. Naïve Bayes

The classification of gene pairs into functionally similar and non-similar pairs was not successful when using the naïve Bayes method with either the binary or frequency encodings of the feature set (Figure 4.8). For the binary encoding the predictions made by the naïve Bayes classifier gave AUC values <0.55 for all threshold values. The frequency encoding gave AUC values <0.55 for threshold values between 0.10 and 0.70. The AUC value increased to between 0.55 and 0.60 (Figure 4.9) a threshold value of 0.80 or 0.90. Despite this slight increase, the AUC values are <0.70 and the classifier cannot be considered successful.

Reducing the number of TFs used as features in the classifier did not significantly affect the AUC values. A slight drop in AUC values can be seen when using only 20 features at the extreme threshold values of 0.90 and 0.20.

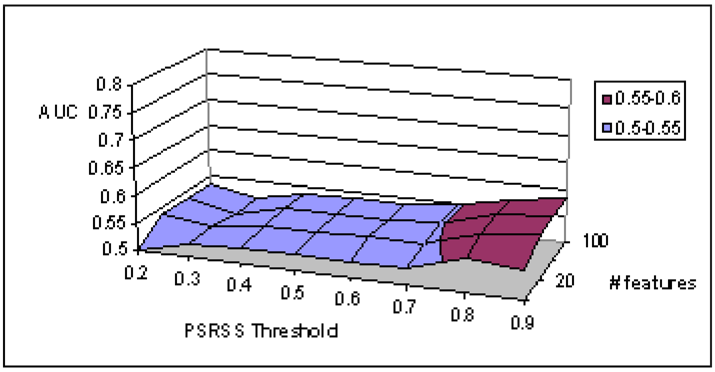


Figure 4.8 AUC values for the five fold cross validation of naïve bayes using the IFNγregions dataset and the frequency encoded feature set. AUC values were recorded for 100, 50, 20 and 10 features as chosen by the χ^2 feature selection ranking. PSRSS threshold value between 0.2 and 0.9 were used for each of the numbers of features. The *a priori* values were set to 0.5.

4.3.3. Graph analysis

The number of nodes and edges present in the graph, where an edge represents at least 1 TF in common, was very high. 94% of the genes from the IFNγ-regions dataset were found to have an edge connecting it to at least one other gene (Figure 4.9, resulting in 38985 edges (Figure 4.10. The inclusion of the PSRSS score ≤0.4 criteria, reduced the number of nodes to 86% of the total, and reduced the number of edges by a much greater amount to 6650.

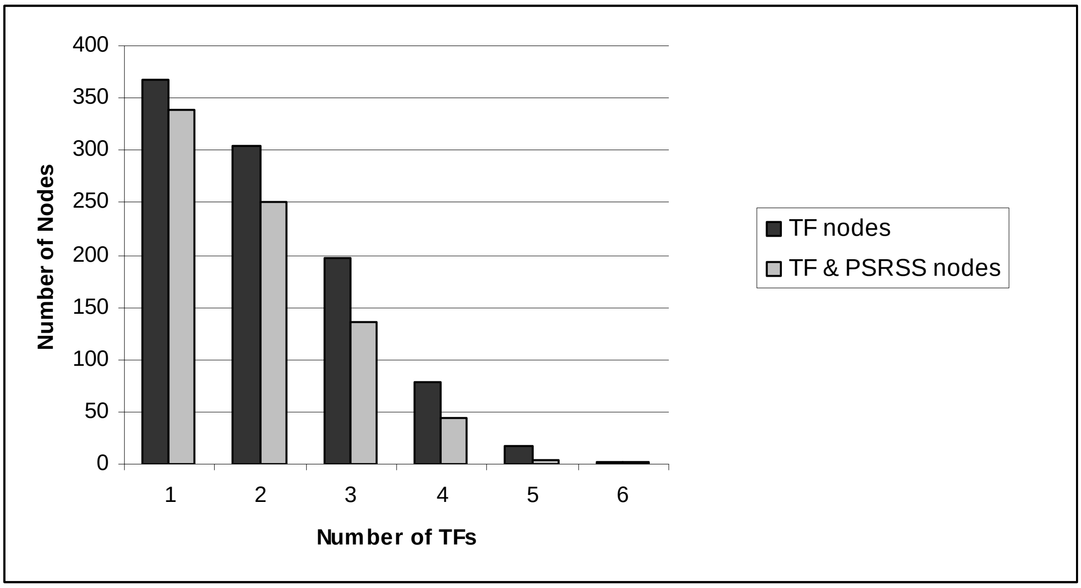


Figure 4.9 Histogram showing the number of nodes present on graphs created from the genes in the IFNγ-regions dataset. Black bars represent the number of nodes (genes) connected by at least one edge where edges represent a certain number of TFs in common between genes. Grey bars represent the number of nodes connected by at least one edge where edges represent a certain number of TFs in common and a PSRSS score ≤0.4 between the genes.

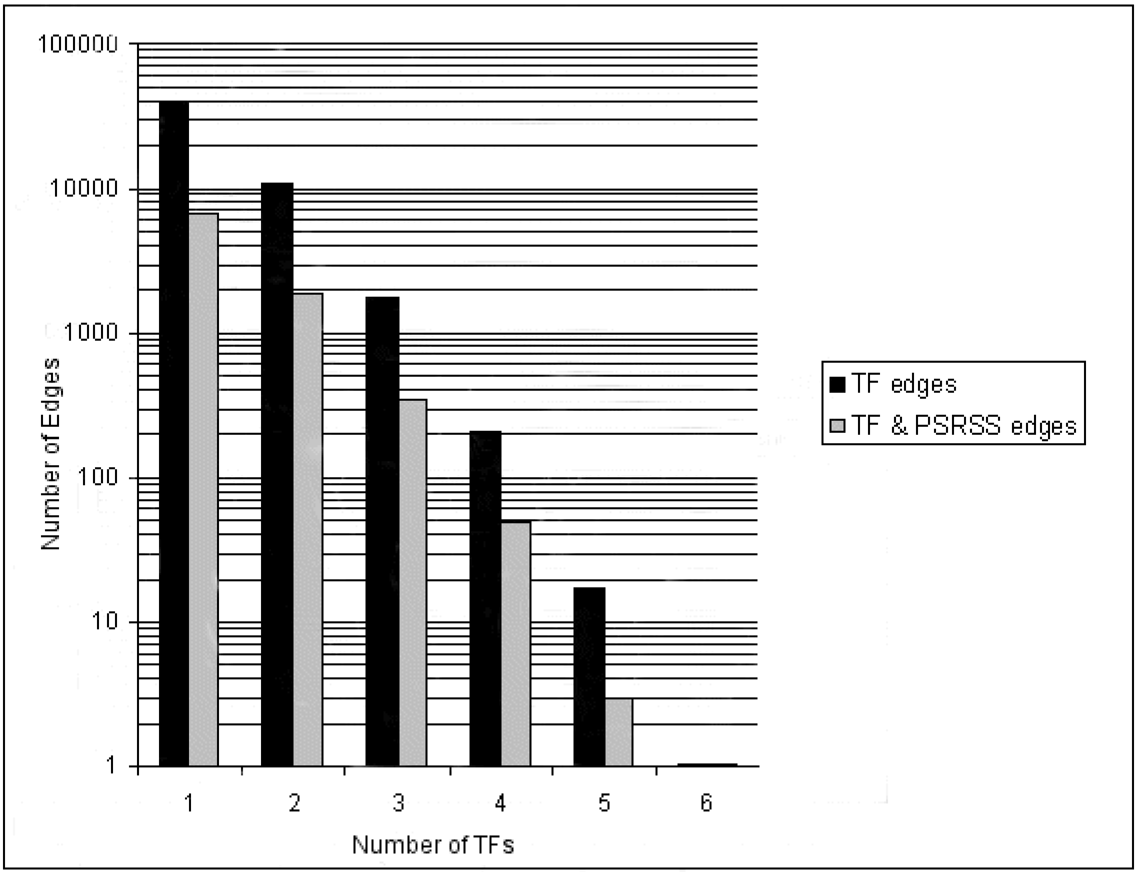


Figure 4.10 A semi-log histogram showing the number of edges present in the graphs when the criteria is either; a number of TFs in common (x axis) between a pair of genes (black), or both a number of TFs in common between a pair of genes and a PSRSS score ≤0.4 (grey).

The graph constructed using a criteria of 6 TFs in common, contained only 2 genes, *COPS5* (COP9 constitutive photomorphogenic homolog subunit 5) from chromosome 8 and glutamate receptor, *GRM5* (metabotropic 5) from chromosome 11. COPS5 is a subunit of the cop9 signalosome and is involved in the degredation of CDKN1B (cyclin-dependent kinase inhibitor 1B), a protein which controls the cell cycle progression at G1 (Tomoda *et al.*, 2005). The COP9 signalosome involved in the regulation of multiple signalling pathways (Huang *et al.*, 2007; Liu, X. *et al.*, 2008; Wang *et al.*, 2008) and has recently been implicated in T cell development (Panattoni *et* *al*., 2008). GRM5 is a glutamate receptor which has been shown to activate phospholipase C, a major component in calcium signalling pathways involved in diverse functions such as apoptosis (Giorgi *et al*., 2008; Pinton *et al*., 2008) and metabolism (Murgia *et al*., 2009). Calcium signalling has been shown to be crucial for the development and function of regulatory T-cells (Vig & Kinet, 2009), responsible for the secretion of IFN-γ. It is possible that *GRM5* is linked to IFN-γ through this pathway.

Using the the criteria for an edge to represent 5 TFs in common, a graph containing 18 nodes and 17 edges is produced (Figure 4.11). Only one complete 3 node subgraph, containing nodes representing genes from one of each of the IFN-γ linked regions, was found. This contained both genes found to contain 6 TFs in common, *GRM5* and *COPS5*, and also *HPSE2* (heparanese 2). Heparanase is secreted by T cells to promote T cell adhesion to the extracellular matrix (Sotnikov *et al*., 2004). Although little is known of the exact mechanisms of *HPSE2*, a known interaction with T cells increases the likelihood of a functional relationship between these three genes and the likelihood of the genes being linked to the IFN-γ response.

The TFs found in common between the genes were; AP1 (activator protein 1), CREB (cAMP response element binding protein), GATA1 (globin transcription factor 1), GATA2 (globin transcription factor 2) and USF (upstream transcription factor). NRF2 (Nuclear factor (erythroid-derived 2)-like 2) was shared by both *COPS5* and *GRM5* but not *HPSE2*.

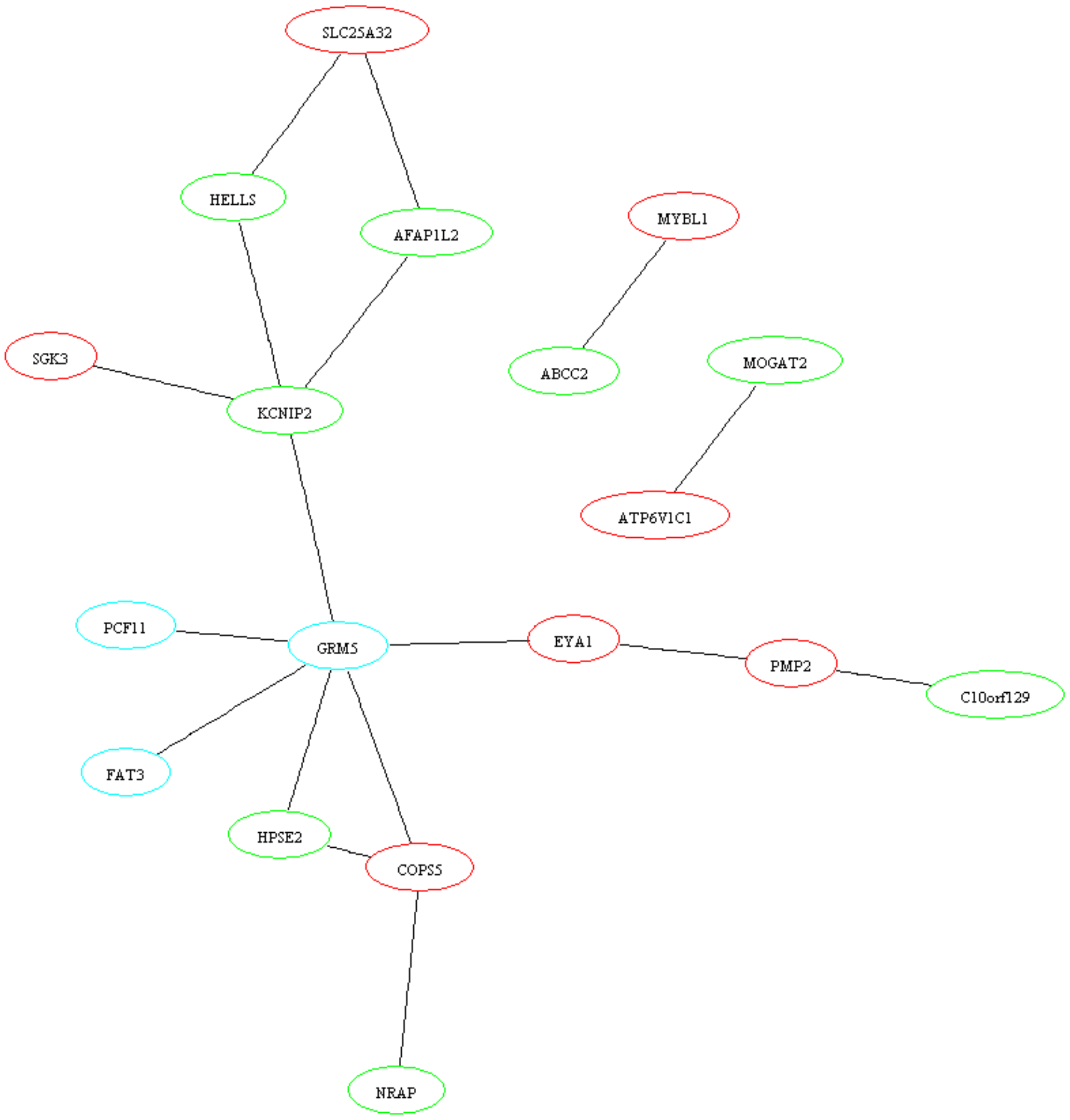


Figure 4.11 The graph produced from the genes in the IFN-γ linked regions when the criteria for an edge is defined by the genes (nodes) containing TFBSs for at least 5 TFs in common. Cyan nodes represent genes from chromosome 11, green nodes represent genes from chromosome 10 and red nodes represent genes from chromosome 8.

A possible cis-regulatory region has been discovered in each of the candidate genes from this analysis (Figure 4.12). The region is found between -300 and -250 nucleotides from the TSS. The region contains a GATA1 motif followed by an AP1 motif with a distance of between 34 and 36 base pairs between the TFBS. The small conserved distance between TFBS and the occurrence of the motif at a similar distance from the TSS in all genes are indicators that these genes could be co-regulated by this motif.

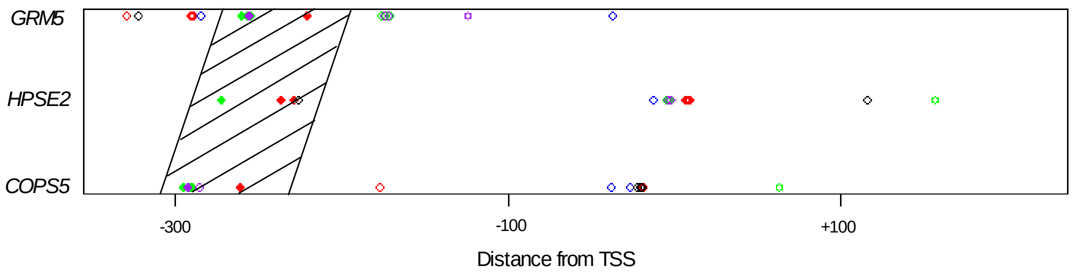


Figure 4.12 Diagram describing the positions, predicted by PMNB, of the binding sites for those TFs found in common between *COPS5*, *HPSE2* and *GRM5*. Red points = AP1 TFBS, blue points = CREB TFBS, green points = GATA1 TFBS, purple points=GATA2 TFBS and black points = USF TFBS. The shaded region indicates a possible Cis regulatory region containing a putative CRM consiting of a GATA1 and an AP1 motif separated by 34-36 nucleotides.

A search of all genes in the IFNγ-regions dataset found only three other genes which contained possible examples of this CRM. These genes were RUNX1T1 (runt-related transcription factor 1), MYBL1 (v-myb myeloblastosis viral oncogene homolog (avian)-like 1) from chromosome 8 and RPS28 (ribosomal protein S28) from chromosome 11.

4.4. Discussion

Neither the PCO nor naïve Bayes analyses were successful in predicting genes with functional similarities; this may be due to a number of factors: the simplicity of the methods, the predicted nature of the TFBS (which will include a large number of false positive results) and the definition of function for the purposes of the analysis. These factors are explored further in Chapter 6.

The candidate genes suggested by the graph analysis had not been detected in the Section 2.3.3 analysis. This was despite all three genes containing binding sites for two of the TFs analysed in Chapter 2; CREB and USF. It was expected, however, that the analysis in Section 2.3.3 would only identify a subset of candidate genes. The known functions of the three genes suggest an involvement in immune response processes involving T cells, two in development of T cells and one in the adhesion of the cells to the extracellular matrix. This lends weight to the hypothesis that these genes are linked to the INF-γ immune response. A further factor is the identification of a possible CRM containing AP1 and GATA1. AP1 has been shown to regulate gene expression in response to a number of different stimuli including bacterial and viral infections (Shen *et al.*, 2006; Tu *et al.*, 2009). The GATA family of transcription factors are known to regulate cells involved in hematopoiesis (Harigae, 2006; Shimizu & Yamamoto, 2005). GATA1 is usually involved in erythropoiesis, the formation of red blood cells, however GATA2 is expressed in a wider range of hamatopoietic progenitors (Lowry & Mackay, 2006; Ohneda & Yamamoto, 2002), and GATA3 is restricted almost exclusively to T cells (Ho *et al.*, 2009; Rothenburg & Scripture-Adams, 2008). The GATA family of transcription factors bind to similar motifs, therefore it is possible the GATA1 motif actually binds to GATA2 or GATA3.

5. Analysis of Epstein Barr Virus ZRE3 TFBSs in the human genome (Flower *et al.*, 2010)

5.1. Introduction

The Epstein Barr virus (EBV) is a member of the gamma-Herpes virus family, and infects 90% of humans by adulthood. In the majority, infection is sub-clinical or results in a self-limiting illness from which there is full recovery. However, the virus remains latent, i.e. it is never cleared by the host immune system; and EBV is associated with a number of malignant diseases including lymphoma and nasopharyngeal carcinoma (Young & Rickinson, 2004). EBV can affect the growth of infected cells, and is used in vitro to transform B lymphocytes such that they are capable of indefinite growth. It has been postulated that the probability of developing cancer is increased by viral replication and a number of EBV genes that may regulate these processes have been investigated (Thorley-Lawson & Allday, 2008). In particular, interest has focused on the transcription factor Zta, given its role in the switch from latent to lytic phases of the EBV cycle. The aim of this work was to apply TFBS prediction methods to identify Zta binding sites in the human genome, thereby identify genes that could be involved in activation of the EBV lytic cycle.

Zta, an Epstein Barr transcription factor, binds to Zta response elements (ZREs) in order to activate the EBV lytic gene promoters. This starts a cascade of over 50 genes involved in the lytic life cycle of the virus (Petosa *et al*., 2006). Although there are a number of ZREs which have been discovered (Lehman *et al*., 1998), we have investigated ZRE3, a binding site with a conserved 7 base nucleotide sequence; TCGCGAA. Zta binds to ZRE3 primarily in a methylated state (El Guindy *et al*., 2006).

Zta has been is known to bind to human promoter regions and interact with cellular factors. This can activate genes in a manner which is beyond the control of the human host. An example of this is the up-regulation of TRK-related tyrosine kinase in Zta transfected cells, but not in control cells (Lu *et al.*, 2000). The up-regulation of a kinase suggests that host signalling cascades could be initiated by Zta. Zta is also able to regulate AP1 protein expression and to compete with Fos-Jun heterodimers for AP1 sites. This suggests that Zta has the potential to interfere with AP1-mediated cell proliferation and differentiation (Speck *et al*., 1997). It has also been shown that Zta is able to induce cell cycle arrest in the G0/G1 phase of epithelial tumour cell lines through the activation of p53, p21 and p27 (Cayrol & Flemington, 1996). A further effect of the Zta protein is an inhibition of the IFN-γ signalling pathway, altering host immune responses and suggesting a mechanism through which EBV may avoid host responses during initial infection (Morrison *et al*., 2001).

The identification of the ZRE3 motifs in the human genome would be a useful step in the elucidation of the Zta mode of action. By determining which genes contain a ZRE3 motif we will identify which human genes Zta may bind and activate; thus yielding information about the functional mechanisms of Zta.

5.1.1. ZRE3 Motif Prediction Methods

Two methods of TFBS prediction were used to predict the binding of ZRE3 motifs in this analysis. Firstly a simple pattern matching technique was used, implemented using perl scripts, searching for the exact 7 base ZRE3 sequence. Exact pattern matching was used because the starting point for the analysis was a single ZRE3 motif. As additional ZRE3 motifs were identified, PWMs were created and used to search the promoter sequences. A number of methods evaluated and developed in Chapters 2 and 3 could not be used for this analysis due to one of two limitations: (a) they do not allow for user defined PWMs to be used and (b) they assume TFBS are conserved. The ZRE3 motif does not appear to be well conserved evolutionarily; preliminary research showed that the PPMNB method did not predict any binding sites for the ZRE3 motif. Hence the TFBS perl modules were used in the current analysis, as this method allows for user defined PWMs, and was found to be the most successful TFBS prediction algorithm in the analysis carried out in Chapter 2.

5.1.2. Methods for the analysis of ZRE3 binding site function

Identification of a ZRE3 motif using computational methods gives no indication of whether the site is a functional one. Here, we have used the location of CpG islands to infer a functional site. Sixty percent of all promoters are found to co-localise with CpG islands (Antequera, 2003), therefore the occurrence of a ZRE3 motif in a CpG island lends weight to the hypothesis that the ZRE3 motif is occurring in a functional promoter. Many of the CpG dependent promoters regulate constitutively expressed housekeeping genes. These genes are expressed throughout all tissues and are involved in core processes, such as the cell cycle (Rozenberg, 2008; Zhu *et al.*, 2008). The occurrence of ZRE3 motifs in these housekeeping genes is of particular interest, signalling that Zta may be able to disrupt the regulation of these core processes. Finally, the presence of ZRE3 motifs in CpG islands is particularly interesting because of the preferential binding of Zta to methylated ZRE3 motifs. CpG islands in the human genome are normally un-methylated; however it is known that in cancer cells these regions can become hypermethylated, altering normal gene expression (Li, M. *et* *al*, 2008). It is possible that the tumourigenic properties of EBV are linked to this process.

In addition to the identification of CpG islands, we have used GO term annotations to analyse the functions of genes containing ZRE3 motifs in their promoters. Comparisons of the GO term annotations associated with a group of genes containing ZRE3 in their promoters, against the annotations of all known human genes, will lead to the identification of functions enriched in the group. The hypothesis for this work assumes that the binding of ZRE3 by Zta is specific to certain functional processes and the identification of over represented GO terms may identify these processes. Another method of identifying genes which are involved in similar processes is by comparing functional annotations using a statistical measure. The measure we have used is PSRSS, a statistic introduced and discussed in Chapters 2 and 4. The comparison of sets of GO term annotations using the PSRSS statistic allows for the identification of small groups of genes in related processes, which may not be over-represented in relation to the annotations of all genes, but which may still be interesting.

