Transcriptional regulation of bidirectional promoters: role of NF-Y

Master's Thesis
Institute of Medical Technology
University of Tampere
Gabriel Ndipagbornchi Teku
March 2010

ACKNOWLEDGEMENTS

I am grateful to Professor Jacobs H.T. for designing this project and taking time to assist and clarify biological concepts that were central to the study, as well as, painfully checking data created as a result of the project.

Many thanks to Professor Vihinen M., whose encouragement and valuable resources facilitated the accomplishment of this project.

I am indebted to Zanotto E., Tolvanen M., Shen B. and Rikkunen P., whose explanations clarified many biological and bioinformatics concepts, and whose words of encouragement assisted me at different phases of the project execution.

I would like to thank Mesue N. for the interesting discussions on statistics. Many thanks to Professor Vihinen M., Assistant Professor Tolvanen M. and Frooghi S. for reading the manuscript. Thanks also to Mesue N. and Aspatwar A. for reading parts of the manuscript.

To all friends and extended family, who have been very supportive and who encouraged me to finish this work, thank you very much.

Lastly, to my wife Dibo, my daughter Naseem, and my son Oben, my heart goes to you for all the times I had to abandon you in order to progress a little further with this work.

Tampere, March 2010

Gabriel Ndipagbornchi Teku

MASTER'S THESIS

Place: UNIVERSITY OF TAMPERE

Faculty of Medicine

Institute of Medical Technology

Author: GABRIEL NDIPAGBORNCHI TEKU

Title: Transcriptional regulation of bidirectional promoters: role of NF-Y

Pages: 62 pp.

Supervisor: Prof. Jacobs H.T., Prof. Vihinen M., Shen B., Tolvanen M., Zanotto E.

Reviewers: Professor Vihinen M. and Assistant Professor Tolvanen M.

Date: March 2010

ABSTRACT

BACKGROUND: About 11% of human genes occur in divergent pairs such that both genes are located on opposite strands of DNA, and their immediate promoters are overlapping. The overlapping proximal promoter of both genes form an intergenic region called a bidirectional promoter which is less than 1000 base pairs in length. There is evidence that some *cis*-regulatory elements in bidirectional promoters control the transcription of both flanking genes. CCAAT boxes are one of the most abundant *cis*-regulatory elements in the human genome. NF-Y, a heterotrimeric transcription factor, activates CCAAT boxes and requires both the CCAAT box and specific flanking nucleotides for DNA binding.

RESULTS: Using sequence analysis approach, data on the incidence of the NF-Y type CCAAT boxes in bidirectional promoters of both human and mouse genomes was used to deduce the functional and biological significance of NF-Y factor in the transcription mechanism of bidirectional gene pairs. In this study, four major findings were made. Firstly, a considerable number of bidirectional promoters consisted of at least an NF-Y type CCAAT box. This shows a critical role of NF-Y in the underlying bidirectional promoter regulation mechanism. Secondly, forward and reverse orientation of NF-Y type CCAAT boxes occurred in similar proportions in both bidirectional and unidirectional promoters, demonstrating NF-Y's ability to bind its recognition sequence in either orientation. Thirdly, a considerable number of NF-Y type CCAAT boxes were found in their functional position in bidirectional promoters, associating NF-Y to the recruitment of the transcription machinery. Lastly, NF-Y type CCAAT boxes were also significantly distributed further upstream to their functional position, suggesting that NF-Y is potentially connected to transactivating bidirectional promoters.

CONCLUSION: These results are in support of NF-Y's essential role in the general and activated transcriptional regulation of bidirectional gene pairs. This work provides an important contribution in understanding the regulatory mechanism of bidirectional promoters and, in turn, their subsequent biomedical applications.

Contents

Abbreviations

5’UTR	Five prime untranslated region
bp	Base pair(s)
BRE	B-responsive element
C/ebp	CCAAT/enhancer binding protein
CBF	CCAAT binding factor
ChIP	Chromatin immunoprecipitation
CP1	CCAAT binding factor 1
EMSA	Electrophoretic Mobility Shift Assay
GABP	GA-binding protein
GO	Gene Ontology
HFM	Histone fold motif
Inr	Initiator element
kb	kilobase
mRNA	Messenger RNA
Mrps12	Mitoribosomal protein S12
NF-Y	Nuclear factor-Y (alias of CBF)
NF-Y1	Datasets of NF-Y sites created from pattern search
NF-Y2	Datasets of NF-Y sites created from matrix and integrated search
NRF-1	Nuclear respiratory factor 1
NRF-2	Nuclear respiratory factor 2
PIC	Preintiation Complex
RNA Pol II	RNA polymerase II
Sarsm	Seryl-tRNA ligase
Sp1	Specificity protein 1
TF	Transcription factor
TSS	Transcription start site

1. Introduction

In higher eukaryotes the packaging of double stranded DNA into aggregate structures presents the cell with important levels of regulation including transcription and DNA replication. In order for the cell to regulate transcription it has to access genetic information at the level of the nucleotide bases (Jiang and Pugh, 2009). The processes used by the cell to access DNA sequences for transcription includes epigenetic modifications, protein-protein and protein-DNA interactions. These processes lead either to positive or negative regulation (Sandelin *et al.*, 2007).

Mammalian genes are arranged in tandem, divergent, convergent and overlapping manner. Tandemly arranged genes are those located on a chromosome, one after the other, in the same orientation, and on the same DNA strand (Fig 1). Divergent genes are those arranged head-to-to-head on opposite DNA strands. Convergent genes are arranged tail-to-tail on the same strand of DNA. Some genes in mammalian genomes occur within other genes in overlaying arrangements. Both genes may occur on the same or opposite strands of DNA (Engström *et al.*, 2006).

Whatever the structural arrangement of genes, they are made of coding (exon) and non-coding (intron) sequence segments. Generally, during transcription the genetic information from the exonic DNA is transcribed to messenger RNA (mRNA). Non-coding introns are spliced off moments after transcription.

In order for basal transcription to take place in higher eukaryotes, various protein-protein and protein-DNA interactions recruit RNA polymerase (RNA pol II) at the transcription start site (TSS). At the immediate five prime untranslated region (5' UTR) of each gene is the proximal promoter that consists of short stretches of DNA cognate sequences called *cis*-regulatory elements. General factors are ubiquitous proteins that bind *cis*-regulatory sequences just upstream and/or downstream to the TSS to enable the recruitment of RNA polymerase II (Sperling, 2007). Preinitiation complex (PIC) consists of general factors that form protein interacting complexes to direct RNA polymerase II to the TSS and organize DNA for transcription to commence (Lee and Young, 2000). These complexes include the general transcription factors IIA – H, i.e., TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH (Kornberg, 2007).

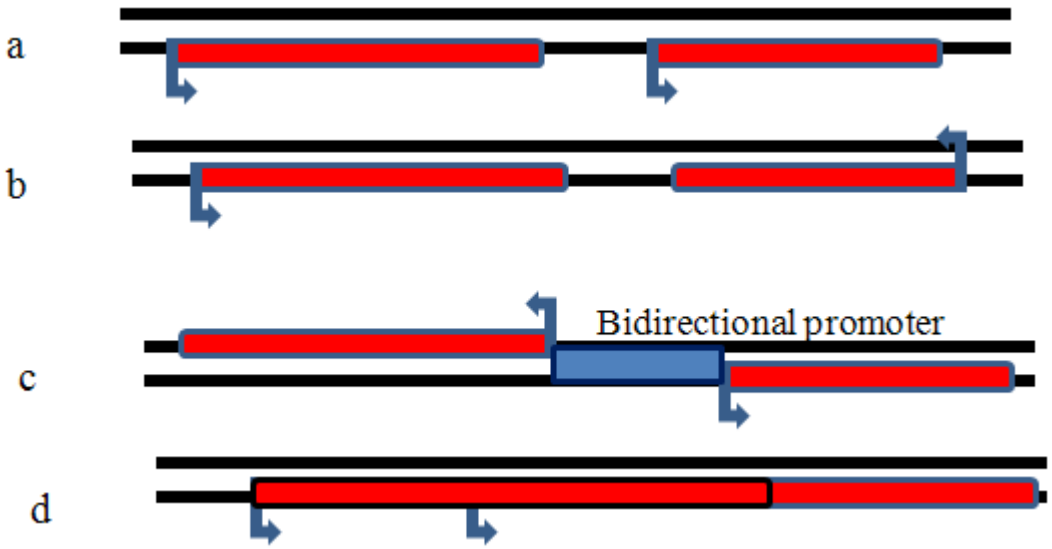

Figure 1. G**enomic arrangements of neighboring genes.** DNA strands are colored black while genes are colored red. **(a)** Genes arranged in tandem. (b) Convergent or tail-to-tail genes. (c) Bidirectional gene pair; the blue box in (c) is a bidirectional promoter between head-to-head genes. (d) Overlapping genes; the TSSs of both genes are far apart and their coding sequence segments overlap.

TFIID recognizes and binds to TATA box, initiator (Inr) sequence, and/or downstream promoter element (DPE), facilitating the recruitment of the PIC (Kornberg, 2007; Sperling, 2007).

Regulated transcription takes place after the basal transcription machinery has been recruited. Transcription factors that are gene-specific and act like switches turning on or off genes in different cell-types, tissue-types, development stages, or in response to a stimuli or stress are known as regulatory factors (Sperling, 2007). They are not as ubiquitous as the general factors and interact with their *cis*-regulatory sequences also called transcription factor binding sites (TFBSs) to turn a gene off or on.

Transcription factors often interact with other proteins called helper molecules or cofactors to activate or deregulate transcription. Cofactors are indirect transcriptional regulators acting as intermediaries to set off the assembly of the PIC and to enable the interaction between the regulatory factors and RNA pol II. The most ubiquitous cofactor is a large complex of polypeptides called the mediator (Sperling, 2007).

In order for transcription to take place the “locked” chromatin has to be “open” to allow transcription to ensue. Chromatin structure is intrinsically condensed and inactive in

higher eukaryotes. To set off the transcription of a gene an activator, due to its high affinity, binds DNA even in chromatin's condensed state, causing other factors to also bind the DNA (Kornberg, 2007). The bound factors then interact with chromatin remodeling agents to allow chromatin to become accessible for transcription. Chromatin remodeling includes postranscriptional modifications of histones by histone acetyltransferases and other ATP-dependent protein complexes, nucleosome sliding or displacement. Nucleosome sliding or displacement and histone acetylation correlates with a loosed chromatin structure that is needed for transcription factor binding and regulation (Dion *et al.*, 2005; Rada-Iglesias *et al.*, 2008).

Following the binding of multiple factors and chromatin remodeling the transcription factors interact with the mediator complex and, possibly, other cofactors, triggering the assembly of the PIC. The PIC in turn recruits RNA pol II to the TSS which then initiates transcription downstream. Transcription elongation proceeds with the help of elongation factors until a termination signal is reached at which point RNA polymerase dissociates from the DNA.

The close proximity of regulatory elements presents other levels of transcriptional regulation. Various functional genomic studies point to the evolutionary conservation of some genes that are biologically or functionally connected to each other. Most of these genes are conserved across species and are generally located in close proximity to each other. These proximal arrangements facilitate the combinatorial regulation of genes in clusters.

In lower organisms like prokaryotes, most of the genes are clustered into operons and the genes of each cluster (or a combination of clusters) are combinatorially regulated, i.e., most transcripts are multicistronic (Koonin, 2009; Rocha, 2008). Mammalian genomes also harbor genes in clusters, most of whose products are functionally or biologically related. Of particular interest are divergent genes that occur in close proximity so that they share promoter elements (Adachi and Lieber, 2002), as well as, the occurrence of a number of genes that are known to have undergone tandem duplication and are closely linked biochemically (Niimura and Nei, 2005). Most of these gene clusters have been the topic of many studies, but the mechanisms of the transcriptional control of closely located divergent genes is still under investigation.

The focus of this study is on the head-to-head arrangement of gene pairs whose TSSs are separated by at most 1000 base pairs (bp) and located at opposite strands of the DNA complex. The proximal promoter region of the gene pair overlaps. They are known as bidirectional promoters and are identified as the intergenic region between the TSSs of the flanking genes (Fig 1). A number of studies have been carried out to probe various aspects of bidirectional promoters (Adachi and Lieber, 2002; Dhadi *et al.*, 2009; Franck *et al.*, 2008; Yang and Elnitski, 2008a).

The functions and biological role of bidirectional promoters is still the topic of various studies. Most investigators propose that bidirectional genes are so arranged to enable coregulation of both flanking genes (Adachi and Lieber, 2002; Lin *et al.*, 2007; Trinklein *et al.*, 2004). In this model *cis*-regulatory sequences necessary for regulating both genes are shared. Some findings have indicated that the divergent arrangement leads to the mutually exclusive expression of the gene pairs (Knutson *et al.*, 2009; Trinklein *et al.*, 2004). Others have suggested cell-type specific expression of some of the bidirectional genes (Trinklein *et al.*, 2004). In addition, Lin and colleagues have proposed that the head-to-head architecture provides a means for regulating genes that are always transcribed (Lin *et al.*, 2007).

Only a few transcription factor binding sites are overrepresented in bidirectional promoters, *viz.*, GA binding protein (GABP), MYC, E2F1, E2F4, NRF-1, CCAAT, YY1, and ACTACAnnTCC (Lin *et al.*, 2007). GABP is the only factor whose recognition sequence occurs in most bidirectional promoters and whose role in regulating bidirectional promoters has been investigated in detail (Collins *et al.*, 2007). The CCAAT box, one of the most overrepresented *cis*-regulatory motifs in bidirectional promoters (Lin *et al.*, 2007), is activated solely by NF-Y. NF-Y is a ubiquitous heterotrimeric protein that requires the intact CCAAT motif for transactivation (Mantovani, 1998; Mantovani, 1999). NF-Y is also capable of bidirectional binding and activation of the CCAAT element. Both NF-Y promoters and bidirectional promoters are usually TATA-less (Mantovani, 1998; Yang and Elnitski, 2008a). CpG islands are regions of the DNA enriched with C+G dinucleotides at a frequency higher than that of the surrounding DNA. Most bidirectional promoters contain a CpG island, are usually TATA-less and consist of multiple TSSs (see section 2.4. below). In a study of 239

known or predicted human and mouse genes using chromatin immunoprecipitation (ChIP) on chip assay Testa *et al.* showed that there is evidence that NF-Y promoters are correlated with genes whose promoters consist of at least a CpG site (Testa *et al.*, 2005). Their data also suggested that NF-Y may play an important role in the regulation of a considerable fraction of divergent genes in the human and mouse genomes. All these prompted this investigation to address the occurrence, function and biological significance of CCAAT elements in the genome wide regulation of bidirectional promoters using sequence analysis.

