CHALMERS | UNIVERSITY OF GOTHENBURG

Analysis of bidirectional promoters in vertebrates

Master of Science Thesis in Bioinformatics and Systems Biology

TABASSUM FARZANA JAHAN

Supervisor and Examiner: Prof. Tore Samuelsson

Co-Supervisor: Marcela Davila Lopez

Chalmers University of Technology
Gothenburg
Sweden

August 2012

Table of Contents

Abstract

The order of genes in eukaryotes is by and large random as a result of recombination events during evolution. However, there is a certain element of non-random gene order. For instance, genes of similar expression tend to cluster more commonly than by chance and functionally related genes tend to colocalize. Genome wide analyses of mammalian genomes have demonstrated an abundance of divergently transcribed genes in short intergenic regions of approximately 1000 bp. This means that the genes of such pairs have transcription start sites in close proximity. The gene pairs are thought to share an intervening regulatory sequence, a *bi-directional promoter*. There is evidence that bidirectional gene pairs are evolutionarily conserved and this may imply a functional significance. They are often associated with genes involved in DNA repair. Interestingly, expression profiles of ovarian and breast cancer show an enrichment of bidirectional gene pairs that include DNA repair genes, such as BRCA1, BRCA2, CKEK1 and FANC family members. The two genes of a bidirectional promoter are likely to be related in terms of transcriptional control. Therefore, through analyses of such gene pairs in eukaryotes we may obtain important information regarding transcriptional control mechanisms.

In this project, intergenic regions of bidirectional gene pairs were explored by sequence analysis. The aim was to examine whether promoters of such pairs have characteristics that are different from the promoter regions of other genes. A number of such pairs were therefore collected from a set of mammalian species, including human. Then these regions were analyzed in a profile based approach with respect to known transcription factor binding sites (TFBSs) and with respect to the TATA box, one of the core promoter elements. Furthermore, in a more unbiased approach MEME was used to identify motifs characteristic of bidirectional promoters. The results reveal a number of over-represented TFBSs as well as motifs identified by MEME. The overlap of these two datasets reveals previously identified TFBSs as well as motifs of potential biological interest.

Acknowledgements

I would like to express my deepest gratitude to my supervisor and examiner Tore Samuelsson for assisting me in developing and designing this thesis. I thank him for being understanding and patient during the entire span of time. Tons of thanks to my co-supervisor Marcela Davila Lopez, for helping me with the technical aspects and for simplifying the complex issues in this project.

I would also like to honor the MPBIS programme in Chalmers University of Technology for providing me with all the knowledge required to successfully complete a project. It has been a tough journey but I am glad and thankful to all my friends who have always been there for me. And lastly, thanks to my parents and family for supporting me with their incessant love and encouragement no matter how difficult I was to keep up with.

1 Introduction

Genes are defined as the biological entities responsible for traits of an organism encoded in the DNA (Noble 2008). Expression of a gene typically involves the transcription from DNA to RNA and translation from RNA to proteins. Regulation of gene expression may occur at different levels in the flow of genetic information; transcription, splicing or translation. Our focus in this thesis is on the transcriptional regulation of gene expression where malfunction can lead to various diseases in human such as Asthma (Burchard, Silverman et al. 1999), Beta thalassemia (Kulozik, Bellan-Koch et al. 1991), Rubinstein-Taybi syndrome (Petrij, Giles et al. 1995) as well as various cancer types (Vlahopoulos, Logotheti et al. 2008). More specifically, transcriptional regulation of bidirectional genes which cover ~10% of all human genes will be elaborated for eight mammalian species (Trinklein, Aldred et al. 2004).

Various elements and steps of the transcriptional machinery in eukaryotes are explained below.

1.1 Transcriptional Machinery in Eukaryotes

Although they both lead to a specific RNA product, prokaryotic and eukaryotic transcription is distinct from each other. Our focus of attention here will be on eukaryotic transcription. Through a linear cascade of events, the eukaryotic transcriptional machinery involves the decondensation of a locus on the chromatin form of DNA, rearrangement of the nucleosome complex, modification of histone proteins, binding of transcriptional activators and co-activators to enhancers and promoters and finally the incorporation of the basal transcriptional initiation complex to the core promoter (Kornberg 2001). A promoter is a region of DNA near genes, located upstream of a particular gene that facilitates the initiation of transcription. The core promoter is the minimal portion of a promoter region and it accommodates a transcription start site (TSS), an RNA polymerase binding site and a general transcription factor binding site such as the TATA box (Butler and Kadonaga 2002). The basal transcriptional initiation complex includes a TATA box as an essential core promoter element in 10-20% of all human genes (Gershenzon and Ioshikhes 2005). A transcription pre-initiation complex forms through the sequential assembly onto a TATA-dependent core promoter region of the polymerase as such in the respective order of following components: TFIID/TFIIA, TFIIB, RNA polymerase II/TFIIF and TFIIH (Kornberg 2007).

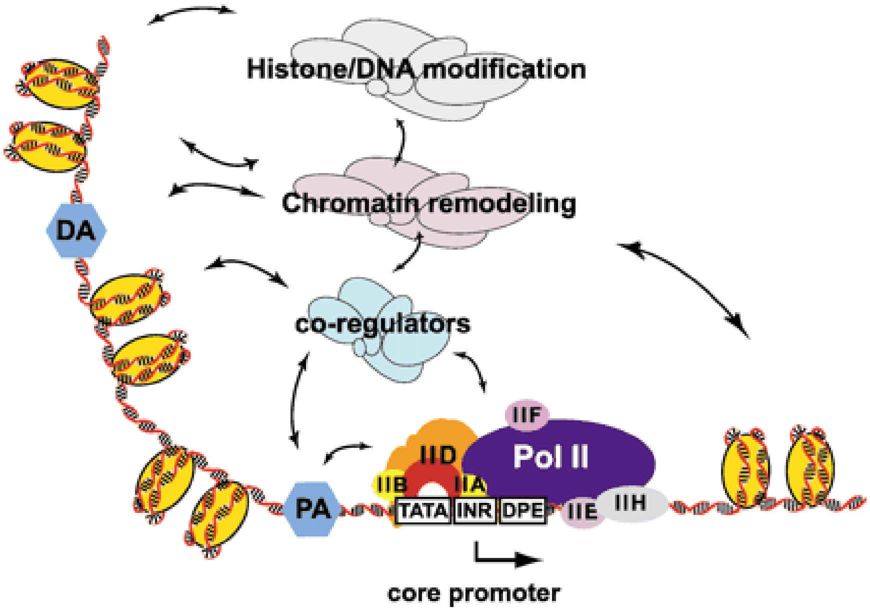


Figure 1 Transcription Machinery. The transcription apparatus is an ensemble of multilayered subunits. This includes covalent modification of Histone/DNA, chromatin remodeling which prepares the DNA template for transcription factor binding. Core promoter elements direct the formation of pre.initiation complex and defines the transcription start site(Hochheimer and Tjian 2003).Ttranscription pre-initiation complex forms through the sequential assembly onto a TATA-dependent core promoter region of the polymerase as such in the respective order of following components: TFIID/TFIIA, TFIIB, RNA polymerase II/TFIIF and TFIIH (Kornberg 2007).

1.2 Core promoter motifs

Figure 2 illustrates some of the sequence elements that can contribute to basal transcription from a core promoter. Only a subset of core promoters contains each of these individual sequence motifs. Not all core promoters contain all the sequence elements. For instance, the TATA box can function in the absence of BRE, INR, and DPE motifs whereas the DPE motif can only proceed as a pair with an INR. Moreover, the BRE is usually located in the upstream site of a subset of TATA box motifs (Smale and Kadonaga 2003).

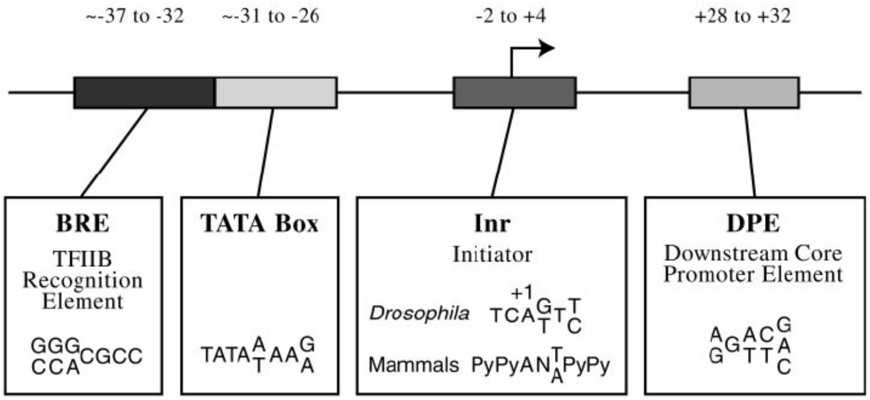


Figure 2 Core promoter elements. The figure shows some of the core promoter motifs that can contribute to basal transcription from a core promoter(Smale and Kadonaga 2003)

As indicated above promoters are structurally and functionally diverse (Smale 1998). In addition the different elements of promoters are essential in combinatorial regulation of gene expression (Butler 2002).

1.3 Bidirectional gene promoters and their characteristics

Gene order is not entirely random in eukaryotes and this observation may be related to the control of gene expression. For instance, clustering of genes from the same metabolic pathway may be one means of regulating gene expression (Lee and Sonnhammer 2003). Another example where we see a conserved ordering is as a consequence of gene duplication events giving rise to paralogous genes. In mammalian genome we frequently observe gene pairs with a short intergenic distance and where the genes are divergently transcribed. (Adachi and Lieber 2002; Yang, Koehly et al. 2007; Yang, Taylor et al. 2008).

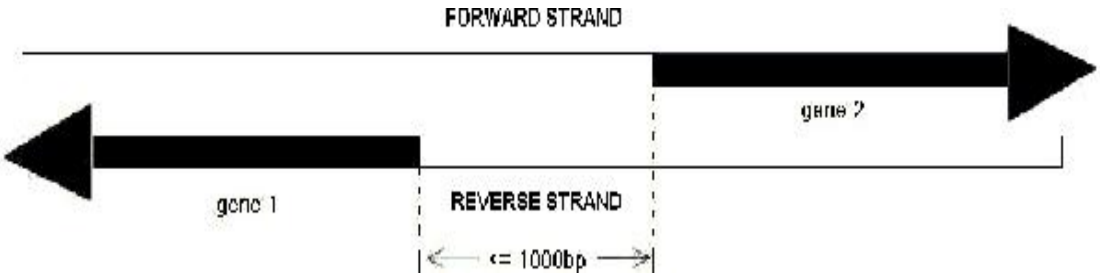


Figure 3 Sketch of Bidirectional promoter. Tw gene is head to head orientation. Gene 1 in the reverse strand and gene 2 in the forward strand. The distance between their Transcriptions start sites also referred as intergenic distance being less than equal to 1k bp. The two gene pair is called bidirectional genes and the intergenic region is bidirectional promoter (Wang, Wan et al. 2009).

Gene pairs with an intergenic distance less than 1000 base pairs which are divergently transcribed we define here as bidirectional genes and they are assumed to be sharing a promoter region called a bidirectional gene promoter (Adachi and Lieber 2002). Bidirectional gene pairs often encode two different peptide subunits that share similar structure and function as in the example of collagen (Burbelo, Martin et al. 1988). In addition, as in the case of the TAP1/LMP2 genes, the gene products can be involved in the same cellular pathway (Wright, White et al. 1995).