5.1.3. Methylation dependent Zta binding

Methylation of ZRE3 influences the ability of Zta to bind. Two amino acid residues (Cys-189 and Ser-186) in Zta are found to be crucial for methylated binding to ZRE3, although not ZRE1 or ZRE2 (Karlsson *et al.*, 2008a). These residues interact with methyl cytosines in the ZRE3 motif. A cysteine to serine substitution (C189S) within the Zta protein motif abolishes ZRE3 binding (Karlsson *et al.*, 2008a). This preferential binding to methylated motifs is an integral part of the involvement of Zta in switching EBV from a latent to a lytic phase (Karlsson *et al.*, 2008b). In the latent phase the EBV genome is silenced by host driven methylation of CpG motifs. Zta preferentially binds to methylated sites and is therefore able to start the cascade leading to the lytic phase.

In this chapter, we have analysed the methylation-dependent binding of Zta by analysing the occurrence of ZRE3 motifs in CpG islands, and by determining, through Electrophoretic mobility shift assay (EMSA) analysis, whether motifs in selected candidate genes are preferentially bound when methylated. The EMSA analysis was carried out by Kirsty Flower, in the laboratory of Dr Alison Sinclair (unpublished data).

5.2. Methods

5.2.1. Exact Pattern Match and First Round electrophoretic mobility shift assays (EMSAs)

An exact pattern search was carried out for the 7 nucleotide sequence of ZRE3; TCGCGAA. The search was conducted in sequences consisting of 500bp upstream of the transcription start site in all 22,740 protein coding genes retrieved from Ensemble 49 (Hubbard *et al*., 2007). Five of the genes predicted to contain a ZRE3 motif were chosen to be tested for binding ability using electrophoretic mobility shift assays (EMSAs). These genes were: max-binding protein (*MNT*), histone deacetylase 2 (*HDAC2)*, Zinc finger CCCH domain-containing protein 8 (*ZC3H8*), cyclin L2 *(CCNL2)* and xeroderma pigmentosum, complementation group C *(XPC)*. These genes were chosen because they were likely to be involved in transcriptional processes, and therefore were the most interesting of the genes found. Both methylated and unmethylated motifs were tested for binding ability.

5.2.2. Creation of PWMs and ZRE3 Motif Prediction

The pattern match search described in section 5.2.1 was repeated using all human protein coding genes from Ensembl version 50 (Flicek *et al*., 2008). Ensembl 50 was used rather than Ensembl 49 as used in section 5.2.1 to ensure that the promoter information represented the most recent knowledge. This analysis gave the total possible number of ZRE3s, given the assumption that the motif is exactly conserved in all ZRE3s. This set of ZRE3 results is designated total_ZRE3.

Two position weight matrices (PWM) were created from genes shown to bind ZRE3 in the preliminary EMSA experiments (Figure 5.1). Both the PWMs were 27 nucleotides in length, the same length as the DNA sequences which had been tested by EMSA (section 5.2.1). The sequence consisted of the seven base conserved motif with the 10 residues found either side in each of the ZRE3 containing genes. The first PWM was created using sequences from all the genes found to bind Zta (ZRE3_A), five from human genes and one from EBV (RpZRE3). In the second PWM (ZRE_B), *XPC* was removed due to aberrant behaviour, weak binding of Zta when unmethylated as well as when methylated. ZRE3_B, therefore, consisted of motifs from four human genes plus RpZRE3.

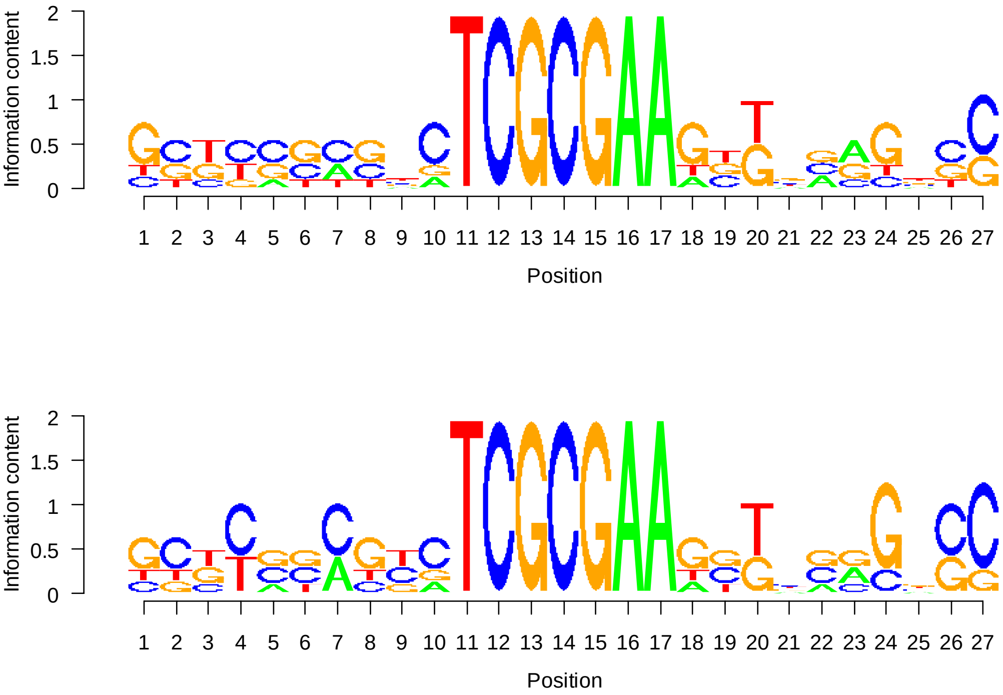


Figure 5.1. Position weight matrices ZRE3_A (top) and ZRE3_B (below). ZRE3_A consists of 5 human sequences plus RpZRE3 from EBV, ZRE3_B consists of 4 human sequences, where one binding sequence, not conforming to the behaviour of the other binding sequences, has been removed, plus RpZRE3 from EBV. PWMs were created using the TFBS perl modules (Lenhard & Wasserman, 2002).

The 500bp upstream region of all human protein coding genes was searched with both PWMs using the TFBS perl modules (Lenhard & Wasserman, 2002), with a threshold of 80% conservation. The use of PWMs alone would allow for the central motif to contain mismatches, even though the positions in the PWM are completely conserved, due to the statistical nature of the search method. Therefore, to ensure the integrity of the ZRE3 conserved motif, the results were further parsed to allow only predictions containing the exact core motif into the final analysis (Figure 5.2). Using this procedure the PWM search was only able to predict a subset of the total_ZRE3 dataset. These genes were assumed to be more likely to be TP matches because the flanking sequences were similar to those in the ZRE3 motifs known to bind to Zta.

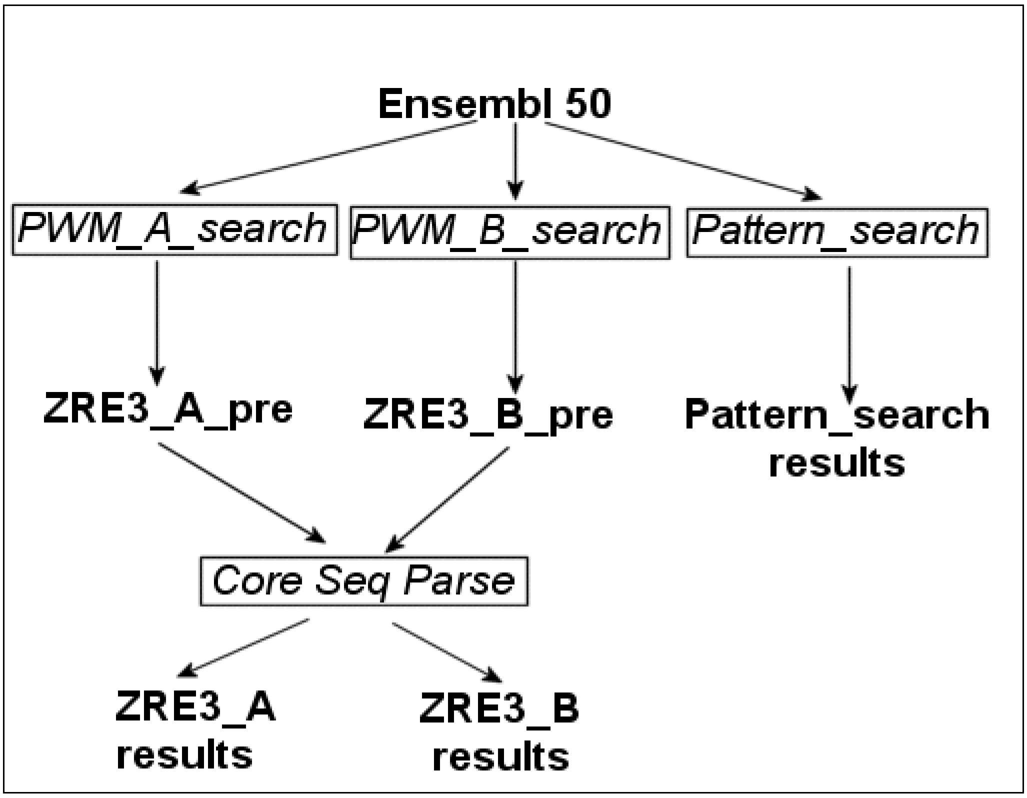


Figure 5.2. The procedure used to search for ZRE3 motifs. All protein coding genes in Ensembl 50 were searched using three methods: PWM search using ZRE3_A (PWM_A_search), PWM search using ZRE3_B (PWM_B_search) and the pattern search for TCGCGAA (Pattern_search). The result of these searches was three sets of ZRE3 predictions; ZRE3_A_pre, ZRE3_B_pre and Pattern_search results. The PWM based sets (ZRE3_A_pre and ZRE3_B_pre) were further parsed to retain only those motifs which contained the core 7 base motif: TCGCGAA. This resulted in a further two sets of ZRE3 predictions ZRE3_A results and ZRE_B results.

5.2.3. Analysis of ZRE3 binding sites

The first method used to predict whether ZRE3 sites were functional, was the prediction of CpG islands. CpG islands were predicted in each promoter sequence using the EMBOSS program CpGPlot (Rice *et al*., 2000). CpGPlot was used with the default options: a window size of 100; a step of 1; a minimum average observed to expected ration of C plus G to CpG in a set of 10 windows of 0.6; a minimum average percentage of G plus C in a set of 10 windows of 50% and a minimum length of CpG island of 200 nucleotides. The positions of the putative CpG islands were compared to the positions of the ZRE3 motifs predicted, and the number of motifs found inside predicted CpG islands was compared for each of the search methods.

The Biological Process and Molecular Function Gene Ontology (GO) term annotations, from Ensembl 50, were retrieved for each of the genes predicted to contain a ZRE3 motif using BioMart (Durinck *et al*., 2005; Smedley *et al*., 2009). The GO annotations were filtered to contain only those annotations which were manually curated or experimentally verified. The number of genes retrieved for each GO term was compared against the total number of human genes in Ensembl with the same annotation. Significance was measure using Chi-squared with Yates correction (Yates, 1934), reducing overestimation of statistical significance when using small amounts of data. Terms which occurred with a significantly higher frequency ($p<0.05$) in the ZRE3 containing datasets than in Ensembl, were retrieved and analysed. Genes that were not annotated with the over-represented terms were removed from the datasets to reduce the number of candidate genes for EMSA tetsing. However, it was noted that genes with no, or sparse, GO annotations may be functional Zta binding genes, but would be excluded from the analysis.

An alternative method of analysing the GO annotations associated with the genes predicted to contain ZRE3 motifs was to create graphs showing genes connected by similar annotations (See section 2.2.4. and 4.3.3). Graphs were created using graphviz v2.20 (www.graphviz.org) and coded in the dot language. Nodes represented genes containing ZRE3 motifs and edges represented a PSRSS score of ≤0.4 between the Biological Process annotations of two genes. A PSRSS score of 0 denotes two genes which are exactly the same, a score of 1 denotes two genes which are completely different. A score of ≤0.4 ensured that only connections between genes with similar functions were retained. Of the three types of GO annotation, only Biological Process annotations were used as these contained the most relevant information about processes in which Zta may be involved. The resulting graphs were analysed to identify for their largest complete subgraphs. For each node, all the nodes to which it had edges were determined. Nodes which did not have edges to all other genes in this set were removed, starting with the node with the least relevant edges. This process was continued until the set of nodes had edges to all other nodes in the set, i.e. it was a complete graph. The subgraphs were annotated with the most specific GO term which was relevant to all nodes in the subgraph.

Text mining was also used as an alternative to GO annotations, as a means of searching for known information about the functions and interactions between genes predicted to contain ZRE3 motifs. PubGene was used to mine for associations in the literature between the ZRE3 containing genes and EBV. The organisms option in PubGene was set to *Homo sapiens*, each gene in turn was used as the search gene and EBV was used as the biological search term. ZRE3 and Zta were not used as biological search terms because they could not be found in the PubGene structured vocabulary. Genes found to have co-occurrences both with EBV and at least one of the genes shown by EMSA to bind Zta (Section 5.2.1) were analysed to look for possible pathways.

5.2.4. Analysis of un-methylated binding

In all previously cases where the Zta transcription factor has been experimentally shown to bind to the ZRE3 binding site the binding has occurred when the site was methylated. However, in a preliminary EMSA experiment (see section 5.2.1) one gene, *XPC*, was also shown to bind to an unmethylated site. This was unexpected and does not occur in the other ZRE3 binding sites. The phenomenon of Zta binding to certain ZRE3 motifs when unmethylated has been further analysed by the Sinclair group (unpublished data). Further EMSA analysis has indicated three nucleotide positions (Left4G, Left7T or Left10A) to be important for the unmethylated binding of Zta by ZRE3. By determining which genes contain these important nucleotides in the flanking region of the ZRE3 motifs, information about the function of the unmethylated binding of ZRE3 motifs by Zta may be elucidated.

XXX**G**XX**T**XX**ATCGCGAA**XXXXXXXXXX

Figure 5.3. The XPC ZRE3 motif, highlighting the nucleotides (black) hypothesised to be important for non-methylated binding to the ZRE3 motif. The ZRE3 motif is included in grey, other positions have been represented with X.

All of the ZRE3 motifs predicted by the PWMs were collated and analysed to find motifs which contained at least one of the three nucleotides (Left4G, Left7T or Left10A) predicted to be important for unmethylated binding. The set of gene containing at least one of these nucleotides was designated as the putative_non-methylated dataset. The location of CpG islands in the the genes found to contain at least one of the nucleotides were identified, and the occurrence of ‘unmethylated binding nucleotide’ containing motifs found in CpG islands was compared to that of the total set of ZRE3 motifs.

The genes in the putative_non-methylated dataset were analysed for functional similarities using Biological Process GO terms. Graphs were created with genes containing ZRE3 motifs as nodes, and edges representing a PSRSS score ≤0.4 between the two genes as described in section 5.2.3.

5.2.5. Second Round EMSA

12 genes were chosen to be tested for Zta binding using EMSAs, analysing both methylated and unmethylated binding. The criteria for choosing the genes were that they were found using both ZRE3_A and ZRE3_B PWMs, were annotated with an over-represented GO term and were positioned within a CpG island. The EMSA experiments were carried out by Kirsty Flower in the laboratory of Dr Alison Sinclair, University of Sussex.

5.2.6. Analysis of results from the second round of EMSAs

New PWMs were created by adding the ZRE3 sequences shown to bind Zta in the second round of EMSA experiments (section 5.2.7) to the existing PWMs; ZRE3_A and ZRE3_B. The first new PWM created, ZRE3_A2, contained all motifs shown to bind to Zta, including the EBV promoter ZRE3, RpZRE3. The second new PWM contained all motifs found to bind Zta when methylated but not when unmethylated i.e. *XPC* was excluded from the PWM.

The genes which were found to contain motifs, when the searched with the ZRE3_A2 and ZRE3_B2 PWMs, were analysed using the CpG analysis, GO term analysis (Section 5.2.3).

5.3. Results

5.3.1. Exact Pattern Match and First Round EMSA

A total of 114 genes were found to have ZRE3 motifs in the 500bp upstream promoter regions found in the Ensembl 49 version of the human genome. This was out of a total of 21,541 protein coding genes in Ensembl 49. Of these genes 5 were chosen for binding analysis carried out by EMSA; *MNT, HDAC2*, *ZC3H8*, *CCNL2*, *XPC.* All five genes chosen were found to bind to Zta at the ZRE3 motif. Four of the five genes only bound when the ZRE3 motif was methylated, showing no unmethylated binding behaviour. One gene, *XPC*, showed both methylated and un-methylated binding behaviour. This was not expected and as such the *XPC* motif will be treated as a possible, but not a definite ZRE3 motif.

5.3.2. ZRE3 motif occurrence

273 genes were found to have ZRE3 motifs in the 500bp promoter regions retrieved from Ensembl v.50 (Figure 5.4), these numbers were reduced even further when using the PWM_A to search with and further still with PWM_B.

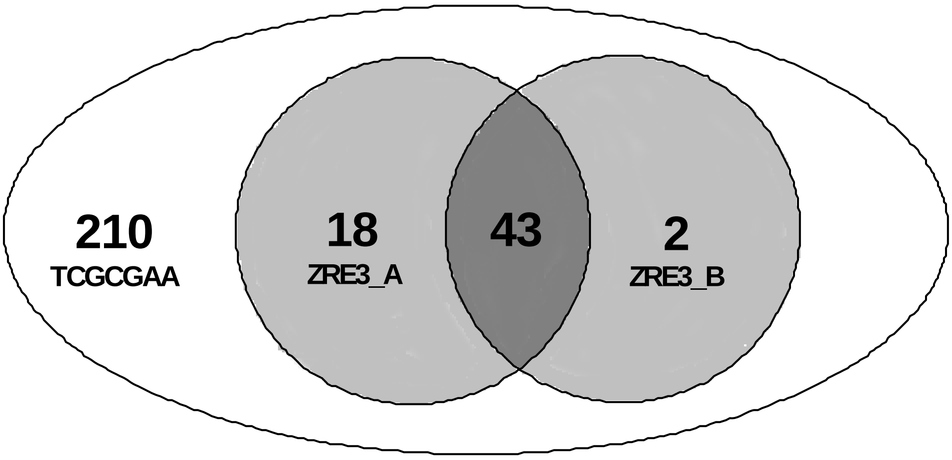


Figure 5.4 Venn diagram showing the numbers of genes found using the TCGCGAA pattern, ZRE3_A PWM and ZRE3_B PWM.

A small number of genes were found to have more than one ZRE3 motif present in the 500bp promoter region (Figure 5.5). Two genes were predicted to have two ZRE3 motifs using both ZRE3_A and ZRE3_B to search; the first, *DNAJB6*, coded for a heat shock protein and was found on chromosome 7, the second was an uncharacterised gene, AC007731.16, located on chromosome 22. The pattern match search for the conserved motif retrieved 48 genes with multiple motifs.

There was a large number of matches where the 7 base motif was found through the pattern matching search, but the flanking sequences did not allow for a prediction to occur using either PWM. ZRE3_A found motifs in 18 genes which were not found using ZRE3_B, these motifs may be more likely to bind Zta when unmethylated, as well as methylated; the only difference between the two PWMs being the inclusion of the *XPC* motif in ZRE3_A which binds when both methylated and un-methylated.

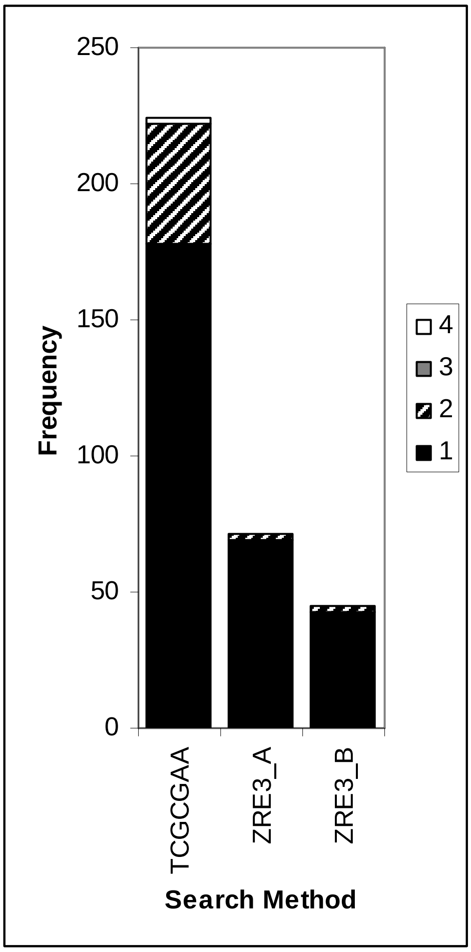


Figure 5.5. The frequency of genes and motifs per gene found by searching using the pattern search, ZRE3_A PWM or ZRE3_B PWM and filtering for only those results containing the exact conserved motif. The proportion of genes containing one, two, three or four motifs in the 500bp promoter region is represented by the shading on the bars.

Of the five known human ZRE3 motifs originally used to construct the PWMs, only two, *MNT* and *HDAC2*, were found in the 500 nucleotide promoter region according to Ensembl 50. Another two of the ZRE3 motifs, ZC3H*8* and *XPC*, were now found to be located inside the coding region of the gene due to changes between Ensembl versions 49 and 50. A third gene, *CCNL2*, had been retired from Ensembl and did not appear to have a new stable mapping in Ensembl 50. *CCNL2* was reintroduced to Ensembl 51.

5.3.3. CpG Islands

142 of the 273 genes were found by searching for the 7 base conserved motif were found in a putative CpG island. A significantly higher percentage of the motifs found using ZRE3_A or ZRE3_B were located inside putative CpG islands than those found using by only the pattern matching of the conserved motif ($p<0.05$; χ^2 with Yates correction). The difference between the number of motifs found in CpG islands by ZRE3_A and ZRE3_B was not shown to be significant.

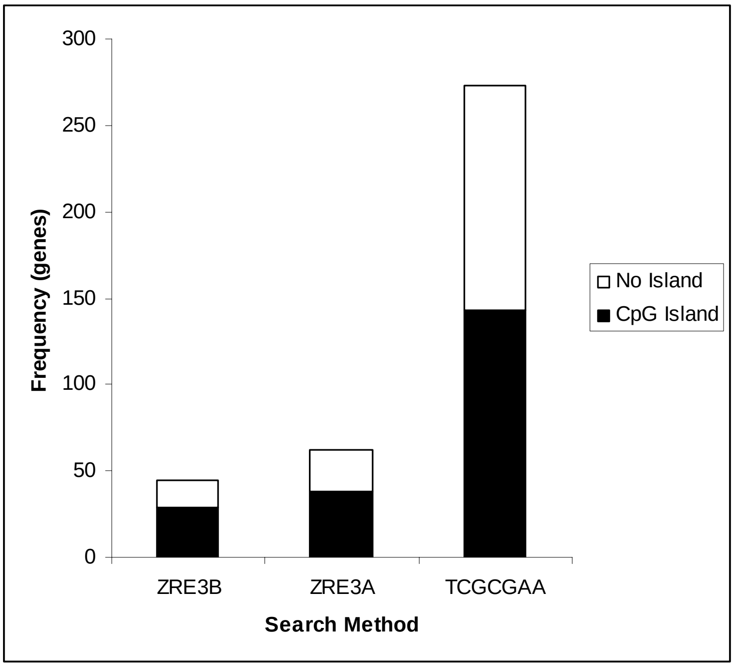


Figure 5.6. Chart showing the number of genes with a ZRE3 motif as predicted by the three search methods that all contained a putative CpG island in the 500bp promoter region.

5.3.4. Statistical over representation of GO terms

The GO terms found to be over-represented in genes with motifs found using both ZRE3_A and ZRE3_B were mainly associated with either gene regulation, e.g. transcription or chromatin processes, or with the cell cycle. This is particularly interesting as Zta involvement in either of these processes could certainly contribute to the tumourogenic properties of EBV.

Over represented GO term	**GO term ID**	**Search method**
Chromatin remodelling Mitosis Mitotic cell cycle checkpoint Oxidation reduction Transcription regulator activity Transcription coactivator activity Transcription corepressor activity Chromatin binding Satellite DNA binding	GO:0006338 GO:0007067 GO:0007093 GO:0055114 GO:0003700 GO:0003713 GO:0003714 GO:0003682 GO:0003696	ZRE3_A ZRE3_B
Cell motion	GO:0006928	ZRE3_A
Chromatin modification Protein amino acid phosphorylation Taurine biosynthetic process Protein kinase C binding Histone deacetylase binding	GO:0016568 GO:0006468 GO:0042412 GO:0005080 GO:0042826	ZRE3_B

Table 5.1. Biological Process and Molecular Function GO terms found to be over represented in the gene sets retrieved using ZRE3_A, ZRE3_B.