2. Bidirectional promoters

In this section I present an assessment of the present state of understanding of divergent genes. In order to understand the importance of bidirectional promoters it is necessary to examine the various genome architectures with respect to genes that occur in close proximity to each other both in lower and in higher organisms as presented below. Thereafter, a summary of important findings on bidirectional promoters in non-eukaryotes is presented. Finally, I present a detail assessment of bidirectional promoters in eukaryotes.

2.1. Genome architectures

Genes are arranged in tandem, in tail-to-tail (convergent) or in a head-to-head (divergent) manner, in the genomes of various organisms (Fig 1). In each of these arrangements many genes can be found overlapping each other at their untranslated 5' or 3' regions. In some cases one gene is said to be located within the other.

Furthermore, various organisms have evolutionarily distributed their genes into different clusters. In the genomes of viruses gene distribution is dense with lots of overlaps and clusters (Firth and Brown, 2006; Rocha, 2008; Weinrich and Hruby, 1986).

Prokaryotic genomes are less compact than viral genomes with fewer gene overlaps. Most of the genes in prokaryotes are arranged in operons that are coregulated. A significant number of these genes are arranged divergently and are conserved within prokaryotes (Korbel *et al.*, 2004).

Operons are absent in eukaryotes, but some eukaryotic genes are arranged in other types of clusters, for instance, pairs of genes organized divergently and as tandem duplicates. Eukaryotic genomes are generally sparse, with many long repetitive, intergenic non-coding regions.

In all, grouping or clustering of genes occurs in prokaryotes and eukaryotes and these groupings are, sometimes, associated with various regulatory mechanisms. The members of these groupings have been shown to involve genes with similar functions,

the use of common promoter regions in the expression of genes that are switched on in a cell type or tissue specific manner, amongst others.

One of the most common of these gene clusters, the bidirectional gene pair arrangement, has been the subject of many recent studies. This gene pair arrangement is significant in occurrence in many organisms, especially mammals. In the next few sections I present insights from previous studies on bidirectional genes and their promoters.

2.2. Divergent genes in viruses

The genome of *Vaccinia virus*, like most viral genomes, is compact with promoters belonging either to the intermediate or late class (Knutson *et al.*, 2009). Many of these promoters are located between divergent genes. Because of the compact nature of the genome promoter elements are close to each other. Consequently, almost all divergent genes in *V. virus* genome share either a transcriptional initiation or core element (Knutson *et al.*, 2009). This arrangement is understood to be a mechanism by which only one of the pair of genes is expressed. The length of the promoter, the symmetric binding of TATA box-binding protein (TBP) close to the transcription initiation site and possible hindrance by TBP and RNA polymerase were proposed to be the reason for this mutually exclusive transactivation mechanism (Knutson *et al.*, 2009).

2.3. Divergent genes in prokaryotes

Prokaryote genomes have genes clustered in operons. Some of these operons and non-operonic genes are bidirectionally transcribed. The extent of conserved coexpressed bidirectional gene pairs in prokaryotes is higher than would be expected at random (Korbel *et al.*, 2004). This observation implies that bidirectional gene architecture is evolutionarily selected for. Using gene context methods across four clades of prokaryotes and a case study in *Escherichia coli* over two-thirds of prokaryotic bidirectional promoters were shown to be arranged such that one of the pair encodes a gene regulatory protein whereas the other encodes a non-regulatory product. It is not yet clear why such an organization exists in prokaryotes (Korbel *et al.*, 2004).

2.4. Divergent genes in eukaryotes

The most detailed studies of bidirectional promoters have been in mammalian genomes. This gene organization is common across unicellular, simple eukaryotes like *Giardia lamblia* (Teodorovic *et al.*, 2007) to higher eukaryotes like mammals. Below, I will present the occurrence, evolutionary and regulatory aspects of bidirectional promoters in eukaryotes.

2.4.1. Occurrence of divergent genes

In plants a genome-wide comparative study of bidirectional promoters of rice (*Oryza sativa*), *Arabidopsis thaliana*, and black cottonwood (*Populus trichocarpa)* was carried out (Dhadi *et al.*, 2009). The proportion of bidirectional promoters in each of the three plant species in the study (Dhadi *et al.*, 2009) was as many as are in the human genome (Trinklein *et al.*, 2004). Another study of bidirectional promoters in *A. thaliana* confirmed this result (Wang *et al.*, 2009). Among insects the analysis of divergent genes showed that a third of the *Drosophila melanogaster* genome is divergently arranged. The average promoter length is less than 400 bp (Yang and Yu, 2009).

Adachi and Lieber showed that a significant proportion of the human genome consists of bidirectional gene pairs (Adachi and Lieber, 2002). Later works estimated the proportion of human bidirectional gene pairs at over 11% (Takai and Jones, 2004; Trinklein *et al.*, 2004; Yang *et al.*, 2007).

2.4.2. Evolutionary conservation of bidirectional gene pairs

Bidirectional gene pairs show evolutionary conservation in plants, though less so when compared to mammals. Conservation of divergent or convergent genes across *O. sativa*, *A. thaliana* and *P. trichocarpa* was seen only among gene pairs that are highly coexpressed or share the same Gene Ontology (GO) classification (Krom and Ramakrishna, 2008). Bidirectional gene pairs in *Saccharomyces cerevisiae* are less conserved than those in mammals (Tsai *et al.*, 2007).

Specific studies on groups of genes with related functions and enriched in bidirectional promoters in mammals have shown evolutionary conservation. An example is the

cancer causing genes in humans; the *cis*-regulatory elements (especially the ETS family of factors) of their promoters are conserved (Yang *et al.*, 2007).

Divergent gene pairs show considerable evolutionary conservation over tail-to-tail or head-to-tail gene pairs. Studies to compare the evolution of divergent, head-to-tail and tail-to-tail gene pairs across vertebrates revealed that selective pressure was towards conserving divergent gene architecture (Franck *et al.*, 2008; Yang *et al.*, 2008). Trinkelin and colleagues identified positive selection for intergenic regions at the 5' ends of bidirectional gene compared to intergenic regions at 3' ends of convergent genes (Trinklein *et al.*, 2004).

Various investigations have been carried out to explain how bidirectional gene pairs came into existence, and the results have been mixed. One proposal states that the overlap of refractory promoter regions as in the case with bidirectional promoters have, through evolution, deterred the insertion of transposable elements, thus conserving bidirectional promoters (Takai and Jones, 2004). According to this proposal mammalian genes were close to each other before being invaded by transposable elements. Another research group suggests that bidirectional gene organization came into existence around the lineage leading to mammals and has since been selected for (Koyanagi *et al.*, 2005). Still, others maintain that due to a considerable number of fully preserved bidirectional gene pairs among mammalian genomes, most bidirectional promoters might have arisen when deuterostomes and protostomes diverged (Yang and Yu, 2009). Therefore, most investigators agree that selection has been towards conserving bidirectional gene pairs (Krom and Ramakrishna, 2008; Piontkivska *et al.*, 2009; Yang and Yu, 2009).

Taken together, the above evolution studies show that bidirectional gene organization plays an important role in the regulation of mammalian genomes. In fact most of the human orthologs of bidirectional gene pairs in mouse are conserved (Yang *et al.*, 2008).

2.4.3. Expression of bidirectional genes

Several studies show that most bidirectional gene pairs are coexpressed and/or coregulated. In plants, the expression of bidirectionally arranged gene pairs in *A. thaliana* genome was more correlated than neighboring and randomly selected gene pairs (Wang *et al.*, 2009). The study also reported that bidirectional promoters with

correlated function are highly coexpressed. An analysis on divergent and convergent genes in *O. sativa*, *A. thaliana* and *P. trichocarpa*, demonstrates that the level of coexpression among divergent and convergent genes is significantly higher than levels of expression among randomly paired genes (Krom and Ramakrishna, 2008). In insects, it is reported that 84.6% of bidirectional gene pairs show either positive or negative correlation to expression in a study of *D. melanogaster* divergent genes (Yang and Yu, 2009).

Most bidirectional gene pairs in human are coexpressed (Trinklein *et al.*, 2004). Bidirectional gene pairs are enriched and coexpressed in human cancer-related genes. Also, a significant number of bidirectional promoters are enriched within human housekeeping genes, especially DNA repair genes (Adachi and Lieber, 2002; Trinklein *et al.*, 2004). Although many bidirectional gene pairs are coexpressed analysis on functional relatedness show mixed results. Examples of bidirectional gene pairs that are coexpressed and are functionally or biologically related include cancer-related genes and house-keeping genes, mentioned above. A comparative analysis of conserved bidirectional gene pairs in vertebrates (Wang *et al.*, 2009) establishes that bidirectional gene pairs have diverse functions, most of them being involved in regulating enzymes.

2.4.4. Computational identification of bidirectional promoters

Bidirectional promoters have been shown to have a general sequence structure that makes its detection possible by machine learning methods. A recent report describes how to use machine learning techniques to distinguish bidirectional promoters from other regulatory elements in the human and mouse genomes (Yang and Elnitski, 2008b). Another report analyzed the potential of transcription in relation to functional correlation between bidirectional gene pairs (Wang *et al.*, 2009). It concluded that bidirectional gene pairs that are functionally correlated had a high transcription regulatory potential and bidirectional gene pairs without functional correlation had low transcriptional regulatory potential, indicating that these latter bidirectional gene pairs might have happened by chance.

2.4.5. Bidirectional promoters operate as a single unit

Variously, earlier studies to characterize bidirectional promoter sequences in both human and mouse, has revealed some insights into whether they function as a single unit or two separate subunits. Bidirectional promoters are made of two subunits that can clearly be demarcated at the sequence level (Engström *et al.*, 2006). One subunit begins at one of the TSSs and spans almost halfway through the promoter length, at which point the other subunit continues on the opposite strand until the other TSS. However, using deletion experiments, most studies reveal that bidirectional promoters operate as a single transcriptional regulatory unit, sharing some of the *cis*-regulatory elements (Lin *et al.*, 2007; Trinklein *et al.*, 2004; Wang *et al.*, 2009).

Some *cis*-regulatory elements maintain their position specificity in bidirectional promoters. In these cases they are located at about one-third the length of the bidirectional promoter with respect to each of the TSSs of the flanking genes (Lin *et al.*, 2007). Other *cis*-acting elements do not maintain their position specificity in bidirectional promoters. This promoter structure was said to allow for sharing of *cis*-regulatory elements. The TSSs of most divergent genes are less than 300 bp apart, ideal for both genes to share transcriptional regulatory elements (Adachi and Lieber, 2002; Trinklein *et al.*, 2004).

2.4.6. Directionality of bidirectional transcription

Because bidirectional promoters control both flanking genes and consist virtually of two shared subunits, it has been interesting to find out the mechanism that controls directionality. Regulation of directionality among bidirectional gene pairs is cell-type specific (Trinklein *et al.*, 2004). In a study using ChIP on chip assay on divergent and unidirectional promoters RNA polymerase II was found to occupy bidirectional promoters by two folds in contrast with unidirectional promoters. This suggests effective recruitment of RNA polymerase II and active transcription in both directions. Moreover, most *cis*-regulatory elements occur close to both TSSs of bidirectional promoters, whereas in most unidirectional promoters they are deficient (Lin *et al.*, 2007). The above results indicate that in bidirectional promoters transcription takes place readily and its directionality may be cell or tissue-specific. Nevertheless, the

structure or context of a promoter might play an important role in determining the direction of bidirectional transcription as shown in two recent studies (Zanotto *et al.*, 2007; Zanotto *et al.*, 2009). Below, the structure of bidirectional promoters is examined via important regulatory elements.

2.4.7. CpG islands and C+G content

CpG islands are regions of the DNA enriched with GC dinucleotides at a frequency higher than that of the surrounding DNA. They are part of the architecture of general promoters and are abundant in house-keeping genes (Farre *et al.*, 2007). In almost all genome wide studies of bidirectional promoters (with the exception of *D. melanogaster*), CpG islands and C+G content are overrepresented and play a major role in the regulation of this class of promoters. Several analyses have reported that CpG islands and a high C+G content is well correlated with bidirectional transcription units in the human and mouse genomes (Adachi and Lieber, 2002; Trinklein *et al.*, 2004; Yang and Elnitski, 2008a). Yang and Elnitski revealed that the average content of C+G nucleotides in human is 64% in bidirectional promoters; 55% in unidirectional promoters (Yang and Elnitski, 2008a). They further showed that CpG islands were found in 90% of bidirectional promoters compared to 45% in unidirectional promoters. In plants, the bidirectional promoter regions of *O. sativa*, *A. thaliana* and *P. trichocarpa* genomes were shown to have a similar C+G content as bidirectional promoters in the human genome (Dhadi *et al.*, 2009). Krom *et al.*, reported that in *O. sativa* genome the GCC-box occurs in the promoter of most gene pairs that are highly correlated (Krom and Ramakrishna, 2008).

Unlike other genomes studied so far, the *D. melanogaster* bidirectional gene pairs are depleted of CpG islands. Yang and Liang proposed that DPE and Inr elements in *D. melanogaster* bidirectional promoters may be the functional equivalents of CpG islands in mammalian bidirectional promoters (Yang and Yu, 2009).

Although CpG islands occur in most bidirectional promoters (Yang and Elnitski, 2008a) and occur in over two-thirds of general promoters (Saxonov *et al.*, 2006), their regulatory role is still being investigated, the most important results of which are as follows. In a study of interspersed repeats in human bidirectional promoters Takai and

Jones suggested that overlapping and overrepresented CpG islands may be a resistance mechanism against invasion by transposable elements (Takai and Jones, 2004). Methylation studies have shown that promoters that are located within CpG sites are regulated by methylation. Bidirectional gene pairs known to regulate tumor suppression were silenced by hypermethylation, implicating CpG islands in the regulation of these promoters (Shu *et al.*, 2006). CpG islands and high C+G content have been involved in the evolution of eukaryote genomes and the reason for positive selection towards bidirectional promoters (Antequera, 2003; Koyanagi *et al.*, 2005). CpG islands are also highly correlated with promoters that consist of multiple TSSs; this is the case with bidirectional promoters (Engström *et al.*, 2006).