DNA repair related functions in a mammalian cell often involve bidirectional gene pairs and thus a potential relationship between these gene pairs and cancer has been hypothesized. For example, expression profiles of ovarian and breast cancers have revealed an enrichment of bidirectional gene pairs that include DNA repair genes, such as BRCA1, BRCA2, CKEK1 and FANC family members (Kleinjan and Lettice 2008).

There are different modes of regulation of bidirectional genes. Thus, they can be coexpressed (Trinklein, Aldred et al. 2004) or anti-regulated where the expression of one gene inversely affect the other one (Ame, Schreiber et al. 2001; Agirre, Roman-Gomez et al. 2006). In addition, the regulation may be exerted at the level of DNA methylation, for instance as in the case of CpG island regions which are shown to silence bidirectional gene expression in different cancer types (Shu, Jelinek et al. 2006). In a recent study by Yang and Elnitski activity of core promoter elements in a more extended set of bidirectional promoters was studied. They identified a high frequency of CpG islands, whereas the TATA boxes were under-represented. Interestingly, the other core elements DPE and INR were not enriched in the data set. A TFIIB recognition element known as BRE was somewhat enriched in bidirectional promoters and the CCAAT box was found to be almost 2 fold enriched(Yang and Elnitski 2008). Lin et al. combined computational analysis with meta-analysis of ChIP-chip experiments, and identified a number of over-represented binding sites including those of MYC, E2F1, E2F4, SP1, SP3 and STAT1 indentified from analysisis of ChIP-chip data. And from computational study, binding sites for NRF-1, CCAAT boxes (similar to NF-Y), YY1 and GA binding protein A (GABPA) were identified (Lin, Collins et al. 2007).

In studies that are consistent with Yang et al, bidirectional promoters show characteristics associated with more active promoters. For instance, they show a higher density of Pol II binding, increased H3 acetylation and increased occupancy of modified histones H3K4me2 and H3K4me3 (Lin, Collins et al. 2007). On the other hand, histone H4 acetylation was under-represented in bidirectional promoters (Wakano, Byun et al. 2012).

The recent studies as described above highlight the regulatory importance of bidirectional promoters and also show how their configuration may take part in diverse mechanisms of transcriptional control. Thus, a closer look into the sequence and structure of these promoters may broaden the existing knowledge on how they contribute in regulating these unique set of genes.

In this thesis our aim was to examine if the promoter region of bidirectional gene have characteristics that are different than other promoters. We used two approaches. First, the human bidirectional promoter region was analyzed with a profile based approach to see if there is any over-representation of transcription factor binding sites (TFBSs). Secondly, we examined homologs of human bidirectional genes in seven other vertebrates to identify conserved motifs using the motif-finding software MEME.

2. Materials and Methods

2.1 Identification of bidirectional gene pair

The genomic sequences of all the chromosomes of the eight species used in this project have been downloaded from the FTP site of Ensembl (www.ensembl.org/info/data/ftp/index.html), release 64 January 2011, in FASTA format. The species include, *Homo sapiens* (Human), *Bos taurus*

(Cow), *Canis familiaris* (Dog), *Loxodonta africana* (Elephant), *Mus musculus* (Mouse), *Pteropus vampyrus* (Mega bat), *Sus scrofa* (Pig), *Tursiops truncates*(Dolphin). The FASTA files contain unmasked genomic sequence for each chromosome in separate FASTA files, and the header contains either the name of the chromosome, scaffold or contig, as well as the location. The protein sequences are also downloaded from the same FTP site as FASTA files which contain all protein translations resulting from Ensembl that are known or novel gene predictions. The header in the files contains a unique Ensembl protein id, chromosome name, position, Swissprot accession followed by the peptide sequence. These protein sequences are then mapped to the genome of each species using correlation data tables (map files). The map files can be customized and downloaded using a data mining tool called BioMart, a support provided in the same website (www.ensembl.org/biomart/martview).

The map files were queried according to Ensembl protein id, description, chromosome name, gene start, gene end and strand. The map files and the peptide files are mapped according to the unique Ensembl id and user given local id. Then the bidirectional or divergent, convergent, and co-directional genes are identified using the chromosome, position, and strand information. The divergent genes are identified with the sign '← →', convergent genes '→←' and co-directional either '←←' or '→→' respectively (Davila Lopez, Martinez Guerra et al. 2010) .

2.2 Extracting bidirectional genes with 1kbp intergenic region

Having transcription direction identified for all protein coding genes in all the species, the bidirectional pairs, ('← →') are extracted according to this information. The size of the intergenic region is identified by subtracting the gene start location of the gene in the reverse strand from the start location of the gene residing in the forward strand, keeping in mind that they are both from the same chromosome. Those pairs that have 1000 bp or less intergenic region are extracted, while overlapping genes are excluded. The human bidirectional promoter sequences are then extracted and saved separately as input set for transcription factor binding site (TFBS) analyses as explained in section 2.7.

2.3 Predicting orthologous genes among all the species

Homologous genes are divided into two groups, orthologs and paralogs, according to common ancestry (Jensen 2001). Orthologs, homologs separated by speciation event, are crucial to this study. The OrthoMCL algorithm Version 2.0 has been used to predict orthologous proteins of the eight mammalian species of interest. The algorithm first interlinks the related proteins in a similarity graph. The graph is based on the output of All vs all BLAST (Basic Local Alignment Search Tool). OrthoMCL then uses a Markov Model Cluster algorithm to categorize the potential orthologs, co-orthologs and inparalogs (Li, Stoeckert et al. 2003). The program was fed with a set of proteomes (of the eight mammals used in project) as input to find the orthologous proteins. The program produces two outputs. One is the pair wise relationship between the proteins and the potential orthologs and paralogs along with their similarity scores. The other is produced by the Mcl program and lists the clusters of the orthologous genes (Li, Stoeckert et al. 2003). From this result we extracted the OrthoMCL clusters of the genes being part of the bidirectional gene pairs using a perl script.

2.4 Algorithm to identify orthologous bidirectional gene pairs and extract their sequence

Perl

Perl is a widely used interpreted scripting language in bioinformatics as well in other areas such as networking, graphic programming etc. It was originally developed for Unix systems programming (Jamison 2003).

In combination with shell scripts, Perl has been widely used in this project. Below is the algorithm programmed with Perl that was used to extract orthologous bidirectional gene pairs.

In OrthoMCL all genes in a genome are assigned an OrthoMCL cluster id. This means that all bidirectional genes will have a specific cluster id, and the homologous genes in other species have the same id. We firstly extracted a pair of bidirectional human gene eg ←Ortho001 and Ortho004→ (arrows indicate the bidirectionality of the genes). Next we identified a pair of bidirectional genes from any other species that have equivalent cluster ids (See figure 4). Finally, we extracted all the orthologous pairs from all the eight species.

Human bidirectional pair:	Equivalent pair from any vertebrate:
←Ortho001 and Orth004→	←Ortho001 and Otho004→

Figure 4 An example of a human bidirectional gene pair with its assigned OrthoMcl Id, and its equivalent homologous bidirectional pair from another species having the same pair of ids.

2.5 Analyzing base counts

We calculated the base content (number of A, G, T, C nucleotides) of all promoter sequences; i.e, the percentage of each nucleotide over the total number of nucleotides, using a simple Perl script. The result is discussed in section 3.5.

2.6 ClustalW alignment and extracting conserved regions

ClustalW is a multiple alignment program that is used to find out the best possible sequence match between two or more sequences using gap penalties, sequence weighting and weight matrix choices (Chenna, Sugawara et al. 2003). We did a multiple alignment on the output of Section 2.4. That means we aligned the promoter regions corresponding to a specific pair of OrthoMCL clusters, and then remove the sections having gaps. The result is conserved regions from the alignment to be used as input for MEME (motif identification tool section 2.8).

2.7 Prediction of transcriptional factor binding sites in human bidirectional promoter regions

JASPER CORE Database

The Database contains transcription factor binding sites modeled as weight matrices. JASPER CORE is an open source database that contains curated and non-redundant set of profiles collected from published scientific papers (Bryne, Valen et al. 2008). Being non-redundant, one of the goals of the database is to have the best model for a specific factor and thus there are not many models for a single factor (Sandelin, Alkema et al. 2004).

To computationally predict possible binding sites, a set of 130 profiles for vertebrates representing TFBS were downloaded from the JASPER CORE database. (http://jaspar.genereg.net/html/DOWNLOAD/jaspar_CORE/non_redundant/all_species/matrix_only/). The profiles were downloaded as position weight matrices (PWM). An example profile for TATA specific binding site is:

```
>MA0108.2 TBP
A  [ 61  16 352   3 354 268 360 222 155  56  83  82  82  68  77 ]
C  [145  46   0  10   0   0   3   2  44 135 147 127 118 107 101 ]
G  [152  18   2   2   5   0  20  44 157 150 128 128 128 139 140 ]
T  [ 31 309  35 374  30 121   6 121  33  48  31  52  61  75  71 ]
```

The promoter sequences that we had extracted earlier (See section 2.2) are scored according to each of the matrices of type shown above. If the score meets a certain threshold score the location and the sequence is saved as potential binding site. We have only analyzed the human bidirectional promoters and have compared the outcome to a collection of human co-directional genes.

2.7.1 Algorithm

The idea is to find all matches to the matrices of length n that pass a certain threshold.

Each of the PWMs is first saved in different files, so we have 130 files. Each of the counts in the matrices is increased with 1 to avoid the occurrence of zero values (i.e. adding pseudocounts).

The sequences are scored window-wise where each window length n is the matrix length .For the above matrix n= 15.

We also define a background matrix (called back) giving the background frequency of each nucleotide which is defined to be 25% for each nucleotide.

A sliding window of length n calculates the score moving one base in each loop.

$$\text{Score} = \sum_{j}^{n} \square \log\left(\frac{\text{Matrix j Seqj}}{\text{backj Seq j}}\right)$$

Then log odds score is calculated, normalizing by the length of the matrix.

$$\text{Log}_{\text{odd}} = \frac{\text{Score}}{n}$$

A maximum score is calculated max_score= Max(Log_{odd})

A minimum score is calculated min_score= Min (Log_{odd})

These two are then used to calculate the final raw Log_{odd} score so that the range is between 0 and 1.

$$\text{Raw}_{\downarrow}\text{Score} = (\text{Log}_{\downarrow}\text{odd} - \text{min}_{\downarrow}\text{score})/(\text{Log}_{\downarrow}\text{odd} - \text{max}_{\downarrow}\text{score})$$

At this point Z score is calculated to be used as a cut off for each binding site hit to a certain matrix M.

To calculate the Z_{Score} the formula is:

$$\frac{Z_{\text{Score}}\text{M} = \left(\text{Raw}_{(\text{score})}\right) - \text{meanM}}{\text{stdM}}$$

The mean and the standard deviation is calculated from randomized human promoter sequences, scoring it according to the same procedure; then finding the mean and standard deviation of the scores derived from each window which is specific for each profile. Hence the Z_{Score} is interpreted as the number of standard deviations above the mean raw score of a certain binding matrix across randomized sequences of human bidirectional promoter regions. Then a Z_{Score} cutoff of 2.33 is used and that corresponds to a p-value of 0.01 (Weirauch and Raney 2007)

Next, all the hits are recorded, including profile name, gene ID, chromosome, window, start site, end site, log odd score, raw score, sequence and finally the Z_{Score} .