5.3.5. GO term graphs

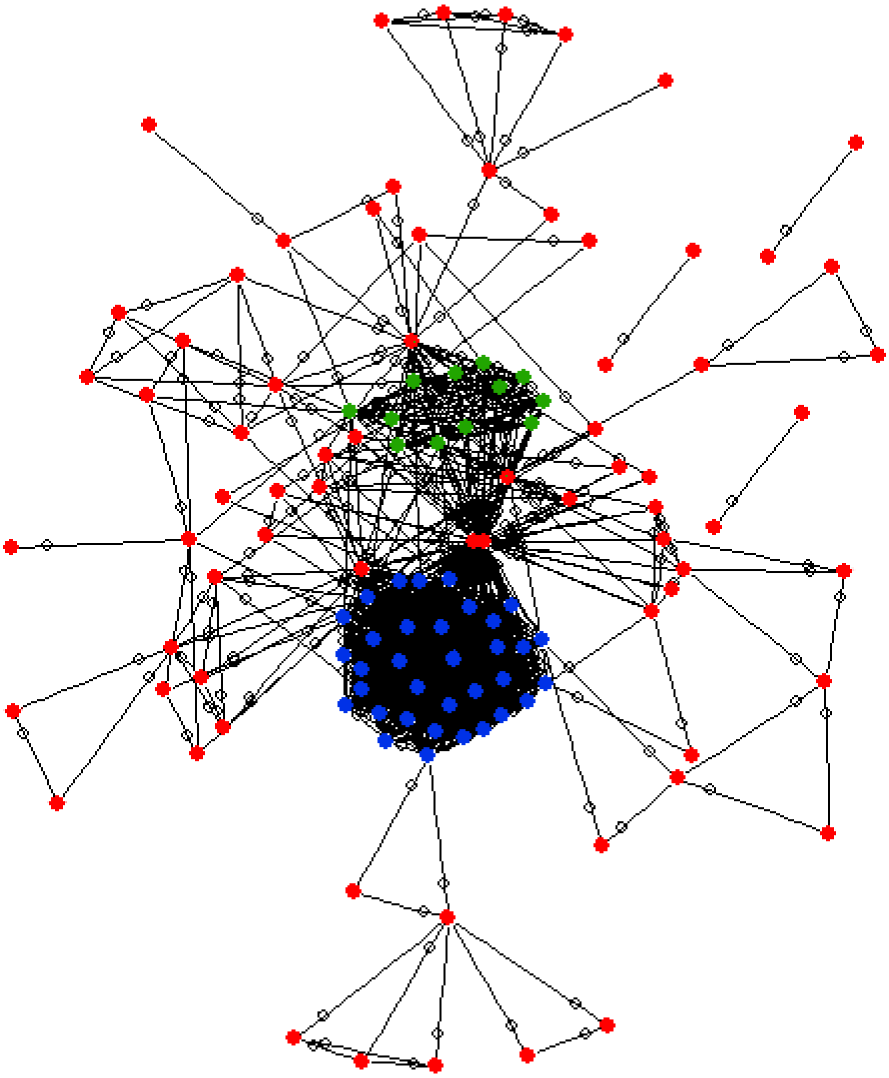


Figure 5.7. Graph created with genes as nodes and edges representing a PSRSS score ≤ 0.4. All genes represented by nodes contain a ZRE3 motif, however of the genes with ZRE3 motifs, only those with a PSRSS ≤0.4 to another ZRE3 containing gene were included in the graph. Blue nodes represent genes with the GO term annotation ‘transcription’, green nodes represent genes with the GO term annotation ‘protein amino acid phosphorylation’, red nodes represent all other genes without either of these two annotations.

The graph created using genes with ZRE3 motifs as nodes and a PSRSS score between the genes of ≤0.40 contained two separate regions of high interconnectivity. The largest of these regions represented genes which are involved in transcription, mainly as transcription factors. This is further evidence of the possible involvement of Zta in regulating or interfering with human transcription. The second group of genes contained protein amino acid phosphorylation genes: these genes are particularly interesting as they are likely to be involved in signalling cascades. Examples are the Serine/threonine-protein kinases 10 and 25 (STK10, STK25); serine/threonine kinases are integral parts of many signalling cascades (Choi, H.S. *et al*., 2008; Craig *et al*., 2008; Henmi *et al*., 2009) and have been recently introduced as targets for cancer therapies (Montagut & Settleman, 2009).

5.3.6. Second Round EMSAs

13 genes with ZRE3 motifs found by both ZRE3_A and ZRE3_B motifs, were found in CpG islands, and were annotated with overrepresented GO terms (Figure 5.8). One of these genes, *MNT*, had already been tested in the prelimary EMSAs (section 5.2.1). The 12 remaining genes are: calpain 2, (m/II) large subunit (*CAPN2),* cysteine dioxygenase, type I (*CDO1)*, bromodomain PHD finger transcription factor *(FALZ)*, kinesin family member 1B (*KIF1B),* lethal giant larvae homolog 1 (*LLGL1),* LIM domain only 4 (*LMO4),* methyl-CpG binding domain protein 4 (*MBD4),* pleckstrin homology domain containing, family J member 1 (*PLEKHJ1),* protein kinase D1 (*PRKD1)*, SEC14-like 1 (*SEC14L1),* transcriptional adaptor 3 (NGG1 homolog, yeast)-like *(TADA3L),* topoisomerase (DNA) II beta 180kDa (*TOP2B).*

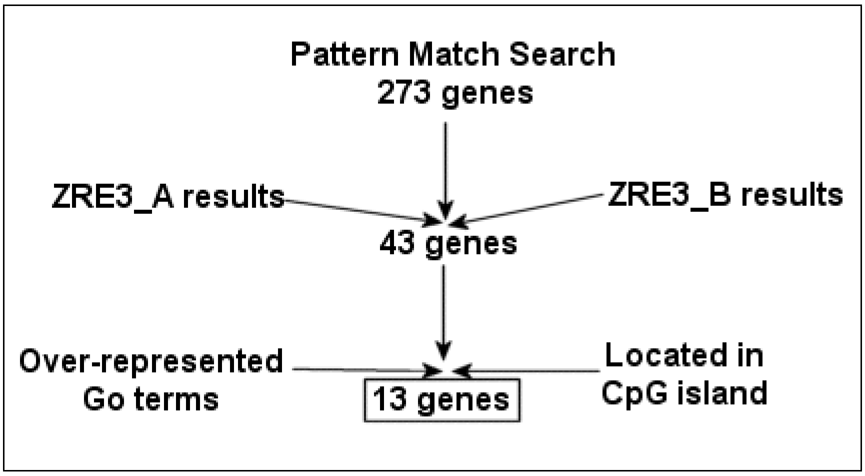


Figure 5.8. Diagram describing the selction process for ZRE3 candidate genes. Genes retrieved from the pattern matching search must be found by both ZRE3_A and ZRE3_B PWMs. Genes must also occur in a CpG island and be annotated with an over represented GO term.

In the EMSAs performed by the Sinclair laboratory, the ZRE3 motifs from all 12 genes were shown to bind preferentially to the methylated versions of the motif (Figure 5.9). Two of the motifs, from the *CDO1* and *LMO4* genes, appear to also bind to the unmethylated form of the ZRE3; however this amount of binding is so reduced from the methylated binding that this may be noise caused by unspecific binding of the motif.

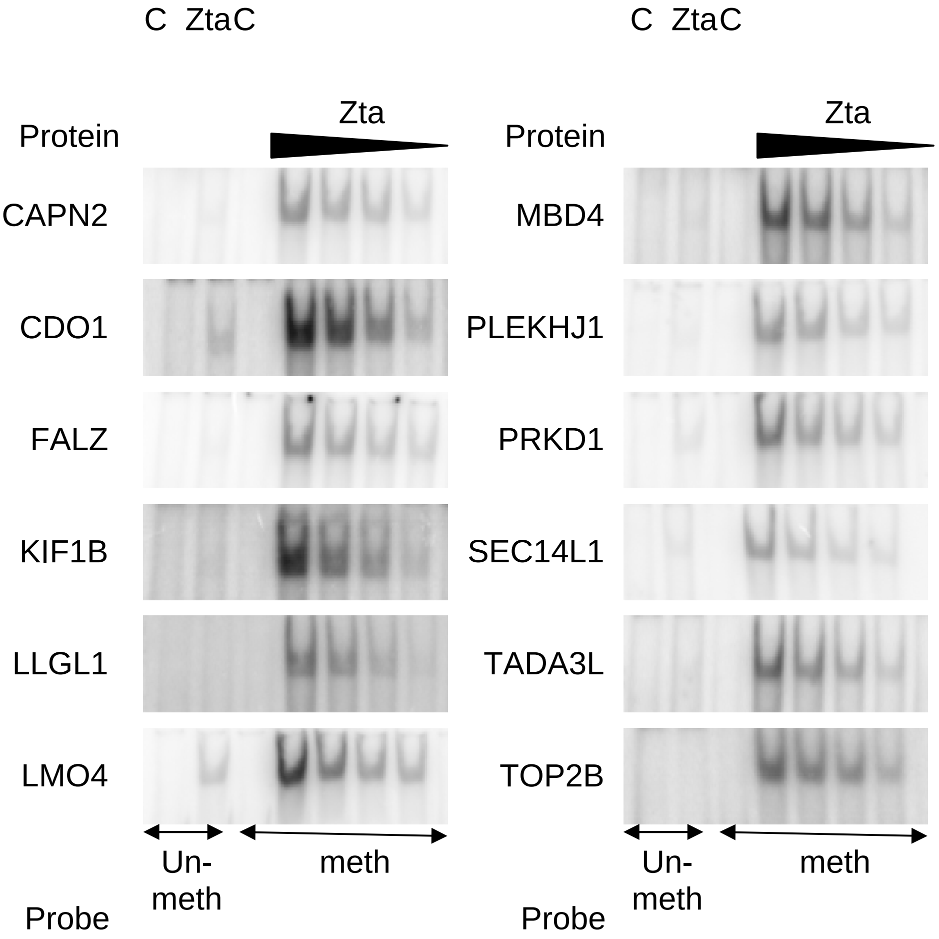


Figure 5.9. EMSA binding assay results for the 12 candidate genes predicted by ZRE_A and ZRE3_B PWMs. In all genes Zta is shown to bind preferentially to the methylated form of ZRE3. Two controls are shown (C), the 1st lane on the gel is the unmethylated control and the 3rd lane on the gel is the methylated control. The 2nd lane shows the binding of Zta to the unmethylated ZRE3 motif. Lanes 4 – 7 show the binding of Zta to the methylated ZRE3 motif, the amount of binding should reduce in proportion to the reduction in the amount of protein shown by the gradient at the top of the gel.

5.3.7. Text Mining

Candidate Gene	*PWM3_A*	*PWM3_B*	*CpG island Co-occurrence*	*Over-represented GO term*
BCL2L11		1	1	transcription coactivator activity transcription corepressor activity
CAPN2	1	1	1	transcription coactivator activity transcription corepressor activity
FRAP1			1	
GTF2A2				
HDAC2	1	1		Chromatin remodelling transcription regulator activity transcription coactivator activity transcription corepressor activity
MAP3K7				transcription coactivator activity transcription corepressor activity
PRKD1	1	1	1	transcription coactivator activity transcription corepressor activity

Table 5.2. The candidate genes identified by the PubGene text mining search and the features from the bioinformatics analysis associated with each gene. False positive results have been removed.

Of the 273 genes with the exact 7 base ZRE3 sequence, 11 genes were found by PubGene to have co-occurrences in the literature with the term ‘EBV’ (Table 5.2). Of these, 4 genes were found due to false positive co-occurrences: *CYCL1* (Countryman, 1994), *F2R* (Suzuki *et al*., 2004), *KCNA4* (Du *et al*., 2007), *NAPB* (Chung *et al*., 2000). False positive results arose due to gene names also representing other biological or scientific entities. For example, the *NAPB*-EBV co-occurrence is derived from an experimental paper which uses Sodium phenylbutyrate (NAPB). Two of the genes found by PubGene, *CAPN2* and *PRKD1*, are in the group of candidate genes that were tested by EMSA. PubGene was not able to use either Zta or ZRE3 as a co-occurrence search term so more specific interactions could not be mined for. The candidate genes found through the text mining search were not tested for Zta binding ability using EMSAs, unless they conformed to the candidate gene criteria: motif predicted by ZRE3_A and ZRE3_B, motif found in a CpG island and gene annotated with a GO term found to be over-represented in the ZRE3 containing gene set.

5.3.8. Analysis of un-methylated binding

EMSA results, carried out by Kirsty Flower (University of Sussex), showed the flanking sequences of the conserved motif to have an important role in the unmethylated binding of Zta in the ZRE3 motif found in XPC. When the ten nucleotides on the left flank of the motif were removed, all binding ability was lost. Removal of the right flank only reduced the binding strength implying that while both flanks have some importance in the unmethylated binding of Zta, the positions on the left flank appear to be crucial. The important residues in the left flank have been ascertained by mutating each nucleotide in turn and performing EMSA analyses on the resulting sequence. The residues on the left flank which have been found to lead to unmethylated binding in the

XPC ZRE3 motif are the G at position 4 (Left4G), the T at position 7 (Left7T) and the A at position 9 (Left9A).

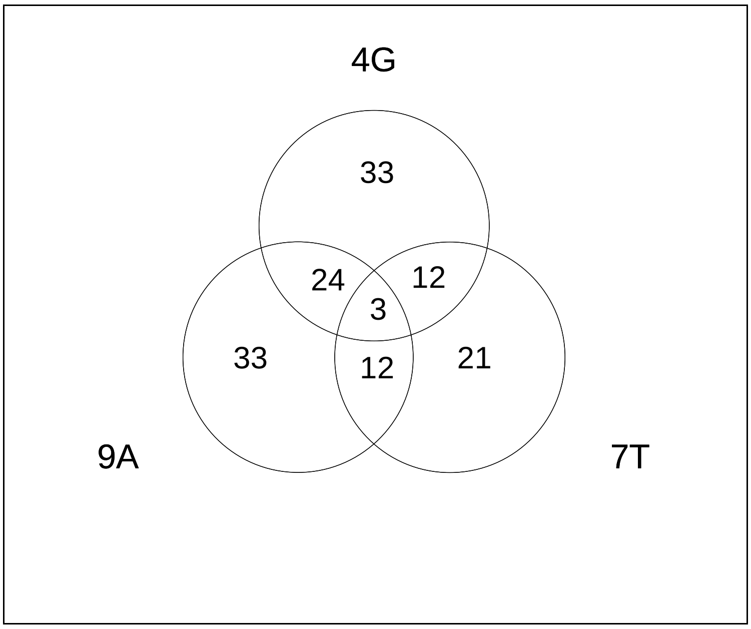


Figure 5.10. Venn diagram showing the number of genes found which contain a ZRE3 motif, in the 500 nucleotide promoter region, with the nucleotides found to be important in non-methylated binding of Zta by ZRE3. The 4G set includes all genes found to have a G at the 4th position on the left flank of the ZRE3 motif. The 7T set includes all genes found to have a T at the 7th position on the left flank of the ZRE3 motif. The 9A set includes all genes found to have an A at the 9th position on the left flank of the ZRE3 motif.

Only three genes, a gene coding for an uncharacterised protein (*C1orf109*), T-complex protein 1 subunit zeta-2 (*CCT6B*) and proteinase-activated receptor 1 Precursor (*F2R*), were found to contain all three of the nucleotides involved in unmethylated binding of the ZRE3 motif (Figure 5.10). However, a larger number of genes were found with some combination of the nucleotides as can be seen from Figure 5.8. There were a larger number of genes found with the LeftG4 and the Left9A than there were with the LeftT7.

No significant difference was found in the positioning of ZRE3 motifs in CpG islands when the non-methylated binding nucleotides were taken into account (Figure 5.11). This was the case when analysing the genes with nucleotides at each position singly and when analysing the entire group.

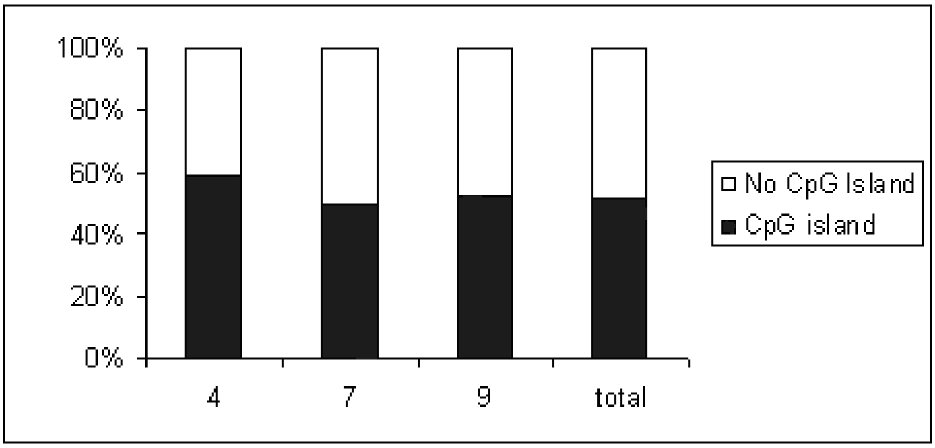


Figure 5.11. Graph showing the percentage of motifs containing each important nucleotide occurring in a CpG island compared to the total percentages of ZRE3 containing genes within or without a CpG island. No significant difference is observed (Chi-squared; p>0.23).

Graphs were created in which the nodes represented those genes which had a ZRE3 motif containing a specific nucleotide, Left4G, Left7T or Left10A, and had a set of GO biological process term annotations. The edges in the graph represented PSRSS scores between genes of ≤0.4. In each of the graphs created from genes containing one of the residues, Left4G, Left7T or Left10A, there is one larger subgraph which contains the majority of the nodes (Figure 5.12).

In each of the three graphs the function annotating the largest group of nodes was ‘transcription’. This was to be expected as it is known from the general ZRE3 analysis that a large number of the genes which contain a ZRE3 motif are involved in transcription (Figure 5.5). Chi squared tests show that in the Left4G graph the proportion of transcription related genes is significantly higher than in the total set of ZRE3 containing genes ($p<0.05$; χ^2 with Yates correction). The proportion of transcription related genes in the set of genes with Left7T or Left10A was not found to be significantly higher than in the general ZRE3 containing gene population ($p>0.05$; χ^2 with Yates correction).

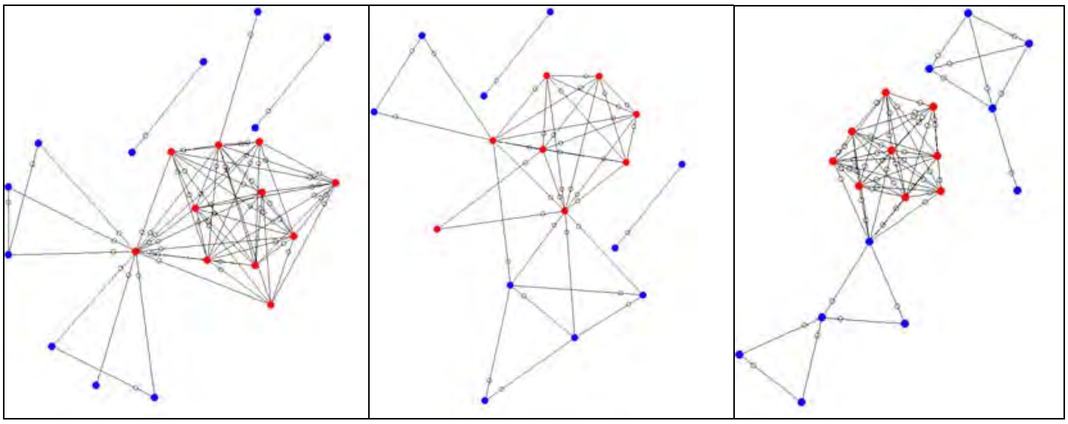


Figure 5.12. Graphs representing similarities between biological process annotations for genes with ZRE3 motifs containing Left4G, Left7T, Left10A. Edges represent a PSRSS score between nodes ≤ 0.2. Nodes coloured red represent genes involved in transcription, the blue nodes represent genes involved in other processes.

5.3.9. Analysis of results from the second round of EMSAs

The new PWMs were built with larger numbers of example motifs than the original PWMs. PWM_A2 contained all 18 motifs found to bind Zta in EMSA analysis; RpZRE3, 6 from the first round of EMSA analysis and 12 from the second round of analysis. PWM_B contained 17 motifs found to bind Zta in EMSA analysis: RpZRE3, 5 from the first round of EMSA analysis, excluding *XPC* (see section 5.2.6) and 12 from the second round of analysis. This should have given a more accurate idea of the flanking regions found next to the ZRE3 motifs. A visual analysis of the PWMs showed obvious differences between the two motifs in all flanking positions (Figure 5.13, Figure 5.14).

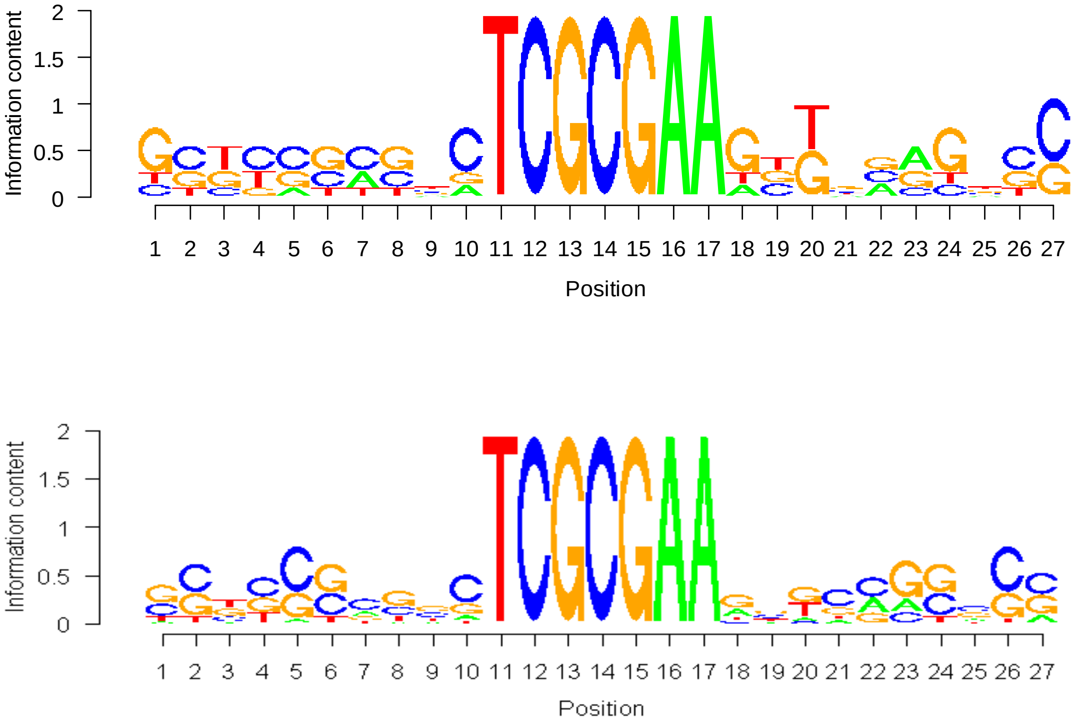


Figure 5.13 Position Weight Matrices ZRE3_A (top) and ZRE3_A2 (bottom). ZRE3_A is created from the original 6 motifs used to search for ZRE3 motifs.

ZRE3_A2 is created from 18 motifs, the original 6 motifs and the 12 motifs found to bind Zta through further EMSA experiments.