2.4.8. TATA box

For many years TATA box has been known as one of the most ubiquitous *cis*-regulatory elements in the general promoter architecture. Promoters are usually classified as TATA-dependent or TATA-less. Each of these promoter types has different characteristics. TATA box promoters are well studied whereas the role of TATA-less promoters is still unclear. Most promoters that consist of multiple TSSs are TATA-less and are enriched with CpG islands (Sandelin *et al.*, 2007). Interestingly, most bidirectional promoters in mammals are TATA-less and located within CpG islands (Engström *et al.*, 2006; Trinklein *et al.*, 2004; Yang and Elnitski, 2008a).

A similar result has been found in other eukaryotes. Most of the bidirectional promoters in the genomes of *O. sativa*, *A. thaliana* and *P. trichocarpa* are TATA-less (Dhadi *et al.*, 2009). Also, TATA-boxes are depleted in *D. melanogaster* bidirectional promoters (Yang and Yu, 2009).

In contrast, the number of unidirectional promoters with at least a TATA-box is significant compared to what would be found by chance (Yang and Elnitski, 2008a). No equivalent to the TATA-box exists within bidirectional promoters. TATA-boxes, though scarce in bidirectional promoters, are overrepresented in histone promoters. Most investigations on bidirectional promoters, so far, show that most of them lack the TATA box, are located on CpG sites, and have multiple TSSs.

2.4.9. Enriched TFBSs in bidirectional promoters

In a study of core promoter elements in human bidirectional promoters Yang and colleague reported the following results (Yang and Elnitski, 2008a). Half of both bidirectional and unidirectional promoters consist of DPE at their functional positions of +30 bp. Inr elements occur in about a third of bidirectional and unidirectional promoters. TFIIB recognition elements (BRE) are found in 16.5% of bidirectional and 11.1% of unidirectional promoters. CCAAT motifs searched at their functional position (75 – 80 bp upstream to the TSS) occur in 12.9% of bidirectional and 6.9% of unidirectional promoters, respectively.

A recent study on human bidirectional promoters showed that the recognition sequence of a small set of transcription factors regulates bidirectional promoters as opposed to just one or myriads of them (Krom and Ramakrishna, 2008; Lin *et al.*, 2007; Piontkivska *et al.*, 2009). Various studies have been made on the most enriched of these motifs in bidirectional promoters. Among them are the recognition sequences for ELK1, GABP, SP1, and CCAAT-boxes (Lin *et al.*, 2007; Yang and Elnitski, 2008a).

Of all the overrepresented regulatory transcription factors only GABP has been studied in detail. ETS-related transcription factor binding sites are significantly enriched within the bidirectional promoters of cancer genes. It was proposed that the regulation of these genes may require other ETS-related factors (Yang *et al.*, 2007). In one study GABP was shown to bind to over half of bidirectional promoters (Lin *et al.*, 2007). Later studies showed that GABP regulates 80% of all bidirectional promoters of at least one cell type and generally promotes genes in both transcriptional directions (Collins *et al.*, 2007).

Like in the human bidirectional promoters, a small set of recognition sequences were enriched in the genomes of *O. sativa*, *A. thaliana* and *P. trichocarpa* (Dhadi *et al.*, 2009). Tsai *et al.* revealed that the number of bidirectional promoters in *S. cerevisiae* that share TFBSs is low in contrast to those of the human genome. They proposed that other mechanisms, for instance, the sharing of TFBSs, must be associated with regulating bidirectional promoters in *S. cerevisiae* (Tsai *et al.*, 2007).

2.4.10. Chromatin structure at bidirectional promoters

Another important layer of gene control is at the chromatin level. Lin *et al.* demonstrated that the chromatin structure in bidirectional promoters is more open relative to other promoters, indicating that bidirectional promoters are readily and generally transcribed (Lin *et al.*, 2007). They showed that this feature is not common among unidirectional promoters.

3. NF-Y structure and functions

NF-Y is a heterotrimeric protein involved in the transcriptional regulation of many genes in eukaryotes. Its homologs are conserved across eukaryotes from *S. cerevisiae* to human (Serra *et al.*, 1998; van Huijsduijnen *et al.*, 1990). It is also known as CCAAT-box binding factor (CBF) and CCAAT binding protein-1 (CP1).
CBF-A, CBF-B and CBF-C are equivalent to NF-YB and NF-YA, and NF-YC, respectively. NF-YB and NF-YA were first characterized from the promoters of rat and mouse genes (Hatamochi *et al.*, 1988; Maity *et al.*, 1988; Maity *et al.*, 1990).
Later experiments isolated and characterized NF-YC which was found to be identical to CBF-C (Sinha *et al.*, 1995).
Each NF-Y subunit has distinct conserved domains for subunit-subunit interactions and NF-Y/DNA interactions (Sinha *et al.*, 1996). The NF-Y factor is ubiquitous and essential for the regulation of many genes in mammalian genomes. To understand the role of NF-Y in the regulation of transcription it is necessary to understand its structure and functional interactions. The assessment below concerns mostly studies on NF-Y in mammalian cells.

3.1. The NF-Y subunits

3.1.1. NF-YA

The NF-YA subunit in human has a DNA binding domain between residues 298 – 318 (Fig 2). The subunit association domain is located between amino acids 266 – 282 and a Q-rich domain between residues 14 – 161 (Mantovani, 1999; Sinha *et al.*, 1996). The subunit association and DNA binding domains have been tightly conserved throughout evolution and form amphipatic α-helices (Xing *et al.*, 1993). NF-YA binds to the dimerized NF-YB/C at its subunit association domain to form the NF-Y factor (Kim *et al.*, 1996; Liang and Maity, 1998). The inherently high specificity of NF-Y recognition of its binding surfaces depends on interacting domains in both NF-YA and NF-YB/C dimer (Liberati *et al.*, 1999; Xing *et al.*, 1993).

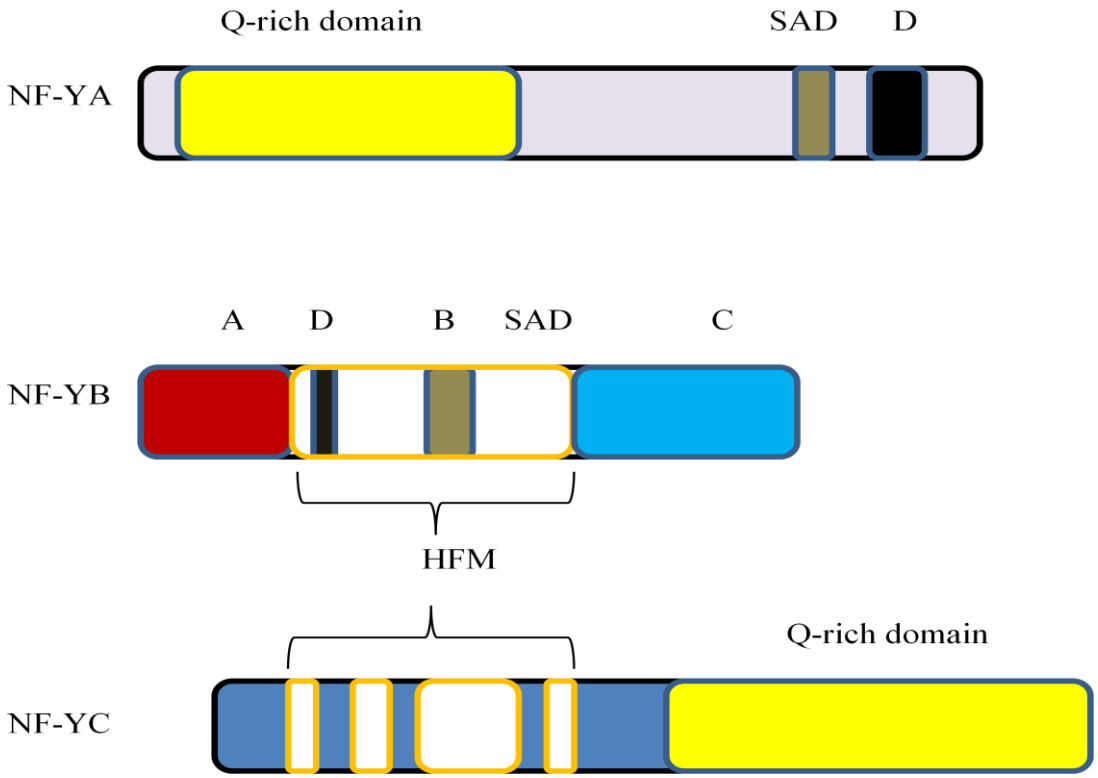

Figure 2. **NF-Y subunits and organization of their subdomains.** NF-YA consists of a Q-rich activation domain colored yellow; a subunit activation domain (SAD) for forming NF-Y with NF-YB/C complex; and a DNA binding domain (D). NF-YB and NF-YC consists of a HFM by which they interact to form NF-YB/C complex. Within the HFM of NF-YB there is a DNA binding domain (D), a subunit interacting domain (SAD) and an NF-Y interacting domain (B); sites A and C have no particular biological function. The Q-rich domain in NF-YC, like that of NF-YA, is used for DNA and protein-protein interactions.

3.1.2. NF-YB and NF-YC

NF-YB and NF-YC subunits are members of a class of proteins with the histone fold motif (HFM) (Baxevanis *et al.*, 1995). Proteins with HFMs are highly evolutionarily conserved with low sequence identity but high structural similarities. They are usually made of an α-helix (α1) followed by a short loop (L1), a long α-helix (α2) followed by another short loop (L2) and a short α-helix (α3). These proteins are often found dimerized in an anti-parallel manner. HFMs form dimers through interaction of the α3 helix of one with the α1 helix of the other, and the α2 helices exhibit extensive

interaction with each other. It is known that the short loops, L1 and L2, are responsible for interaction with DNA complex (Arents and Moudrianakis, 1995; Kim *et al.*, 1996). Detailed study of NF-YC and NF-YB shows that they have conserved α1, L1, α2, L2 HFMs with some divergence in α3 (Arents and Moudrianakis, 1995; Coustry *et al.*, 1996; Mantovani, 1999). At a region necessary for interacting with other proteins NF-YB and NF-YC have high similarity to each other. These regions have been named αC and αN in NF-YB and NF-YC, respectively (Sinha *et al.*, 1996). By comparing NF-YB and NF-YC to core HFMs, NF-YB is related to H2B, and NF-YC to H2A families of proteins (Baxevanis *et al.*, 1995).

Mutation studies have showed that the HFMs are necessary for NF-YB/C heterodimerization (Sinha *et al.*, 1996; Xing *et al.*, 1993), NF-YA interaction to form the NF-Y factor (Sinha *et al.*, 1995; Zemzoumi *et al.*, 1999), interaction with other regulatory factors (Bellorini *et al.*, 1997; Coustry *et al.*, 1998), and DNA sequence-specific binding (Kim *et al.*, 1996; Romier *et al.*, 2003). The NF-YB/C dimer is also capable of interacting with other proteins and specific CCAAT boxes *in vitro* (Bellorini *et al.*, 1997; Hatamochi *et al.*, 1988; Hu *et al.*, 2006).

3.1.3. NF-Y assembly from subunits

All three subunits (NF-YA, NF-YB and NF-YC) are necessary for interacting and forming the heterotrimeric factor (Liang and Maity, 1998). NF-YB and NF-YC form the NF-YB/C dimer (Liberati *et al.*, 1999) that interacts with NF-YA to form the NF-Y heterotrimeric protein. Binding of the dimerized NF-YB/C to NF-YA occurs at the subunit interaction domains of both entities. There exist two regions, both located at the C termini of NF-YC and NF-YA (Fig 2) that has been shown to contain the DNA activation domains for NF-Y (Coustry *et al.*, 1996; Romier *et al.*, 2003). These subdomains show similarity in residues to each other; both consist predominantly of hydrophobic and glutamine residues. Nakshatri *et al.* showed that the assembly of NF-Y from its subunits is affected by cellular redox processes (Nakshatri *et al.*, 1996), but the role of other factors and physical conditions in the cell is still unclear.

3.2. Interactions of NF-Y with the Preinitiation Complex

One of the *cis*-acting elements present in general promoter architectures is the CCAAT box. NF-YB/C dimer interacts with and enables the PIC to achieve an indirect positive regulatory effect (Bellorini *et al.*, 1997; Sun *et al.*, 2009). A study showed that NF-Y, through interaction with the CCAAT box and the Inr element on the Ea promoter was able to increase the affinity of TBP to the Ea promoter for transcription initiation (Bellorini *et al.*, 1996). Later, a region made of acidic and hydrophobic residues in NF-YB and another in NF-YC (Fig 2) were delineated as NF-YB/C domains of interaction with TBP and its associated factors (Bellorini *et al.*, 1997). Furthermore, studies on the interaction of TBP-associated factors (TAF_{II}s) with NF-Y and its subunits reveal that NF-Y is capable of interacting with different combinations of the TAF_{II}s HFM protein units *in vitro* (Frontini *et al.*, 2002). NF-Y is also essential in the recruitment of RNA polymerase II and TBP in some promoters (Kabe *et al.*, 2005)

3.3. NF-Y binding to CCAAT box elements

CCAAT box is present in promoters of many gene families including tissue specific genes (Kreuter *et al.*, 1999; Tomita and Kimura, 2008), as well as housekeeping and cell cycle-regulated genes (Hu *et al.*, 2006; Kabe *et al.*, 2005; Radomska *et al.*, 1999). Studies show that in active promoters the location of the CCAAT element at the proximal promoter is invariable (exceptions are cell cycle-regulated promoters), indicating functional importance. Several mutagenesis studies of the CCAAT motif show decreased activity in many promoters (Maity and de Crombrugghe, 1998). TATA-less promoters require the CCAAT motif for efficient recruitment of the PIC (Mantovani, 1999; Steffen *et al.*, 1999; Testa *et al.*, 2005).

The sole activator of the intact CCAAT box is the ubiquitous NF-Y factor. NF-Y binds to its recognition sequence at the core of which is the CCAAT box element. The bases C, A/G, A/G on the 5' side of the CCAAT box, and C/G, A/G, G,A/C, G on the 3' side, are required for adequate binding (Mantovani, 1998). NF-Y also binds and distorts the minor groove of DNA. This binding is also assisted by the flanking nucleotides of NF-Y recognition sequence (Ronchi *et al.*, 1995).

Another protein capable of binding to the CCAAT box is the CCAAT/enhancer binding protein (C/ebp) (Osada *et al.*, 1996; Xu *et al.*, 2006). Like NF-Y, it is ubiquitous and is an important activator protein in many genes involved in growth, differentiation and various cellular responses (Ramji and Foka, 2002). But most of the *cis*-acting sequences activated by C/ebp, unlike those of the NF-Y factor, are variable and in most cases do not consist of the intact CCAAT motif. C/ebp consensus sequence is RTTGCGYAAY (Elizondo *et al.*, 2009; Osada *et al.*, 1996).