As we scored intergenic regions of bidirectional genes, both strands of the DNA were scored.

2.7.2 Mann-Whitney U test to identify over-represented sites

Mann-Whitney U test is a non parametric test which is used when the distribution of the variables does not follow a normal distribution or the sample size is too small to predict the distribution (Fay and Proschan 2010). In our case the hypothesis test is carried out to compare the Z scores of TFBSs of bidirectional and co-directional promoters.

The null and alternative hypotheses considered for the test are as below:

H_0 : The Z scores of bidirectional promoters are not significantly different from the Z score of co-directional promoters.

H_A : The Z scores are significantly different.

To avoid Type I error (i.e. to reject null hypothesis when it is true) after the multiple hypothesis testing, we performed Bonferroni correction and identified the significantly different motifs with p value less than 0.01 (Fay and Proschan 2010). See table 3 in the Result section. We made a boxplot (see Fig. 13 in the Appendix section) of the significantly differing Z scores presenting a statistical summary of scores from the two data sets. The plot shows the differences between the Z scores of the over-represented TFBSs in bidirectional promoters that have a higher mean value than that of the co-directional ones.

2.8 Identification of motifs using MEME

MEME

MEME or Multiple EM for Motif Elicitation is a tool to search for motif in a group of related DNA or protein sequences. Its algorithm uses a number of functions including Expectation maximization (EM) based on heuristics in order to choose EM start point, it also uses maximum likelihood ratio, and greedy search technique to find multiple motifs (Bailey, Williams et al. 2006) . The promoter region of orthologous bidirectional gene pairs was used as input for this analysis.

Below is an example command line, illustrating the parameters used.

```
/meme_4.6.1/bin/meme AllHomologous_genes.fna -mod zoops -dna -minw 5 -maxw 8 -maxsize 350000 -nmotifs 10 -maxiter 90
```

Parameters are:

-mod used to describe the distribution of motifs across the database. We used zoops which is zero or one occurrence per sequence.

-dna used to specify that the input sequences are DNA sequences.

-minw is the minimum motif width set to 5.

-maxw is the maximum motif width set to 8.

-maxsize is the maximum data set size which is set to 350000 characters.

-maxiter is the number of iteration the EM algorithm will run until it converges.

2.9 Mapping identified motifs to JASPER CORE Database profiles using TOMTOM

The output of Meme was a set of ten significant motifs. We mapped those ten motifs against known transcriptional factor binding sites using a tool called TOMTOM. This is a tool that is used to compare an input motif with a database of known motifs. The tool compares the motifs and produces a table ordered according to a q value which means the minimum number of false matches among the motifs (Bailey, Williams et al. 2006). Table 5 shows the motifs mapped to JASPER profile Ids which we later annotated from JASPER Core documentation (See table 6 in Appendix).

3. Results

The aim of this work was to identify sequence elements that are characteristic of bidirectional promoters. The data analyzed here is a set of promoters from human as well as from a number of other mammals; *Bos taurus*, *Canis familiaris*, *Loxodonta africana*, *Mus musculus*, *Pteropus vampyrus*, *Sus scrofa* and *Tursiops truncates*. In order to analyze promoter sequences two different approaches were used; one where transcription factor bindings sites were identified using a profile-based approach, and one where a more unbiased approach was taken to identify over-represented sequence motifs by making use of MEME.

3.1 Analysis of genome maps with information on gene order and relative orientation of genes -extraction of bidirectional promoter sequences.

From previous work we know that divergently transcribed genes that have an intergenic region less than 1000 bp are likely to have a bidirectional promoter (Adachi and Lieber 2002). Therefore, our first step was downloading from ENSEMBL, genome maps for the different species considered with information on gene localisation and on gene orientation. These maps could then be used to extract all bidirectional genes having an intergenic distance no greater than 1000 bp (See section 2.2). In addition, we extracted for reference a collection of co-directional gene pairs. Table 1 shows the number of divergent, convergent and codirectional genes we were able to identify in the human genomic data we acquired from Ensembl. Table 2 shows the number of divergent genes identified in the eight species.

Human Genes	Number
Divergent ←→	**1574**
Convergent →←	**1804**
Co-directional→→ and ←←	**1520**

Table 1 Number of the three different gene types residing within 1000bp

Name of Species	Divergent ←→
Homo sapiens	1574
Bos taurus	1066
Canis familiaris	280
Loxodonta africana	560
Mus musculus	755
Pteropus vampyrus	1510
Sus scrofa	526
Tursiops truncates	1120

Table 2 Number of divergent genes in different species with intergenic distance within 1000 bp.

Adachi and Lieber, who were the first to examine head to head arrangement of genes, also noted that many of these genes share an intergenic region where the distance between the transcription start sites is less than 300 bp (Adachi and Lieber 2002). Later a number of studies were done by Trinklein and colleagues, who were also the first in doing genome wide computational analysis of bidirectional promoters. They concluded that bidirectional genes share an intervening region having TSSs of the two genes approx. 1000 bp apart (Wakano, Byun et al. 2012). Figure 5 shows the lengths of all the bidirectional promoters from the eight species studied here. From the graph it can be observed that most bidirectional promoters have a size less than 500 bp.

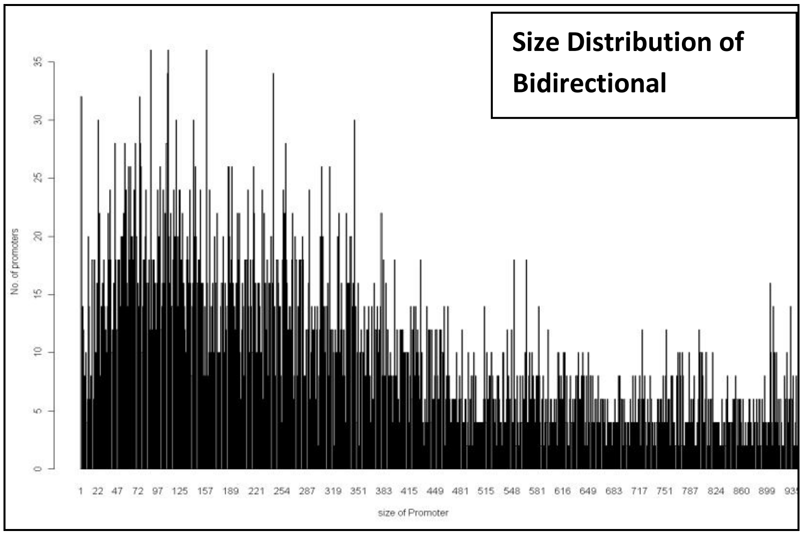


Figure 5 shows the length of bidirectional promoters on the x axis and their frequency in y axis. Bidirectional promoter from all the species are plotted here.

3.3 Identification of homology with OrthoMCL

In order to compare bidirectional promoters in the different mammalian species we wanted to identify for each of the human promoters the corresponding homologous promoter in the other species. This is to say that for a human bidirectional gene pair A-B we need to identify all pairs in the other species where the genes of the pair are homologous to A and B, respectively. In order to identify homology we made use of OrthoMCL. This program does a clustering of protein sequences that is based on an all to all BLAST analysis as explained in the Method section. All protein sequences from the different species were used as input to OrthoMCL. A total of 1876 orthologous clusters were identified .Figure 6 shows the number of genes contained in each cluster.

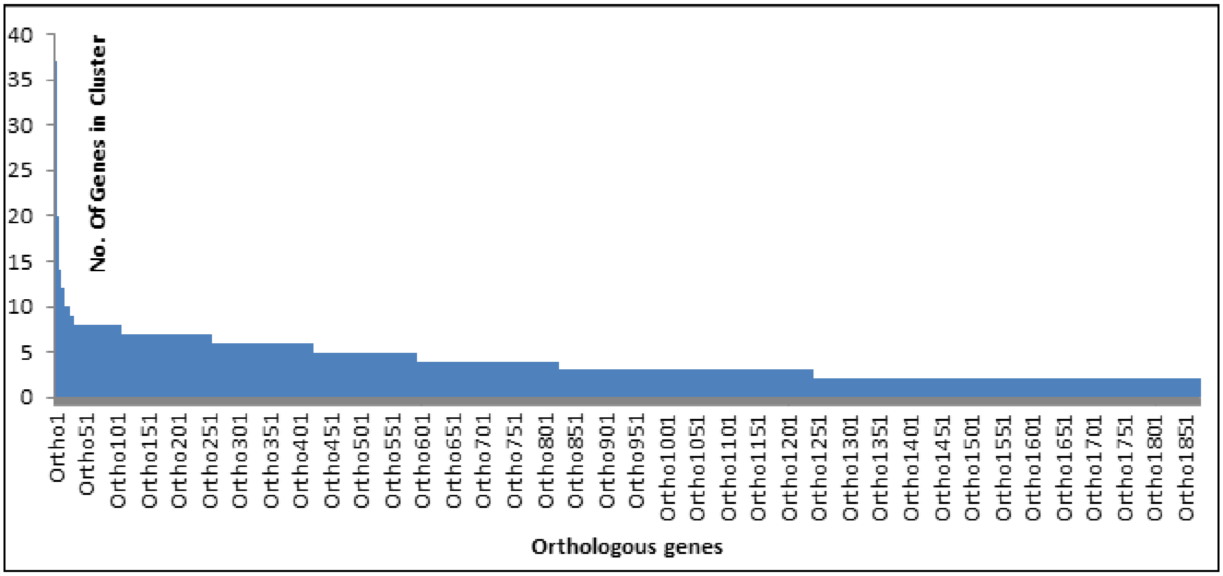


Figure 6 No of genes in each cluster generated by OrthoMCL. In the horizontal axis is the number of clusters and in the vertical axis is the number of genes in them.

3.4 Identification of orthologous gene pairs

Using the results from OrthoMCL we could assign each gene in all the gene order maps a unique OrthoMCL cluster ID. With this information it was in turn possible to identify all non-human homologs to the different human bidirectional pairs. The statistics of this analysis is shown in Fig 7, which compares the number of human bidirectional pairs to the number of orthologous bidirectional pairs in the other species. The results show that for many species we identify a comparatively low number of pairs. This is presumably because these genomes are less complete with respect to assembly and less well annotated with respect to protein genes. On the other hand the well-studied mouse genome is characterized by a larger number of orthologous gene pairs.