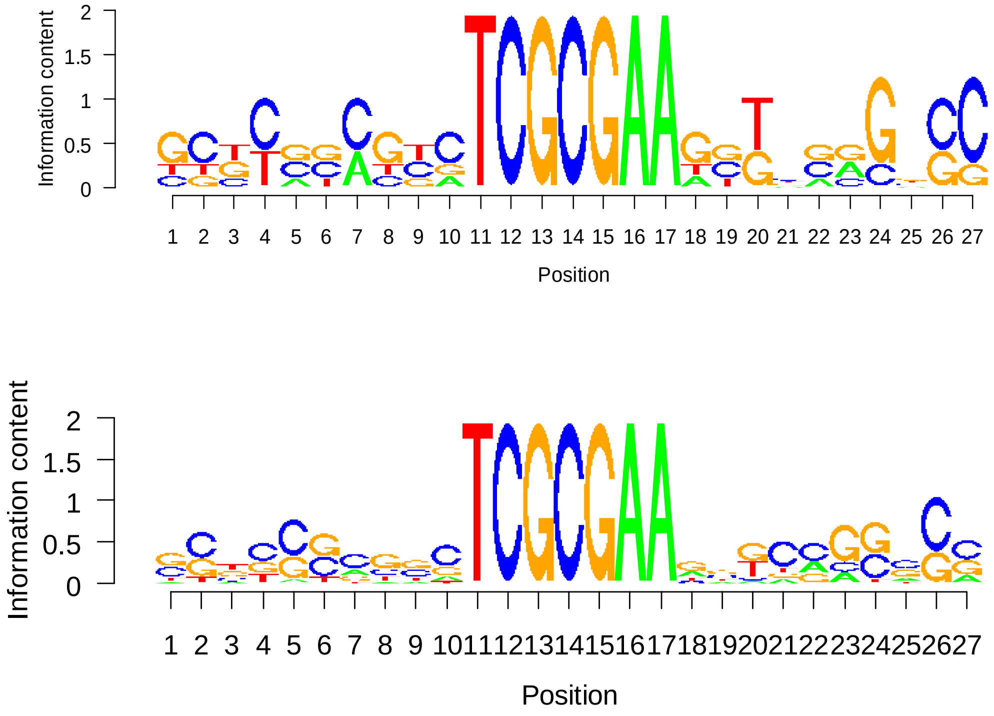


Figure 5.14 Position Weight Matrices ZRE3_B (top) and ZRE3_B2 (bottom). ZRE3_B is created from 5 of the original motifs used to search for ZRE3 motifs, *XPC* is not included due to aberrant behaviour. ZRE3_A2 is created from 17 motifs, the original ZRE3_B 5 motifs and the 12 motifs found to bind Zta through further EMSA experiments.

The new PWMs predicted more ZRE3 motifs than the original PWMs (Figure 5.15). This is most likely to be due to the PWMs becoming more general as a consequence of more diverse sequences being added to the PWM. This implies that either the flanking regions of the ZRE3 motifs are not highly conserved, or that due to the small number of sequences included in the PWM, sequences which are more diverse are being predicted as ZRE3 motifs.

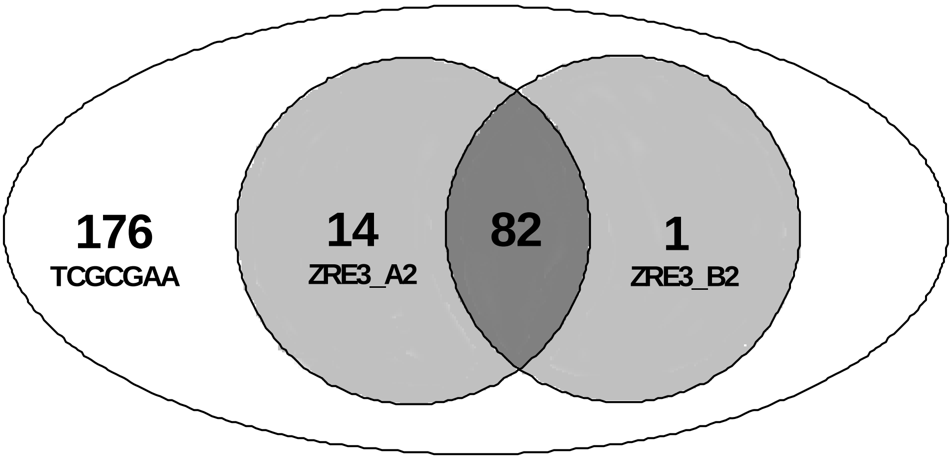


Figure 5.15. Venn diagram showing the genes with ZRE3 motifs predicted in the second round ZRE3 predictions using the TCGCGAA pattern matching, the ZRE3_A2 PWM and the ZRE3_B2 PWM prediction methods.

The ZRE3 predictions made with ZRE3_B2 were found significantly more often in CpG islands than in the set of ZRE3 prediction made by pattern matching (χ^2 with Yates correction; p=0.0420) (Figure 5.16). There was no significant difference found between the number of motifs found in CpG islands in the ZRE3_A2 set and the pattern matching set (χ^2 with Yates correction; p=0.1076).

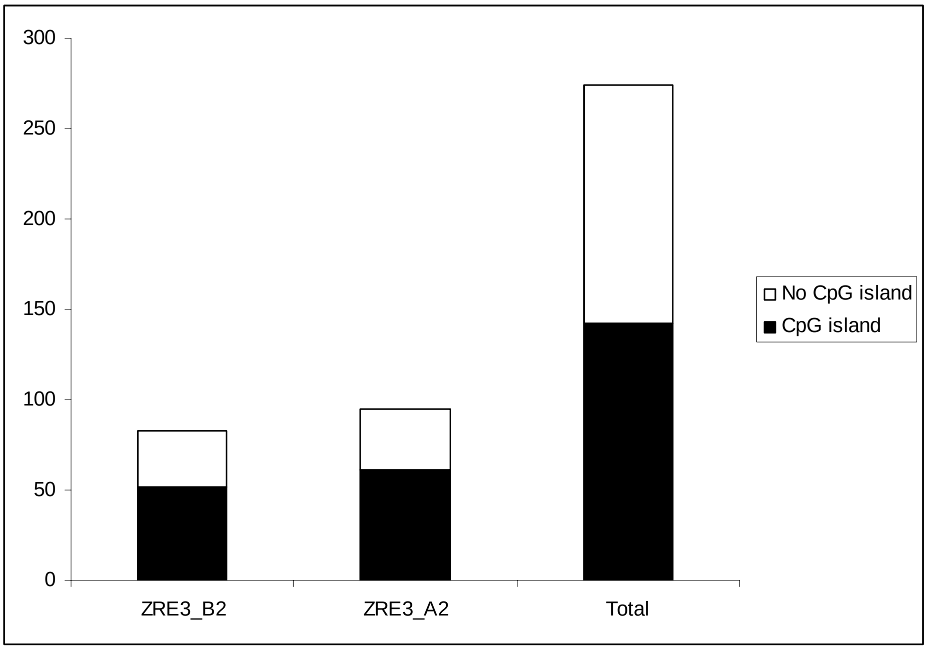


Figure 5.16. Comparison of the number of motifs found in a CpG island against those not found in a CpG island. Significantly more motifs are found in CpG islands in the results found by ZRE3_A2 than out of the total ZRE3 motifs (χ^2 with Yates correction; p=0.0420).

The GO terms which were overrepresented in both the PWM_A2 and PWM_B2 genes were the same as those found using PWM_A and PWM_B (Table 5.1) and no further PWM were created.

5.4. Conclusions

5.4.1. ZRE3 predictions

Due to the small number of sequences included in the ZRE3_A and ZRE3_B PWM, it was to be expected that the inclusion of one more sequence in ZRE3_A would alter the set of genes predicted. To determine whether there is a substantial difference between the flanking sequences of motifs which only bind when methylated and those which also bind when unmethylated, a larger number of experimentally verified sequences would need to be analysed. It would be interesting to test some ZRE3 motifs which were found only with the ZRE3_A PWM and not the ZRE3_B PWM to determine whether these motifs do bind Zta when unmethylated.

If there are differences in the flanking sequences of the ZRE3 motif of methylated and unmethylated binding motifs, it would be beneficial to have more sequences of both types to enable the creation of two separate PWM, one to predict motifs with the classical behaviour and one to predict the motifs with the aberrant behaviour. Regardless of whether the differences in the genes predicted using ZRE3_A and those predicted using ZRE3_B are real, a larger number of sequences would be required to create a robust and representative PWM which predicts binding sites accurately.

The CpG island predictions imply that the PWMs may be finding more functional motifs than the pattern matching search. The presence of a CpG island at the same location as the motif was highly conserved in genes retrieved by both PWMs, indicating that there may be a larger number of housekeeping genes in these datasets than would be expected by chance. Studies in EBV related gastric carcinoma (Chong *et al.*, 2003; Kang *et al.*, 2002; *Vo et al.*, 2002) have shown that hypermethylation occurs frequently in tumour related genes, reducing gene expression in these genes. The presence of ZRE3 in the promoters of genes with CpG islands implies that Zta may be able to activate these genes in a similar way to its involvement in the switch from latent to lytic EBV phases (Karlsson *et al*., 2008b).

Out of the total number of GO terms, that were associated with the genes containing putative ZRE3 sites, only a small number were found to be over represented compared to the annotations in Ensembl 50. Some of the GO terms retrieved provide likely processes with which Zta may be involved, in particular, transcription regulator and repressor activities, the cell cycle related terms and negative regulation of cell proliferation.

5.4.2. Candidate genes

Candidate genes were selected by combining the methods used in this analysis, PWM matches, CpG island co-localisation and annotation with a GO term found to be enriched in the set of ZRE3 containing genes (Figure 5.8). The genes which met these criteria consisted of motifs in 12 genes which had not been tested in the preliminary EMSAs and *MNT*, which had already been shown to bind Zta in Section 5.2.1. These genes are further discussed in Section 5.4.2.1. Further experiments (Section 5.3.6, Figure 5.9) have shown that ZRE3 motifs in all of these candidate genes bind to Zta when the motif is methylated. An analysis of the GO term annotations of these candidate genes finds 5 genes which are related to transcription; *BPTF* (*FALZ*), *LMO4*, *MBD4*, *SEC14L1* AND *TADA3L*. A further 2 genes are annotated with the term ‘negative regulation of cell proliferation’ *PRKD1* and *STI1*.

A text mining analysis of all genes predicted through the PWM analysis to bind to Zta, through a ZRE3 binding site, was carried out (Section 5.3.7). A further 5 candidate genes have been identified through this analysis, the evidence found in the literature for an involvement between these genes and EBV is further described in Section 5.4.2.2.

5.4.2.1. Experimentally verified candidate genes

(1) *BPTF*

Bromodomain PHD finger transcription factor (*BPTF* or *FALZ*), is a bromodomain transcription factor with similarity to *FAC1* (Jones *et al*., 2000). The bromodomain is a conserved 110 amino acid structural region associated with signal dependent, but not basal, transcription regulators. The bromodomain is associated with chromatin mediated transcription regulation and is also often associated with proteins which have histone acetyltransferase activity. A number of other bromodomain proteins, although not BPTF, have implicated in tumorigenesis: RING3 (Denis & Green, 1996), HRX/ALL-1 (Tkachuk *et al.,* 1992; Gu *et al.*, 1992), TIF1 (Le Douarin *et al.,* 1995), RGF7 (Klugbauer & Rabes, 1999), CBP (Lill *et al*., 1997), BRG1 (Dunaief *et al.,* 1994) and P/CAF (Yang *et al.,* 1996).

(2) *LMO4*

LMO4 encodes the nuclear Lim-only protein 4 (LMO4) which is up-regulated in breast cancer and experiments involving over expression in mice have shown LMO4 to cause hyperplasia and tumour formation. It has been suggested that LMO4 regulated the expression of the bone morphogenic protein 7 (BMP7) through the inhibition of HDAC2 recruitment (Wang *et al.,* 2007). Interestingly, *HDAC2* has a predicted ZRE3 motif, found with both ZRE3_A and ZRE3_B, although it was not included in this list of candidate genes due to its lack of a CpG island surrounding the motif. *HDAC2* has been included in the list of candidate genes generated through the text mining analysis (Section 5.4.2.2.) where it has been shown to be linked to EBV processes.

(3) *MBD4*

MBD4 encodes one of a family of methyl-CpG binding proteins which all contain a methyl-CpG binding domain. MBD4 is a thymine glycosylase which recognises the product of de-amination at methyl CpG sites and has been shown to be mutated in human carcinomas with microsatellite instability (Bellacosa *et al.*, 1999; Hendrich *et al.*, 1999; Riccio *et al.*, 1999). The methyl-CpG binding properties of this protein are particularly interesting if one assumes that many genes involved in a cascade starting from ZRE3 are methylated in a similar way to the genes in the cascade causing the switch from latent to lytic states in EBV. None of the other MBD family proteins were found to contain a ZRE3 motif by the PWMs or by the pattern matching search.

(4) *SEC14L1*

SEC14L1 has been mapped to a specific region of 17q25 which contains at least one putative breast and ovarian tumour suppressor gene. *SEC14L1* contains a CRAL/TRIO which is also found in cellular retinaldehyde-binding protein. Loss of this domain may contribute to the formation of breast tumours as retinoids have previously been shown to inhibit the growth of cancerous breast tissue cells (Kalikin *et al.*, 2001).

(5) *TADA3L*

TADA3L, encoding a homolog of the yeast ADA3 protein, has been shown to act as a cofactor for p53 activity through direct interaction between the N-termini of the two proteins while the c-terminus has been shown to bind to p300 and TADA2L, components of histone acetyltransferase complexes (Wang *et al.*, 2001).

TADA3L has also been shown to bind to the E6 protein of human papillomavirus, although only in proteins with large numbers of cancer associated mutations, and not to E6 proteins which were associated with benign lesions (Kumar *et al.,* 2002). Co-expression of TADA3L with E6 has been shown to inhibit TADA3L/p53-mediated transactivation.

(6) *PRKD1*

Serine/threonine protein kinase D1 (*PRKD1*), previously known as 'Protein Kinase C, mu', has been implicated in many processes including apoptosis, immune regulation and cell proliferation (Jaggi *et al.,* 2007; Li *et al.,* 2006). *PRKD1* is down-regulated in advanced prostate cancer and influences cell adhesion and motility of prostate cancer cells in vitro (Jaggi *et al.,* 2007; Jaggi *et al.,* 2003). The protein has also been shown to influence androgen receptor function in prostate cancer cells although the exact mechanism is yet to be elucidated (Mak *et al.,* 2008).

(7) *CAPN2*

The *CAPN2* gene codes for the calpain 2, (m/ll) large subunit, one subunit in the m-calpain proteases. Calpains are calcium activated neutral proteases, cysteine proteases which are intracellular and non-lysosomal (Hosfield *et al*., 1999). Calpains have been shown to promote either apoptosis or survival signals in response to different stimuli (Tan *et al*., 2006). Calpain has also been shown to be involved in caspase-independent cell death and necrosis (Goll *et al*., 2003).

(8) *CDO1*

Cysteine dioxygenase type I, CDO1, initiates several important metabolic pathways related to pyruvate and several sulfurate compounds including sulfate, hypotaurine and taurine. CDO1 is a critical regulator of cellular cysteine concentrations and has an important role in maintaining the hepatic concentation of intracellular free cysteine (Joseph & Maroney, 2007; Stipanuk *et al*., 2009).

There are no known interactions between EBV or EBV related genes and *CDO1*. However, *CDO1* has been shown to be overexpressed in Sezary syndrome, a rare and aggressive CD4$^+$ cutaneous T-cell lymphoma and therefore may be involved in tumourogenesis. This is most likely to be through the inhibition of apoptosis through taurine which is catalysed by CDO1 (Booken *et al*., 2008).

(9) *KIF1B*

KIF1B, the kinesin family member 1B, encodes a motor protein involved in the transportation of mitochondria and synaptic vesicle precursors. KIF1B has been shown to induce apoptotic cell death and may act as a haploinsufficient tumor suppressor gene (Munirajan *et al*., 2008). No direct relationship between KIF1B and EBV has been discovered.

(10) *LLGL1*

LLGL1 is a human homolog to the lethal giant larvae gene, a tumour suppressor gene, found in *Drosophila*. Recent experiments (Lu *et al.*, 2009) have indicated that mutations in LLGL1 are involved in hepatocellular carcinoma progression and that the gene has similar tumour suppressor properties as that sound in the *Drosophila* homolog.

Other experiments have shown that loss of *LLGL1* expression in endometrial cancer patients may contribute to lymph node metastasis and may be one of the factors lead to a poor prognosis (Tsuruga *et al*., 2007).

(11) *MNT*

MNT was the only one of the original 5 genes for EMSA binding assays that was also found as one of the bioinformatics pipeline candidate genes.

The Max binding protein (MNT) is though to be a transcriptional repressor of Myc dependent transcription activation and cell growth. MNT is involved in repressing transcription by binding to DNA binding proteins at its N terminal Sin3-interaction domain. Like *CDO1, MNT* has been shown to be involved in Sezary syndrome. Loss of *MNT* was observed in 40-55% of patients and was associated with deregulated gene expression (Vermeer *et al*., 2008).

(12) *PLEKHJ1*

PLEKHJ1 encodes the pleckstrin homology domain containing family J member 1 also referred to as the guanine nucleotide releasing protein. There does not appear to be any literature available on the function of PLEKHJ1. Although Ensembl 50 has annotated this gene with the terms ‘transcription coactivator activity’ and ‘transcription corepressor activity’ it is unclear how this annotation was inferred.

(13) *TOP2B*

TOP2B encodes a DNA topoisomerase, an enzyme able to control and alter the topological state of DNA during transcription. *TOP2B* involvement in the cell cycle has been inferred through it modification by mitogens. It has also been suggested that TOP2B may be involved in an apoptotic response seen in response to doxorubicin in peripheral blood cells (Kersting *et al*., 2006).

5.4.2.2. Text Mining candidate genes

(1) *BCL2L11*

BCL2L11 is a proapoptotic protein. It has been shown that EBV infection leads to post transcriptional down-regulation of BCL2L11 by degradation through the proteasome pathway. The signal for degredation is given by phosphoryation of BCL2L11 by the constitutive EBV-activated kinase ERK1/2. EBV-mediated resistance to growth factor deprivation in human B lymphocytes has been shown to be dependent on BCL2L11, suggesting an important contribution to the oncogenic potential of EBV (Clybouw *et al*., 2005).

BCL2L11 was not predicted to contain a ZRE3 motif by the ZRE3_A PWM, precluding it from being included in the candidate genes for EMSA analysis. However, it was predicted to be contain a ZRE3 motif by ZRE3_B, ZRE3_B2 and by ZRE3_A2. *BCL2L11* was also found to coincide with a predicted CpG island and was annotated with the over-represented GO terms: ‘transcription coactivator activity’ and ‘transcription corepressor activity’. This evidence coupled with the known link to EBV makes it an ideal candidate for further analysis.

(2) *FRAP1*

FRAP1, FK506 binding protein 12-rapamycin associated protein 1, also know as the mammalian target of papamycin (mTOR) is a serine/threonine protein kinase which regulates cell growth, proliferation, motility, and survival as well as protein synthesis and transcription. Expression of *LMP2A* by EBV activates phosphatidylinositol 3-kinase/Akt located upstream of mTOR. In carcinoma genes, LMP2A has been shown to activate mTOR, leading to wortmannin and rapamycin sensitive inhibition of 4E binding protein 1, a negative regulator of transcription, and increased c-Myc protein translation (Moody *et al.*, 2005).

Although a link through the literature has been discovered, this is a fairly tenous connection. The bioinformatics analysis did not identify the ZRE3 motif in *FRAP1* and a true motif when using any of the PWM and it was not annotated with any of the over-represented GO terms. Although it is entirely possible that *FRAP1* is involved in the EBV-host interaction, it seems unlikely that it is through the mechanism of Zta binding to ZRE3.

(3) *GTF2A2*

GTF2A2 or TFIIA is an ubiquitous transcription factor involved in the formation of the RNA polymerase II pre-initiation complex. Experiments into the functional binding of Zta to DNA have shown that the involvement of TFIIA is essential (Lieberman, 1994). Firstly the architectural proteins HMG-1 and HMG-2 mediate the formation of an enhanceosome (Panne, 2008), a protein complex binding to the enhancer region of a gene causing an acceleration of transcription, consisting of Zta and cellular Sp1. Following this the TFIIA and TFIID proteins are recruited to the promoter to form a complex (Ellwood *et al.*, 1999).

Although GTF2A2 has been shown to be an essential component of the mechanism allowing the binding of Zta to DNA, it does not seem likely that it is itself regulated through a ZRE3 binding site. The ZRE3 binding site found in *GTF2A2* was only found through the pattern matching search and not using any of the PWMs. GTF2A2 was not found to be annotated with any of the over-represented GO terms. However, this literature search has shown a role for GTF2A2 in transcription, highlighting one of the limitations inherent in the reliance on GO term annotations.

(4) *<u>HDAC2</u>*

HDAC2 encodes histone deacetylase 2 which has been shown to interact with the EBV transcription factor EBNA3C (Epstein-Barr virus nuclear antigen 3C). *EBNA3C* is one of the small number of gene expressed during the latent phase of the EBV life cycle (Bajaj *et al.*, 2008; Saha *et al.*, 2009; Subramanian *et al.*, 2002; Yi *et al.*, 2009). It has been shown that a complex containing both HDAC1 and HDAC2 allows EBNA3C to recruit deacetylase activity (Knight *et al.*, 2003). HDAC2 may also be involved in the EBV DNA replication process at OriP through a complex with SNF2h and HDAC1 which co-ordinated G1-specific chromatin remodelling (Zhou *et al.*, 2005). HDAC2 has also been shown to interact with *LMO4*, another ZRE3 containing gene (Wang *et al.*, 2007).

HDAC2 was one of the original 5 ZRE3 containing genes confirmed to bind to Zta, it was not discovered as a candidate gene via the bioinformatics pipeline because it does not occur in a CpG island. The occurrence of a ZRE3 in a CpG island as one of the criteria for a candidate gene was added to allow the possibilities to be restricted to those which were most likely to be of interested. It is likely that this criteria has prevented the method from identifying likely candidate genes such as *HDAC2*, however in the interests of practicality it was necessary to reduce the number of candidate genes in this way.

(5) *MAP3K7*

EBV protein LMP1 has been shown to activate the TRAF6, TAB1, MAP3K7, IKKalpha/IKKbeta/IKKgamma mediated NF-KB pathway (Soni, 2007). The LMP1 protein contains 6 transmembrane proteins, 2 C-terminal activation regions (CTAR1 and CTAR2) and 2 transformation effector sites (TES1 and TES2). LMP1 TES2/CTAR2 has been shown to activate the TRAF6, TAB1, MAP3K7, IKKalpha/IKKbeta/IKKgamma mediated NF-KB pathway. LMP1 TES1/CTAR1 has been shown to activate the TRAF2, NIK, IKKalpha and p52 mediated noncanonical NF-KB pathway (Soni, 2007). Removal of MAP3K7 has been shown to result in the loss of LMP1-induced JNK activation. It has been suggested that a LMP1-associated complex consisting of TRAF6, TAB2 and MAP3K7 has an essential role in the activation of JNK (Uemura, 2006). The bioinformatics analysis of *MAP3K7* showed an annotation of the gene with the GO terms ‘transcription coactivator activity’ and ‘transcription corepressor activity’, these terms are confirmed by the literature search. However MAP3K7 is not a strong candidate for regulation by Zta as the ZRE3 motif was not identified by any of the PWMs.

5.4.3. Future analysis

The search for ZRE3 motifs using an exact pattern match, and both the PWMs was carried out in the 500bp upstream of the TSS. Analyses carried out in Section 2.1, Section 3.9 of this thesis have shown that the position of a TFBS regulating a gene can be both upstream and downstream and at a greater further distance from the TSS than 500bp. It would be interesting to do a whole genome search for ZRE3 and to compare the distances of the motifs found to the start sites of genes, both to discover more genes which may be regulated by ZRE3, and to assess the variation in the position of the ZRE3 motifs.