3.4. NF-Y promoters

As mentioned above, NF-Y recognizes strictly its DNA binding sequences and these sequences consist of the intact CCAAT box at the core (Mantovani, 1998). This binding sequence is present in both the forward and reverse orientation in NF-Y promoters. NF-Y promoters are usually TATA-less (Mantovani, 1998). In a study of 239 known or predicted human and mouse genes using ChIP on chip approach, Testa *et al.* showed that there is evidence that NF-Y promoters are correlated with CpG island genes (Testa *et al.*, 2005). Their data also suggested that NF-Y may play an important role in the regulation of a considerable fraction of bidirectional gene pairs in the human and mouse genomes. NF-Y type CCAAT boxes generally show position specificity within 1 kilobase (kb) upstream to the TSS (Blanchette *et al.*, 2006). However, various studies show that the functional position of NF-Y type CCAAT boxes is between -100 and -60 relative to the TSS (Mantovani, 1998; Vardhanabhuti *et al.*, 2007). On the other hand, cell cycle regulated promoters with multiple CCAAT boxes and no TATA box usually have functional CCAAT motifs at locations closer to and, sometimes, overlapping the TSS (Haugwitz *et al.*, 2002; Salsi *et al.*, 2003).

3.5. NF-Y modifications

Various isoforms of NF-YC have been isolated. The NF-YC gene generates four products, each of which differs from the other at the evolutionarily loosely conserved Q-rich subdomain. The isoforms result from two alternative promoters of the NF-YC gene and alternative splicing of exons of the Q-rich subdomain. This enables NF-Y to maintain different isoforms in different cell types with, possibly, varying levels of

activity and functions within the different cell types (Ceribelli *et al.*, 2009). This might mean that NF-Y is well adapted to its multiple roles as a ubiquitous *trans*-acting, repressing, inducing and coactivating factor.

NF-YA is a protein made of 347 amino acids (the long isoform). Seven isoforms of NF-YA have been identified, all of which are alternatively spliced with deletions at the highly evolutionarily conserved and functional domain that is glutamine- and serine-rich (Ge *et al.*, 2002). Further, among the mRNAs of the three subunits of NF-Y that of NF-YA is the only one whose level fluctuates indicating that it is the regulatory subunit of NF-Y (Manni *et al.*, 2008).

3.6. Epigenetic interactions of NF-Y

NF-YA bends DNA, presetting its 3D structure and organizes promoters. This facilitates the recruitment of PIC and binding of other factors to the promoter (Liberati *et al.*, 1998; Ronchi *et al.*, 1995). NF-Y/DNA binding correlates with or recruits histone acetyltransferase, favoring NF-Y transactivation (Gurtner *et al.*, 2008; Li *et al.*, 2009). NF-Y/DNA binding also correlates with positive transcription markers (Gurtner *et al.*, 2008).

4. Activation of bidirectional promoters by NF-Y

NF-Y is the only known DNA-binding protein to require the intact CCAAT motif for binding (Mantovani, 1998). The C/ebp family of proteins recognizes the intact CCAAT box but does not require it for binding. Their binding sites are flexible, most of which lack the intact CCAAT motif (Osada *et al.*, 1996).

The present knowledge on bidirectional promoters and NF-Y transactivation support the hypothesis that a significant number of NF-Y promoters are enriched in bidirectional promoters and that these could be crucial in the regulation of bidirectional promoters. Of all the overrepresented *cis*-regulatory elements in bidirectional promoters only the role of GABP factor has been investigated in detail (Collins *et al.*, 2007; Lin *et al.*, 2007).

NF-Y is capable of activating promoters by binding to its recognition sequences in either the forward or reverse orientation (Acosta *et al.*, 2007; Hewetson and Chilton, 2003; Mantovani, 1998). This binding sometimes involves interactions with coactivators (Huang *et al.*, 2005), transcriptional inducers (Finch *et al.*, 2001), mutually exclusive and cooperative binding with other factors (Dong *et al.*, 2006).

NF-Y type CCAAT elements are one of the most enriched in bidirectional promoters. A sequence analysis of the bidirectional promoters in human revealed a significant number of NF-Y-binding CCAAT elements (Yang and Elnitski, 2008a). Earlier studies of *cis*-regulatory elements overrepresented in the entire bidirectional promoter using various approaches consistently showed that NF-Y CCAAT boxes were one of the most overrepresented (Lin *et al.*, 2007; Testa *et al.*, 2005). The high incidence of NF-Y type CCAAT boxes suggests its important role in transactivating bidirectional promoters.

Bidirectional promoters are generally known to be TATA-less, CpG island, and consisting of multiple TSSs. There is evidence that the NF-Y promoter structure also consist of these features (Dolfini *et al.*, 2009; Testa *et al.*, 2005). All the above suggests that NF-Y may be essential in regulating bidirectional promoters.

NF-Y was experimentally shown to be necessary to enable bidirectional transcription of a pair of divergent human and mouse genes, *viz.*, mitoribosomal protein S12 (Mrps12) and mitochondrial seryl-tRNA ligase also known as Sarsm (Zanotto *et al.*, 2007;

Zanotto *et al.*, 2009). These genes share a bidirectional promoter that contains four CCAAT motifs that NF-Y interacts with. Selective binding of NF-Y to two of the CCAAT boxes in both forward and reverse orientations controlled directionality of transcription. It was suggested that the presence of the array of CCAAT boxes in the bidirectional promoter and its interaction with NF-Y guarantees effective bidirectional activation.

NF-Y has also been found to cooperate with several other factors to regulate the bidirectional promoter of the fragile X mental retardation-1 gene and an alternative albumin-binding protein (a ABP) in humans (Mahishi and Usdin, 2006). However, only a few experimental studies have been reported of NFY's role in regulating bidirectional promoters.

Earlier, we carried out a computational analysis on the incidence of NF-Y and C/ebp recognition sequences using pattern search (Zanotto *et al.*, 2009). The results revealed the preferential occurrence of NF-Y sites in bidirectional promoters above the expected level by over three folds. However, most bidirectional promoters do not have these NF-Y CCAAT boxes. It was therefore concluded that bidirectional promoters are not broadly activated by the NF-Y type CCAAT boxes.

Notwithstanding, it cannot be ruled out that the NF-Y type CCAAT boxes may represent an essential *cis*-regulatory element of the basal transcription machinery of bidirectional promoters. TATA-less promoters that consist of a CCAAT element require the CCAAT element for basal transcription (Maity and de Crombrugghe, 1998; Mantovani, 1999; Steffen *et al.*, 1999). Incidentally, bidirectional promoters are TATA-less and a revelation of many bidirectional promoters with at least one CCAAT box shall be a step forward in unraveling the transcription mechanism of these promoters.

Another possibility may be that NF-Y, along with other factors, cooperatively regulates bidirectional promoters. The synergistic activation or repression of promoters by NF-Y and other activators commonly found in bidirectional promoters (Lin *et al.*, 2007) has been reported by various studies (Huang *et al.*, 2004; Sato *et al.*, 1996; Sitwala *et al.*, 2002; Taira *et al.*, 1999; Ueda *et al.*, 1998). In all, the detail mechanism by which bidirectional promoters control the transcription of divergent genes is still being researched.

In this work I further analyzed the NF-Y binding sites that we reported previously (Zanotto *et al.*, 2009). Using sequence analysis, the occurrence, distribution and biological significance of NF-Y type bidirectional promoters were studied. This together with future research might shed more light on the mechanism of regulation of bidirectional promoters. This study represents the first study that focuses on details on the NF-Y factor and its role in transactivating bidirectional promoters in human and mouse genomes.

5. The aims of this study

This study aims at investigating the occurrence and distribution of NF-Y type CCAAT box-containing bidirectional promoters of the human and mouse genomes. The resulting data is then used to deduce the functional and biological significance of NF-Y in the transcriptional regulation of bidirectional promoters in both genomes. The investigation is meant to reveal the role of NF-Y in regulating bidirectional promoters and whether these *cis*-acting elements potentially activate both divergent genes.

6. Methods

6.1. Source of genomic annotations

Human and mouse genome annotation data were downloaded from the "knowngenes" datasets of the UCSC genome browser (Karolchik *et al.*, 2004): the human genes from the hg18 dataset and the mouse genes from mm9 dataset (http://genome.ucsc.edu/; August 30, 2009). In both datasets the UCSC gene name, chromosome name, strand, start and end loci columns of each gene annotation was retained and other fields flushed out. Gene records were sorted in order based on genomic loci in each chromosome. Since the knowngenes datasets include isoforms and alternatively spliced variants, the downloaded data were sanitized. The data were separated with respect to gene strand, and then sorted by start locus. When multiple records had the same start site the longest variant or isoform was retained and all others rejected. Because all the variants would have the same or similar promoter the retained gene record was a representative of the rest. This prevented the use of multiple copies of the same promoter when *cis*-acting promoter elements were search. Datasets were then sorted in the order they occur in each chromosome after this sanitization.

6.2. Identifying bidirectionally and unidirectionally transcribed genes

In this study, bidirectional gene pairs are defined as those arranged head-to-head to each other. The distance between the transcription start sites of each bidirectional pair was limited to less than 1 kb. Also, intergenic regions between both TSSs were constrained to non-overlapping. On the other hand, a gene was unidirectional if it was ≥1 kb from an upstream gene of the same orientation, or ≥10 kb from an oppositely oriented gene.

6.3. Bidirectional and unidirectional promoter sequences

The entire intergenic region upstream to both TSSs of each flanking head-to-head gene pair was considered as the bidirectional promoter. Unidirectional promoters were considered as 1 kb upstream to the transcription start site of each undirectional gene.

The genomic sequences of each promoter region were then downloaded from the UCSC genome browser based on the chromosomal loci of the defined promoters using UCSC genome browser's galaxy tools (Giardine *et al.*, 2005; Karolchik *et al.*, 2004).

6.4. Random promoter datasets and statistical significance

Two datasets of promoters were created through random sequence generation; a 1000 and a 10,000 random sequence datasets to represent random bidirectional and unidirectional promoters, respectively. To keep noise in the analysis to a minimum, average lengths and base frequencies of human and mouse bidirectional and unidirectional promoters were used to generate the corresponding random sequences. The length of corresponding bidirectional and unidirectional random sequences was considered to be the same as the average length of the promoter datasets: 350 bp for bidirectional and 1 kb for unidirectional. Also, base frequencies of random sequences were the same as the average base frequencies of their corresponding promoter datasets. In random bidirectional datasets A = 19.3%, C = 29.9%, G = 32.9% and T = 17.9%, while in random unidirectional datasets all the bases were equiprobable (i.e., 25% for each base). These random sequence datasets were used as a control for calculating the expected frequency of occurrence of NF-Y transcription factor binding sites and their overrepresentation in bidirectional promoters.

6.5. Search for CCAAT boxes

Two datasets of NF-Y type and one C/ebp type CCAAT boxes were created. One from simple pattern matching and the other from an integrated bioinformatics tool, as explained below.

6.5.1. Pattern search for NF-Y type CCAAT boxes

NF-Y requires the intact CCAAT box as well as flanking nucleotides on both the 5' and 3' sides of its DNA-binding domain (Bi *et al.*, 1997; Mantovani, 1998). Its consensus binding site in higher eukaryotes is YRRCCAATCA, where, Y represents C or T and R represents A or G. It also activates the CCAAT box in both oerientation: CCAAT in the forward orientation and ATTGG in the reverse orientation. The sequence yrrCCAATca

was used to search NF-Y type CCAAT boxes in both bidirectional and unidirectional promoter sequences. Lowercase letters in the sequence indicate the nucleotides that were allowed to have mismatches. The search result consisted of sites with a perfect match, as well as, one, two and three mismatches. For the purpose of consistency and clarity these datasets were named NF-Y1.

6.5.2. Search for NF-Y type CCAAT boxes using Genomatix resources

Since *cis*-regulatory sequences found using simple pattern matches, as above, include false positives (Wasserman and Sandelin, 2004), I decided to create another NF-Y putative binding sites with further functional lines of evidence.

The promoter sequences were loaded unto the Genomatix server (Cartharius *et al.*, 2005; Quandt *et al.*, 1995). MatInspector intergrated algorithms were used to search the promoters for the NF-Y type veterbrate family matrix libraries in Matbase (Cartharius *et al.*, 2005). Matches retained were those that have been characterized as fitting at least one functional model, created using comparative genomics and Genomatix's ModelInspector, FrameWorker, and Bibliosphere (Cartharius *et al.*, 2005; Frech *et al.*, 1997; Scherf *et al.*, 2005). These datasets were named NF-Y2.

NF-Y type CCAAT boxes were searched on each promoter sequence in both the forward and reverse orientations for both the NF-Y1 and NF-Y2 datasets due to their inherent bidirectional activation of the CCAAT box. The functional position specificity of NF-Y type CCAAT boxes was -100 to -50 bp with respect to the TSS.

Together with the observed frequencies of occurrence of NF-Y type CCAAT boxes, the expected frequencies obtained from the random background datasets were used to compute chi-square statistic. The corresponding P-value was used to test the hypothesis that NF-Y type CCAAT boxes in the promoter sets occur only by chance.

A method that uses approximation of binomial to normal distribution was used to compute z-scores that in turn were used to determine overrepresentation of NF-Y type CCAAT boxes in bidirectional promoters (Ho Sui *et al.*, 2005). Z-scores less than or equal to -2 and more than or equal to +2 were considered statistically significantly overrepresented.

6.5.3. Pattern search for C/ebp type CCAAT boxes

C/ebp family of transcription factors (TFs) are also CCAAT box activators and their consensus binding site is the RTTGCGYAAY sequence (Osada *et al.*, 1996). To use a more inclusive recognition sequence based on the data from Osada and colleagues, the recognition sequence for C/ebp TFs was adjusted to VTTGCGYAAY, where V represents A or C or G (Osada *et al.*, 1996). Also, to carry out a representative occurrence of CCAAT type binding sites, the recognition sequences of both factors were adjusted to maintain the intact CCAAT motif. The TFBS consensus for C/ebp was then constrained to include the intact CCAAT box, *viz.*, VTTRCCCAAT sequence. Furthermore, vTTrCCCAAT was used to search C/ebp type CCAAT boxes. Lowercase letters in the sequence indicate the nucleotides allowed to have mismatches. These adjustments would still reflect the functional sites of the C/ebp family due to their ability to flexibly recognize their DNA-binding domains. Also, nucleotide bases out of the intact CCAAT region and required for binding (i.e., the TT at -3 and -4 with respect to the CCAAT motif) were retained. C/ebp TFBSs were searched from promoters in only the forward orientations due to C/ebp's inability to recognize its binding sites in the reverse orientation.