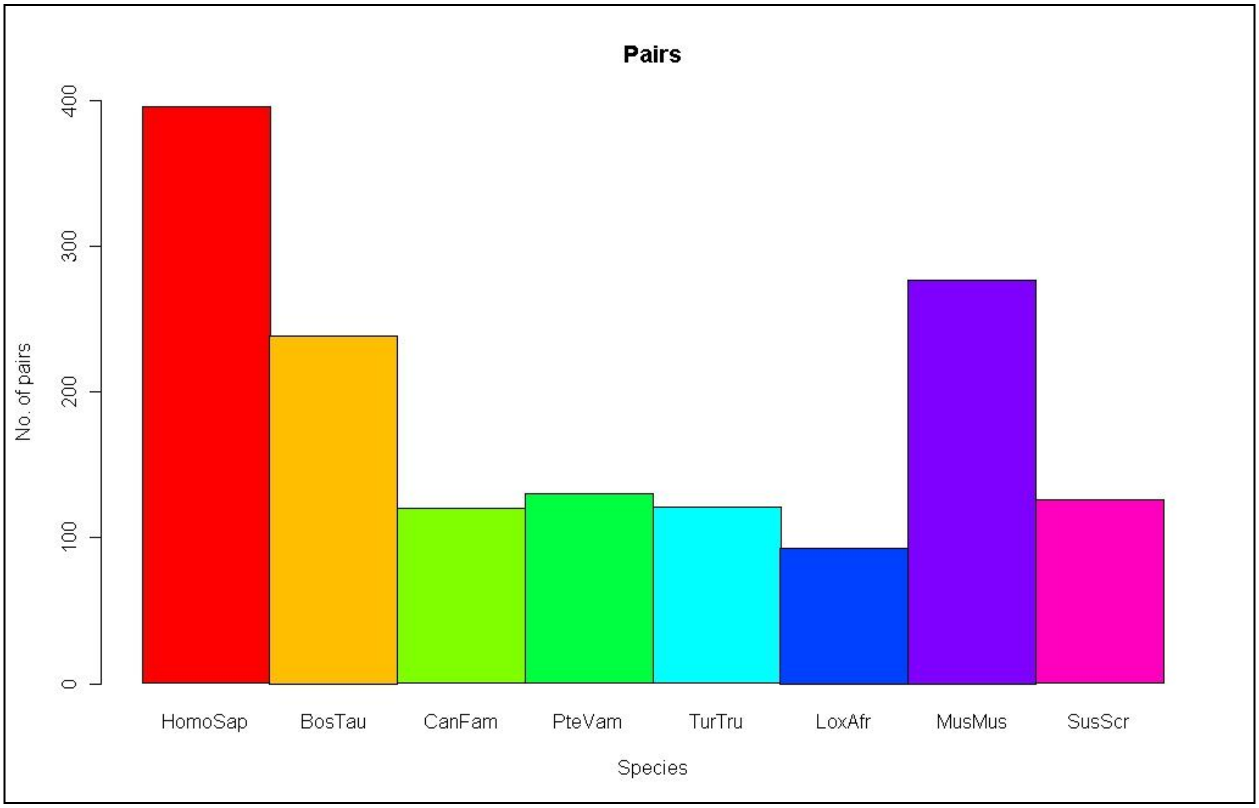


Figure 7 The histogram shows the number of pairs of genes of different mammals orthologous to human bidirectional gene pairs.

The percentage of ‘A’, ‘G’, ‘C’ and ‘T’ over the total number of bases in the bidirectional promoter region of orthologous genes in the eight species is shown in the pie chart see fig 8. This shows that G+C content in the sequences is quite high relative to other bases. This result is consistent with previous studies by Adachi and Lieber who identified high GC counts as well as enrichment of CpG islands (Adachi and Lieber 2002). Trinklein and colleagues arrived at similar results and concluded that the higher frequency of CpG islands is a major factor responsible for higher basal level of transcription (Trinklein, Aldred et al. 2004). Analysis of genome wide Pol II chromatin immune-precipitation studies shows that the CpG islands of bidirectional promoters are characterized by a higher Pol II occupancy (Barski, Cuddapah et al. 2007; Yang and Elnitski 2008) as compared non bidirectional promoters.

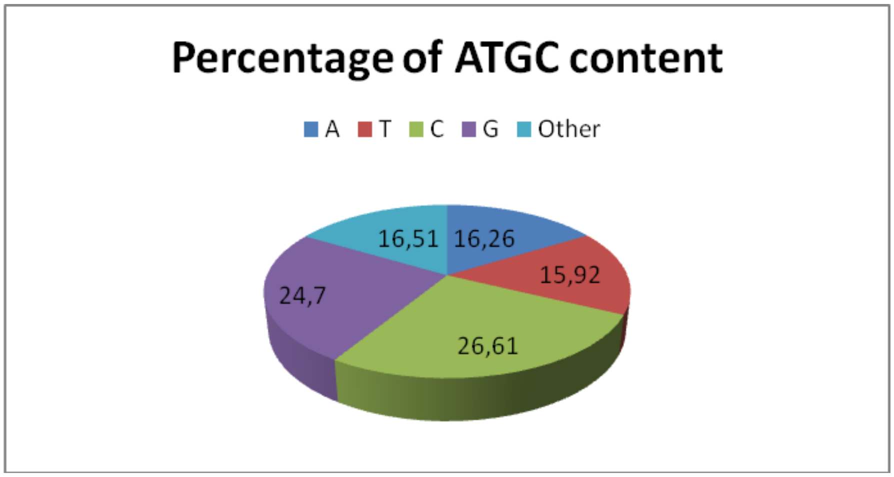


Figure 8 Pie chart showing percentage of bases in the promoter region of all bidirectional genes extracted from eight species. The figure depicts high percentage of G+C content compared to other nucleotides

3.5 Prediction of TFBSs in human bidirectional promoter regions

Predicting TFBSs has always been a challenge. Different kinds of experimental and computational techniques have been used to detect these sites. In this project we used a profile based (or position specific scoring matrix-based) identification technique to predict TFBSs in human bidirectional promoters sequences. The algorithm used here is explained in detail in the Materials and method section 2.3.1. Position specific score matrices (PSSMs) were downloaded from the JASPER core database. There were in all 130 JASPER core profiles. The JASPER profiles are all based on published material.

Figures 9, 10 and 11 shows graphs of possible binding sites identified in both the strands of human bidirectional promoters. The graphs contain TFBSs based on the profiles 1-40, 41-80 and 81- 130, respectively. The vertical axes shows the different profiles and the horizontal axis shows their counts.