Further experimental work is planned to confirm whether the bound ZRE3s are functional transcription factors. Quantatative polymerase chain reactions (qPCR) can be used to amplify, as in classical PCR, and also quantify a target DNA/RNA molecule. Where classical PCR can be used to determine whether a gene is expressed by targeting the gene transcripts, qPCR can be used to quantify the level of expression (VanGuilder *et al*., 2008). The use of small interfering RNA (siRNA) is another experimental technique which could be used. SiRNA takes advantage of naturally occurring post-transcriptional gene silencing which induces the degradation of homologous mRNA transcripts and hence causes the suppression of post-transcriptional gene expression (Hammond *et al*., 2001; Mello & Conte, 2004). This technique could be used to knock down the ZRE3 candidate genes, to allow for the function of the genes to be determined.

6. Discussion

6.1. TFBS prediction methods

A number of methods for the prediction of TFBSs have been developed and analysed in this thesis. However, it is clear from both the current work and from recent publications in this field, that the problem of predicting functional TFBSs remains a complex and, as yet, unsolved problem (e.g. Hawkins *et al*, 2009).

Methods using PWMs and more complex machine learning techniques all produce large numbers of false positive TFBS, and the problem lies in differentiating between functional and non-functional sites. Using consensus methods that combine predictions from more than one source (Chapter 2, Chapter 3), using measures of evolutionary conservation (Chapter 3), integrating epigenetic factors such as CpG island location (Chapter 5) and including known functional information (Chapter 2, Chapter 5) can all lead to a reduction in the number of false positive TFBSs.

The success of the PPMNB method (Chapter 2) as well as numerous other phylogeny based TFBS prediction methods (Hu *et al*., 2007; Liu *et al*., 2008; Struckmann *et al*., 2008) suggests that the inclusion of a measure of evolutionary conservation is useful in the prediction of TFBSs. However, such methods assume that all TFBS follow the same pattern of evolutionary conservation which is likely to be a model that is too simplistic. A recent study into the function, expression and evolution of human transcription factors shows that while the majority of human TFs are vertebrate specific, some are found only in primate species, and others can be found in organisms as evolutionarily distant as *Saccharomyces cerevisiae* (Vaquerizas *et al*., 2009). The PPMNB method presented in Chapter 3 does not account for this variability in expression as the type of TF is not included as a feature in the classifier. Recent work has addressed this problem by using probabilistic models of loss and gain of TFBSs as a part of a CRM (He *et al*., 2009). The issue of different rates of evolution is further complicated by the tissue specific nature of most TFs, which allows for differential expression to take place (Jiang *et al*., 2004; Johnston *et al*., 2007; Vaquerizas *et al*., 2009). The modelling of different rates of evolution of TFBS and the generation and use of tissue specific TFBS data will lead to more accurate prediction methods in the future.

6.2. Identification of IFN-gamma linked gene targets

Three regions on chromosomes 8, 10 and 11 were found to contain genes linked to the IFN-γ response to *M. tuberculosis,* as modelled by the response to the BCG vaccine (Newport *et al*., see appendix). The genetic linkage studies located 3 large regions on each chromosome, comprising 532 genes in total, but only a small number of these were expected to be target genes actually linked to the IFN-γ response. The aim of the current work was to identify these target genes by identifying those that shared common TFBSs. The hypothesis was that the target genes linked to the response were likely to be co-regulated, and co-regulated genes are more likely to share TFs (Allocco *et al*., 2004; Marco *et al*., 2009).

Analyses that compare the occurrence of shared TFs and the level of co-expression between genes almost exclusively concentrate on determining the TFs in common in genes already known to be co-expressed (Eisermann *et al.*, 2008; Hatanaka *et al.*, 2008; Sarkar & Maitra, 2008; Veerla *et al*., 2006; Zadissa *et al*., 2007). A small number of studies have shown that the higher number of shared transcription factors indicates a higher level of co-expression (Allocco *et al*., 2004; Marco *et al*., 2009). However, these analyses also imply that in higher mammals this is a complex relationship. In the Allocco *et al*. (2004) analysis, 100% of *Sacharromyces cerevisae* genes which had an expression correlation of ≥0.9 were found to share at least one TF in common (Allocco *et al*., 2004). Although genes which shared a TF were still shown to be significantly more likely to be co-expressed in the *Drosphila melanogaster* analysis carried out by Marco *et al*. (2009) than genes which did not share a TF, 76% of genes with expression correlations ≥0.9 did not share a TF in common (Marco *et al*., 2009). In the current analysis we are searching for functionally linked genes in a large dataset of genes that were not known to be co-expressed. Hence the relationship between co-expression and function needs to be considered.

The definition of function is in itself a complex area. In this thesis we have defined function using GO term annotations. These methods rely on the accuracy and completeness of the annotations. However, it is likely that the annotations are incomplete, with many proteins having multiple and as yet unknown functions. A further problem, inherent in the use of GO terms to determine function, is that while some annotations may denote a function likely to involve co-regulated genes, other annotations may describe functions where the genes are less likely to be co-regulated. In the analyses by Marco *et al*. (2009) and Allocco *et al*. (2004), it is co-expression that is shown to be correlated with the number of TFs found in common, not function. While the level of co-expression exhibited between genes is a good indicator of similar functions, many genes involved in similar functions are differentially expressed and equally many genes which are co-expressed are found to be involved in divergent functions (Montaner *et al*., 2009). Therefore, the relationship between co-expression, co-regulation and function requires further evaluation.

The use of a consensus method in Chapter 2 identified 5 TFs and 21 candidate genes. The use of the PMANB or PPMNB classifiers (Chapter 3), revealed a total of 84 and 53 TFs respectively and the percentage of genes in the dataset containing binding sites for the TFs ranged from less than 5% to 100%. The consensus method (Chapter 2) results are likely to include a large number of false negatives, whilst the Bayes methods include a large number of false positives. Both the consensus and Naïve Bayes methods use known PWM as the basis for the TFBS search. This means that novel and uncharacterised TFBS will not be predicted. Using methods to identify novel TFBS (e.g. Siddharthan, 2008) would increase the complexity of the predictions still further. The use of such methods is a further area of study that could be addressed and could lead to the identification of additional and novel CRMs occurring in genes from the genetic linkage data.

The graph analysis in Chapter 4 identified 3 likely candidate genes and a putative cis-regulatory motif by which they are regulated. These genes were *GRM5* (metabotropic 5), *COPS5* (COP9 constitutive photomorphogenic homolog subunit 5) and *HPSE2* (heparanese 2). The CRM consisted of a GATA1 motif followed by an AP1 motif with between 34 and 36 nucleotides between them. The motifs were found within the region -300 and -250 nucleotides from the TSS. The known functions of these genes and their link to T-cell processes, increases the likelihood of the genes being linked to the IFN-γ immune response. The discovery of two TFBSs at similar distances from each other and from the TSS in each of the genes is likely to correspond to a cis-regulatory motif

(CRM). This would suggest that the three genes are co-regulated, lending even more weight to the hypothesis that these are the IFN-γ linked genes.

Other work has also shown a link between GATA-like and AP-1 TFs. A murine study (Roger *et al*., 2005) identified a cis regulatory region within 350bp upstream of the TSS critical for regulating the immune response gene *tlr4* (toll-like receptor 4). This region contains binding sites for: Ets, AP-1, Oct and GATA-like TFs. The Ets and AP-1 binding motifs were determined to be positive regulators of *tlr4*, the GATA-like and Oct motifs were determined to be negative regulators of *tlr4* (Roger *et al*., 2005). The possibility of the AP-1 and the GATA motifs acting antagonistically is an interesting proposition and should be analysed in the experimental confirmation of our putative CRM. Other studies, however, have shown AP-1 and members of the GATA family to act synergistically. The endothelin 1 gene has been shown to contain a GATA-2 motif followed between 34 and 23 base pairs later by an AP-1 motif (Kawana *et al*., 1995). The CRM is closer to the TSS than that found in our genes but the distance between the motifs is comparable. The study shows that the AP-1 and the GATA-2 motifs act co-operatively to greatly enhance gene expression. It was also shown that both GATA-1 and GATA-3 are also able to co-operate with AP-1 (Kawana *et al*., 1995).

Electrophoretic mobility shift assays (EMSA) were used to confirm Zta binding to ZRE3 motifs in Chapter 5. This technique can be used to verify the binding of any protein to any DNA motif, provided that the protein can be obtained and that the DNA motif is suitable for the design of primers. To analyse the entire set of TFBS predictions made by the methods developed in Chapters 2 and 3 would be extremely time consuming and expensive due to the large numbers of transcription factors and

DNA primers that the analysis would entail. However, EMSA analysis would be a suitable technique for the confirmation of the constituent TFBSs in the putative CRM found in *GRM5*, *COPS5* and *HPSE2*.

The functionality of the Gata-1 / AP-1 CRM suggested by our analysis could be determined through experimental means. Similar studies (Kawana *et al*., 1995; Roger *et al*., 2005) have used PCR-based site mutagenesis to alter the TFBS and hence knock out its ability to bind. Quantitative real time polymerase chain reaction (qPCR) is one method which would allow for the prediction of the functional binding of a TF to a TFBS. As with classical PCR, the method allows for the amplification of a DNA or RNA sequence through the use of specific primers. However, the advantage to using qPCR is that the amount of amplification is quantified, from which the starting amount of DNA/RNA of that sequence can be calculated. This is particularly useful for determining the amount of mRNA present for a specific gene. The amount of mRNA present can be used to infer the level expression exhibited by the gene.

6.3. Prediction of EBV transcription factor, Zta, regulated host genes

A combination of PWM and exact pattern matches revealed an initial 273 ZRE3 genes of which 18 were shown experimentally to bind Zta. The EMSAs showed Zta to bind strongly to all the ZRE3 sites selected when methylated, and less strongly when un-methylated. This is the expected behaviour, although XPC was shown to bind strongly when un-methylated as well as when methylated. Some un-methylated binding was seen in two of the other genes tested by EMSA (*CDO1* and *LMO4)*; however, this was not strong and may have been due to unspecific binding.

The binding of Zta to all the ZRE3 sites suggests one of two possibilities. The first is that the methods used successfully removed non functional ZRE3 sites from the analysis. The second is that all ZRE3 sites bind Zta *in vivo* through the strong consensus binding site, and the flanking regions make very little difference to the binding ability. If this second hypothesis is correct, further experimentation to determine the functionality of the candidate genes is required. The Sinclair group plan to carry out qPCR experiments to determine the level of gene expression reliant on the ZRE3 motifs

An alternative method which could be used to confirm the binding of TFBSs is chromatin immunoprecipitation sequencing (ChIP-Seq). ChIP-Seq consists of two parts; a chromatin precipitation step (ChIP) and a sequencing step (Seq). The ChIP step first cross-links a DNA binding protein, e.g. a transcription factor, to the DNA motif using an agent such as formaldehyde or DTBP (Di-tert-butyl peroxide). The DNA bound to the proteins is then lysed and broken up in pieces 0.2-1 kb in length via sonication. Purification of the protein-DNA complexes is carried out by immunoprecipitation; the cross-linking of the protein-DNA complex is reversed, allowing the molecules to be separated. At this point the second step, sequencing, is used to determine the sequence of the protein binding DNA motifs.

The ChIP-Seq method is preferential to the more traditional ChIP-chip method which requires large sets of tiling arrays to determine the binding sequence and has a much lower resolution. ChIP-Seq is a particularly useful technique when looking for binding sites for only one or a small number of transcription factors. For this reason it may be useful for the IFN-γ linked genes regions if a small number of TFs were chosen for analysis; AP1 and GATA1 would be suitable candidate for this as they have been found in a putative CRM. A particularly suitable application for this method would be to determine whether further putative ZRE3 binding sites from the analysis in Chapter 5 are able to bind to the EBV Zta transcription factor. This method would allow for all of the predicted ZRE3 sites to be analysed.

6.4. Further development of computational methods

Experimental analyses have shown that genetic regulation is dependent on co-operation between TFs with TFBSs clustered into CRMs (Berman *et al.*, 2002; Clyde *et al.*, 2003; Harbison *et al.*, 2004). In complex systems, such as the immune system, this can involve hundreds of genes and TFs. In these cases a systems biology approach as applied in a number of recent papers (e.g. Janky *et al*., 2009; Ray *et al*., 2008; Segal *et al*, 2008) is required to understand of the complex spatial and temporal relationships within the system.

To enable the successful application of systems biology to the problem, additional information on TFBS and CRM occurrence will be required. This information would include a) knowledge of the system in which the co-regulated genes are assumed to act, b) knowledge of the transcription factors or CRMs which are known to regulate the system; in our case the immune system. This knowledge would enable the creation of an immune specific network allowing for an analysis of how our predicted candidate genes fit into the whole system.

To achieve this aim, a set of immune specific genes could be collated; but this in itself is a complex task. The human immune system traverses many different organs and cell types, and responds to many different stimuli. Some information is known about specific TFs which are known to express genes involved in the immune system (Pan *et al*., 2009; Yu *et al*., 2009; Zhou *et al*., 2009). For example, the GATA family of genes are known to be involved in the differentiation of blood cells (Ho *et al.,* 2009; Lowry & Mackay, 2006; Ohneda & Yamamoto, 2002; Rothenburg & Scripture-Adams, 2008). This and other information from the literature and from organism specific databases could be used as a starting point to collate a dataset of immune response genes.

The use of experimental methods to determine genes which are up-regulated during an immune response would be useful in the creation of an immune specific dataset. However, the generation of experimental data is expensive, and is often restricted to biological systems in model organisms. One solution would be to use one model system to develop the methods, before applying them to human genes as in the recent work conducted using *Drosophila* microarray data compiled by Li *et al*. and used in the Marco *et al*. analysis of co-regulation (Li *et al*., 2008; Marco *et al*., 2009).

One method for using human genes directly would be to use a database of co-expressed genes, such as COXPRESdb (Obayashi *et al*., 2008). This database collates the GeneChip data found in NCBI GEO into sets of co-expressed genes. The database consists of coexpressed gene networks for 19,777 human genes, 1820 GO terms and 62 human tissues. A large number of sets of co-expressed genes could be retrieved from the database, and used as training and testing datasets for predictions of co-regulation and functional similarity. The COXPRES database has been used to carry out a large scale search for TFBS in co-expressed genes (Hatanaka *et al*., 2008); however no attempt has been made to predict other co-expressed genes from this information.

6.5. Conclusion

This thesis has seen the analysis of existing, and the development of novel, methods for TFBS and functional similarity prediction. The, as yet unsolved problem, is the ability to differentiate between functional and non-functional binding sites. All current methods are currently limited by the availability of experimentally validated TFs and their binding sites. As new high-throughput techniques for experimental validation of TFBS become available, and the data is stored in public databases, then the computational methods in this area will advance significantly.

Reference List

Abou-Sleiman, P.M., Healy, D.G., Wood, N.W. (2004) Genetic approaches to solving common diseases. *J. Neurol.* 251 (10), 1169-1172.

Aerts, S., Vilain, S., Hu, S., Tranchevent, L.C., Barriot, R., Yan, J., Moreau, Y., Hassan, B.A., Quan, X.J. (2009) Integrating computational biology and forward genetics in Drosophila. *PLoS Genet.,* 5(1), e1000351.

Akiyama, Y. (1998) TFSEARCH: Searching Transcription Factor Binding Sites. Available: http://www.cbrc.jp/research/db/TFSEARCH.html. Last accessed 19 September 2009.

Allocco, D.J., Kohane, I.S. and Butte, A.J. (2004) Quantifying the relationship between co-expression, co-regulation and gene function. *BMC Bioinformatics*, 5, 18.

Antequera, F. (2003) Structure, function and evolution of CpG island promoters. *Cell Mol. Life. Sci.*, 60(8), 1647-1658.

Altman, R.B. and Raychaudhuri, S. (2001) Whole-genome expression analysis: challenges beyond clustering. *Curr. Opin. Struct. Biol.*, 11, 340-347.

Altshuler, D., Daly, M.J., Lander, E.S. (2008) Genetic mapping in human disease. *Science,* 322(5903), 881-888.

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

Awomoyi, A.A., Marchant, A., Howson, J.M., McAdam, K.P., Blackwell, J.M., Newport, M.J. (2002) Interleukin-10, polymorphism in SLC11A1 (formerly NRAMP1) and susceptibility to tuberculosis. *J. Infect. Dis*. 186(12), 1808-1814

Ayadi, W., Karray-Hakim, H., Feki, L., Khabir, A., Boudawara, T., Ghorbel, A., Daoud, J., Frikha, M., Hammami, A. (2009) IgA antibodies against the Epstein-Barr nuclear antigen1 as a valuable biomarker for the diagnosis of nasopharyngeal carcinoma in Tunisian patients. *J. Med. Virol*., 81(8), 1412-1421.

Baghdadi, J.E., Rolova, M., Alter, A., Rangue, B., Chentoufi, M., Lazrak, F., Archane, M.I., Casanova, J.L., Benslimane, A., Schurr, E., Abel, L. (2006) An autosomal dominant major gene confers predisposition to pulmonary tuberculosis in adults. *J. Exp. Med.*, 203, 1679-1684.

Bajaj, B.G., Murakami, M., Cai, Q, Verma, S.C., Lan, K., Robertson, E.S. (2008) Epstein-Barr virus nuclear antigen 3C interacts with and enhances the stability of the c-Myc oncoprotein. *J. Virol.*, 8, 4082-4090.

Balaram, P., Kien, P.K., Ismail, A. (2009) Toll-like receptors and cytokines in immune responses to persistent mycobacterial and Salmonella infections. *Int. J. Med. Microbiol.* 299(3), 177-185.

Baldwin, N.E., Chesler, E.J., Kirov, S., Langston, M.A., Snoddy, J.R., Williams, R.W., Zhang, B. (2005) Computational, integrative and comparatice methods for the elucidation of genetic coexpression networks. *J. Biomed. Biotechnol.,* 2005(2), 172-180.

Bellacosa, A., Cicchillitti, L., Schepis, F., Riccio, A., Yeung, A.T., Matsumoto, Y., Golemis, E.A., Genuardi, M., Neri, G. (1999) MED1, a novel human methyl-CpG binding endonuclease, interacts with DNA mismatch repair protein MLH1. *Proc. Natl. Acad. Sci. USA*, 96(7), 3969-3974.

Barrett, J.H., Sheehan, N.A., Cox. A., Worthington, J., Cannings, C. Teare, M.D. (2007) Family based studies and genetic epidemiology: theory and practice. *Hum. Hered.* 64(2), 146-148.

Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I.F., Soboleva, A., Tomashevsky, M., Edgar, R. (2007) NCBI GEO: mining tens of millions ofexpression profiles – database and tools update. *Nucleic Acids Res*., 35, D760-D765.

Bates, D.L., Chen, Y., Kim, G., Guo, L., Chen, L. (2008) Crystal structures of multiple GATA zinc fingers bound to DNA reveal new insights into DNA recognition and self-association by GATA. *J. Mol. Biol.* 281, 1292-1306.

Bellamy, R., Ruwende, C., Corrah, T., McAdam, K.P., Whittle, H.C., Hill, A.V. (1998) Variations in the NRAMP1 gene and susceptibility to tuberculosis in West Africans. *N. Engl. J. Med.,* 338 (10), 640-644.

Bellamy, R., Beyers, N., McAdam, K.P., Ruwende, C., Gie, R., Samaai, P., Bester, D., Meyer, M., Corrah, T., Collin, M., Camidge, D.R., Wilkinson, D., Hoal-Van Helden, E., Whittle, H.C., Amos, W., van Helden, P, Hill, A.V. (2000) Genetic susceptibility to tuberculosis in Africans: a genome-wide scan. *Proc. Natl. Acad. Sci. U. S. A*, 97, 8005-8009.

Berg, O.G., von Hippel, P.H. (1987) Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.*, 193, 723-750.

Berman, B.P., Nibu, Y., Pfeiffer, B.D., Tomancak, P., Celniker, S.E., Levine, M., Rubin, G.M., Eisen, M.B. (2002) Exploiting transcription factor binding sites clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *Proc. Natl. Acad. Sci. U.S.A.*, 99(2), 757-762.

Besson, C., Amiel, C., Le-Pendeven, C., Plancoulaine, S., Bonnardel, C., Rangue, B., Abbed, K., Brice, P., Fermé, C., Carde, P., Hermine, O., Raphael, M., Bresson, J.L., Nicolas, J.C., Gessain, A., Dethe, G., Abel, L. (2009) Strong correlations of anti-viral capsid antigen antibody in first-degree relatve from families with Epstein-Barr virus-related lymphomas. *J. Infect. Dis.*, 199(8), 1121-1127.

Bhaduri-McIntosh, S., Landry, M.L., Mikiforow, S., Rotenberg, M., El-guindy, A., Milller, G. (2007) Serum IgA antibodies to Epstein-Barr virus (EBV) early lytic antigens are present in primary EBV infection. *J.Infect. Dis.*, 195(4), 483-492.

Birkhaug, K. (2005) Vaccination against tuberculosis with BCG. *Psych. Quart.*, 21(3), 453-473.

Booken, N., Gratchev, A., Utikal, J., Weiss, C., Yu, X., Oadoumi, M., Schmuth, M., Sepp, N., Nashan, D., Rass, K., Tüting, T., Assaf, C., Dippel, E., Stadler, R., Klemke, C.D., Goerdt, S. (2008) Sézary syndrome is a unique cutaneous T-cell lymphoma as identified by an expanded gene signature including diagnostic marker molecules CDO1 and DNM3. *Leukemia*, 22(2), 393-399.

Borg, I., Groenen, P. (2005). MDS models and measures of fit. In: *Modern Multidimensional Scaling: theory and applications*. 2nd ed. New York: Springer-Verlag. 29-48.

Bornkamm, G.W., Hammerschmidt, W. (2001) Molecular virology of Epstein-Barr virus. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, 356(1408), 437-459.

Bornkamm, G.W. (2009) Epstein-Barr virus and the pathogenesis of Burkitt's lymphoma: more questions than answers. *Int. J. Cancer*, 124(8), 1745-1755.

Brady, G., Macarthur, G.J., Farrell, P.J. (2008) Epstein-Barr virus and Burkitt lymphoma. *Postgrad Med. J.*, 84(993), 372-377.

Brazma, A., Jonassen, I., Vilo, J., Ukkonen, E. (1998) Predicting gene regulatory elements in silico on a genomic scale. *Genome Res.*, 8, 1202-1215.

Brivanlou, A.H., Darnell, J.E., Jr. (2002) Signal transduction and the control of gene expression. *Science*, 295, 813-818.

Brightbill, H.D., Libraty, D.H., Krutzik, S.R., Yang, R.B., Belisle, J.T., Bleharski, J.R., Maitland, M., Norgard, M.V., Plevy, S.E., Smale, S.T., Brennan, P.J., Bloom, B.R., Godowski, P.J., Modlin, R.L. (1999) Host defense mechanisms triggered by microbial lipoproteins through toll-like receptors. *Science*, 285(5428), 732-736.

Brustolini, O.J., Fietto, L.G., Cruz, C.D., Passos, F.M. (2009) Computational analysis of the interaction between transcription factors and the predicted secreted proteome of the yeast Kluyveromyces lactis. *BMC Bioinformatics*, 10, 194.

Bryne, J.C., Valen. E., Tang, M.H., Marstrancd, T., Winther, O., da Piedade, I., Krogh, A., Lenhard, B., Sandelin, A. (2008) JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.*, 36, D102-D106.

Burd, A.L., Ingraham, R.H., Goldrick, S.E., Kroe, R.R., Crute, J.J., Grygon, C.A. (2004) Assembly of major histocompatability complex (MHC) class II transcription factors: association and promoter recognition of RFX proteins. *Biochemistry*, 43(40), 12750-12760.

Calvo, B., Larranaga, P., Lozano, J.A. (2007a) Learning Bayesian classifiers from positive and unlabeled examples. *Patttern Recognition Letters*, 28, 2375-2384.

Calvo, B., López-Bigas, N., Furney, S.J., Larrañaga, P., Lozano, J.A. (2007b) A partially supervised classification approach to dominant and recessive human disease gene prediction. *Comput. Methods Programs Biomed.*, 85, 229-237.