6.6. NF-Y TFBSs spacing and distribution analysis

A tabulated frequency data of occurence of all NF-Y type CCAAT boxes from NF-Y2 datasets were plotted on histograms to analyze their distribution and position specificity. The spacing between matches of NF-Y type CCCAAT boxes from NF-Y2 datasets was also plotted on histograms.

6.7. CpG islands, C+G content, TATA boxes and NF-Y in bidirectional promoters

The CpG islands annotated data was downloaded from the UCSC genome browser (http://main.g2.bx.psu.edu/; January 16, 2010). Files of both the CpG island data and promoter data were created. Both files were searched for intersecting chromosome loci. Intersecting records in the promoter datasets were considered as genes whose promoter

consists of a CpG island. The C+G content and CpG island content of each class of promoter was also computed. [C or G or A] TATA [T or A][T or A] and [C or G or T] TATA [T or A] [T or A] sequences were used to search for TATA boxes at all positions in each promoter.

7. Results

In this report a set of NF-Y type promoters in both bidirectional and unidirectional datasets were created. With these data I intended to carry out detailed study of structural features, including positional and spatial specificity of NF-Y type bidirectional promoters and the incidence of TATA boxes and CpG islands. The aim was to demonstrate whether NF-Y controls basal transcription and potentially transactivates both divergent gene pairs flanking bidirectional promoters.

7.1. Characterizing bidirectional and unidirectional datasets

A total of 66804 and 49410 gene records were downloaded from the human and mouse knowngenes datasets of the UCSC genome browser, respectively. The data was then sanitized (section 6.2), resulting in 43854 gene records for human and 36958 for mouse. Further characterization of each gene into bidirectional or unidirectional was based on these datasets.

Based on the definitions of bidirectional and unidirectional genes datasets were created consisting of 3876 bidirectional genes (i.e., 1717 gene pairs) from human and 2892 from mouse (i.e., 1446 gene pairs). On the other hand, the sanitized data yielded 17918 and 14138 unidirectional genes for human and mouse, respectively.

7.2. NF-Y type CCAAT boxes

The search for NF-Y binding sites was done within entire promoter sequences downloaded from the UCSC genome browser. NF-Y recognition sequences without mismatches were statistically over three folds than would be expected to occur by chance in bidirectional promoters. This result was based on observed frequency of occurrence of NF-Y type CCAAT boxes against the random sequence dataset (Table 1). Also, using the binomial approximation of the normal distribution (section 6.5.2), NF-Y type CCAAT boxes were statistically overrepresented in the NF-Y2 bidirectional promoter datasets (z-scores > +2.30). The proportion of bidirectional promoters that consisted of at least an NF-Y site in the NF-Y1 datasets -including mismatches- represented 34.83 % and 21.72% in human and mouse genomes, respectively (Table 1, regular boldface).

Table 1. **Frequency of CCAAT boxes in the NF-Y1 datasets, in % (2 dp). bi** represents bidirectional promoters and **uni** represents unidirectional promoters

Search condition	Human		Mouse	
	bi	**uni**	**Bi**	**uni**
>= 1 NF-Y[a] CCAAT site, 0 mismatch	***4.78***[c]	2.5	***7.81***	3.76
>= 1 NF-Y CCAAT site, <= 1 mismatch	16.83	13.51	21.16	15.67
>= 1 NF-Y site, <= 2 mismatches	27.66	36.93	32.43	38.82
>= 1 NF-Y site, <= 3 mismatches	**34.83**[d]	58.53	**21.72**	61.77
>= 2 NF-Y sites, <= 1 mismatch	4.54	1.92	6.57	2.26
>= 2 NF-Y sites, <= 2 mismatches	10.77	10.03	13.07	10.65
>= 2 NF-Y sites, <= 3 mismatches	16.19	25.03	19.5	27.43
>= 1 C/EBP[b] site, <= 3 mismatches	20.73	37.36	21.72	40.11
>= 1 CCAAT box of each , <= 3 mismatches	18.87	32.89	20.75	35.1
>= 2 NF-Y and 1 C/ebp sites, <= 3 mismatches	10.83	17.82	13.14	19.49

a. yrrCCAATca was used to search for NF-Y sites; lowercase bases were allowed to have mismatches.
b. vTTrCCCAAT was used to search for C/ebp sites; lowercase bases were allowed to have mismatches.
c. three-fold statistically significantly different from random expectation of NF-Y sites without mismatches. $\chi 2$ test; P-value < 0.0001; 1 df.
d. frequency of bidirectional promoters with at least one NF-Y site; 3 mismatches allowed

In the NF-Y2 datasets, 31.14% and 35.75% of human and mouse bidirectional promoters, respectively, consisted of at least an NF-Y binding site (Table 2).

Table 2. **NF-Y2 data: Frequency of NF-Y binding sites, in % (2 dp).**

Search condition	Human		Mouse	
	bi	**Uni**	**bi**	**Uni**
≥ 1 NF-Y sites	**31.14**[a]	30.68	**35.75**	35.08
1 NF-Y sites	18.54	22.89	21.72	73.99
≥ 2 NF-Y sites	12.59	7.79	14.04	35.15
2 NF-Y sites	7.52	5.9	8.85	75.81
NF-Y sites in -100 and -50 bp	***17.14***[b]	20.29	***20.68***	341.1
≥ 1 NF-Y sites in -100 and -50 bp	15.57	17.91	17.98	87.86
2 NF-Y sites in -100 and -50 bp	1.57	2.03	2.42	7.47
≥ 2 NF-Y sites in -100 and -50 bp	1.57	2.19	2.56	163.01

a. Number of promoters with at least an NF-Y binding site
b. Number of NF-Y binding sites that occur at NF-Y functional position of -100 and -50 bp with respect to the TSS

In bidirectional promoters, 24.9% in human and 19.7% in mouse occurred at their functional positions of -100 to -50 in the NF-Y1 dataset. Similarly, 17.14% and 20.68%

of human and mouse bidirectional promoters, respectively, in the NF-Y2 datasets occurred at functional position (Fig. 3). In the NF-Y1 datasets 20.73% of human and 21.72% of mouse bidirectional promoters consisted of at least one C/ebp type CCAAT box.

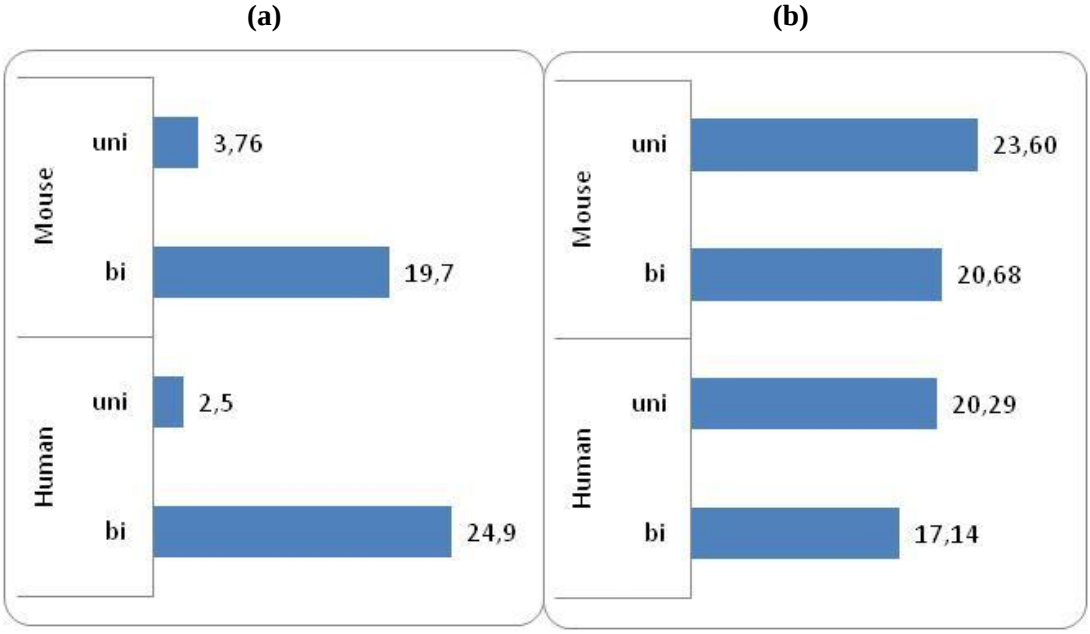

Figure 3. **Percentage occurrence of NF-Y binding sites at their functional position. (a)** NF-Y1 dataset. **(b)** NF-Y2 datasets

7.3. Orientation and position specificity of NF-Y binding sites

Orientation and positional specificity is important for *cis*-regulatory elements to control transcription in promoters. The percentage of forward and reverse oriented NF-Y binding sites was similar in all NF-Y2 promoter categories (Fig. 4). NF-Y sites were found in similar proportions on both sides of the bidirectional promoter center-line (Fig. 5). There were 46.98% and 42.76% of NF-Y sites in human bidirectional promoters located within 100 bp of the TSS of the minus and positive strand genes, respectively. In mouse, 46.49% and 43.68% of NF-Y sites in bidirectional promoters were located within 100 bp of the minus and positive strand genes, respectively. Hence, about half of the NF-Y sites in bidirectional promoter datasets were located beyond 100 bp of both TSSs.

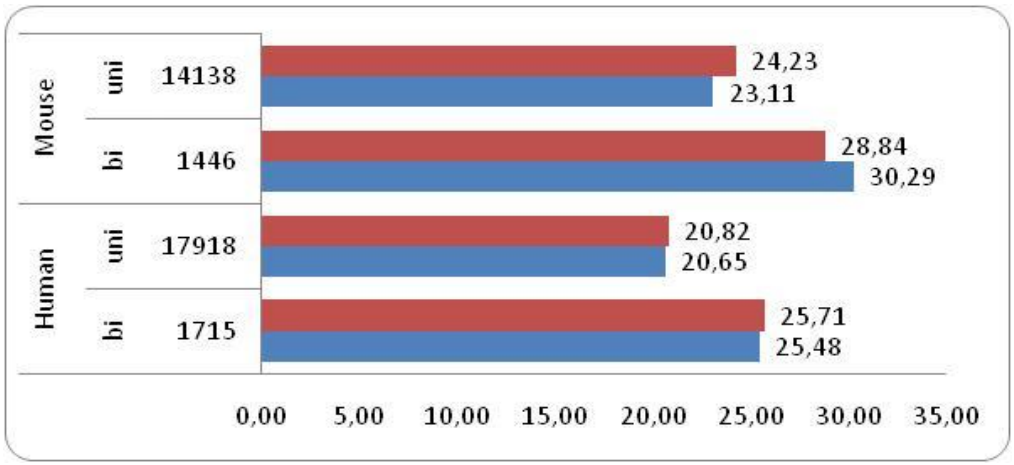

Figure 4. Orientation of NF-Y binding sites in the NF-Y2 datasets. The bars colored in orange are reverse strand, and those colored blue are forward strand NF-Y type CCAAT boxes. This shows a similar distribution of the forward and reverse NF-Y binding sites in all datasets. All values are in percent and given to 2 dp.

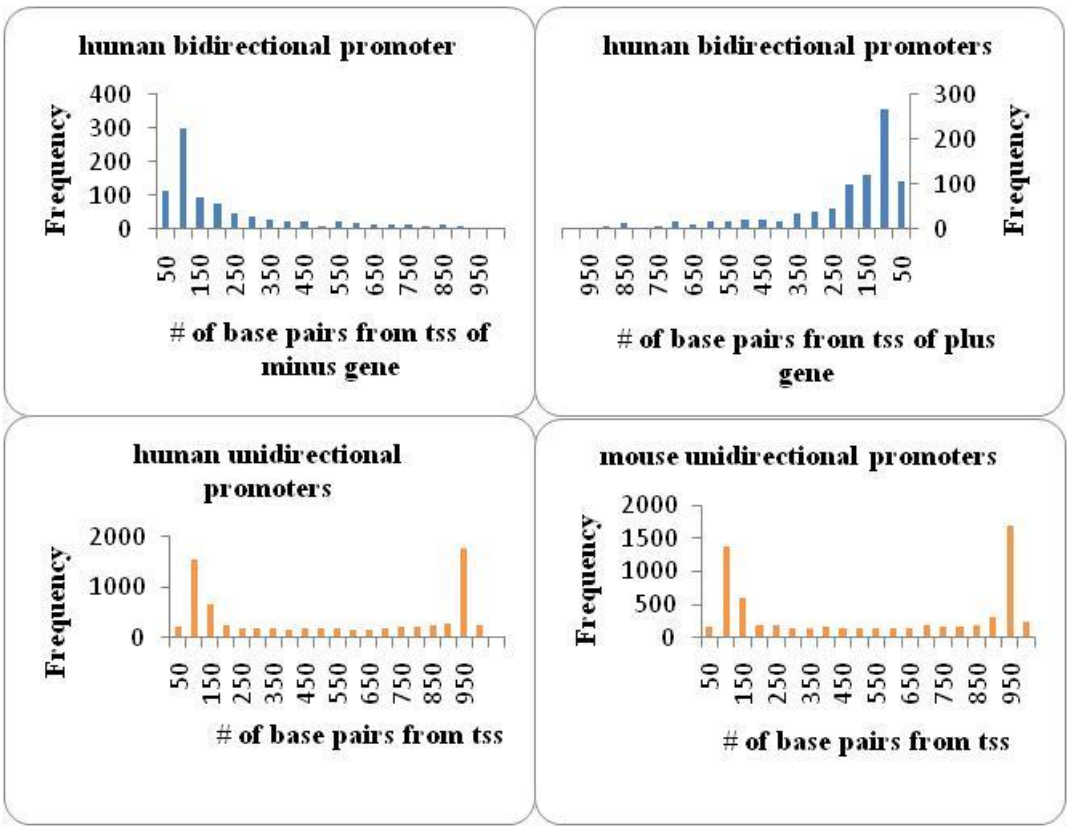

Figure 5. Distribution of NF-Y binding sites with respect to the TSS in NF-Y2 dataset. The bidirectional promoter histograms indicate the most prominent peak at NF-Y's functional position (-100 to -51 bp from the TSSs). The unidirectional promoter histograms in orange, below, show two main peaks; one at NF-Y's functional position and the other between -950 and -900 from the TSS.