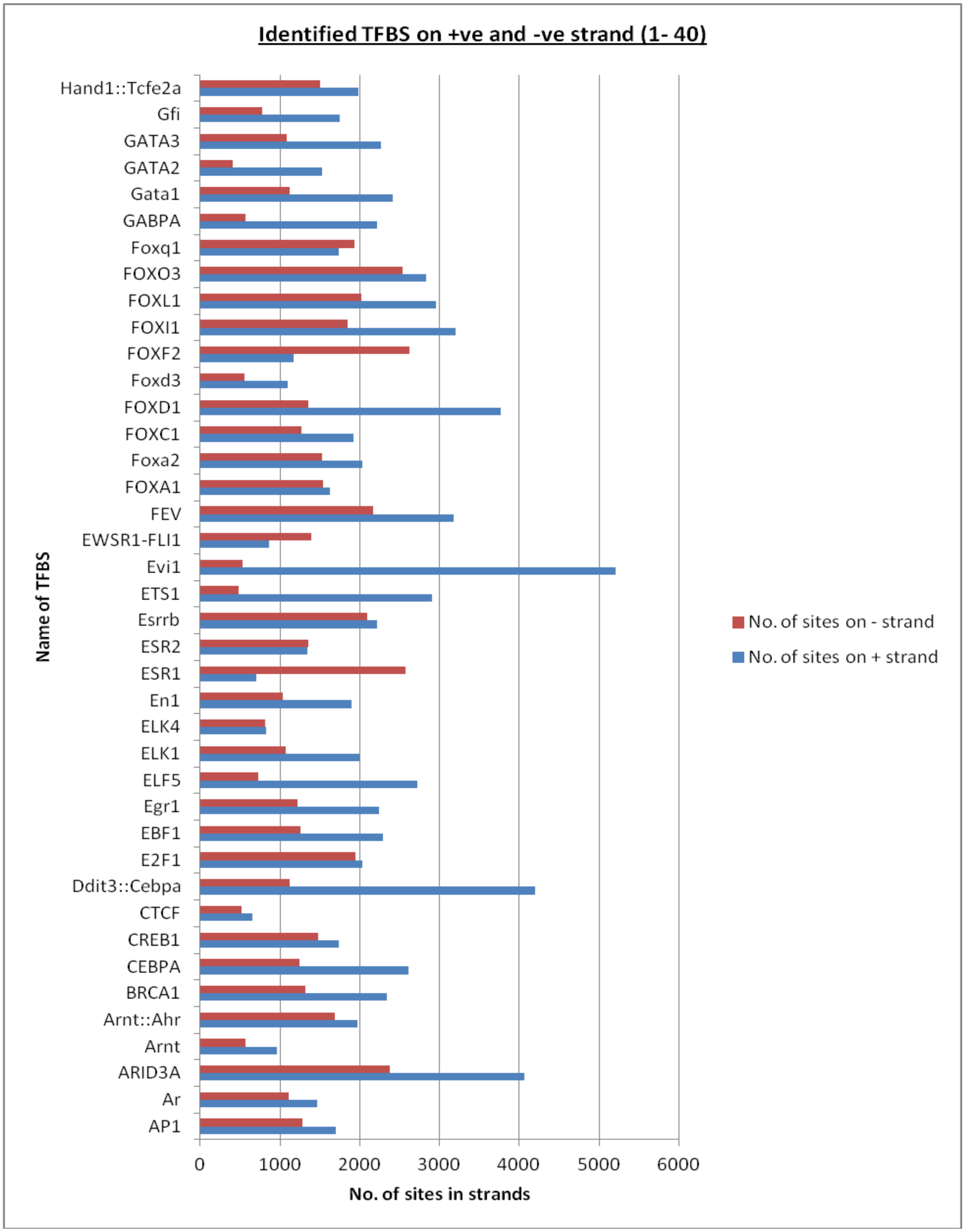


Figure 9 TBFSs identified in human bidirectional promoters, profiles 1-40

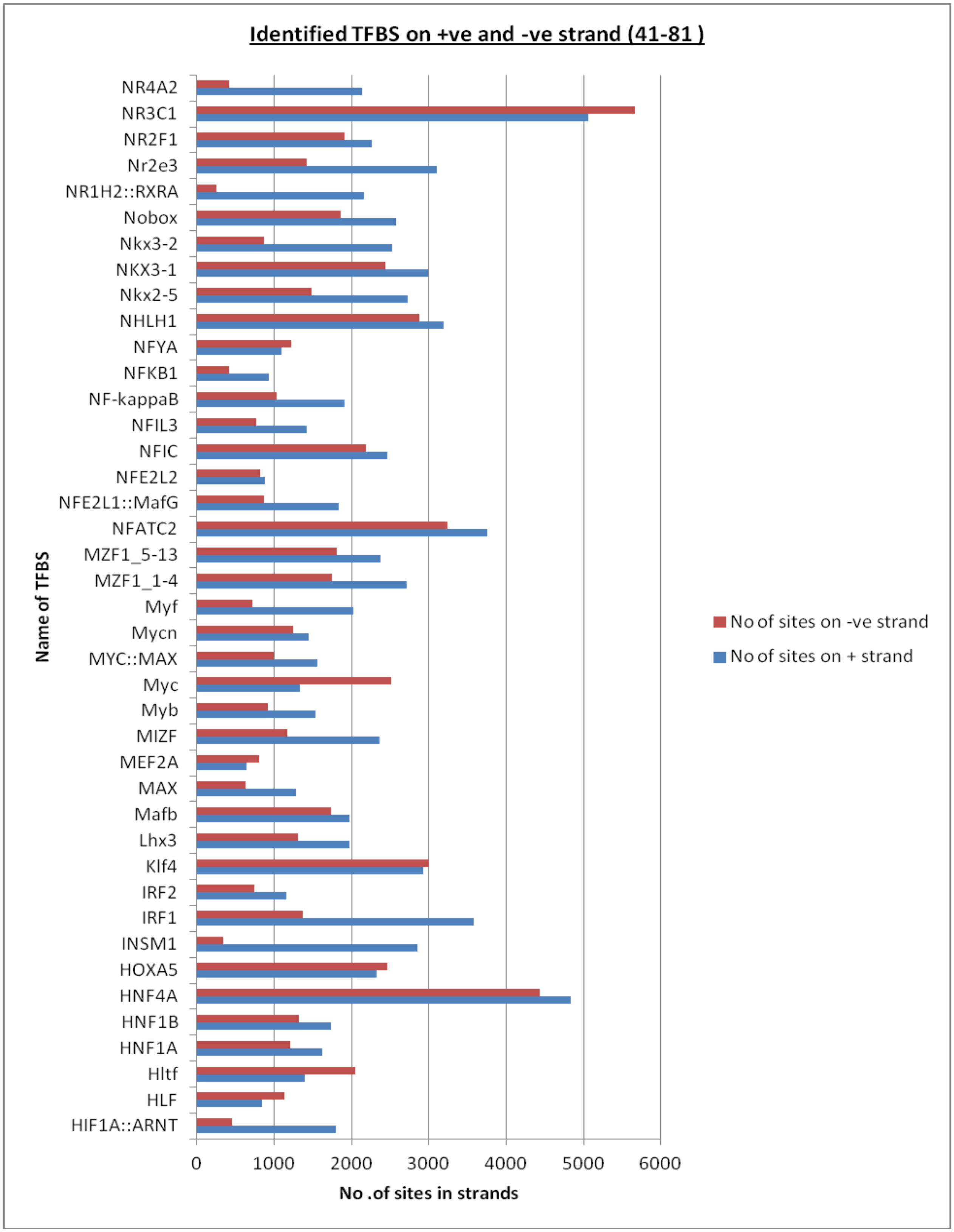


Figure 10 TBFSs identified in human bidirectional promoters, profiles 41-81

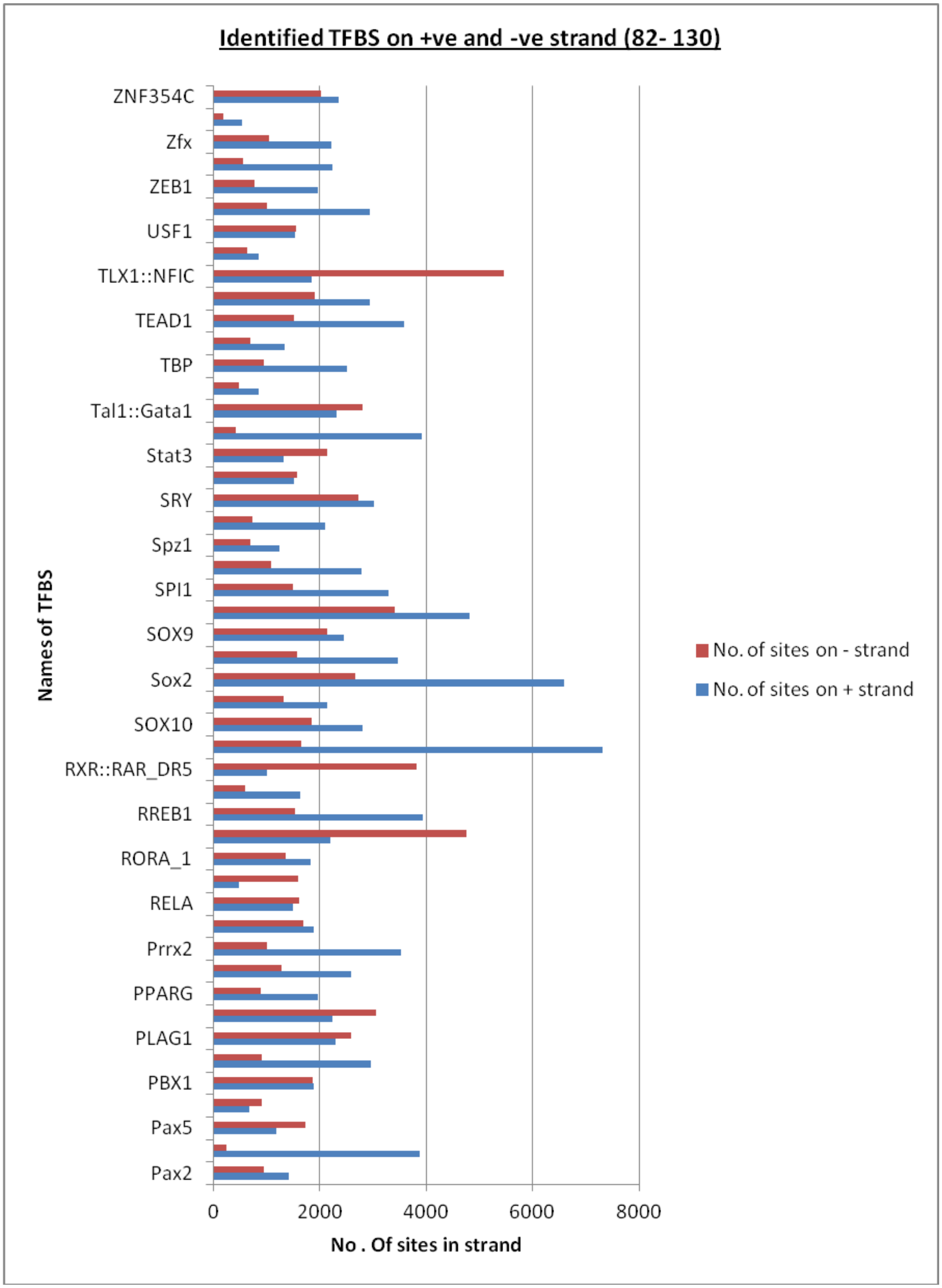


Figure 11 TBFSs identified in human bidirectional promoters, profiles 82-130

3.6 TFBSs that are over-represented

When considering the data shown in the previous section, we wanted to know what TFBSs were over-represented as compared to co-directional promoters. We therefore analyzed also co-directional promoters with respect to TFBSs. The non-parametric Mann-Whitney U test was applied to test the null hypothesis that the Z scores of bidirectional TFBSs are not significantly different than the co-directional ones. After a multiple hypothesis testing correction by the Bonferroni correction method, significantly differing motifs with a p-value smaller than 0.01 were identified. These are listed in Table 3. Figure 11 in the Appendix section shows a box plot that compares the distribution of Z-scores in each group. Finally the JASPER IDs were used to extract the corresponding annotation in the JASPER database (See appendix Table 6). We found that a large number of over-represented motifs are sequence motifs recognized by zinc finger proteins.

JASPER Profile Name	p-value	JASPER Profile Name	p-value
MA0259.1 HIF1A::ARNT	1.70E-09	MA0019.1 Ddit3::Cebpa.	3.20E-39
MA0099.2 AP1	1.60E-09	MA0125.1 Nobox.	3.20E-10
MA0004.1 Arnt	1.10E-13	MA0029.1 Evi1.matrix.score	1.70E-44
MA0152.1 NFATC2	1.00E-09	MA0088.1 znf143.matrix.score	8.60E-05
MA0081.1 SPIB	1.60E-10	MA0035.2 Gata1.matrix.score	5.30E-10
MA0087.1 Sox5	3.90E-70	MA0009.1 T.matrix.score	1.60E-11
MA0157.1 FOXO3	3.80E-34	MA0113.1 NR3C1.matrix.score	4.80E-56
MA0140.1 Tal1::Gata1	1.20E-12	MA0164.1 Nr2e3.	1.00E-193
MA0047.2 Foxa2.	1.20E-08	MA0106.1 TP53.	5.20E-06
MA0100.1 Myb	1.90E-05	MA0136.1 ELF5.	6.40E-14
MA0028.1 ELK1	3.00E-44	MA0031.1 FOXD1	1.90E-17
MA0039.2 Klf4	3.90E-05	MA0143.1 Sox2.	1.30E-09
MA0092.1 Hand1::Tcfe2a	2.80E-06	MA0080.2 SPI1.	7.60E-28
MA0098.1 ETS1	2.70E-17	MA0084.1 SRY.	8.00E-06
MA0132.1 Pdx1.	6.50E-92	MA0102.2 CEBPA.	2.00E-11
MA0153.1 HNF1B.	1.90E-28	MA0036.1 GATA2.	6.20E-11
MA0141.1 Esrrb.	7.40E-13	MA0077.1 SOX9.	7.00E-06
MA0089.1 NFE2L1::MafG	2.50E-19	MA0063.1 Nkx2-5.	8.00E-15
MA0101.1 REL.	2.80E-11		

Table 3: Over represented profiles and their P-values. The first column is the profile IDs and the second column shows their corresponding p-value (less than 0.01). The TFBSs that had been identified having significantly differing Z scores in comparison to co-directional promoters are listed here.

3.7 Comparison of TATA binding sites in bidirectional and co-directional promoters of the human genome

In addition to different transcription factor binding sites, the JASPAR database also contains a profile for the TATA promoter element. We used this profile with the same algorithm as described for the TFBSs above to score both bidirectional and co-directional promoters in the human genome. A total of 1574 divergent genes were analyzed and the results show that there were a total of 1690 sites reported and that occur in 449 unique genes. A total of 1224 co-directional genes were analyzed, resulting in 9694 sites in 556 unique genes.

The total length of the co-directional and bidirectional promoter sequences was 493217 bp and 156963 bp, respectively. We calculated the number of TATA sites in proportion to the total length of both gene sequences. We also calculated the percentage of TATA sites in both divergent and co-directional gene sequences. Figure 12 shows the portions of genes having TATA binding sites in bidirectional promoters in comparison to codirectional. The results are consistent with previous work on bi-directional promoters, showing that TATA binding sites are under-represented in bi-directional promoters (Yang and Elnitski 2008).