Carmack, C.S., McCue, L.A., Newberg, L.A., Lawrence, C.E. (2007) PhyloScan: identification of transcription factor binding sites using cross-species evidence. *Algorithms Mol. Biol.*, 2, 1.

Cartharius, K., Frech, K., Grote, K., Klocke, B., Haltmeier, M., Klingenhoff, A., Frisch, M., Bayerlein, M., Werner, T. (2005) MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics*, 21, 2933-2942.

Casanova, J-L., Abel, L. (2002) Genetic dissection of immunity to mycobacteria: the human model. *Ann. Rev. Immunol.*, 20, 581-620.

Castelo, R., Guigó, R. (2004) Splice site identification by idlBNs. *Bioinformatics*, 20(1), i69-76.

Cayrol, C., Flemington, E.K. (1996) The Epstein-Barr virus bZIP transcription factor Zta causes G0/G1 cell cycle arrest through induction of cyclin-dependent kinase inhibitors. *EMBO J.*, 15, 2748-2759.

Chen, L., Lu, L., Feng, K., Li, W., Song, J., Zheng, L., Yuan, Y., Zeng, Z., Feng, K., Lu, W., Cai. Y. (2009)Mulitple classifier integration for the prediction of protein structural classes. *J. Comput. Chem.,* Epub ahead of print.

Chan, J., Xing, Y., Magliozzo, R.S., Bloom, B.R. (1992) Killing of virulent Mycobacterium tuberculosis by reactive nitrogen intermediates produced by activated murine macrophages. *J. Exp. Med.* 175(4), 1111-1122.

Chan, J., Tanaka, K., Carroll, D., Flynn, J., Bloom, B.R. (1995) Effects of nitric oxide sythase inhibitors on murine infection with Mycobacterium tuberculosis. *Infect. Immun.* 63(2), 736-740.

Chang, Y.N., Dong, D.L., Hayward, G.S., Hayward, S.D. (1990) The Epstein-Barr virus Zta transactivator: a member of the bZIP family with unique DNA-binding specificity and a dimerization domain that lacks the characteristic heptad leucine zipper motif. *J. Virol.*, 64, 3358-3369.

Chevallier-Greco, A., Manet, E., Chavrier, P., Mosnier, C., Daillie, J., Sergeant, A. (1986) Both Epstein-Barr virus (EBV)-encoded trans-acting factors, EB1 and EB2, are required to activate transcription from an EBV early promoter. *EMBO J.*, 5, 3243-3249.

Chekmenev, D.S., Haid, C., Kel, A.E. (2005) P-match: transcription factor binding site search by combining patterns and weight matrices. *Nucleic Acids Res.*, 33, W432-7.

Chi, T., Carey, M. (1993) The ZEBRA activation domain: modular organization and mechanism of action. *Mol. Cell Biol.*, 13, 7045-7055.

Choi, D., Appukuttan, B., Binek, S.J., Planch, S.R., Stout, J.T., Rosenbaum, J.T., Smith, J.R. (2008) Prediction of Cis-Regulatory Elements controlling Genes Differentially Expressed by Retinal and Choroidal Vascular Endothelial Cells. *J. Occul. Bio. Dis. Infor.*, 1(1), 37-45.

Choi, H.S., Kim, J.R., Lee, S.W., Cho, K.H. (2008) Why have serine/threonine kinases been evolutionarily selected in eukaryotic signalling cascades? *Comput. Biol. Chem.*, 32(3), 218-221.

Chung, Y.L., Lee, Y.H., Yen, S.H., Chi, K.H. (2000) A novel approach for nasopharyngeal carcinoma treatment uses phenylbutyrate as a protein kinase C modulatorL implications for radiosensitization and EBV-targeted therapy. *Clin. Cancer Res.* 6(4), 1452-1458.

Clements, M., van Someren, E.P., Knijnenburg, T.A., Reinders, M.J. (2007) Integration of known transcription factor binding site information and gene expression data to advance from co-expression to co-regulation. *Genomics Prot. Bioinf.*, 5, 86-101.

Clevidence, D.E., Overdier, D.G., Tao, W., Oian, X., Pani, L., Lai, E., Costa, R.H. (1993) Identification of nine tissue-specific transcription factors of the hepatocyte nuclear factor 3/forkhead DNA-binding-domain family. *Proc. Natl. Acad. Sci. U. S. A*, 90, 3948-3952.

Clybouw, C., McHichi, B., Mouhamad, S., Auffredou, M.T., Bourgeade, M.F., Sharma, S., Leca, G., Vazguez, A. (2005) EBV infection of human B lymphocytes leasds to down-regulation of Bim expression: relationship to resistance to apoptosis. *J. Immunol.*, 175(5), 2968-2973.

Clyde, D.E., Corado, M.S., Wu, X., Paré, A., Papatsenko, D., Small, S. (2003) A self-organizing system of repressor gradients establishes segmental complexity in Drosophila. *Nature*, 426 (6968), 849-853.

Cohen, J. (1977) Statistical power analysis for the behavioural sciences. In: *The significance of a single product moment r_S*. 2nd ed. New Jersey: Lawrence Erlbaum Associates Inc. 75-108.

Collins, A. (2009) Approaches to the identification of susceptibility genes. *Parasite Immunol.* 31(5), 225-233.

Collins, J.F., Hu, Z. (2007) Promoter analysis of intestinal genes induced during iron-deprivation reveals enrichment of conserved SP1-like binding sites. *BMC Genomics*, 8, 420.

Comstock, G.W. (1978) Tuberculosis in twins: a re-analysis of the Prophit survey. *Am. Rev. Respir. Dis.*, 117, 621-624.

Corà, D., Herrmann, C., Dieterich, C., Di Cunto, F., Provero, P., Caselle, M. (2005) Ab initio identification of putative human transcription factor binding sites by comparative genomics. *BMC Bioinformatics*, 6, 110.

Corbett, E.L. (2003) The growing burden of tuberculosis: global trends and interactions with the HIV epidemic. *Arch. Intern. Med.*, 163(9), 1009-1021.

Countryman, J., Miller, G. (1985) Activation of expression of latent Epstein-Barr herpesvirus after gene transfer with a small cloned subfragment of heterogeneous viral DNA. *Proc. Natl. Acad. Sci. U. S. A*, 82, 4085-4089.

Craig, E.A., Stevens, M.V., Vaillancourt, R.R., Camenisch, T.D. (2008) MAP3Ks as central regulators of cell fate during development. *Dev. Dyn.*, 237(11), 3102-3114.

Crooks, G.E., Hon, G., Chandonia, J.M., Brenner, S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, 14, 1188-1190.

Dai, X., He, J., Zhao, X. (2007) A new systematic computational approach to predicting target genes of transcription factors. *Nucleic Acids Res.*, 35, 4433-4440.

Daniel, T.M., Sippola., A.A., Okwera, A., Kabengera, S., Hatanga, E., Aisu, T., Nyole, S., Byekwaso, F., Vjecha, M., Ferguson, L.E. (1994) Reduced sensitivity of tuberculosis serodiagnosis in patients with AIDS in Uganda. *Tuber. Lung Dis.*, 75, 33-37.

Danko, C.G., Pertsov, A.M. (2009) Identification of gene co-regulatory modules and associated cis-elements involved in degenerative heart disease. *BMC Med. Genomics*, 2, 31.

Delbridge, L.M., O-Riordan, M.X. (2007) Innate recognition of intracellular bacteria. *Curr. Opin. Immunol.* 19(1), 10-16.

Denis, F., Laurent, A., Gilleron, R., Tommasi, M. (2003) Text classification and co-training from positive and unlabeled examples. *Proceedings of the ICML 2003 Workshop: The Continuum from Labeled to Unlabeled Data*, 80-87.

Denis, G.V., Green, M.R. (1996) A novel, mitogen-activated nuclear kinase is related to a Drosophila developmental regulator. *Genes Dev.*, 10(3), 261-271.

Dietrich, J., Doherty, T.M. (2009) Interaction of Mycobacterium tuberculosis with the host: consequences for vaccine development. *APMIS*, 117(5-6), 440-457.

Dolin, P.J., Raviglione, M.C., Kochi, A. (1994) Global tuberculosis incidence and mortality during 1990-2000. *Bull. World Health Organ.*, 72, 213-220.

Dong, Y., Rohn, W.M. and Benveniste ,E.N. (1999) IFN-gamma regulation of the type IV class II transactivator promoter in astrocytes. *J. Immunol.*, 162, 4731-4739.

Doraisamy, S., Golzari, S., Norowi, N.M. (2008) A Study on Feature Selection and Classification Techniques for Automatic Genre Classification of Traditional Malay Music. *ISMIR 2008*, 331-336.

Du, H.Y., Olivo, M., Mahendran, R., Huang, Q., Shen, H.M., Ong, C.N., Bay, B.H. (2007) Hypericin photoactivation triggers down-regulation of matrix metalloproteinase-9 expression in well-differentiated human nasopharyngeal cancer cells. *Cell Mol. Life Sci.,* 64(7-8), 979-988.

Dunaief, J.L., Strober, B.E., Guha, S., Khavari, P.A., Alin, K., Luban, J., Begemann, M., Crabtree, G.R., Goff, S.P. (1994) The retinoblastoma protein and BRG1 form a complex and cooperate to induce cell cycle arrest. *Cell*, 79(1), 119-30.

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., Huber, W. (2005) BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21(16), 3439-3440.

van der Eijk, E.A., van de Vosse, E., Vandenbroucke, J.P., van Dissel. J.T. (2007) Heredity versus environment in tuberculosis in twins: the 1950s United Kingdom Prophit Survey Simonds and Comstock revisited. *Am. J. Respir. Crit. Care Med.*, 176, 1281-1288.

Eisermann, K., Tandon, S., Bazarov, A., Brett, A., Fraizer, G., Piontkivska, H. (2008) Evolutionary conservation of zinc finger transcription factor binding sites in promoters of genes co-expressed with WT1 in prostate cancer. *BMC Genomics*, 9, 337.

El-Guindy, A.S., Paek, S.Y., Countryman, J., Miller, G. (2006) Identification of constitutive phosphorylation sites on the Epstein-Barr virus ZEBRA protein. *J. Biol. Chem.*, 281, 3085-3095.

Elnitski, L., Jin, V.X., Farnham, P.J., Jones, S.J. (2006) Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques. *Genome Res.*, 16, 1455-1464.

Ellwood, K., Huang, W., Johnson, R., Carey, M. (1999) Multiple layers of co-operativity regulate enhancesome-responsive RNA polymerase II transcription complex assembly. *Mol. Cell. Biol.* 19(4), 2613-2623.

Epstein, M.A., Achong, B.G., Barr, Y.M. (1964) Virus particles in cultured lymphoblasts from burkitt’s lymphoma. *Lancet*, 1(7335), 702-703

Erill, I., O'Neill, M.C. (2009) A reexamination of information theory-based methods for DNA-binding site identification. *BMC Bioinformatics*, 10, 57.

Farrell, P.J., Rowe, D.T., Rooney, C.M., Kouzarides, T. (1989) Epstein-Barr virus BZLF1 trans-activator specifically binds to a consensus AP-1 site and is related to c-fos. *EMBO J.*, 8, 127-132.

Ferretti, V., Poitras, C., Bergeron, D., Coulombe, B., Robert, F., Blanchette, M. (2007) PReMod: a database of genome-wide mammalian cis-regulatory module predictions. *Nucleic Acids Res*., 35, D122-D126.

Ferry, J.A. (2006) Burkitt's lymphoma: clinicopathologic features and differential diagnosis. *Oncologist.*, 11, 375-383.

Fine, P.E. (1995) Variation in protection by BCG: implications of and for heterologous immunity. *Lancet*, 346, 1339-1345.

Flicek, P., Aken, B.L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., Down, T., Dyer, S.C., Eyre, T., Fitzgerald, S., Fernandez-Banet, J., Gräf, S., Haider, S., Hammond, M., Holland, R., Howe, K.L., Howe, K., Johnson, N., Jenkinson, A., Kähäri, A., Keefe, D., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Megy, K., Meidl, P., Overduin, B., Parker, A., Pritchard, B., Prlic, A., Rice, S., Rios, D., Schuster, M., Sealy, I., Slater, G., Smedley, D., Spudich, G., Trevanion, S., Vilella, A.J., Vogel, J., White, S., Wood, M., Birney, E., Cox, T., Curwen, V., Durbin, R., Fernandez-Suarez, X.M., Herrero, J., Hubbard, T.J., Kasprzyk, A., Proctor, G., Smith, J., Ureta-Vidal, A., Searle, S. (2008) Ensembl 2008. *Nucleic Acids Res.*, 36, D707-D714.

Flower, K., Hellen, E., Newport, M.J., Jones, S., Sinclair, A. (2010) Evaluation of a Prediction Protocol to Identify Potential Targets of Epigenetic Reprogramming by the Cancer Associated Epstein Barr Virus. *Plos One*, 5(2), e9443.

Flynn, J.L., Chan, J. (2001) Immunology of tuberculosis. *Annu. Rev. Immunol.*, 19, 93-129.

Friberg, M.T. (2007) Prediction of transcription factor binding sites using ChIP-chip and phylogenetic footprinting data. *J. Bioinform. Comput. Biol.*, 5, 105-116.

Fried, M., Crothers, D.M. (1981) Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis. *Nucleic Acids Res.*, 9, 6505-6525.

Friedman, N., Geiger, D., Goldszmidt, M. (1997) Bayesian Network Classifiers. *Machine Learning*, 29, 131-163.

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

Frousios, K.K., Iconomidou, V.A., Karletidi, C.M., Hamodrakas, S.J. (2009) Amyloidogenic determinants are usually not buried. *BMC Struct. Biol.*, 9(1), 44.

Fu, Y., Weng, Z. (2005) Improvement of TRANSFAC matrices using multiple local alignment of transcription factor binding site sequences. *Genome Inform.*, 16, 68-72.

Ganley, A.R., Kobayashi, T. (2007) Phylogenetic footprinting to find functional DNA elements. *Methods Mol. Biol.*, 395, 367-380.

Garner, M.M., Revzin, A. (1981) A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system. *Nucleic Acids Res.*, 9, 3047-3060.

Giorgi, C., Romagnoli, A., Pinton, P., Rizzuto, R. (2008) Ca2+ signalling, mitochondria and cell death. *Curr. Mol. Med.*, 8(2), 119-130.

Girardin, S.E., Hugot, J.P., Sansonetti, P.J. (2003) Lessons from Nod2 studies: towards a link between Crohn's disease and bacterial sensing. *Trends Immunol.*, 24(12), 652-658.

Goll, D.E., Thompson, V.F., Li, H., Wei, W., Cong, J. (2003) The calpain system. *Physiol Rev.*, 83, 731-801.

Gower, J.C. (1966) Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53, 325-328.

Grabe, N. (2000) Alibaba2.1. Available: http://www.gene-regulation.com/pub/programs/alibaba2/index.html. Last accessed 13 September 2009.

Gross, P., Oelgeschläger, T. (2006) Core promoter-selective RNA polymerase II transcription. *Biochem. Soc. Symp.*, 73, 225-236.

Gu, Y., Cimino, G., Alder, H., Nakamura, T., Prasad, R., Canaano, O., Moir, D.T., Jones, C., Nowell, P.C., Croe, C.M. (1992) The (4;11)(q21;q23) chromosome translocations in acute leukemias involve the VDJ recombinase. *Proc. Natl. Acad. Sci. USA*, 89(21), 10464-10468.

Gutierrez, M.G., Master, S.S., Singh, S.B., Taylor, G.A., Colombo, M.I., Deretic, V. (2004) Autophagy is a defense mechanism inhibiting BCG and Mycobacterium tuberculosis survival in infected macrophages. *Cell*, 119(6), 753-766.

Guyon, I., Weston, J., Barnhill, S., Vapnik, V. (2002) Gene selection for cancer classification using support vector machines. *Machine Learning*, 46, 389-422.

Halfon, M.S., Gallo, S.M., Bergman, C.M. (2008) REDfly 2.0: an integrated database of cis-regulatory modules and transcription factor binding sites in Drosophila. *Nucleic Acids Res.*, 36, D594-598.

Hall, M.A., Smith, L.A. (1998) Practical feature subset selection for machine learning. *Proceesings of the 21st Australasian Computer Science Conference*, 181-191.

Hammond, S.M., Caudy, A.A., Hannon, G.J. (2001) Post-transcriptional gene silencing by double-stranded RNA. *Nature Rev. Genetics*, 2, 110-119.

Hanley, J.A., McNeil, B.J. (1982) The Meaning and Use of the Area under a Reciever Operating Characteristic (ROC) Curve. *Radiology*, 143, 29-36.

Hanley, J.A., McNeil, B.J. (1983) A Method of Comparing Areas under ROC curves derived from same cases. *Radiology*, 148, 839-843.

Harbison, C.T., Gordon, D.B., Lee, T.I., Rinadldi, N.J., Macisaac, K.D., Danford, T.W., Hannett, N.M., Tagne, J.B., Reynolds, D.B., Yoo, J., Jennings, E.G., Zeitlinger, J., Pokholok, D.K., Kellis, M., Rolfe, P.A., Takusagawa, K.T., Lander, E.S., Gifford, D.K., Fraenkel, E., Young, R.A. (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431(7004), 99-104.

Harigae, H. (2006) GATA transcription factors and haematological diseases. *Tohoku. J. Exp. Med.*, 210(1), 1-9.

Haseman, J.K., Elston, R.C. (1972) The investigation of linkage between a quantitative trait and a marker locus. *Behav. Genet.*, 2, 3-19.

Hatanaka, Y., Nagasaki, M., Yamaguchi, R., Obayashi, T., Numara, K., Fujita, A., Shimamura, T., Tamada, Y., Imoto, S., Kinoshita, K, Nakai, K., Mayano, S. (2008) A novel strategy to search conserved transcription factor binding sites among coexpressing genes in human. *Genome Inform*., 20, 212-221.

Hawiger, J. (2001) Innate immunity and inflammation: a transcriptional paradigm. *Immunol. Res.*, 23, 99-109.

Hawkins, J., Grant, C., Noble, W.S., Bailey, T.L. (2009) Assessing phylogenetic motifs for predicting transcription factor binding sites. *Bioinformatics*, 25(12), i339-i347.

He, X., Ling, X., Sinha, S. (2009) Alignment and prediction of cis-regulatory modules based on a probabilistic model of evolution. *PLoS. Comput. Biol.*, 5, e1000299.

Hendrich, B., Hardeland, U., Ng, H.H., Jiricny, J., Bird, A. (1999) The thymine glycosylase MBD4 can bind to the product of deamination at methylated CpG sites. *Nature*, 410(6750), 301-304.

Heym, B., Honoré, N., Truffot-Pernot, C., Banerjee, A., Schurra, C., Jacobs, W.R.Jr., van Embden., J.D., Grosset, J.H., Cole, S.T. (1994) Implications of multidrug resistance for the future of short-course chemotherapy of tuberculosis: a molecular study. *Lancet*, 344, 293-298.

Ho, I.C., Tai, T.S., Pai, S.Y. (2009) GATA3 and the T-cell lineage: essential functions before and after T-helper-2cell differentiation. *Nat. Rev. Immunol.*, 9(2), 125-135.

Hosfield, C.M., Elce, J.S., Davies, P.L., Jia, Z. (1999) Crystal structure of calpain reveals the structural basis for Ca(2+)-dependent protease activity and a novel mode of enzyme activation. *EMBO J.*, 18, 6880-6889.

Hu, Z., Hu, B., Collins, J.F. (2007) Prediction of synergistic transcription factors by function conservation. *Genome Biol.*, 8(12), R257.

Huang, J., Yuan, H., Lu, C., Cao, X., Wan, M. (2007) Jab1 mediates protein degredation of the Rad9-Rad1-Hus1 checkpoint complex. *J. Mol. Biol.* 371(2), 514-527.

Hubbard, T.J., Aken, B.L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., Down, T., Dyer, S.C., Fitzgerald, S., Fernandez-Banet, J., Graf, S., Haider, S., Hammond, M., Herrero, J., Holland, R., Howe, K., Johnson, N., Kahari, A., Keefe, D., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Melsopp, C., Megy, K., Meidl, P., Ouverdin, B., Parker, A., Prlic, A., Rice, S., Rios, D., Schuster, M., Sealy, I., Severin, J., Slater, G., Smedley, D., Spudich, G., Trevanion, S., Vilella, A., Vogel, J., White, S., Wood, M., Cox, T., Curwen, V., Durbin, R., Fernandez-Suarez, X.M., Flicek, P., Kasprzyk, A., Proctor, G., Searle, S., Smith, J., Ureta-Vidal, A., Birney, E. (2007) Ensembl 2007. *Nucleic Acids Res.*, 35, D610-D617.

Hughes, J.D., Estep, P.W., Tavazoie, S., Church, G.M. (2000) Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J. Mol. Biol.,* 296(5), 1205-1214.

Ivanciuc, O. (2008) Weka machine learning for predicting the phospholipidosis inducing potential. *Curr. Top. Med. Chem.*, 8(18), 1691-1709.

Jacobs, R.F. (1994) Multiple-drug-resistant tuberculosis. *Clin. Infect. Dis.*, 19, 1-8.

Jaggi, M., Rao, P.S., Smith, D.J., Hemstreet, G.P., Balaji, K.C. (2003) Protein kinase C mu is down-regulated in androgen-independent prostate cancer. *Biochem. Biophys. Res. Commun.,* 307(2), 254-260.

Jaggi, M., Du, C., Zhang, W., Balaji, K.C. (2007) Protein kinase D1: a protein of emerging translational interest. *Front Biosci.,* 12, 3757-3767.

Jakobsdottir, J., Gorin, M.B., Conley, Y.P., Ferrell, R.E., Weeks, D.E. (2009) Interpretation of genetic association studies: markers with replicated highly significant odds ratios may be poor classifiers. *PLoS Genet.*, 5(2), e1000337.

Janky, R., Helden, J., Babu, M.M. (2009) Investigating transcriptional regulation: from analysis of complex networks to discovery if cis-regulatory elements. *Methods*., 48(3), 277-286.

Jenssen, T.K., Laegreid, A., Komorowski, J. and Hovig, E. (2001) A literature network of human genes for high-throughput analysis of gene expression. *Nat. Genet.*, 28, 21-8.

Jiang, Z., Wu, X.L., Garcia, M.D., Griffin, K.B., Michal, J.J., Ott, T.L., Gaskins, C.T., Wright Jr, R.W. (2004) Comparative gene-base in silico analysis of transcriptomes in different bovine tissues and (or) organs. *Genome*, 47(6), 1164-1172.

Jin, V.X., O'Geen, H., Ivengar, S., Green, R., Farnham, P.J. (2007) Identification of an OCT4 and SRY regulatory module using integrated computational and experimental genomics approaches. *Genome Res*., 17(6), 807-817.

Jo, E.K. (2008) Mycobacterial interaction with innate receptors: TLRs, C-type lectins, and NLRs. *Curr. Opin. Infect. Dis.* 21(3), 279-286.

Johnston, D.S., Turner, T.T. Finger, J.N., Owtcharuk, T.L., Kopf, G.S., Jelinsky, S.A. (2007) Identifiaction of epididymis-specific transcriptions in the mouse and rat by transcriptional profiling. *Asian J. Androl.,* 9(4), 522-527.

Jones, M.H., Hamana, N., Shimane, M. (2000) Identification and characterization of BPTF, a novel bromodomain transcription factor. *Genomics,* 63, 35-39.

Joseph, C.A., Maroney, M.J. (2007) Cystein dioxygenase: structure and mechanism. *Chem. Commun. (Camb)*, 32, 3338-3349.

Kalikin, L.M., Bugeaud, E.M., Palmbos, P.L., Lyons, R.H.Jr., Petty, E.M. (2001) Genomic characterization of human SEC14L1 splice variants within a 17q25 candidate tumour suppressor gene region and identification of an unrelated embedded expressed sequence tag. *Mamm. Genome.*, 12(12), 925-929.