Fig. 5 also shows two clearly demarcated and dominating peaks in the unidirectional promoter datasets; between -150 to -100 bp and between -950 and 900 bp from the TSS. NF-Y binding sites were approximately equi-distributed over the rest of the unidirectional promoters.

I further investigated the spatial specificity of promoters with multiple NF-Y binding sites. In all promoter categories, most NF-Y binding sites were located within 100 bp of each other (Fig. 6).

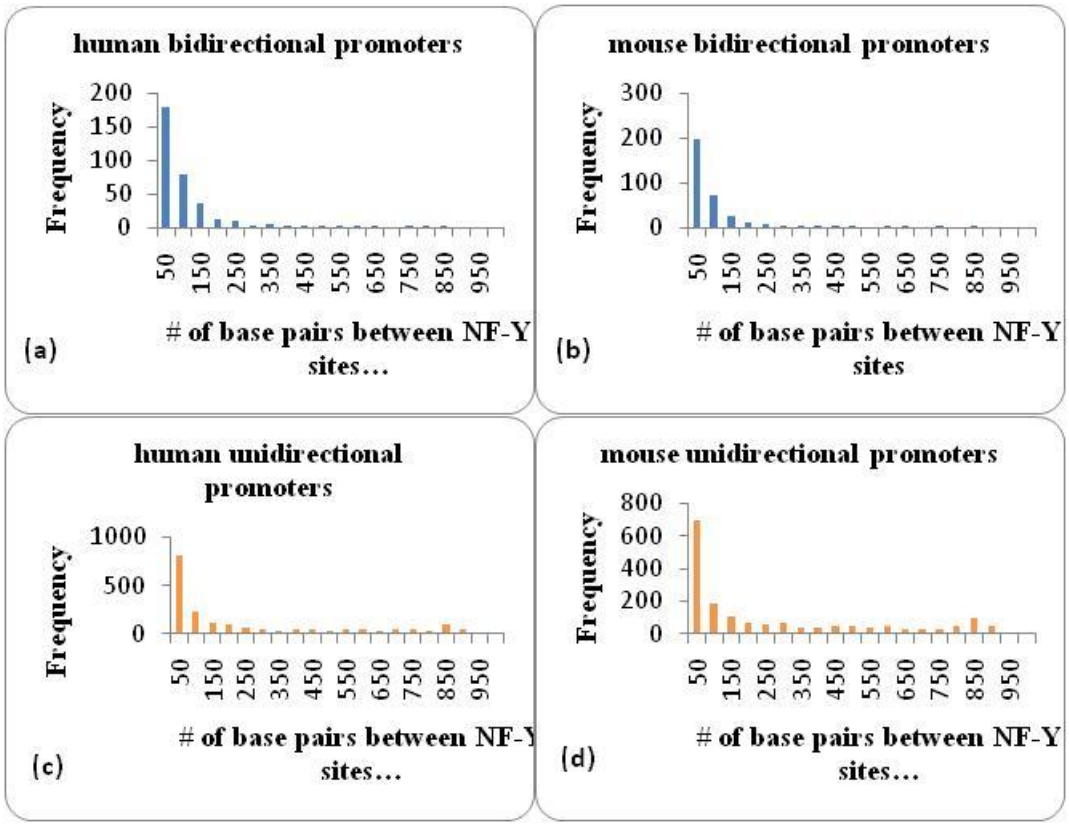

Figure 6. Spatial specificity of multiple NF-Y binding sites in the NF-Y2 datasets. All datasets show that the spacing of most pairs of NF-Y binding sites were within 100 bp

7.4. NF-Y binding sites and bidirectional promoters

Here, I probed the structure of bidirectional promoters in the NF-Y2 datasets that consist of at least an NF-Y binding site with respect to TATA boxes, C+G content and CpG islands. First, the distribution of TATA boxes, CpG islands and C+G content in all

promoter categories was examined. Promoters of each type were searched for the occurrence of at least one TATA box. In the bidirectional promoters, 7.51% and 9.06% consisted of at least a TATA box in human and mouse, respectively (χ^2 = 2.01; P-value > 0.05; 1 df). In contrast, TATA boxes were enriched by 60.63% in human, and 64.27% in mouse unidirectional promoters (Table 3).

Table 3. Frequency of TATA boxes, CpG islands and C+G content in percent (2 dp)

Search condition	Human		Mouse	
	Bi	uni	bi	Uni
TATA	7.51	60.83	9.06	64.27
CpG islands	85.79	32.56	82.02	30.24
C+G content	63.81	48.95	61.34	47.22

The C+G content in bidirectional promoters was 63.81% and 61.34%, whereas, those in unidirectional promoters were 48.95% and 47.22% in human and mouse, respectively. Also, 85.79% and 32.56% bidirectional and unidirectional promoters were co-located within CpG islands in the human datasets and 82.02% and 32.24% in the mouse datasets.

I then investigated the distribution of TATA boxes and CpG islands in bidirectional promoters that consist of at least one NF-Y binding site. Bidirectional promoters that consisted of at least an NF-Y binding site, lacked TATA box and are located within a CpG island represented 80% in human and 74% in mouse datasets (Fig. 7).

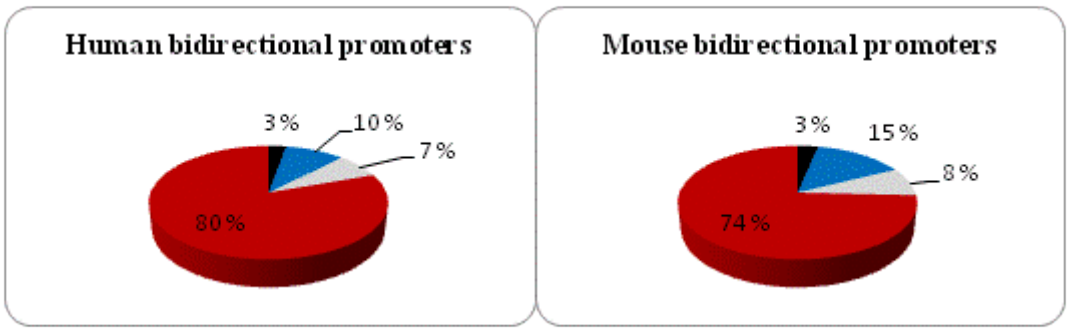

Figure 7. **Co-occurrence of NF-Y sites, TATA boxes and CpG islands in bidirectional promoters**

8. Discussion

Using sequence analyses, I studied the incidence and distribution of NF-Y type CCAAT boxes in bidirectional promoters. To make predictions from these analyses, two NF-Y type CCAAT box datasets were created: one from simple pattern search and the other using Genomatix's MatInspector integrated tools (see section 6.5). I demonstrated, through these sequence analyses, that NF-Y type CCAAT box-containing bidirectional promoters occur preferentially and extensively. This and other findings suggest that NF-Y could potentially be central in bidirectional promoter regulatory mechanism.

In this study divergent gene pairs whose TSSs overlapped were excluded. In cases where a gene had alternative TSSs, the promoter whose TSS is upstream relative to the others was retained. By this approach, only one TSS was used as representative of all alternatives for each promoter locus. Although there might be NF-Y type CCAAT boxes downstream to the TSSs only those found upstream to the TSSs were analyzed; most functional NF-Y sites are located upstream to -49 bp from the TSS with the exception of cell cycle-regulated genes (Mantovani, 1998). Also, NF-Y sites show preference within 1 kb of the TSS (Blanchette *et al.*, 2006).

The three folds over expectation of NF-Y sites without mismatches in bidirectional promoters of the NF-Y1 datasets coupled with its statistical overrepresentation in NF-Y2 datasets demonstrates that NF-Y is critical in regulating bidirectional promoters. Over a third of bidirectional promoters in NF-Y2 datasets consisted of at least one NF-Y site (Table 2). Previous studies show that CCAAT boxes occur in a substantial proportion of human gene promoters (Bucher, 1990; Suzuki *et al.*, 2001). This is consistent with the results in Table 2, in which the proportion of NF-Y sites in each promoter type is about one-third of the total. A previous study found CCAAT boxes to be one of the most overrepresented *cis*-regulatory elements in bidirectional promoters (Lin *et al.*, 2007), although the exact proportion was not reported. However, another study found 12.9% of bidirectional promoters consist of at least one CCAAT box within -75 to -80 base pairs of the TSSs in a human bidirectional promoter set and 6.9 % in unidirectional promoters (Yang and Elnitski, 2008a). The difference could be as a result of different NF-Y putative recognition sequences and different methods. Yang and Elnitski used DSWYVY, where D represents [A or G or T]; S represents [C or G]; W

represents [A or T]; Y represents [C or T] and V represents [A or C or G] (Yang and Elnitski, 2008a). On the other hand, two approaches were used in this investigation: pattern search using the yrrCCAATca sequence as putative NF-Y binding sequence and an integrated approach that uses matrices and other evidence to characterize functional sites (see section 6.5). Thus, NF-Y2 datasets (Table 2) created using the integrated approach is a fair representation of potential functional NF-Y sites. Chromatin immunoprecipation (ChIP) on chip assays would be needed to confirm which of the potential functional sites exhibit functionality *in vivo.* Nevertheless, the data in Table 2 suggests a highly significant occurrence and role for NF-Y in regulating bidirectional promoters.

In contrast to earlier report, the distribution of NF-Y sites of forward orientation was similar to those of the reverse orientation in all datasets (Fig. 4). In a study of 178 human CCAAT promoters, 60% of CCAAT boxes were found in the forward and 40% in the reverse orientation (Mantovani, 1998). The differences between these results could be due to the size of datasets. Notwithstanding, the result of a similar number of forward and reverse NF-Y type CCAAT boxes is in accordance with its inherent bidirectional DNA-binding nature.

The location of a considerable number of NF-Y sites within its functional position with respect to both TSSs of bidirectional promoters (Table 2) shows that these sites could be critical to the basal transcription machinery of bidirectional promoters. A similar result was obtained previously (Lin *et al.*, 2007). *Cis*-regulatory elements that occur at positions other than their functional sites, but within the promoter, are likely connected to the transcription of both divergent gene pairs (Lin *et al.*, 2007). The distribution of significant NF-Y sites elsewhere than their functional position is in agreement with this model. Also, NF-Y is capable of activating the CCAAT box in both orientations. Taken together, these observations demonstrate that NF-Y sites that occur within their functional position could be involved in controlling basal transcription. On the other hand, NF-Y sites upstream to their functional position could potentially be associated with transactivating both flanking genes.

Like bidirectional promoters, unidirectional promoters show a significant number of NF-Y sites within 150 bp to the TSS. Interestingly, another prominent peak was found

between -901 to -950 bp from the TSS (Fig. 5). It is not clear why this biphasic nature exist in the distribution of NF-Y sites in unidirectional promoters. One possible hypothesis could be that these sites represent enhancer regions (Yu *et al.*, 2005). Further analysis by ChIP assay is needed to shed light on the role of NF-Y sites beyond its functional position in both promoter types.

NF-Y and C/ebp TFs are capable of cooperatively activating promoters (Milos and Zaret, 1992; Xu *et al.*, 2006). The over 20% incidence of C/ebp binding sites in all bidirectional promoters (Table 1), coupled with over 30% occurrence of NF-Y sites does not rule out cooperative or mutually exclusive control of a considerable proportion of these bidirectional promoters by both TFs (Zanotto *et al.*, 2009). Although NF-Y and C/ebp activation is possible, more experiments targeting bidirectional promoters with the recognition sequences of both factors could be interesting.

Most NF-Y promoters consist of a single NF-Y type CCAAT box (Salsi *et al.*, 2003), which is consistent with the results in Fig 6. There was no significant difference in the incidence and distribution of multiple NF-Y sites in both bidirectional and unidirectional promoters of both mouse and human genome. Most genes with multiple CCAAT boxes are cell cycle-regulated. In the promoters of the cell cycle-regulated genes the functional CCAAT boxes are often located within 31 or 32 bp apart from each other. This conformation permits the CCAAT boxes to bind on the same side of the DNA double strand and interact with each other, with or without the help of cofactors in a time-dependent manner (Salsi *et al.*, 2003). The similar distribution of spacing between NF-Y sites (Fig. 6) indicates that this mechanism could well be the same for both bidirectional and unidirectional promoters.

Bidirectional promoters are generally located within CpG islands and lack a TATA-box. Most bidirectional promoters that consisted of at least one NF-Y site occurred in a CpG island and are TATA-less (Fig. 7). This strongly supports the above mentioned proposal that NF-Y may be crucial in the underlying regulatory mechanism of bidirectional promoters.

NF-Y binds to the minor (Ronchi *et al.*, 1995) and major groove of DNA, bending and distorting the 3D structure of DNA (Coustry *et al.*, 2001), and leading to the initiation and promotion of transcription (Matuoka and Chen, 2002). Genes with bidirectional

promoters are readily transcribed, have open chromatin structure and consist of positive histone motifs (Lin *et al.*, 2007). NF-Y may be essential in keeping the chromatin structure open by binding, bending and distorting it, and therefore enabling bidirectional promoters to be broadly and generally transcribed. NF-Y also serves a dual function, associating with both positive and negative histone markers. It will be interesting to probe the correlation between the NF-Y sites and positive histone markers using ChIP-on-chip assays in bidirectional promoters.

NF-Y also recruits the PIC by interacting with TBP and TBP-associated factors (Frontini *et al.*, 2002). The symmetrical and extensive distribution of NF-Y sites about the mid-line and close to both TSSs of bidirectional promoters, the paucity of TATA boxes in bidirectional promoters and the requirement of CCAAT elements by TATA-less promoters to drive basal transcription underpins the essential role of NF-Y in regulating basal transcription in bidirectional promoters.

The roles played by NF-Y as a general, inducible and cell cycle or context-dependent regulatory factor are well understood. I propose that the considerable number of NF-Y sites in bidirectional promoters, as demonstrated here, could be crucial in the basal regulation of bidirectional promoters. Also, the significant proportion of NF-Y sites beyond the functional position could potentially be associated with transactivating both divergent genes. However, NF-Y's role in activating both divergent genes of bidirectional promoters is still unclear. Further investigation using ChIP assay or the study of more NF-Y type CCAAT-containing bidirectional promoters using electrophoretic mobility shift assay (EMSA) is required to reveal NF-Y's effect on the regulatory mechanism of bidirectional promoters.