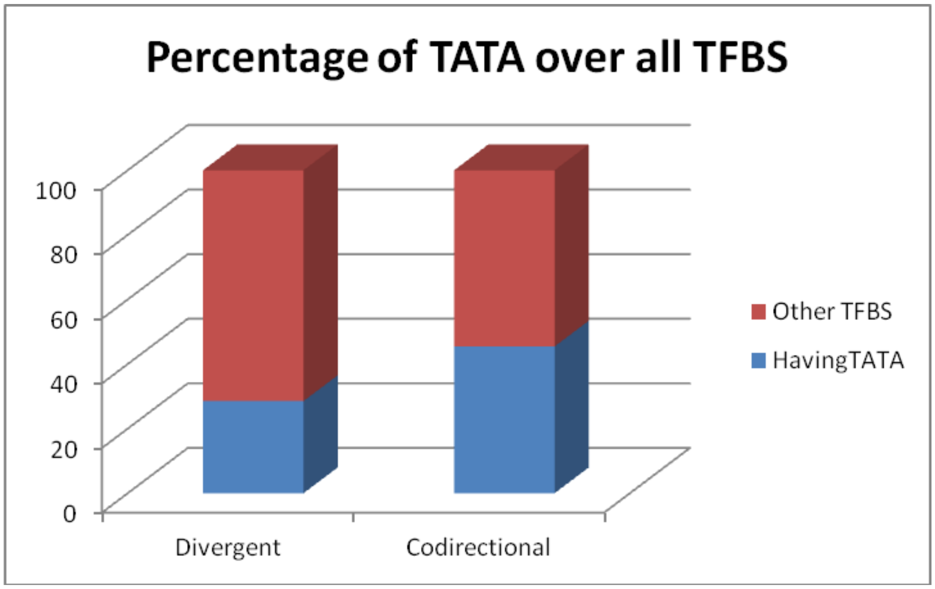
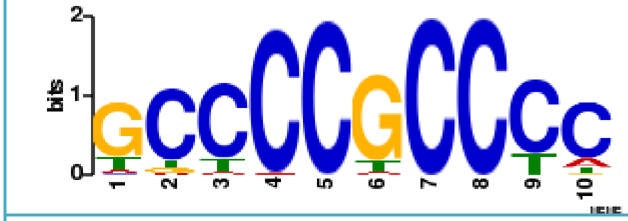
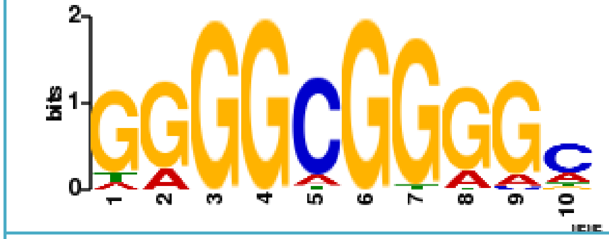
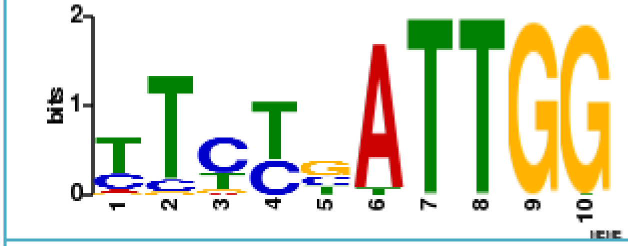
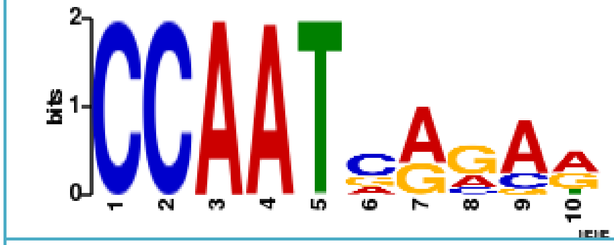
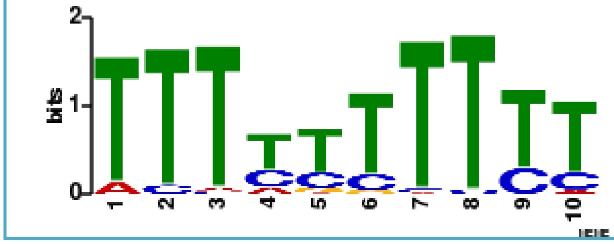


Figure 12 Percentage of TATA box in bidirectional promoter compared to co-directional promoters. TATA box is seen to be significantly under-represented in bidirectional promoters in comparison to co-directional.

3.8 Motifs identified using MEME suite

In addition to the approach using profiles to examine TFBSs and TATA box sites, we also used MEME to find motifs that might be characteristic of bidirectional promoters. The input data was a set of orthologous bidirectional promoter sequences. For each such promoter (pertaining to a certain pair of genes) we aligned the sequences from the different species with ClustalW. From these alignments we removed any regions with gaps, resulting in a set of alignments for each pair, containing one or more alignments with no gaps. The methods involved are described in sections 2.3, 2.4 and 2.5.

The resulting sequences were then analyzed with MEME as described in the Method section. The results are shown in Table 4. The most significant motifs are G and C-rich sequences.

Motif logo	Motif Number	Regular Expression	Number of sites	E-value
	1	GCCCCGCC[CT]C	594	3.2e-089
	2	GGGGCGGGG[CA]	591	6.5e-098
	3	[TC]T[CT][TC][GCT]ATTGG	243	2.9e-055
	4	CCAAT[CG][AG][GA][AC][AG]	280	5.7e-042
	5	GA[GA][TA]TGTAGT	113	1.7e-021

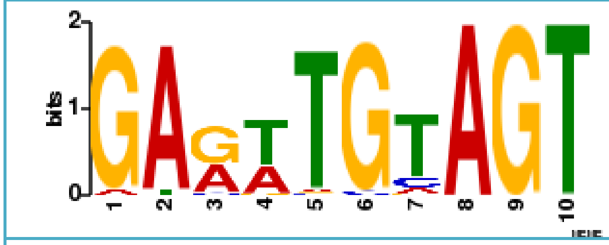
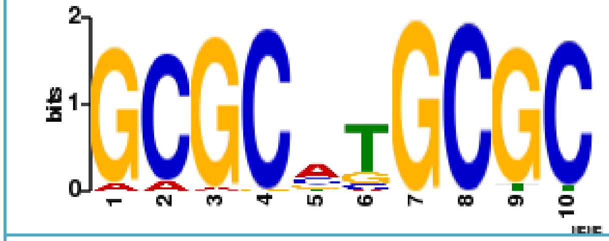
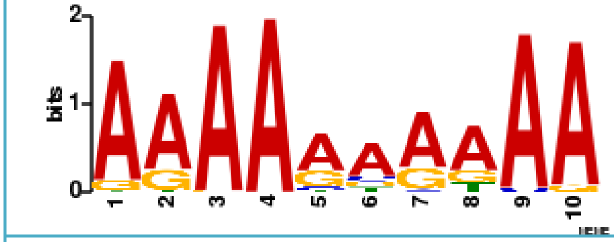
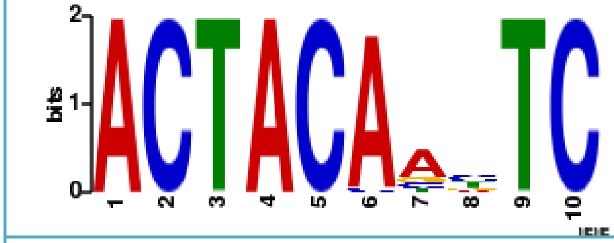
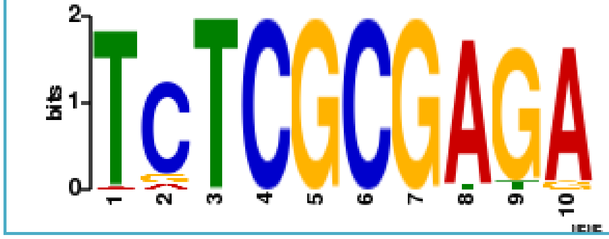
	6	GA[GA][TA]TGTA GT	95	1.3e-024
	7	GCGC[AC]TGCGC	297	8.1e-015
	8	A[AG]AA[AG]A[A G]AAA	109	4.5e-004
	9	ACTACAA[CT]TC	66	1.2e-016
	10	TCTCGCGAGA	48	3.6e-005

Table 4 Meme Output data. Results of MEME analysis of bi-directional promoters. The columns show 1) the motif sequence logs, 2) Motif number, 3) consensus sequence, 4) the number of sites where the motif occurs and 5) the Expect value. The algorithm of MEME and the parameters used is explained in the method section 2.7.

The motifs as identified above with MEME were finally compared to the motifs available in the JASPER CORE database. For this comparison I used the TOMTOM tool (a package within the MEME suite, see the methods section), where the motif profile is matched against the profile of the known binding sites in Jasper.

Table 5 shows the Jasper IDs profiles matched with p-value <0.05. Annotations from the JASPER database show that motif 1, 2 and 7 correspond to a DNA sequence motif recognized by zinc finger protein binding domains in eukaryotes (see Table 6 in the Appendix). Motifs 3 and 4 most likely correspond to the CCAAT box which is a well known core promoter element. Interestingly, the motif 6 is a TAAT core; a motif which is essential in DNA binding activity and the nucleotides flanking this core sequence directs binding specificity.

Hakkinen et al previously observed an enrichment of CCAT in bidirectional promoters (Hakkinen, Healy et al. 2011). In addition, they found that there is a correlation between multiple tandem arrangement and presence of this motif showing co-operative interactions within the binding sites. The Staf/ZNF143 zinc finger protein is a gene which is believed to control a number of genes that take part in DNA repair and genome stability and the bidirectional promoter region has potential binding sites for this specific protein (Izumi, Wakasugi et al. 2010).

Motif 1	**Motif2**	**Motif 3**	**Motif 4**	**Motif 6**	**Motif 7**
MA0079.2	MA0079.2	MA0316.1	MA0060.1	MA0002.1	MA0404.1
MA0039.2	MA0039.2	MA0060.1	MA0316.1	MA0027.1	MA0048.1
MA0443.1	MA0079.1	MA0314.1	MA0314.1		MA0375.1
MA0338.1	MA0443.1	MA0315.1	MA0315.1		MA0357.1
MA0283.1	MA0338.1	MA0188.1	MA0038.1		MA0162.1
MA0339.1	MA0283.1	MA0038.1	MA0188.1		MA0449.1
MA0431.1	MA0339.1	MA0331.1	MA0070.1		
MA0399.1	MA0431.1	MA0078.1	MA0180.1		
MA0450.1	MA0450.1	MA0127.1	MA0235.1		
MA0425.1	MA0399.1	MA0070.1	MA0078.1		
MA0146.1	MA0425.1	MA0180.1	MA0331.1		
MA0285.1	MA0285.1	MA0229.1	MA0125.1		
MA0337.1	MA0146.1		MA0229.1		
MA0410.1	MA0410.1				
MA0344.1	MA0337.1				
MA0456.1	MA0323.1				
MA0014.1	MA0123.1				
MA0323.1	MA0344.1				
MA0123.1	MA0014.1				
MA0441.1	MA0456.1				
MA0436.1	MA0441.1				

MA0268.1	MA0436.1
MA0068.1	MA0268.1
MA0362.1	MA0068.1
MA0449.1	MA0375.1
MA0375.1	MA0362.1
MA0394.1	MA0449.1
MA0395.1	MA0162.1
MA0162.1	MA0394.1
MA0270.1	MA0373.1
MA0290.1	MA0270.1
	MA0139.1
	MA0290.1

Table 5 Meme motifs mapped to Jasper Core. Motifs identified using MEME was mapped against JASPER Core database to find a match within known TFBSs. Each motif was matched with one or more JASPER profiles and below are the Ids associated with each of them.

4. Conclusions

In order to examine mechanisms of transcriptional control in human and other animals we may take advantage of comparative genomics in order to identify features that are conserved during evolution. We here used such an approach to examine the sequence properties of bidirectional promoters. Bidirectional promoters are of interest as the transcriptional control signals of the two different genes are overlapping and from a biological perspective they are interesting as genes of such bidirectional pairs are related to DNA repair and to the development of cancer.