Kallman, E.J., Reisner, D. (1943) Twin studies on genetic variation in resistance to tuberculosis. *J. Hered.* 34, 269-276.

Karlsson, Q.H., Schelcher, C., Verrall, E., Petosa, C., Sinclair, A.J. (2008a) Methylated DNA recognition during the reversal of epigenetic silencing is regulated by cysteine and serine residues in the Epstein-Barr virus lytic switch protein. *PLoS Pathog.*, 4(3), e1000005.

Karlsson, Q.H., Schelcher, C., Verrall, E., Petosa, C., Sinclair, A.J. (2008b) The reversal of epigenetic silencing of the EBV genome is regulated by viral bZIP protein. *Biochem. Soc. Trans.*, 36(4), 637-639.

Kawana, M., Lee, M-E., Quertermous, E.E., Quertermous, T. (1995) Cooperative interaction of GATA-2 and AP1 regulates transcription of the endothelin-1 gene. *Mol. Cell. Biol.*, 15(8), 4225-4231.

Kel, A.E., Gössling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O.V., Wingender, E. (2003) MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.*, 31, 3576-3579.

Kel-Margoulis, O.V., Kel, A.E., Reuter, I., Deineko, I.V., Wingender, E. (2002) TRANSCompel: a database on composite regulatory elements in eukaryotic genes. *Nucleic Acids Res.*, 30, 332-334.

Kersting, G., Tzvetkov, M.V., Huse, K., Kulle, B., Hafner, V., Brockmöller, J., Wojnowski, L. (2006) Topoisomerase II beta expression level correlated with doxorubicin-induced apoptosis in peripheral blood cells. *Naunyn Schmiedebergs Arch. Pharmacol.*, 374(1), 21-30.

Kieff, E. (1996) Epstein-Barr virus and its replication. In: Fields BN, Knipe DM, Howley PM, eds. *Fields virology*, 3rd ed. Philadelphia: Lipincott-Raven Publishers, 2343–96.

Kim, T.M., Jung, M.H. (2006) Identification of transcriptional regulators using binding site enrichment analysis. *In Silico Biol.*, 6(6), 531-544.

Knight, J.S., Lan, K., Subramanian, C., Robertson, E.S. (2003) Epstein-Barr virus nuclear antigen 3C recruits histone deacetylase activity and associates with the corepressors mSin3A and NCoR in human B-cell lines. *J. Virol.*, 77(7), 4261-4272.

Kornberg, R.D. (1974) Chromatin structure: a repeating unit of histones and DNA. *Science*, 184, 868-871.

Koudritsky, M., Domany, E. (2008) Positional distribution of human transcription factor binding sites. *Nucleic Acids Res.,* 36, 6795-6805.

Kullback, S. (1987) The Kullback-Leibler distance. *The American Statistician,* 41, 340-341.

Kumar, A., Zhao, Y., Meng, G., Zeng, M., Srinivasan, S., Delmolino, L.M., Gao, Q., Dimri, G., Weber, G.F., Wazer, D.E., Band, H., Band, V.(2002) Human papillomavirus oncoprotein E6 inactivates the transcriptional coactivator human ADA3. *Mol. Cell Biol.,* 22(16), 5801-5812.

Kumar, S., Carugo, O. (2008) Consensus prediction of protein conformational disorder from amino acidic sequence. *Open Biochem. J.,* 2, 1-5.

Lam, L.V., Thanh, T., Chi, V.T., Tuy le, M. (2009) Genetic diversity of Hevea IRRDB'81 collection assessed by RAPD markers. *Mol. Biotechnol*., 42(3), 292-298.

Lardenois, A., Chalmel, F., Bianchetti, L., Sahel, J.A., Léveillard, T., Poch, O. (2006) PromAn: an integrated knowledge-based web server dedicated to promoter analysis. *Nucleic Acids Res.,* 34, W578-583

Larsson, E., Lindahl, P., Mostad, P. (2007) HeliCis: a DNA motif discovery tool for colocalized motif pairs with periodic spacing. *BMC. Bioinformatics.*, 8, 418.

Le Douarin, B., Zechel, C., Garnier, J.M., Lutz, Y., Tora, L., Pierrat, P., Gronemever, H., Chambon, P., Losson, R. (1995) The N-terminal part of TIF1, a putative mediator of the ligand-dependent activation function (AF-2) of nuclear receptors, is fused to B-raf in the oncogenic protein T18. *EMBO J.*, 14(9), 2020-2033.

Lee, S.J., Choi, D., Rhim, H., Cho, H.J., Ko, Y.G., Kim, C.G., Kang, S. (2008) PHB2 interacts with RNF2 and represses CP2c-stimulated transcription. *Mol. Cell Biochem.,* 319(1-2), 69-77.

Lee, T.I., Young, R.A. (2000) Transcription of eukaryotic protein-coding genes. *Annu. Rev. Genet.*, 34, 77-137.

Lehman, A.M., Ellwood, K.B., Middleton, B.E., Carey, M. (1998) Compensatory energetic relationships between upstream activators and the RNA polymerase II general transcription machinery. *J. Biol. Chem.*, 273, 932-939.

Lenhard, B., Wasserman, W.W. (2002) TFBS: Computational framework for transcription factor binding site analysis. *Bioinformatics*, 18, 1135-1136.

Lenhard, B., Sandelin, A., Mendoza, L., Engström, P., Jareborg, N., Wasserman, W.W. (2003) Identification of conserved regulatory elements by comparative genome analysis. *J. Biol.,* 2, 13.

Levitsky, V.G. (2004) RECON: a program for prediction of nucleosome formation potential. *Nucleic Acids Res.*, 32, W346-W349.

Li, M., Paik, H.I., Balch, C., Kim, Y., Li, L., Huang, T.H., Nephwe, K.P., Kim, S. (2008) Enriched transcription factor binding sites in hypermethylated gene promoters in drug resistance cancer cells. *Bioinformatics*, 24(16), 1745-1748.

Li, X.Y., MacArthur, S., Bourgon, R., Nix, D., Pollard, D.A., Iver, V.N., Hechmer, A., Simirenko, L., Stapleton, M., Luengo Hendricks, C.L., Chu, H.C., Ogawa, N., Inwood, W., Sementchenko, V., Beaton, A., Weiszmann, R., Celniker, S.E., Knowles, D.W., Gingeras, T., Speed, T.P., Eisen, M.B., Biggin, M.D. (2008) Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. *PLoS Biol.*, 6(2), e27.

Li, Z., Wang, C., Prendergast, G.C., Pestell, G.C. (2006) Cyclin D1 functions in cell migration. *Cell Cycle*, 5(21), 2440-2442.

Lieberman, P. (1994) Identification of functional targets of the Zta transcription activator by formation of stable preinitiaion complex intermediates. *Mol. Cell. Biol.*, 14(12), 8365-8375.

Lill, N.L., Grossman, S.R., Ginsberg, D., DeCaprio, J., Livingston, D.M. (1997) Binding and modulation of p53 by p300/CBP coactivators. *Nature*, 387(6635), 823-827.

Lin, C.T. (2009) Relationship between Epstein-Barr virus and nasopharyngeal carcinoma pathogenesis. *Chinese J. Cancer*, 28(8), 791-804.

Lin, H., Xiao, J., Luo, X., Chen, G., Wang, Z. (2009) Transcriptional control of pacemaker channel genes HCN2 and HCN4 by Sp1 and implications in re-expression of these genes in hypertrophies myocytes. *Cell Physiol. Biochem.*, 23(4-6), 317-326.

Ling, Y., West, A.G., Roberts, E.C., Lakey, J.H., Sharrocks, A.D. (1998) Interaction of transcription factors with serum response factor. Identification of the Elk-1 binding surface. *J. Biol. Chem.*, 273, 10506-10514.

Liu, G., Weirauch, M.T., Van Tassell, C.P., Li, R.W., Sonstegard, T.S., Matukumalli, L.K., Connor, E.E., Hanson, R.W., Yang, J. (2008) Identification of conserved regulatory elements in mammalian promoter regions: a case study using the PCK1 promoter. *Gen. Prot. Bioinf.*, 6(3-4), 129-143.

Liu, H., Setiono, R. (1995) Chi2: Feature Selection and Discretization of Numeric Attributes. *Proceesings of the IEEE 7th International Conference on Tools with Artificial Intelligence*, 338-391.

Liu, P.T., Stenger, S., Li, H., Wenzel. L., Tan, B.H., Krutzik, S.R., Ochoa, M.T., Schauber, J., Wu, K., Meinken, C., Kamen, D.L., Wagner. M., Bals, R., Steinmeyer, A., Zügel, U., Gallo, R.L., Eisenberg, D., Hewison, M., Hollis, B.W., Adams, J.S., Bloom, B.R., Modlin, R.L. (2006) Toll-like receptor triggering of a vitamin D-mediated human antimicrobial response. *Science*, 311 (5768), 1770-1773.

Liu, X., Pan, Z., Zhang, L., Sun, O., Wan, J., Tian, C., Xing, G., Yang, J., Liu, X., Jiang, J., He, F. (2008) JAB1 accelerates mitochondrial apoptosis by interaction with proapoptotic BclGs. *Cell Signal,* 20 (1), 230-240.

Lönnroth, K., Jaramillo, E., Williams, B.G., Dye, C., Raviglione, M. (2009) Drivers of tuberculosis epidemics: the role of risk factors and social determinants. *Soc. Sci. Med.,* 68(12), 2240-2246.

Lorch, Y., LaPointe, J.W., Kornberg, R.D. (1987) Nucleosomes inhibit the initiation of transcription but allow chain elongation with the displacement of histones. *Cell*, 49, 203-210.

Lowry, J.A., Mackay, J.P. (2006) GATA-1: one protein, many partners. *Int. J. Biochm. Cell. Biol.,* 38(1), 6-11.

Lu, J., Chen, S.Y., Chua, H.H., Liu, Y.S., Huang, Y.T., Chang, Y., Chen, J.Y., Sheen, T.S., Tasi, C.H. (2000) Upregulation of tyrosine kinase TKT by the Epstein-Barr virus transactivator Zta. *J. Virol.*, 74, 7391-7399.

Lu, X., Feng, X., Man, X., Yang, G., Tang, L., Du, D., Zhang, F., Yuan, H., Huang, Q., Zhang, Z., Liu, Y., Strand, D., Chen, Z. (2009) Aberrant splicing of Hughl-1 is associated with hepatocellular carcinoma progression. *Clin. Cancer Res.,* 15(10), 3287-3296.

Luger, K., Mäder, A.W., Richmod, R.K., Sargent, D.F., Richmond, T.J. (1997) Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature*, 389 (6648), 251-260.

Ma, P.C., Rould, M.A., Weintraub, H., Pabo, C.O. (1994) Crystal structure of MyoD bHLH domain-DNA complex: perspectives on DNA recognition and implications for transcriptional activation. *Cell,* 77(3), 451-459.

Maeder, M.L., Polansky, B.J., Robson, B.E., Eastman, D.A. (2007) Phylogenetic footprinting analysis in the upstream regulatory regions of the Drosophila enhancer of split genes. *Genetics*, 177, 1377-1394.

MacMicking, J.D., North, R.J., LaCourse, R., Mudgett, J.S., Shah, S.K., Nathan, C.F. (1997) Identification of nitric oxide synthase as a protective locus against tuberculosis. *Proc. Natl. Acad. Sci. U.S.A.*, 94(10), 5243-5248.

Maeda, E., Akahane, M., Kiryu, S., Kato, N, Yoshikawa, T., Hayashi, N, Aoki, S., Minami, M., Uozaki, H., Fukayama, M., Ohtomo, K. (2009) Spectrum of Epstein-Barr virus-related diseases: a pictorial review. *Jpn. J. Radiol.,* 27(1), 4-19.

Mak, P., Jaggi, M., Syed, V., Chauhan, S.C., Hassan, S., Biswas, H., Balaji, K.C. (2008) Protein kinase D1 (PKD1) influences androgen receptor (AR) function in prostate cancer cells. *Biochem. Biophys. Res. Commun.*, 373(4), 618-623.

Makita, Y., de Hoon, M.J., Ogasawara, N., Miyano, S., Nakai, K. (2005) Bayesian joint prediction of associated transcription factors in Bacillus subtilis. *Pac. Symp. Biocomput.,* 507-518.

Manning, C.D., Raghavan, P., Schutze, H. (2008) χ^2 Feature selection. In: *Introduction to Information Retrieval.*, 1st ed. Cambridge University Press.

Manson, A., Whitten, S.T., Ferreon, J.C., Fox, R.O., Hilser, V.J (2009) Characterizing the role of ensembl modulation in mutation-induced changes in binding affinity. *J. Am. Chem. Soc.*, 131(19), 6785-6793.

Marco, A., Konikoff, C., Karr, T.L., Kumar, S. (2009) Relationship between gene co-expression and sharing of transcription factor binding sites in Drosophila melanogaster. *Bioinformatics,* Epub ahead of print.

Matteelli, A., Migliori, G.B., Cirillo, D., Centis, R., Girard, E., Raviglion, M. (2007) Multidrug-resistant and extensively drug-resistance Mycobacterium tuberculosis: epidemiology and control. *Expert Rev. Anti. Infect. Ther.,* 5(5), 857-871.

Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Checkmenev, D., Krull, M., Hornischer, K., Voss, N., Stegmaier, P., Lewicki-Potapov, B., Saxel, H., Kel, A.E., Wingender, E. (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, 34, D108-D110.

McNeil, B.J., Hanley, J.A. (1984) Statistical Approaches to the Analysis of ROC curves. *Medical Decision Making,* 4, 136-149.

Mello, C.C., Conte, D. (2004) Revealing the world of RNA interference. *Nature,* 431, 338-342.

Mitchell, P.J., Tjian, R. (1989) Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science,* 245, 371-378.

Mo, W.N., Tang, A.Z., Zhou. L., Huang, G.W., Wang, Z., Zeng, Y. (2009) Analysis of Epstein-Barr viral DNA load, EBV-LMP2 specific cytotoxic T-lymphocytes and levels of CD4+CD25+ T cells in patients with nasopharyngeal carcinomas positive for IgA antibody to EBV viral capsid antigen. *Chin. Med. J. (Engl.),* 122(10), 1173-1178.

Montagut, C., Settleman, J. (2009) Targeting the RAF-MEK-ERK pathway in cancer therapy. *Cancer Lett.*, 283(2), 125-34.

Montaner, D., Minquez, P., Al-Shahrour, F., Dopazo, J. (2009) Gene set internal coherence in the context of functional profiling. *BMC Genomics*, 10, 197

Moody, C.A., Scott, R.S., Amirghahari, N., Nathan, C.A., Young, L.S., Dawson, C.W., Sixbey, J.W. (2005) Modulation of the cell growth regulator mTOR by Epstein-Barr virus-encoded LMP2A. *J. Virol.,* 79(9), 5499-5506.

Morrison, T.E., Mauser, A., Wong, A., Ting, J.P., Kenney, S.C. (2001) Inhibition of IFN-gamma signaling by an Epstein-Barr virus immediate-early protein. *Immunity*, 15, 787-799.

Muller, C.W., Rey, F.A., Sodeika, M., Verdine, G.L., Harrison, S.C. (1995) Structure of the NF-kappa B p50 homodimer bound to DNA. *Nature*, 373, 311-317.

Munirajan, A.K., Ando, K., Mukai, A., Takahashi, M., Suenaga, Y., Ohira, M., Koda, T., Hirota, T., Ozaki, T., Nakagawara, A. (2008) KIF1Bbeta functions as a haploinsufficient tumor suppressor gene mapped to chromosome 1p36.2 by inducing apoptotic cell death. *J. Biol. Chem.*, 283(36), 24426-24434.

Murray, C.J., Styblo, K., Rouillon, A. (1990) Tuberculosis in developing countries: burden, intervention and cost. *Bull. Int. Union Tuberc. Lung Dis.*, 65, 6-24.

Nathan, C. (2002) Inducible nitric oxide synthase in the tuberculous human lung. *Am. J. Respir. Crit. Care Med.*, 166(2), 130-131.

Neph, S., Tompa,M. (2006) MicroFootPrinter: a tool for phylogenetic footprinting in prokaryotic genomes. *Nucleic Acids Res.*, 34, W366-W368.

Newport, M.J., Goetghebuer, T., Weiss, H.A. Whittle, H., Siegrist, C.A., Marchant, A., MRC Gambia Twin Study Group. (2004) Genetic regulation of immune responses to vaccines in early life. *Genes Immun.*, 5(2), 122-129.

Newport, M.J., Goetghebuer, T., Marchant, A. (2005) Hunting for immune response regulatory genes: vaccination studies in infant twins. *Expert Rev. Vaccines*, 4(5), 739-746.

Newport, M.J. (2009) Why hasn’t human genetics told us more about TB? *Int. J. Tuberc. Lung Dis.*, in press.

Niida, A., Smith, A.D., Imoto, S., Aburatani, H., Zhang, M.O., Akiyama, T. (2009) Gene set-based module discovery in the breast cancer transcriptome. *BMC Bioinformatics*, 10, 71.

O'Keefe, G.M, Nguyen, V.T., Pig Tang, T.T., Benveniste, E.N. (2001) IFN-gamma regulation of class II transactivator promoter IV in macrophages and microglia: involvement of the suppressors of cytokine signaling-1 protein. *J. Immunol.*, 166, 2260-2269.

Obayashim, T., Hayashi, S., Shibaoka, M., Saeki, M., Ohta, H., Kinoshita, K. (2008) COXPRESdb: a database of coexpressed gene networks in mammals. *Nucleic Acids Res.*, 36, D77-D82.

Ogata, K., Morikawa, S., Nakamura, H., Sekikawa, A., Inoue, T., Kanai, H., Sarai, A., Ishii, S., Nishimura, Y. (1994) Solution structure of a specific DNA complex of the Myb DNA-binding domain with cooperative recognition helices. *Cell*, 79, 639-648.

Oh, T.M., Kim, J.K., Choi, Y., Choi, S., Yoo, J.Y. (2009) Prediction and experimental validation of novel STAT3 target genes in human cancer cells. *PLoS One*, 4(9), e6911.

Ohneda, K., Yamamoto, M. (2002) Roles of hematopoietic transcription factors GATA-1 and GATA-2 in the development of red blood cell lineage. *Acta Haematol.*, 108(4), 237-245.

Ott, J. (1991) Basics of linkage analysis. In: *Analysis of Human Genetic Linkage*. 2nd ed. Baltimore: John Hopkins University Press. 53-83.

Palyanov, A.Y., Krivov, S.V., Karplus, M., Chevmarev, S.F. (2007) A lattice protein with an amyloidogeneic latent state: stability and folding kinetics. *J. Phys. Chem. B.*, 111(1), 2675-2687.

Pan, F., Yu, H., Dang, E.V., Barbi, J., Pan, X., Grosso, J.F., Jinasena, D., Sharma, S.M., McCadden, E.M., Getnet, D., Drake, C.G., Liu, J.O., Ostrowski, M.C., Pardoll, D.M. (2009) Eos mediates Foxp3-dependent gene silencing in CD4+ regulatory T cells. *Science,* 325(5944), 1142-1146.

Panattoni, M., Sanvito, F., Basso, V., Doglioni, C., Casorati, G., Montini, E., Bender, J.R., Mondino, A., Pardi, R. (2008) Targeted inactivation of the COP9 signalosome impairs multiple stages of T cell development. *J. Exp. Med.*, 205(2), 465-477.

Panne, D. (2008) The enhanceosome. *Curr. Opin. Struct. Biol*., 18(2), 236-242.

Park, S.H., Goo, J.M., Jo, C.H. (2004) Receiver operating characteristic (ROC) curve: practical review for radiologists. *Korean J. Radiol.,* 5(1), 11-8.

Parkinson, H., Kupushesky, M., Kolesnikov, N., Rustici, G., Shojatalab, M., Abeygunawardena, N., Berube, H., Dylag, M., Emam, I, Farne, A., Holloway, E., Lukk, M., Malone, J., Mani, R., Pilicheva, E., Ryner, T.F., Rezwan, F., Sharma, A., Williams, E., Bradley, X.Z., Adamusiak, T., Brandizi, M., Burdett, T., Coulson, R., Krestyaninova, M., Kurnosov, P., Maguire, E., Neogi, S.G., Rocca-Serra, P., Sansone, S.A.., Sklyar, N., Zhai, M., Sarkans, U., Brazma, A. (2009) ArrayExpress update – from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res.,* 37, D868-D872.

Pasquinelli, V., Townsend, J.C., Jurado, J.O., Alvarez, I.B., Quiroga, M.F., Barnes, P.F., Samten, B., García, V.E. (2009) IFN-gamma production during active tuberculosis is regulated by mechanisms that involve IL-17, SLAM, and CREB. *J.Infect.Dis*., 199(5), 661-5.

Pericak-Vance, M.A. (2001) Analysis of genetic linkage data for Mendelian traits. *Curr. Protoc. Hum. Genet.,* Chapter 1, Unit 1.4.

Perumal, R., Nimmakayala, P., Erattaimuthu, S.R., No, E.G., Reddy, U.K., Prom, L.K., Odvody, G.N., Luster, D.G., Magill, C.W. (2008) Simple sequence repeat markers useful for sorghum downy mildew (Peronosclerospora sorghi) and related species. *BMC Genet*., 9, 77.

Pesquita, C., Faria, D., Bastos, H., Ferreira, A.E., Falcão, A.O., Couto, F.M. (2008) Metrics for GO based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics*, 9(5), S4.

Petosa, C., Morand, P., Baudin, F., Moulin, M., Artero, J.B., Muller, C.W. (2006) Structural basis of lytic cycle activation by the Epstein-Barr virus ZEBRA protein. *Mol. Cell*, 21, 565-572.

Pinton, P., Giorgi, C., Sivero, R., Zecchini, E., Rizzuto, R. (2008) Calcium and apoptosis: ER-mitochondria Ca2+ transfer in the control of apoptosis. *Oncogene*, 27(50), 6407-6418.

Pirooznia, M., Yang, J.Y., Yang, M.O., Deng, Y. (2008) A comparative study of different machine learning methods on microarray gene expression data. *BMC Genomics*, 9 Suppl 1, S13.

Pontoglio, M. (2000) Hepatocyte nuclear factor 1, a transcription factor at the crossroads of glucose homeostasis. *J. Am. Soc. Nephrol.*, 11(16), S140-143.

Prasanna Kumar, S., Thippeswamy, G., Sheela, M.L., Prabhakar, B.T., Salimath, B.P. (2008) Butyrate-induced phosphatise regulates VEGF and angiogenesis via Sp1. *Arch. Biochem. Biophys.*, 478(1), 85-95.

Pruitt, K.D., Tatusova, T., Klimke, W., Maglott, D.R. (2009) NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res.*, 3, D32-36.

Ptashne, M., Gann, A. (1997) Transcriptional activation by recruitment. *Nature*, 386, 569-577.

Quandt, K., Frech, K., Karas, H., Wingender, E., Werner, T. (1995) MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res.*, 23, 4878-4884.

Rabinovich, A., Jin, V.X., Rabinovich, R., Xu, X., Farnham, P.J. (2008) E2F in vivo binding specificity: comparison of consensus versus nonconsensus binding sites. *Genome Res*., 18(11), 1763-1777

Rahmann, S., Muller, T., Vingron, M. (2003) On the power of profiles for transcription factor binding site detection. *Stat. Appl. Genet. Mol. Biol.*, 2, Article7.

Ratanamahatana, C., Gunopulos, D. (2002) Scaling up the Naive Bayesian Classifier: Using Decision Trees for Feature Selection. *Proceedings of Workshop on Data Cleaning and Preprocessing*.

Ray, M., Ruan, J., Zhang, W. (2008) Variations in the transcriptome of Alzheimer's disease reveal molecular networks involved in cardiovascular diseases. *Genome Biol*., 9(10), R148.