The activation of CCAAT box by NF-Y facilitates the recruitment of other transcription factors (Iwano *et al.*, 2001; Yamada *et al.*, 2000; Yu and Luo, 2009). It is possible that in NF-Y type CCAAT box-containing bidirectional promoters, NF-Y activation of CCAAT box enhances the binding of other factors, which in turn, modulates the expression of both the divergent genes. Transcription factors whose *cis*-acting elements are enriched in bidirectional promoters are of particular interest. Among them are the recognition sequences for ELK1, GABP, SP1, and CCAAT-boxes (Lin *et al.*, 2007). GABP is the only TF whose role in transactivating bidirectional promoters has been investigated in

detail. It was shown that GABP regulates most bidirectional promoters (Collins *et al.*, 2007; Lin *et al.*, 2007). The *cis*-acting elements of the bidirectional promoter of mrps12 and sarsm divergent gene pair in mouse and human, includes four NF-Y CCAAT boxes, a GABP site and an AP-1 site. Although studies on these bidirectional promoters did not show clear evidence of cooperativity of NF-Y and GABP, the results do not rule this out either (Zanotto *et al.*, 2009). However, very few NF-Y/GABP bidirectional promoters have been experimentally studied so far. It will be interesting to perform further experimental work on NF-Y/GABP-containing bidirectional promoters to reveal any cooperativity that may exist in their regulation of bidirectional promoters. Nonetheless, it is very plausible that NF-Y and GABP could exhibit cooperative activation given their extensive distribution and modes of activation in bidirectional promoters.

9. Conclusions

In this thesis, using sequence analysis methods, NF-Y type CCAAT boxes were shown to be significantly overrepresented in bidirectional promoters. This showed a critical role of NF-Y in the underlying bidirectional promoter regulation mechanism. Forward and reverse orientations of NF-Y type CCAAT boxes occurred in similar proportions in both bidirectional and unidirectional promoters, demonstrating NF-Y's ability to bind its recognition sequence in either orientation.

A considerable number of NF-Y type CCAAT boxes were found in their functional position in bidirectional promoters, implicating NF-Y in recruiting the PIC and controlling basal transcription. However, NF-Y type CCAAT boxes were also significantly distributed beyond their functional position, suggesting that NF-Y is potentially connected to transactivating bidirectional promoters. These findings, taken together, are important contributions in understanding the regulatory mechanism of bidirectional promoters and their subsequent biomedical applications. It would also be essential to investigate the role played by NF-Y in conferring an active chromatin structure to bidirectional promoters, as evidenced in NF-Y promoters (Donati *et al.*, 2008) and bidirectional promoters (Lin *et al.*, 2007).

10. References

Acosta, A., Zarinan, T., Macias, H., Pasapera, A.M., Perez-Solis, M.A., Olivares, A., Ulloa-Aguirre, A., and Gutierrez-Sagal, R. (2007). Regulation of Clara cell secretory protein gene expression by the CCAAT-binding factor NF-Y. Arch. Biochem. Biophys. *459*, 33-39.

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

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

Arents, G., and Moudrianakis, E.N. (1995). The histone fold: a ubiquitous architectural motif utilized in DNA compaction and protein dimerization. Proc. Natl. Acad. Sci. U. S. A. *92,* 11170-11174.

Baxevanis, A.D., Arents, G., Moudrianakis, E.N., and Landsman, D. (1995). A variety of DNA-binding and multimeric proteins contain the histone fold motif. Nucleic Acids Res. *23,* 2685-2691.

Bellorini, M., Dantonel, J.C., Yoon, J.B., Roeder, R.G., Tora, L., and Mantovani, R. (1996). The major histocompatibility complex class II Ea promoter requires TFIID binding to an initiator sequence. Mol. Cell. Biol. *16,* 503-512.

Bellorini, M., Lee, D.K., Dantonel, J.C., Zemzoumi, K., Roeder, R.G., Tora, L., and Mantovani, R. (1997). CCAAT binding NF-Y-TBP interactions: NF-YB and NF-YC require short domains adjacent to their histone fold motifs for association with TBP basic residues. Nucleic Acids Res. *25*, 2174-2181.

Bi, W., Wu, L., Coustry, F., de Crombrugghe, B., and Maity, S.N. (1997). DNA binding specificity of the CCAAT-binding factor CBF/NF-Y. J. Biol. Chem. *272,* 26562-26572.

Blanchette, M., Bataille, A.R., Chen, X., Poitras, C., Laganiere, J., Lefebvre, C., Deblois, G., Giguere, V., Ferretti, V., Bergeron, D., Coulombe, B., and Robert, F. (2006). Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. Genome Res. *16,* 656-668.

Bucher, P. (1990). Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. J. Mol. Biol. *212,* 563-578.

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

Ceribelli, M., Benatti, P., Imbriano, C., and Mantovani, R. (2009). NF-YC complexity is generated by dual promoters and alternative splicings. J. Biol. Chem.

Collins, P.J., Kobayashi, Y., Nguyen, L., Trinklein, N.D., and Myers, R.M. (2007). The ets-related transcription factor GABP directs bidirectional transcription. PLoS Genet. *3,* e208.

Coustry, F., Hu, Q., de Crombrugghe, B., and Maity, S.N. (2001). CBF/NF-Y functions both in nucleosomal disruption and transcription activation of the chromatin-assembled topoisomerase II alpha promoter. Transcription activation by CBF/NF-Y in chromatin is dependent on the promoter structure. J. Biol. Chem. *276,* 40621-40630.

Coustry, F., Maity, S.N., Sinha, S., and de Crombrugghe, B. (1996). The transcriptional activity of the CCAAT-binding factor CBF is mediated by two distinct activation domains, one in the CBF-B subunit and the other in the CBF-C subunit. J. Biol. Chem. *271,* 14485-14491.

Coustry, F., Sinha, S., Maity, S.N., and Crombrugghe, B. (1998). The two activation domains of the CCAAT-binding factor CBF interact with the dTAFII110 component of the Drosophila TFIID complex. Biochem. J. *331 (Pt 1),* 291-297.

Dhadi, S.R., Krom, N., and Ramakrishna, W. (2009). Genome-wide comparative analysis of putative bidirectional promoters from rice, Arabidopsis and Populus. Gene *429*, 65-73.

Dion, M.F., Altschuler, S.J., Wu, L.F., and Rando, O.J. (2005). Genomic characterization reveals a simple histone H4 acetylation code. Proc. Natl. Acad. Sci. U. S. A. *102*, 5501-5506.

Dolfini, D., Zambelli, F., Pavesi, G., and Mantovani, R. (2009). A perspective of promoter architecture from the CCAAT box. Cell. Cycle *8*,

Donati, G., Gatta, R., Dolfini, D., Fossati, A., Ceribelli, M., and Mantovani, R. (2008). An NF-Y-dependent switch of positive and negative histone methyl marks on CCAAT promoters. PLoS ONE *3*, e2066.

Dong, S., Kanno, T., Yamaki, A., Kojima, T., Shiraiwa, M., Kawada, A., Mechin, M.C., Chavanas, S., Serre, G., Simon, M., and Takahara, H. (2006). NF-Y and Sp1/Sp3 are involved in the transcriptional regulation of the peptidylarginine deiminase type III gene (PADI3) in human keratinocytes. Biochem. J. *397*, 449-459.

Elizondo, G., Medina-Diaz, I.M., Cruz, R., Gonzalez, F.J., and Vega, L. (2009). Retinoic acid modulates retinaldehyde dehydrogenase 1 gene expression through the induction of GADD153-C/EBPbeta interaction. Biochem. Pharmacol. *77*, 248-257.

Engström, P.G., Suzuki, H., Ninomiya, N., Akalin, A., Sessa, L., Lavorgna, G., Brozzi, A., Luzi, L., Tan, S.L., Yang, L., *et al.* (2006). Complex Loci in human and mouse genomes. PLoS Genet. *2*, e47.

Farre, D., Bellora, N., Mularoni, L., Messeguer, X., and Alba, M.M. (2007). Housekeeping genes tend to show reduced upstream sequence conservation. Genome Biol. *8*, R140.

Finch, J.S., Rosenberger, S.F., Martinez, J.D., and Bowden, G.T. (2001). Okadaic acid induces transcription of junB through a CCAAT box and NF-Y. Gene *267*, 135-144.

Firth, A.E., and Brown, C.M. (2006). Detecting overlapping coding sequences in virus genomes. BMC Bioinformatics *7*, 75.

Franck, E., Hulsen, T., Huynen, M.A., de Jong, W.W., Lubsen, N.H., and Madsen, O. (2008). Evolution of closely linked gene pairs in vertebrate genomes. Mol. Biol. Evol. *25*, 1909-1921.

Frech, K., Danescu-Mayer, J., and Werner, T. (1997). A novel method to develop highly specific models for regulatory units detects a new LTR in GenBank which contains a functional promoter. J. Mol. Biol. *270*, 674-687.

Frontini, M., Imbriano, C., diSilvio, A., Bell, B., Bogni, A., Romier, C., Moras, D., Tora, L., Davidson, I., and Mantovani, R. (2002). NF-Y recruitment of TFIID, multiple interactions with histone fold TAF(II)s. J. Biol. Chem. *277*, 5841-5848.

Ge, Y., Jensen, T.L., Matherly, L.H., and Taub, J.W. (2002). Synergistic regulation of human cystathionine-β-synthase-1b promoter by transcription factors NF-YA isoforms and Sp1. Biochimica Et Biophysica Acta (BBA) - Gene Structure and Expression *1579*, 73-80.

Giardine, B., Riemer, C., Hardison, R.C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., *et al.* (2005). Galaxy: a platform for interactive large-scale genome analysis. Genome Res. *15*, 1451-1455.

Gurtner, A., Fuschi, P., Magi, F., Colussi, C., Gaetano, C., Dobbelstein, M., Sacchi, A., and Piaggio, G. (2008). NF-Y dependent epigenetic modifications discriminate between proliferating and postmitotic tissue. PLoS One *3*, e2047.

Hatamochi, A., Golumbek, P.T., van Schaftingen, E., and de Crombrugghe, B. (1988). A CCAAT DNA binding factor consisting of two different components that are both required for DNA binding. J. Biol. Chem. *263*, 5940-5947.

Haugwitz, U., Wasner, M., Wiedmann, M., Spiesbach, K., Rother, K., Mossner, J., and Engeland, K. (2002). A single cell cycle genes homology region (CHR) controls cell

cycle-dependent transcription of the cdc25C phosphatase gene and is able to cooperate with E2F or Sp1/3 sites. Nucleic Acids Res. *30,* 1967-1976.

Hewetson, A., and Chilton, B.S. (2003). An Sp1-NF-Y/progesterone receptor DNA binding-dependent mechanism regulates progesterone-induced transcriptional activation of the rabbit RUSH/SMARCA3 gene. J. Biol. Chem. *278,* 40177-40185.

Ho Sui, S.J., Mortimer, J.R., Arenillas, D.J., Brumm, J., Walsh, C.J., Kennedy, B.P., and Wasserman, W.W. (2005). oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes. Nucleic Acids Res. *33,* 3154-3164.

Hu, Q., Lu, J.F., Luo, R., Sen, S., and Maity, S.N. (2006). Inhibition of CBF/NF-Y mediated transcription activation arrests cells at G2/M phase and suppresses expression of genes activated at G2/M phase of the cell cycle. Nucleic Acids Res. *34,* 6272-6285.

Huang, D.Y., Kuo, Y.Y., Lai, J.S., Suzuki, Y., Sugano, S., and Chang, Z.F. (2004). GATA-1 and NF-Y cooperate to mediate erythroid-specific transcription of Gfi-1B gene. Nucleic Acids Res. *32,* 3935-3946.

Huang, W., Zhao, S., Ammanamanchi, S., Brattain, M., Venkatasubbarao, K., and Freeman, J.W. (2005). Trichostatin A induces transforming growth factor beta type II receptor promoter activity and acetylation of Sp1 by recruitment of PCAF/p300 to a Sp1.NF-Y complex. J. Biol. Chem. *280,* 10047-10054.

Iwano, S., Saito, T., Takahashi, Y., Fujita, K., and Kamataki, T. (2001). Cooperative regulation of CYP3A5 gene transcription by NF-Y and Sp family members. Biochem. Biophys. Res. Commun. *286,* 55-60.

Jiang, C., and Pugh, B.F. (2009). Nucleosome positioning and gene regulation: advances through genomics. Nat. Rev. Genet. *10,* 161-172.

Kabe, Y., Yamada, J., Uga, H., Yamaguchi, Y., Wada, T., and Handa, H. (2005). NF-Y is essential for the recruitment of RNA polymerase II and inducible transcription of several CCAAT box-containing genes. Mol. Cell. Biol. *25,* 512-522.

Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D., and Kent, W.J. (2004). The UCSC Table Browser data retrieval tool. Nucleic Acids Res. *32*, D493-6.

Kim, I.S., Sinha, S., de Crombrugghe, B., and Maity, S.N. (1996). Determination of functional domains in the C subunit of the CCAAT-binding factor (CBF) necessary for formation of a CBF-DNA complex: CBF-B interacts simultaneously with both the CBF-A and CBF-C subunits to form a heterotrimeric CBF molecule. Mol. Cell. Biol. *16,* 4003-4013.

Knutson, B.A., Drennan, M., Liu, X., and Broyles, S.S. (2009). Bidirectional transcriptional promoters in the vaccinia virus genome. Virology *385*, 198-203.

Koonin, E.V. (2009). Evolution of genome architecture. Int. J. Biochem. Cell Biol. *41*, 298-306.

Korbel, J.O., Jensen, L.J., von Mering, C., and Bork, P. (2004). Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. Nat. Biotechnol. *22*, 911-917.

Kornberg, R.D. (2007). The molecular basis of eucaryotic transcription. Cell Death Differ. *14*, 1989-1997.

Koyanagi, K.O., Hagiwara, M., Itoh, T., Gojobori, T., and Imanishi, T. (2005). Comparative genomics of bidirectional gene pairs and its implications for the evolution of a transcriptional regulation system. Gene *353,* 169-176.

Kreuter, R., Soutar, A.K., and Wade, D.P. (1999). Transcription factors CCAAT/enhancer-binding protein beta and nuclear factor-Y bind to discrete regulatory elements in the very low density lipoprotein receptor promoter. J. Lipid Res. *40,* 376-386.

Krom, N., and Ramakrishna, W. (2008). Comparative analysis of divergent and convergent gene pairs and their expression patterns in rice, Arabidopsis, and populus. Plant Physiol. *147,* 1763-1773.

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

Li, T., Huang, H., Huang, B., Huang, B., and Lu, J. (2009). Histone acetyltransferase p300 regulates the expression of human pituitary tumor transforming gene (hPTTG). J. Genet. Genomics *36,* 335-342.