In terms of comparative genomics, a challenge from a technical point is to identify relationships of orthology. Here we solved this problem with the help of OrthoMCL, such that each gene was assigned a cluster ID and having this information we could assign to every pair of bidirectional genes the homologous pair in other species. Using this information in turn we could identify all "homologues" of all human bidirectional promoters.

The resulting promoter sequences were then analyzed with on the one hand profiles of the Jasper database and on the other hand the MEME motif finding tool. One of the profiles was the TATA box motif, and we were able to confirm previous observations that the TATA box is somewhat under-represented in bidirectional promoters as compared to other promoter regions. In addition, we identified a set of TFBSs that are over-represented in bidirectional promoters.

The prediction of TFBSs with PSSMs is not always straight-forward and reliable. Such prediction may however be a good approximation that can give rise to candidate binding sites that are biologically interesting. Even though TFBS can be effectively identified in vitro using a large set of experimentally discovered binding sites, such results do not always refer to a direct regulatory function or even reveal that the site actually binds a protein. It has been argued that this it is not because the computational methods are wrong but shows the biological truth: various other factors such as competition, chromatin structure are as important as transcription factor binding affinity (Bulyk 2003).

A number of interesting consensus sequence motifs were obtained with MEME. Examples are a motif presumably related to the CCAAT box and GC-rich sequences that most likely are related to the GC-rich sequences that are known to be present in promoters. A search in the JASPER database shows that all motifs as identified with MEME are consistent with previously known transcription factor binding sites. Further analysis of these motifs may give more insight into the function of the binding sites.

In addition to the methods that we have used here there are a number of other procedures that may be explored. For the prediction of TFBSs one may try probabilistic computational algorithms like Hidden Markov Models (HMMs). For identification of motifs one could use tools which include the Gibbs sampling algorithm. One example of such a tool is the Gibbs motif sampler (Neuwald, Liu et al. 1995; Stormo and Fields 1998).

References

Adachi, N. and M. R. Lieber (2002). "Bidirectional gene organization: a common architectural feature of the human genome." Cell 109(7): 807-809.

Agirre, X., J. Roman-Gomez, et al. (2006). "Abnormal methylation of the common PARK2 and PACRG promoter is associated with downregulation of gene expression in acute lymphoblastic leukemia and chronic myeloid leukemia." International journal of cancer. Journal international du cancer 118(8): 1945-1953.

Ame, J. C., V. Schreiber, et al. (2001). "A bidirectional promoter connects the poly(ADP-ribose) polymerase 2 (PARP-2) gene to the gene for RNase P RNA. structure and expression of the mouse PARP-2 gene." The Journal of biological chemistry 276(14): 11092-11099.

Bailey, T. L., N. Williams, et al. (2006). "MEME: discovering and analyzing DNA and protein sequence motifs." Nucleic Acids Research 34(Web Server issue): W369-373.

Barski, A., S. Cuddapah, et al. (2007). "High-resolution profiling of histone methylations in the human genome." Cell 129(4): 823-837.

Bryne, J. C., E. Valen, et al. (2008). "JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update." Nucleic Acids Research 36(Database issue): D102-106.

Bulyk, M. L. (2003). "Computational prediction of transcription-factor binding site locations." Genome Biology 5(1): 201.

Burbelo, P. D., G. R. Martin, et al. (1988). "Alpha 1(IV) and alpha 2(IV) collagen genes are regulated by a bidirectional promoter and a shared enhancer." Proceedings of the National Academy of Sciences of the United States of America 85(24): 9679-9682.

Burchard, E. G., E. K. Silverman, et al. (1999). "Association between a sequence variant in the IL-4 gene promoter and FEV(1) in asthma." American journal of respiratory and critical care medicine 160(3): 919-922.

Butler, J. E. and J. T. Kadonaga (2002). "The RNA polymerase II core promoter: a key component in the regulation of gene expression." Genes & development 16(20): 2583-2592.

Chenna, R., H. Sugawara, et al. (2003). "Multiple sequence alignment with the Clustal series of programs." Nucleic Acids Research 31(13): 3497-3500.

Davila Lopez, M., J. J. Martinez Guerra, et al. (2010). "Analysis of gene order conservation in eukaryotes identifies transcriptionally and functionally linked genes." Plos One 5(5): e10654.

Fay, M. P. and M. A. Proschan (2010). "Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules." Stat Surv 4: 1-39.

Gershenzon, N. I. and I. P. Ioshikhes (2005). "Synergy of human Pol II core promoter elements revealed by statistical sequence analysis." Bioinformatics 21(8): 1295-1300.

Hakkinen, A., S. Healy, et al. (2011). "Genome wide study of NF-Y type CCAAT boxes in unidirectional and bidirectional promoters in human and mouse." Journal of Theoretical Biology 281(1): 74-83.

Hochheimer, A. and R. Tjian (2003). "Diversified transcription initiation complexes expand promoter selectivity and tissue-specific gene expression." Genes Dev 17(11): 1309-1320.

Izumi, H., T. Wakasugi, et al. (2010). "Role of ZNF143 in tumor growth through transcriptional regulation of DNA replication and cell-cycle-associated genes." Cancer Science 101(12): 2538-2545.

Jamison, D. C. (2003). Perl programming for biologists. Hoboken, N.J., Wiley-Liss.

Jensen, R. A. (2001). "Orthologs and paralogs - we need to get it right." Genome Biology 2(8): INTERACTIONS1002.

Kleinjan, D. A. and L. A. Lettice (2008). "Long-range gene control and genetic disease." Advances in genetics 61: 339-388.

Kornberg, R. D. (2001). "The eukaryotic gene transcription machinery." Biological chemistry 382(8): 1103-1107.

Kornberg, R. D. (2007). "The molecular basis of eukaryotic transcription." Proceedings of the National Academy of Sciences of the United States of America 104(32): 12955-12961.

Kulozik, A. E., A. Bellan-Koch, et al. (1991). "Thalassemia intermedia: moderate reduction of beta globin gene transcriptional activity by a novel mutation of the proximal CACCC promoter element." Blood 77(9): 2054-2058.

Lee, J. M. and E. L. Sonnhammer (2003). "Genomic gene clustering analysis of pathways in eukaryotes." Genome research 13(5): 875-882.

Li, L., C. J. Stoeckert, Jr., et al. (2003). "OrthoMCL: identification of ortholog groups for eukaryotic genomes." Genome Research 13(9): 2178-2189.

Lin, J. M., P. J. Collins, et al. (2007). "Transcription factor binding and modified histones in human bidirectional promoters." Genome Research 17(6): 818-827.

Neuwald, A. F., J. S. Liu, et al. (1995). "Gibbs motif sampling: detection of bacterial outer membrane protein repeats." Protein Science 4(8): 1618-1632.

Noble, D. (2008). "Genes and causation." Philosophical transactions. Series A, Mathematical, physical, and engineering sciences 366(1878): 3001-3015.

Petrij, F., R. H. Giles, et al. (1995). "Rubinstein-Taybi syndrome caused by mutations in the transcriptional co-activator CBP." Nature 376(6538): 348-351.

Sandelin, A., W. Alkema, et al. (2004). "JASPAR: an open-access database for eukaryotic transcription factor binding profiles." Nucleic Acids Research 32(Database issue): D91-94.

Shu, J., J. Jelinek, et al. (2006). "Silencing of bidirectional promoters by DNA methylation in tumorigenesis." Cancer Research 66(10): 5077-5084.

Smale, S. T. and J. T. Kadonaga (2003). "The RNA polymerase II core promoter." Annual review of biochemistry 72: 449-479.

Stormo, G. D. and D. S. Fields (1998). "Specificity, free energy and information content in protein-DNA interactions." Trends in Biochemical Sciences 23(3): 109-113.

Trinklein, N. D., S. F. Aldred, et al. (2004). "An abundance of bidirectional promoters in the human genome." Genome research 14(1): 62-66.

Vlahopoulos, S. A., S. Logotheti, et al. (2008). "The role of ATF-2 in oncogenesis." BioEssays : news and reviews in molecular, cellular and developmental biology 30(4): 314-327.

Wakano, C., J. S. Byun, et al. (2012). "The dual lives of bidirectional promoters." Biochimica Et Biophysica Acta 1819(7): 688-693.

Wang, Q., L. Wan, et al. (2009). "Searching for bidirectional promoters in Arabidopsis thaliana." Bmc Bioinformatics 10 Suppl 1: S29.

Weirauch, M. and B. Raney. (2007). "HMR Conserved Transcription Factor Binding Sites." Retrieved 08/08, 2011, from http://genome.csdb.cn/cgi-bin/hgTrackUi?g=tfbsConsSites&hgsid=252594.

Wright, K. L., L. C. White, et al. (1995). "Coordinate regulation of the human TAP1 and LMP2 genes from a shared bidirectional promoter." The Journal of experimental medicine 181(4): 1459-1471.

Yang, M. Q. and L. L. Elnitski (2008). "Diversity of core promoter elements comprising human bidirectional promoters." Bmc Genomics 9 Suppl 2: S3.

Yang, M. Q., L. M. Koehly, et al. (2007). "Comprehensive annotation of bidirectional promoters identifies co-regulation among breast and ovarian cancer genes." PLoS computational biology 3(4): e72.

Yang, M. Q., J. Taylor, et al. (2008). "Comparative analyses of bidirectional promoters in vertebrates." BMC bioinformatics 9 Suppl 6: S9.

Appendix

Comparison of Z scores of bidirectional TFBSs and co-directional TFBSs

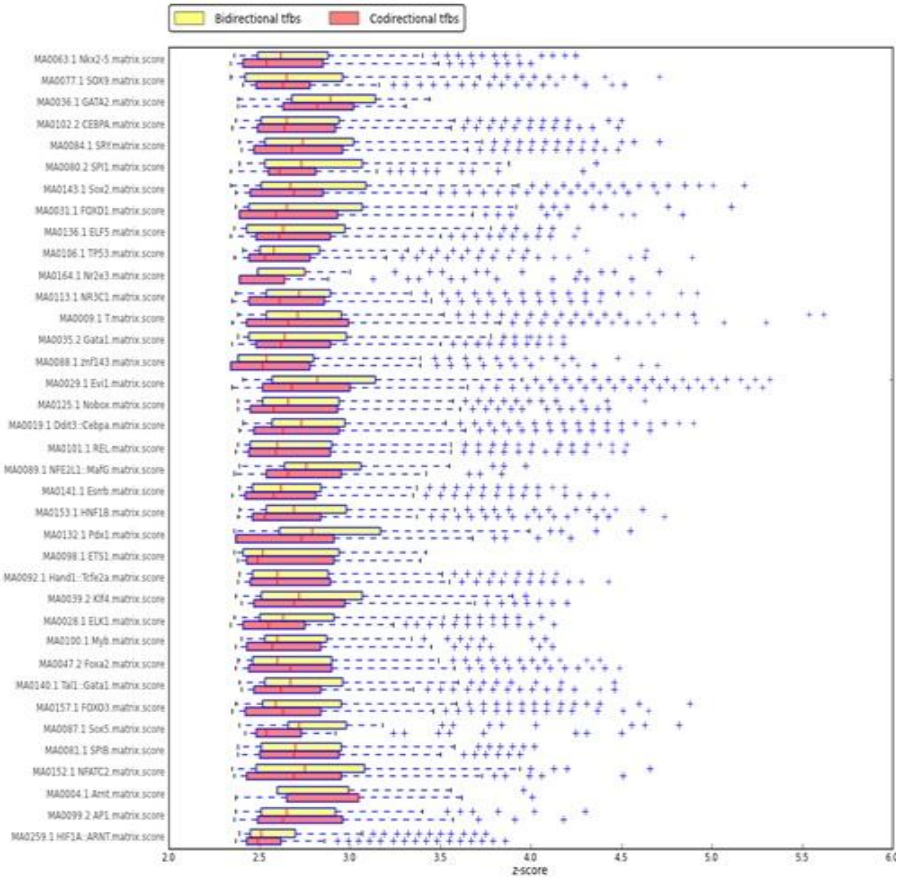


Figure 13: A box plot comparing the distribution of the z-scores of bidirectional TFBSs as compared to co-directional TFBSs.