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

Riccio, A. Aaltonen, L.A., Godwin, A.K., Loukola, A., Percesepe, A., Salovaara, R., Masciullo, V., Genuardi, M., Paravatou-Petsotas, M., Bassi, D.E., Ruggeri, B.A., Klein-Szanto, A.J., Testa, J.R., Neri, G., Bellacosa, A. (1999) The DNA repair gene MBD4 (MED1) is mutated in human carcinomas with microsatellite instability. *Nat. Genet.* 23(3), 266-268.

Rice, P., Longden, I., Bleasby, A. (2000) EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.*, 16(6), 276-7.

Rivals, I., Personnaz, L., Taing, L., Potier, M-C. (2007) Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics*, 23(4), 401-407.

Robertson, G., Bilenky, M., Lin, K., He, A., Yuen, W., Dagpinar, M., Varhol, R., Teague, K., Griffuth, O.L., Zhang, X., Pan, Y., Hassel, M., Sleumer, M.C., Pan, W., Pleasance, E.D., Chuang, M., Hao, H., Li, Y.Y., Robertson, N., Fjell, C., Li, B., Montgomery, S.B., Astakhova, T., Zhou, J., Snader, J., Siddigui, A.S., Jones, S.J. (2006) cisRED: a database system for genome-scale computational discovery of regulatory elements. *Nucleic Acids Res*., 34, D68-D73.

Rodriguez, A., Jung, E.J., Flemington, E.K. (2001) Cell cycle analysis of Epstein-Barr virus-infected cells following treatment with lytic cycle-inducing agents. *J. Virol.*, 75(10), 4482-4489.

Rhode, K., Yates, R.M., Purdy, G.E., Russell. D.G. (2007) Mycobacterium tuberculosis and the environment within the phagosome. *Immunol. Rev.* 219, 37-54.

Roth, F.P., Hughes, J.D., Estep, P.W., Church, G.M. (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.*, 16(10), 939-945.

Rothenburg, E.V., Scripture-Adams, D.D. (2008) Competition and collaboration: GATA-3, PU.1, and Notch signalling in early T-cell fate determination. *Semin. Immunol.*, 20(4), 236-246.

Rozenberg, J.M., Shlyakhtenko, A., Glass, K., Rishi, V., Myakishev, M.V., FitzGerald, P.C., Vinson, C. (2008) All and only CpG containing sequences are enriched in promoters abundantly bound by RNA polymerase II in multiple tissues. *BMC Genomics*, 5(9), 67.

Saha, A., Murakami, M., Kumar, P., Bajaj, B., Sims, K., Robertson, E.S. (2009) Epstein-Barr virus nuclear antigen 3C augments Mdm2-mediated p53 ubiquitination and degredation by deubiquitinating Mdm2. *J. Virol.*, 83(9), 4652-4669.

Salam, M.A., Matin, K., Matsumoto, N., Tsuga, Y., Hanada, N., Senpuku, H. (2004) E2f1 mutation induces early onset of diabetes and Sjögren's syndrome in nonobese diabetic mice. *J. Immunol.*, 173(8), 4908-4918.

Samten, B. Howard, S.T., Weis, S.E., Wu, S., Shams, H., Townsend, J.C., Safi, H., Barnes, P.F. (2005) Cyclic AMP response element-binding protein positively regulates production of IFN-gamma by T cells in response to a microbial pathogen. *J. Immunol.*, 174, 6357-6363.

Samten, B., Townsend, J.C., Weis, S.E., Bhoumik, A., Klucar, P., Shams, H., Barnes, P.F. (2008) CREB, ATF, and AP-1 transcription factors regulate IFN-gamma secretion by human T cells in response to mycobacterial antigen. *J. Immunol.*, 181(3), 2056-2064.

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

Sandelin, A., Wasserman, W.W., Lenhard, B. (2004b) ConSite: web-based prediction of regulatory elements using cross-species comparison. *Nucleic Acids Res.*, 32, W249-W252.

Sandve, G.K., Abul, O., Walseng, V., Drabløs, F. (2007) Improved benchmarks for computational motif discovery. *BMC Bioinformatics.*, 8, 193.

Santalucía, T., Moreno, H., Palacín, M., Yacoub, M.H., Brand, N.J., Zorzano, A. (2001) A novel functional co-operation between MyoD, MEF2 and TRalpha1 is sufficient for the induction of GLUT4 gene transcription. *J. Mol. Biol.*, 314(2), 195-204.

Sarkar, C., Maitra, A. (2008) Deciphering the cis-regulatory elements of co-expressed genes in PCOS by in silico analysis. *Gene*, 408(1-2), 72-84.

Satija, R., Pachter, L., Hein, J. (2008) Combining statistical alignment and phylogenetic footprinting to detect regulatory elements. *Bioinformatics*, 24, 1236-1242.

Schoonjans, F., Zalata, A., Depuydt, C.E., Comhaire, F.H. (1995) MedCalc: a new computer program for medical statistics. *Comput. Methods Programs Biomed.*, 48, 257-262.

Segal, E., Raveh-Sadka, T., Schroeder, M., Unnerstall, U., Gaul, U. (2008) Predicting expression patterns from regulatory sequence in Drosophila segmentation. *Nature*, 451(7178), 535-540.

Shen, F., Hu, Z., Goswami, J., Gaffen, S.L. (2006) Identification of common transcriptional regulatory elements in interleukin-17 target genes. *J. Biol. Chem.*, 281(34), 24138-24148.

Shimizu, R., Yamamoto, M. (2005) Gene expression regulation and domain function of hematopoietic GATA factors. *Semin. Cell Dev. Biol.*, 16(1), 129-136.

Silver, N., Best, S., Jiang, J., Thein, S.L. (2006) Selection of housekeeping genes for gene expression studies in human reticulocytes using real-time PCR. *BMC Mol. Biol.*, 7, 33.

Sinclair, A. (2003) bZIP proteins of human gammaherpesviruses. *J. Gen. Virol.*, 84(8), 1941-1949.

Singh, L.N., Wang, L.S., Hannenhalli, S. (2007) TREMOR--a tool for retrieving transcriptional modules by incorporating motif covariance. *Nucleic Acids Res.*, 35, 7360-7371.

Smedley, D., Haider, S., Ballester, B., Holland, R., London, D., Thorisson, G., Kasprzyk, A. (2009) BioMart – biological queries made easy. *BMC Genomics*, 14(10), 22.

Sobral, M.F., Roberto, C., Navas, D., Palmilha, I, Lima, M.B., Cravador, A. (2009) Identification of decendent of an extinct bovine population from the Algarve region of Portugal using numerical taxonomy analysis of morphological traits. *J. Anim. Breed. Genet.*, 126(4), 319-326.

Soni, V., Cahir-McFarland, E., Kieff, E. (2007) LMP1 TRAFficking activates growth and survival pathways. *Adv. Exp. Med. Biol.*, 597, 173-187.

Sorek, R., Ast, G. (2003) Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Res.*, 13, 1631-1637.

Sotnikov, I., Hershkoviz, R., Grabovsky, V., Ilan, N., Cahalon, L., Vlodavsky, I., Alon, R., Lider O. (2004) Enzymatically quiescent heparanase augments T cell interactions with VCAM-1 and extracellular matrix components under versatile dynamic contexts. *J. Immunol.*, 172(9), 5185-5193.

Speck, S.H., Chatila, T., Flemington, E. (1997) Reactivation of Epstein-Barr virus: regulation and function of the BZLF1 gene. *Trends Microbiol.*, 5, 399-405.

Stegmaier, P., Kel, A.E., Wingender, E. (2004) Systematic DNA-binding domain classification of transcription factors. *Genome Inform.*, 15, 276-286.

Steyvers, M. (2002) Multidimensional Scaling. In: *Encyclopedia of Cognitive Science.* 1st ed. Macmillan Reference Ltd.

Spellman, P.T., Rubin, G.M. (2002) Evidence for large domains of similarity expressed genes in the Drospophila genome. *J. Biol.*, 1(1), 5.

Stipanuk, M.H., Ueki, I., Dominy, J.E.Jr., Simmons, C.R., Hirschberger, L.L. (2009) Cysteine dioxcygenase: a robust system for regulation of cellular cysteine levels. *Amino Acids*, 37(1), 55-63.

Stop TB Partnership and World Health Organization (2006) Global Plan to Stop TB 2006-2015. Available: http://ww.stoptb.org/globalplan/. Last accessed 22 September 2009

Striepe, J., Goessling, E. (unknown date). Patch. Available: http://www.gene-regulation.com/cgi-bin/pub/programs/patch/bin/patch.cgi. Last accessed 19 September 2009.

Struckmann, S., Araúzo-Bravo, M.J., Schöler, H.R., Reinbold, R.A., Fuellen, G. (2008) ReXSpecies—a tool for the analysis of the evolution of gene regulation across species. *BMC Evol. Biol.*, 8, 11.

Su, Y., Wang, T., Sun, Y., Ye, H. (2009) High ISSR variation in 14 surviving individuals of Euryodendron excelsum (Ternstroemiaceae) endemic to China. *Biochem. Genet.*, 47(1-2), 56-65.

Subramanian, C., Knight, J.S., Robertson, E.S. (2002) The Epstein Barr nuclear antigen EBNA3C regulates transcription, cell transformation and cell migration. *Front Biosci.*, 7, d704-716.

Sugawara, H., Iwata, H., Souri, M., Ichinose, A. (2007) Regulation of human protein Z gene expression by liver-enriched transcription factor HNF-4 alpha and ubiquitous factor Sp1. *J. Thomb. Haemost.,* 5(11), 2250-2258.

Suzuki, A., Hirata, M., Kamimura, K., Maniwa, R., Yamanaka, T., Mizuno, K., Kishikawa, M., Hirose, H., Amano, Y., Izumi, N., Miwa, Y., Ohno, S. (2004) aPKC acts upstream of PAR-1b in both the establishment and maintenance of mammalian epithelial polarity. *Curr. Biol*., 14(16), 1425-1435.

Takada, K., Shimizu, N., Sakuma, S., Ono, Y. (1986) Trans activation of the latent Epstein-Barr virus (EBV) genome after transfection of the EBV DNA fragment. *J. Virol.*, 57, 1016-1022.

Takahashi, S., Matsuura, N., Kurokawa, T., Takahashi, Y., Miura, T. (2002) Co-operation of the transcription factor hepatocyte nuclear factor-4 with Sp1 or Sp3 leads to transcriptional activation of the human haem oxygenase-1 gene promoter in a hepatoma cell line. *Biochem. J*., 367(3), 641-652.

Tan, Y., Wu, C., de Veyra, T., Greer, P.A. (2006) Ubiquitous calpains promote both apoptosis and survival signals in response to different cell death stimuli. *J. Biol. Chem.*, 281, 17689-17698.

Tao, Y., Sam, L., Friedman, C., Lussier, Y.A. (2007) Information theory applied to the sparse gene ontology annotation network to predict novel gene function. *Bioinformatics*, 23(13), i529-i538.

Tavazoie, S. Hughes, J.D., Campbell. M.J., Cho, R.J., Church, G.M. (1999) Systematic determination of genetic network architecture. *Nat. Genet.*, 22(3), 281-285.

Thoma-Uszynski, S., Stenger, S., Takeuchi, I., Ochoa, M.T., Engele, M., Sieling, P.A., Barnes, P.F., Rollinghoff, M., Bolcskei, P.L., Wagner, M., Akira, S., Norgard, M.V., Belisle, J.T., Godowski, P.J., Bloom, B.R., Modlin, R.L. (2001) Induction of direct antimicrobial activity through mammalian toll-like receptors. *Science*, 291 (5508), 1544-1547.

Thorley-Lawson, D.A., Allday, M.J. (2008) The curious case of the tumour virus: 50 years of Burkitt's lymphoma. *Nat. Rev. Microbiol., 6*, 913-924.

Tkachuk, D.C., Kohler, S., Cleary, M.L. (1992) Involvement of a homolog of Drosophila trithorax by 11q23 chromosomal translocations in acute leukemias. *Cell*, 71(4), 591-700.

Tomoda, K., Kato, J.Y., Tatsumi, E., Takahashi, T., Matsuo, Y., Yoneda-Kato, N. (2005) The Jab1/COP9 signalosome subcomplex is a downstream mediator of Bcr-Abl kinase activity and facilitates cell-cycle progression. *Blood*, 105(2), 775-783.

Turner, R.B., Smith, D.L., Zawrotny, M.E., Summers, M.F., Posewitz, M.C., Winge, D.R. (1998) Solution structure of a zinc domain conserved in yeast copper-regulated transcription factors. *Nat. Struct. Biol*., 5, 551-555.

Tronche, F., Yaniv, M. (1992) HNF1, a homeoprotein member of the hepatic transcription regulatory network. *Bioessays*, 14(9), 579-587.

Tsai, H.K., Chou, M.Y., Shih, C.H., Huang, G.T., Chang, T.H., Li, W.H. (2007) MYBS: a comprehensive web server for mining transcription factor binding sites in yeast. *Nucleic Acids Res.*, 35, W221-W226.

Tsuruga, T., Nakugawa, S., Watanabe, M., Takizawa, S., Matsumoto, Y., Nagasaka, K., Sone, K., Hiraike, H., Miyamoto, Y., Hiraike, O., Minaguchi, T., Oda, K., Yasugi, T., Yano, T., Taketani, Y. (2007) Loss of Hughl-1 expression associated with lymph node metastasis in endometrial cancer. *Oncol. Res.*, 16(9), 431-435.

Tu, S.P., Chi, A.L., Ai, W., Takaishi, S., Dubeykovskaya, Z., Quante, M., Fox, J.G., Wang, T.C. (2009) p53 inhibition of AP1-dependent TFF2 expression induces apoptosis and inhibits cell migration in gastric cancer. *Am. J. Physiol. Gastrointest. Liver Physiol*., 297(2), G385-396.

Tuteja, G., Jensen, S.T., White, P., Kaestner, K.H. (2008) Cis-regulatory modules in the mammalian liver: composition depends on strength of Foxa2 consensus site. *Nucleic Acids Res.*, 36(12), 4149-4157.

Uemura, N., Kajino, T., Sajo, H., Sato, S., Akira, S., Matsumoto, K., Ninomiya-Tsuji, J. (2006) TAK1 is a component of the Epstein-Barr virus LMP1 complex and is essential for activation of JNK but not of NF-kappaB. *J. Biol. Chem.,* 281(12), 7863-7872.

VanGuilder, H.D., Vrana, K.E., Freeman, W.M. (2008) Twenty-five years of quantitative PCR for gene expression analysis. *Biotechniques*, 44(5), 619-626.

Van Loo, P., Marynen, P. (2009) Computational methods for the detection of cis-regulatory modules. *Brief Bioinform.*, 10(5), 509-524.

Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A., Luscombe, N.M. (2009) A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.*, 10(4), 252-263.

Veerla, S., Höglund, M. (2006) Analysis of promoter regions of co-expressed genes identifies by microarray analysis. *BMC Bioinformatics*, 7, 384.

Vergne, I., Singh, S., Roberts, E., Kyei, G., Master, S., Harris, J., de Haro, S., Naylor, J., Davis, A., Delgado, M., Deretic, V. (2006) Autophagy in immune defense against Mycobacterium tuberculosis. *Autophagy*, 2(3), 175-178.

Vermeer, M.H., van Doorn, R., Dijkman, R., Mao, X., Whittaker, S., van Voorst Vader, P.C., Gerritsen, M.J., Geerts, M.L., Gellrich, S., Söderberg, O., Leuchowius, K.J., Landegren, U., Out-Luiting, J.J., Knijnenburg, J., Ijszenga, M., Szuhai, K., Willemze, R., Tensen, C.P. (2008) Novel and highly recurrent chromosomal alterations in Sézary syndrome. *Cancer Res.*, 68(8), 2689-2698.

Vig, M., Kinet, J.P. (2009) Calcium signalling in immune cells. *Nat. Immunol.*, 10(1), 21-27.

Vilar, S., Karpiak, J., Constanzi, S. (2009) Ligand and structure-based models for the prediction of ligand-receptor affinities and virtual screenings: Development and application to the beta(2)-adrenergic receptor. *J Comput Chem.*, Epub ahead of print.

Visscher, P.M., Hopper, J.L. (2001) Power of regression and maximum likelihood methods to map QTL from sib-pair and DZ twin data. *Ann. Hum. Genet.*, 65, 583-601.

Vlieghe, D., Sandelin, A., de Bleser, P.J., Vleminckx, K., Wasserman, W.W., van Roy, F., Lenhard, B. (2006) A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res.*, 34, D95-D97.

Wang, J., Barnes, R.O., West, N.R., Olson, M., Chu, J.E., Watson, P.H. (2008) Jab1 is a target of EGFR signalling in ERalpha-negative breast cancer. *Breast Cancer Res.*, 10(3), R51.

Wang, N., Lin, K.K., Lu, Z., Lam, K.S., Newton, R., Xu, X., Yu, Z., Gill, G.N., Andersen, B. (2007) The Lim-only factor LMO4 regulated expression of the BMP7 gene through an HDAC2-dependent mechanism, and controls cell proliferation and apoptosis of mammary epithelial cells. *Oncogene*, 26(44), 6431-41.

Wang, T., Kobayashi, T., Takimoto, R., Denes, A.E., Snyder, E.L., el-Deiry, W.S., Brachmann, R.K. (2001) hADA3 is required for p53 activity. *EMBO J.*, 20(22), 6404-6413.

Wang, X., Hu, J., Zhang, M.O., Li, Y. (2008) Identification of phylogenetically conserved microRNA cis-regulatory elements across 12 Drosophila species. *Bioinformatics*, 24, 165-171.

Wasserman, W.W., Sandelin, A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.*, 5, 276-287.

Weigelt, K., Lichtinger, M., Rehli, M., Langmann, T. (2009) Transcriptomic profiling identifies a PU.1 regulatory network in macrophages. *Biochem. Biophys. Res. Commun.*, 380(2), 308-312.

Westholm, J.O., Xu, F., Ronne, H., Komorowski, J. (2008) Genome-scale study of the importance of binding site context for transcription factor binding and gene regulation. *BMC. Bioinformatics*, 9, 484.

Witten, I.H., Frank, E., Trigg, L., Hall, M., Holmes, G., Cunnigham, S.J. (1999) Weka: Practical Machine Learning Tools and Techniques with Java Implementations. *Proc ICONIP/ ANZIIS /ANNES99 Future Directions for Intelligent systems and Information Sciences*, 192-196.

Wolfsberg, T.G., Gabrielian, A.E., Campbell, M.J., Cho, R.J., Spounge, J.L., Landsman, D. (1999) Candidate regulatory sequence elements for cell cycle-dependent transcription in Saccharomyces cerevisiae. *Genome Res.*, 9, 775-792.

World Health Organisation (2009) Global tuberculosis control – epidemiology, strategy, financing. Available: http://www.who.int/tb/publications/global_report/en/. Last accessed 22 September 2009.

World Health Organisation (2008) Viral Cancers, Epstein-Barr virus. Available: http://www.who.int/vaccine_research/diseases/viral_cancers/en/index1.html. Last accessed 22 September.

Wright, N.J., Hesseling, P.B., McCormick, P., Tchintseme, F. (2009) The incidence, clustering and characteristics of Burkitt lymphoma in the Northwest province of Cameroon. *Trop. Doct.*, Epub ahead of print.

Wu, X., Zhu, L., Guo, J., Zhang, D-Y., Lin, K. (2006) Prediction of yeast protein-protein interaction network: insights from the Gene Ontology and annotations. *Nucleic Acids Res.*, 34(7), 2137-2150.

Xia, C., Cheshire, J.K., Patel, H., Woo, P. (1997) Cross-talk between transcription factors NF-kappa B and C/EBP in the transcriptional regulation of genes. *Int. J. Biochem. Cell Biol.*, 29, 1525-1529

Xia, Y., Jiang, B., Zou, T., Gao, G., Shang, L., Chen, B., Liu, O., Gong, Y. (2008) Sp1 and CREB regulate basal transcription of the human SNF2L gene. *Biochem. Biophys. Res. Commun.*, 368(2), 438-444.

Yanashima, R., Kitagawa, N., Matsubara, Y., Weatheritt, R., Oka, K., Kikuchi, S., Tomita, M., Ishizaki, S. (2009) Network features and pathway analysis of a signal transduction cascade. *Front Neuroinformatics*, 3, 13.

Yang, M., Yordanov, B., Levy, Y., Brüschweiler, R., Huo, S. (2006) The sequence-dependent unfolding pathways play a critical role in the amyloidogenicity of transthyretin. *Biochemistry*, 45(39), 11992-12002.

Yang, X.J., Ogryzko, V.V., Nishikawa, J., Howard, B.H., Nakatani, Y. (1996) A p300/CBP-associated factor that competes with the adenoviral oncoprotein E1A. *Nature*, 382(6589), 319-324.

Yang, Y., Tin, C., Pandev, A., Abbott, D., Sassetti, C., Kelliher, M.A. (2007) NOD2 pathway activation by MDP or Mycobaterium tuberculosis infection involves the stable polyubiquitination of Rip2. *J.Biol. Chem.*, 282 (50), 36223-9.

Yi, F., Saha, A., Murakami, M., Kumar, P., Knight, J.S., Cai, Q., Choudhuri, T., Robertson, E.S. (2009) Epstein-Barr virus nuclear antigen 3C targets p53 and modulates its transcriptional apoptotic activities. *Virology*, 388(2), 236-247.

Young, L.S. (1993) Mycobacterial diseases and the compromised host. *Clin. Infect. Dis.*, 17(2), S436-S441.

Young, L.S., Rickinson, A.B. (2004) Epstein-Barr virus: 40 years on. *Nat. Rev. Cancer, 4*, 757-768.

Yu, Q., Sharma, A., Oh., S.Y., Moon, H.G., Hossain, M.Z., Salay, T.M., Leeds, K.E., Du, H., Wu, B., Waterman, M.L., Zhu, Z., Sen, J.M. (2009) T cell factor 1 initiates the T helper type 2 fate by inducing the transcription factor GATA-3 and repressing interferon-gamma. *Nat. Immunol.*, 10(9), 992-999.

Zadissa, A., McEwan, J.C., Brown, C.M. (2007) Inference of transcriptional regulation using gene expression data from the bovine and human genomes. *BMC Genomics*, 8, 265.

Zeng, L., Wu,J., Xie,J. (2008) Statistical methods in integrative analysis for gene regulatory modules. *Stat. Appl. Genet. Mol. Biol.*, 7(1), Article 28.

Zhang, H. (2004) The Optimality of Naive Bayes. In: *FLAIRS Conference*. AAAI Press

Zhang, L.Y., Marchand, S., Tinker, N.A., Belizile, F. (2009) Population structure and linkage disequilibrium in barley assessed by DArT markers. *Theor. Appl. Genet.,* 119(1), 43-52.

Zhang, L., Sheppard, O.R., Shah, A.M., Brewer, A.C. (2008) Positive regulation of the NADPH oxidase NOX4 promoter in vascular smooth muscle cells by E2F. *Free Radic. Biol. Med.*, 45(5), 679-85.

Zheng, J., Wu, J., Sun, Z. (2003) An approach to identify over-represented cis-elements in related sequences. *Nucleic Acids Res.*, 31(7), 1995-2005.

Zhu, J., He, F., Hu, S., Tu, J. (2008) On the nature of human housekeeping genes. *Trends Genet.,* 24(10), 481-484.

Zhu, M., Ghodsi, A. (2005) Automatic dimensionality selection from the screeplot via the use of profile likelihood. *Comp. Stat. & Data Anal.*, 51(2), 918-930.

Zhou, J.X., Lee, C.H., Qi, C.F., Wang, H., Naghashfar, Z., Abbasi, S., Morse, H.C.3rd. (2009) IFN regulatory factor 8 regulates MDM2 in germinal center B cells. *J. Immunol.*, 183(5), 3188-3194.

Zhou, J., Chau, C.M., Deng, Z., Shiekhattar, R., Spindler, M.P., Schepers, A., Lieberman, P.M. (2005) Cell cycle regulation of chromatin at an origin of DNA replication. *EMBO J.*, 24 (7), 1406-1417.

Zhu, M., Ghodsi, A. (2006) Automatic dimensionality selection from the screeplot via the ue of profile likelihood. *Comp. Stat. Data Anal.,* 51, 918-930.