Liang, S.G., and Maity, S.N. (1998). Pathway of complex formation between DNA and three subunits of CBF/NF-Y. Photocross-linking analysis of DNA-protein interaction and characterization of equilibrium steps of subunit interaction and dna binding. J. Biol. Chem. *273,* 31590-31598.

Liberati, C., di Silvio, A., Ottolenghi, S., and Mantovani, R. (1999). NF-Y binding to twin CCAAT boxes: role of Q-rich domains and histone fold helices. J. Mol. Biol. *285,* 1441-1455.

Liberati, C., Ronchi, A., Lievens, P., Ottolenghi, S., and Mantovani, R. (1998). NF-Y organizes the gamma-globin CCAAT boxes region. J. Biol. Chem. *273,* 16880-16889.

Lin, J.M., Collins, P.J., Trinklein, N.D., Fu, Y., Xi, H., Myers, R.M., and Weng, Z. (2007). Transcription factor binding and modified histones in human bidirectional promoters. Genome Res. *17,* 818-827.

Mahishi, L., and Usdin, K. (2006). NF-Y, AP2, Nrf1 and Sp1 regulate the fragile X-related gene 2 (FXR2). Biochem. J. *400,* 327-335.

Maity, S.N., and de Crombrugghe, B. (1998). Role of the CCAAT-binding protein CBF/NF-Y in transcription. Trends Biochem. Sci. *23,* 174-178.

Maity, S.N., Golumbek, P.T., Karsenty, G., and de Crombrugghe, B. (1988). Selective activation of transcription by a novel CCAAT binding factor. Science *241,* 582-585.

Maity, S.N., Vuorio, T., and de Crombrugghe, B. (1990). The B subunit of a rat heteromeric CCAAT-binding transcription factor shows a striking sequence identity with the yeast Hap2 transcription factor. Proc. Natl. Acad. Sci. U. S. A. *87,* 5378-5382.

Manni, I., Caretti, G., Artuso, S., Gurtner, A., Emiliozzi, V., Sacchi, A., Mantovani, R., and Piaggio, G. (2008). Posttranslational regulation of NF-YA modulates NF-Y transcriptional activity. Mol. Biol. Cell *19,* 5203-5213.

Mantovani, R. (1999). The molecular biology of the CCAAT-binding factor NF-Y. Gene *239,* 15-27.

Mantovani, R. (1998). A survey of 178 NF-Y binding CCAAT boxes. Nucleic Acids Res. *26,* 1135-1143.

Matuoka, K., and Chen, K.Y. (2002). Transcriptional regulation of cellular ageing by the CCAAT box-binding factor CBF/NF-Y. Ageing Res. Rev. *1,* 639-651.

Milos, P.M., and Zaret, K.S. (1992). A ubiquitous factor is required for C/EBP-related proteins to form stable transcription complexes on an albumin promoter segment in vitro. Genes Dev. *6,* 991-1004.

Nakshatri, H., Bhat-Nakshatri, P., and Currie, R.A. (1996). Subunit association and DNA binding activity of the heterotrimeric transcription factor NF-Y is regulated by cellular redox. J. Biol. Chem. *271,* 28784-28791.

Niimura, Y., and Nei, M. (2005). Comparative evolutionary analysis of olfactory receptor gene clusters between humans and mice. Gene *346,* 13-21.

Osada, S., Yamamoto, H., Nishihara, T., and Imagawa, M. (1996). DNA binding specificity of the CCAAT/enhancer-binding protein transcription factor family. J. Biol. Chem. *271,* 3891-3896.

Piontkivska, H., Yang, M.Q., Larkin, D.M., Lewin, H.A., Reecy, J., and Elnitski, L. (2009). Cross-species mapping of bidirectional promoters enables prediction of unannotated 5' UTRs and identification of species-specific transcripts. BMC Genomics *10,* 189.

Quandt, K., Frech, K., Karas, H., Wingender, E., and 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.

Rada-Iglesias, A., Ameur, A., Kapranov, P., Enroth, S., Komorowski, J., Gingeras, T.R., and Wadelius, C. (2008). Whole-genome maps of USF1 and USF2 binding and histone H3 acetylation reveal new aspects of promoter structure and candidate genes for common human disorders. Genome Res. *18,* 380-392.

Radomska, H.S., Satterthwaite, A.B., Taranenko, N., Narravula, S., Krause, D.S., and Tenen, D.G. (1999). A nuclear factor Y (NFY) site positively regulates the human CD34 stem cell gene. Blood *94,* 3772-3780.

Ramji, D.P., and Foka, P. (2002). CCAAT/enhancer-binding proteins: structure, function and regulation. Biochem. J. *365,* 561-575.

Rocha, E.P. (2008). The organization of the bacterial genome. Annu. Rev. Genet. *42,* 211-233.

Romier, C., Cocchiarella, F., Mantovani, R., and Moras, D. (2003). The NF-YB/NF-YC structure gives insight into DNA binding and transcription regulation by CCAAT factor NF-Y. J. Biol. Chem. *278,* 1336-1345.

Ronchi, A., Bellorini, M., Mongelli, N., and Mantovani, R. (1995). CCAAT-box binding protein NF-Y (CBF, CP1) recognizes the minor groove and distorts DNA. Nucleic Acids Res. *23,* 4565-4572.

Salsi, V., Caretti, G., Wasner, M., Reinhard, W., Haugwitz, U., Engeland, K., and Mantovani, R. (2003). Interactions between p300 and multiple NF-Y trimers govern cyclin B2 promoter function. J. Biol. Chem. *278*, 6642-6650.

Sandelin, A., Carninci, P., Lenhard, B., Ponjavic, J., Hayashizaki, Y., and Hume, D.A. (2007). Mammalian RNA polymerase II core promoters: insights from genome-wide studies. Nat. Rev. Genet. *8*, 424-436.

Sato, R., Inoue, J., Kawabe, Y., Kodama, T., Takano, T., and Maeda, M. (1996). Sterol-dependent transcriptional regulation of sterol regulatory element-binding protein-2. J. Biol. Chem. *271*, 26461-26464.

Saxonov, S., Berg, P., and Brutlag, D.L. (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc. Natl. Acad. Sci. U. S. A. *103*, 1412-1417.

Scherf, M., Epple, A., and Werner, T. (2005). The next generation of literature analysis: integration of genomic analysis into text mining. Brief Bioinform *6*, 287-297.

Serra, E., Zemzoumi, K., di Silvio, A., Mantovani, R., Lardans, V., and Dissous, C. (1998). Conservation and divergence of NF-Y transcriptional activation function. Nucleic Acids Res. *26*, 3800-3805.

Shu, J., Jelinek, J., Chang, H., Shen, L., Qin, T., Chung, W., Oki, Y., and Issa, J.P. (2006). Silencing of bidirectional promoters by DNA methylation in tumorigenesis. Cancer Res. *66*, 5077-5084.

Sinha, S., Kim, I.S., Sohn, K.Y., de Crombrugghe, B., and Maity, S.N. (1996). Three classes of mutations in the A subunit of the CCAAT-binding factor CBF delineate functional domains involved in the three-step assembly of the CBF-DNA complex. Mol. Cell. Biol. *16*, 328-337.

Sinha, S., Maity, S.N., Lu, J., and de Crombrugghe, B. (1995). Recombinant rat CBF-C, the third subunit of CBF/NFY, allows formation of a protein-DNA complex with CBF-A

and CBF-B and with yeast HAP2 and HAP3. Proc. Natl. Acad. Sci. U. S. A. *92*, 1624-1628.

Sitwala, K.V., Adams, K., and Markovitz, D.M. (2002). YY1 and NF-Y binding sites regulate the transcriptional activity of the dek and dek-can promoter. Oncogene *21*, 8862-8870.

Sperling, S. (2007). Transcriptional regulation at a glance. BMC Bioinformatics *8*, S2.

Steffen, M.L., Harrison, W.R., Elder, F.F., Cook, G.A., and Park, E.A. (1999). Expression of the rat liver carnitine palmitoyltransferase I (CPT-Ialpha) gene is regulated by Sp1 and nuclear factor Y: chromosomal localization and promoter characterization. Biochem. J. *340 (Pt 2)*, 425-432.

Sun, F., Xie, Q., Ma, J., Yang, S., Chen, Q., and Hong, A. (2009). Nuclear factor Y is required for basal activation and chromatin accessibility of fibroblast growth factor receptor 2 promoter in osteoblast-like cells. J. Biol. Chem. *284*, 3136-3147.

Suzuki, Y., Tsunoda, T., Sese, J., Taira, H., Mizushima-Sugano, J., Hata, H., Ota, T., Isogai, T., Tanaka, T., Nakamura, Y., *et al.* (2001). Identification and characterization of the potential promoter regions of 1031 kinds of human genes. Genome Res. *11*, 677-684.

Taira, T., Sawai, M., Ikeda, M., Tamai, K., Iguchi-Ariga, S.M., and Ariga, H. (1999). Cell cycle-dependent switch of up-and down-regulation of human hsp70 gene expression by interaction between c-Myc and CBF/NF-Y. J. Biol. Chem. *274*, 24270-24279.

Takai, D., and Jones, P.A. (2004). Origins of bidirectional promoters: computational analyses of intergenic distance in the human genome. Mol. Biol. Evol. *21*, 463-467.

Teodorovic, S., Walls, C.D., and Elmendorf, H.G. (2007). Bidirectional transcription is an inherent feature of Giardia lamblia promoters and contributes to an abundance of sterile antisense transcripts throughout the genome. Nucl. Acids Res. *35*, 2544-2553.

Testa, A., Donati, G., Yan, P., Romani, F., Huang, T.H., Vigano, M.A., and Mantovani, R. (2005). Chromatin immunoprecipitation (ChIP) on chip experiments uncover a widespread distribution of NF-Y binding CCAAT sites outside of core promoters. J. Biol. Chem. *280,* 13606-13615.

Tomita, T., and Kimura, S. (2008). Regulation of mouse Scgb3a1 gene expression by NF-Y and association of CpG methylation with its tissue-specific expression. BMC Mol. Biol. *9,* 5.

Trinklein, N.D., Aldred, S.F., Hartman, S.J., Schroeder, D.I., Otillar, R.P., and Myers, R.M. (2004). An abundance of bidirectional promoters in the human genome. Genome Res. *14,* 62-66.

Tsai, H.K., Su, C.P., Lu, M.Y., Shih, C.H., and Wang, D. (2007). Co-expression of adjacent genes in yeast cannot be simply attributed to shared regulatory system. BMC Genomics *8,* 352.

Ueda, A., Takeshita, F., Yamashiro, S., and Yoshimura, T. (1998). Positive regulation of the human macrophage stimulating protein gene transcription. Identification of a new hepatocyte nuclear factor-4 (HNF-4) binding element and evidence that indicates direct association between NF-Y and HNF-4. J. Biol. Chem. *273,* 19339-19347.

van Huijsduijnen, R.H., Li, X.Y., Black, D., Matthes, H., Benoist, C., and Mathis, D. (1990). Co-evolution from yeast to mouse: cDNA cloning of the two NF-Y (CP-1/CBF) subunits. EMBO J. *9,* 3119-3127.

Vardhanabhuti, S., Wang, J., and Hannenhalli, S. (2007). Position and distance specificity are important determinants of cis-regulatory motifs in addition to evolutionary conservation. Nucleic Acids Res. *35,* 3203-3213.

Wang, Q., Wan, L., Li, D., Zhu, L., Qian, M., and Deng, M. (2009). Searching for bidirectional promoters in Arabidopsis thaliana. BMC Bioinformatics *10,* S29.

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

Weinrich, S.L., and Hruby, D.E. (1986). A tandemly-oriented late gene cluster within the vaccinia virus genome. Nucleic Acids Res. *14*, 3003-3016.

Xing, Y., Fikes, J.D., and Guarente, L. (1993). Mutations in yeast HAP2/HAP3 define a hybrid CCAAT box binding domain. EMBO J. *12*, 4647-4655.

Xu, Y., Zhou, Y.L., Luo, W., Zhu, Q.S., Levy, D., MacDougald, O.A., and Snead, M.L. (2006). NF-Y and CCAAT/enhancer-binding protein alpha synergistically activate the mouse amelogenin gene. J. Biol. Chem. *281*, 16090-16098.

Yamada, K., Tanaka, T., Miyamoto, K., and Noguchi, T. (2000). Sp family members and nuclear factor-Y cooperatively stimulate transcription from the rat pyruvate kinase M gene distal promoter region via their direct interactions. J. Biol. Chem. *275*, 18129-18137.

Yang, L., and Yu, J. (2009). A comparative analysis of divergently-paired genes (DPGs) among Drosophila and vertebrate genomes. BMC Evolutionary Biology *9*, 55.

Yang, M.Q., and Elnitski, L.L. (2008a). Diversity of core promoter elements comprising human bidirectional promoters. BMC Genomics *9*, 1-8.

Yang, M.Q., and Elnitski, L.L. (2008b). Prediction-based approaches to characterize bidirectional promoters in the mammalian genome. BMC Genomics *9*, 1-11.

Yang, M.Q., Koehly, L.M., and Elnitski, L.L. (2007). Comprehensive annotation of bidirectional promoters identifies co-regulation among breast and ovarian cancer genes. PLoS Comput. Biol. *3*, e72.

Yang, M.Q., Taylor, J., and Elnitski, L. (2008). Comparative analyses of bidirectional promoters in vertebrates. BMC Bioinformatics *9*, 1-8.

Yu, F.X., and Luo, Y. (2009). Tandem ChoRE and CCAAT motifs and associated factors regulate Txnip expression in response to glucose or adenosine-containing Molecules. PLoS One *4*, e8397.

Yu, P., Ma, D., and Xu, M. (2005). Nested genes in the human genome. Genomics *86,* 414-422.

Zanotto, E., Häkkinen, A., Teku, G., Shen, B., Ribeiro, A.S., and Jacobs, H.T. (2009). NF-Y enforces directionality of transcription from the bidirectional *Mrps12*/*Sarsm* promoter in both mouse and human cells. Biochim. Biophys. Acta. *1789*, 432-442.

Zanotto, E., Shah, Z.H., and Jacobs, H.T. (2007). The bidirectional promoter of two genes for the mitochondrial translational apparatus in mouse is regulated by an array of CCAAT boxes interacting with the transcription factor NF-Y. Nucleic Acids Res. *35*, 664-677.

Zemzoumi, K., Frontini, M., Bellorini, M., and Mantovani, R. (1999). NF-Y histone fold α1 helices help impart CCAAT specificity. J. Mol. Biol. *286*, 327-337.