Mapped Annotations of Jasper Profiles from Jasper Core Database

Profile	Description		
MA0004.1	9232	class	Zipper-Type
	9232	comment	-
	9232	family	Helix-Loop-Helix
	9232	medline	7592839
	9232	pazar_tf_id	TF0000003
	9232	tax_group	vertebrates
	9232	type	SELEX
MA0009.1	9237	class	Beta-Hairpin-Ribbon
	9237	comment	-
	9237	family	T
	9237	medline	8344258
	9237	pazar_tf_id	TF0000006
	9237	tax_group	vertebrates
	9237	type	SELEX
MA0019.1	9247	class	Zipper-Type
	9247	comment	dimer between Ddit3 and Cebpa
	9247	family	Leucine Zipper
	9247	medline	8657121
	9247	tax_group	vertebrates
	9247	type	SELEX
MA0028.1	9256	class	Winged Helix-Turn-Helix
	9256	comment	-
	9256	family	Ets
	9256	medline	1425594
	9256	pazar_tf_id	TF0000017
	9256	tax_group	vertebrates
	9256	type	SELEX
MA0029.1	9257	class	Zinc-coordinating
	9257	comment	-
	9257	family	BetaBetaAlpha-zinc finger
	9257	medline	8321231
	9257	pazar_tf_id	TF0000018
	9257	tax_group	vertebrates
	9257	type	SELEX
MA0031.1	9259	class	Winged Helix-Turn-Helix
	9259	comment	-

	9259	family	Forkhead
	9259	medline	7957066
	9259	tax_group	vertebrates
	9259	type	SELEX
MA0035.2	9379	class	Zinc-coordinating
	9379	comment	Data is from Frank Grosveld's Lab.
	9379	family	GATA
	9379	medline	-
	9379	pazar_tf_id	TF0000022
	9379	tax_group	vertebrates
	9379	type	ChiP-seq
MA0036.1	9264	class	Zinc-coordinating
	9264	comment	-
	9264	family	GATA
	9264	medline	8321207
	9264	pazar_tf_id	TF0000023
	9264	tax_group	vertebrates
	9264	type	SELEX
MA0063.1	9291	class	Helix-Turn-Helix
	9291	comment	-
	9291	family	Homeo
	9291	medline	7797561
	9291	pazar_tf_id	TF0000040
	9291	tax_group	vertebrates
	9291	type	SELEX
MA0077.1	9305	class	Other Alpha-Helix
	9305	comment	-
	9305	family	High Mobility Group
	9305	medline	9973626
	9305	pazar_tf_id	TF0000053
	9305	tax_group	vertebrates
	9305	type	SELEX
MA0081.1	9309	class	Winged Helix-Turn-Helix
	9309	comment	-
	9309	family	Ets
	9309	medline	7624145
	9309	pazar_tf_id	TF0000057
	9309	tax_group	vertebrates

	9309	type	SELEX
MA0084.1	9312	class	Other Alpha-Helix
	9312	comment	-
	9312	family	High Mobility Group
	9312	medline	8190643
	9312	pazar_tf_id	TF0000059
	9312	tax_group	vertebrates
	9312	type	SELEX
MA0087.1	9315	class	Other Alpha-Helix
	9315	comment	-
	9315	family	High Mobility Group
	9315	medline	1396566
	9315	pazar_tf_id	TF0000062
	9315	tax_group	vertebrates
	9315	type	SELEX
MA0088.1	9316	class	Zinc-coordinating
	9316	comment	-
	9316	family	BetaBetaAlpha-zinc finger
	9316	medline	9009278
	9316	tax_group	vertebrates
	9316	type	COMPILED
MA0089.1	9317	class	Zipper-Type
	9317	comment	Heterodimer between TCF11 and Mafg
	9317	family	Leucine Zipper
	9317	medline	9421508
	9317	pazar_tf_id	TF0000063
	9317	tax_group	vertebrates
	9317	type	SELEX
MA0092.1	9320	class	Zipper-Type
	9320	comment	-
	9320	family	Helix-Loop-Helix
	9320	medline	7791788
	9320	pazar_tf_id	TF0000066
	9320	tax_group	vertebrates
	9320	type	SELEX
MA0098.1	9326	class	Winged Helix-Turn-Helix
	9326	comment	-
	9326	family	Ets

	9326	medline	1542566
	9326	pazar_tf_id	TF0000070
	9326	tax_group	vertebrates
	9326	type	SELEX
MA0100.1	9328	class	Helix-Turn-Helix
	9328	comment	-
	9328	family	Myb
	9328	medline	1861984
	9328	pazar_tf_id	TF0000072
	9328	tax_group	vertebrates
	9328	type	SELEX
MA0101.1	9329	class	Ig-fold
	9329	comment	-
	9329	family	Rel
	9329	medline	1406630
	9329	pazar_tf_id	TF0000073
	9329	tax_group	vertebrates
	9329	type	SELEX
MA0106.1	9334	class	Zinc-coordinating
	9334	comment	-
	9334	family	Loop-Sheet-Helix
	9334	medline	1588974
	9334	pazar_tf_id	TF0000077
	9334	tax_group	vertebrates
	9334	type	SELEX
MA0113.1	9342	class	Zinc-coordinating
	9342	comment	-
	9342	family	Hormone-nuclear Receptor
	9342	medline	15563547
	9342	tax_group	vertebrates
	9342	type	COMPILED
MA0125.1	9354	class	Helix-Turn-Helix
	9354	comment	-
	9354	family	Homeo
	9354	medline	16997917
	9354	pazar_tf_id	TF0000820
	9354	tax_group	vertebrates
	9354	type	SELEX

MA0132.1	9361	class	Helix-Turn-Helix
	9361	comment	-
	9361	family	Homeo
	9361	medline	14704343
	9361	pazar_tf_id	TF0000824
	9361	tax_group	vertebrates
MA0136.1	9364	class	Winged Helix-Turn-Helix
	9364	comment	-
	9364	family	Ets
	9364	medline	16704374
	9364	pazar_tf_id	TF0000828
	9364	tax_group	vertebrates
	9364	type	SELEX
MA0140.1	9368	class	Zipper-Type
	9368	comment	Heterodimer between TAL1 and GATA1. Data is from Frank Grosveld's Lab.
	9368	family	Helix-Loop-Helix
	9368	medline	-
	9368	pazar_tf_id	TF0000022
	9368	tax_group	vertebrates
	9368	type	ChiP-seq
MA0141.1	9369	class	Zinc-coordinating
	9369	comment	-
	9369	family	Hormone-nuclear Receptor
	9369	medline	18555785
	9369	pazar_tf_id	-
	9369	tax_group	vertebrates
	9369	type	ChiP-seq
MA0143.1	9371	class	Other Alpha-Helix
	9371	comment	-
	9371	family	High Mobility Group
	9371	medline	18555785
	9371	pazar_tf_id	TF0000779
	9371	tax_group	vertebrates
	9371	type	ChiP-seq
	9378	class	Winged Helix-Turn-Helix
	9378	comment	-
	9378	family	Ets

	9378	medline	19160518
	9378	pazar_tf_id	TF0000039
	9378	tax_group	vertebrates
	9378	type	ChiP-seq
	9380	class	Zinc-coordinating
	9380	comment	-
MA0039.2	9380	family	BetaBetaAlpha-zinc finger
	9380	medline	18555785
	9380	pazar_tf_id	TF0000026
	9380	tax_group	vertebrates
	9380	type	ChiP-seq
	9385	class	Winged Helix-Turn-Helix
	9385	comment	-
	9385	family	Forkhead
	9385	medline	19553195
	9385	pazar_tf_id	TF0000029
	9385	tax_group	vertebrates
	9385	type	ChiP-seq
MA0152.1	9390	class	Ig-fold
	9390	comment	Annotations from PAZAR NFAT1_MOUSE + NFAT1_HUMAN + NFAT1_RAT (TF0000191, TF0000193, TF0000195) in the pleiades genes project.
	9390	family	Rel
	9390	medline	17916232
	9390	pazar_tf_id	TF0000193
	9390	tax_group	vertebrates
	9390	type	COMPILED
MA0153.1	9391	class	Helix-Turn-Helix
	9391	comment	Annotations from PAZAR HNF1B_HUMAN + HNF1B_MOUSE (TF0000780, TF0000782) in the TFe project.
	9391	family	Homeo
	9391	medline	17916232
	9391	pazar_tf_id	TF0000780
	9391	tax_group	vertebrates
	9391	type	COMPILED
MA0157.1	9395	comment	Annotations from PAZAR FOXO3_MOUSE + FOXO3_HUMAN (TF0000811, TF0000812) in the TFe project.
	9395	family	Forkhead
	9395	medline	17916232

	9395	pazar_tf_id	-
	9395	tax_group	vertebrates
	9395	type	COMPILED
MA0164.1	9402	class	Zinc-coordinating
	9402	family	Hormone-nuclear Receptor
	9402	medline	15634773
	9402	pazar_tf_id	-
	9402	tax_group	vertebrates
	9402	type	SELEX
	9403	class	Winged Helix-Turn-Helix
	9403	comment	Annotations from PAZAR PU.1 in the pleiades genes project (TF0000134).
	9403	family	Ets
	9403	medline	17916232
	9403	pazar_tf_id	TF0000056
	9403	tax_group	vertebrates
	9403	type	COMPILED
	9405	comment	Dimer. Annotations from PAZAR C-JUN + JUN_RAT + JUN_MOUSE + JUN_HUMAN + FOS/JUN_HUMAN + FOS_HUMAN in the pleiades genes project (TF0000129, TF0000147, TF0000234, TF0000243, TF0000670, TF0000287).
	9405	family	Leucine Zipper
	9405	medline	17916232
	9405	pazar_tf_id	TF0000071
	9405	tax_group	vertebrates
	9405	type	COMPILED
	9407	class	Zipper-Type
	9407	comment	last 3 nt removed
	9407	family	Leucine Zipper
	9407	medline	1672737
	9407	tax_group	vertebrates
	9407	type	COMPILED
MA0259.1	9503	class	Zipper-Type
	9503	comment	dimer between HIF1A and ARNT
	9503	family	Helix-Loop-Helix
	9503	medline	16234508
	9503	tax_group	vertebrates

Table 6 Mapped Annotations of Jasper Profiles from Jasper Core Database (Bryne, Valen et al. 2008).