1. Introduction

1.1 Transcriptional regulation in eukaryotes

Many of the biological processes in eukaryotes are regulated at the level of transcription. Transcription defines the process of making a copy of genetic information stored in a DNA strand and transcribing it into a complementary RNA strand with the aid of many proteins including RNA polymerases. There are three variable RNA polymerases in most eukaryotes. RNA polymerase I is located in the nucleolus and plays a role in the synthesis of 18S, 5.8S and 28S ribosomal RNAs (rRNA); RNA polymerase III is located in the nucleoplasm and functions in synthesizing small RNAs, such as transfer RNA (tRNA) and 5S rRNA, and RNA polymerase II is located in the nucleoplasm and responsible for messenger RNA (mRNA) and small nuclear RNA (snRNA) synthesis (Weaver 2008). In contrast, plants encode two more nuclear polymerases, RNA polymerase IV and V. The two RNA polymerases play specific roles in gene silencing in plants (Herr *et al.* 2005, Onodera *et al.* 2005, Wierzbicki *et al.* 2008).

Each RNA polymerase recognizes a different type of promoter. RNA polymerase II is involved in the transcription of all protein-encoding genes. In most RNA polymerase II promoters, there are at least three common features: the transcription start site, the TATA box, and sequences which serve as the binding sites of transcriptional regulators. The sequences bound by transcriptional regulators include upstream activating sequences, enhancers, upstream repressing sequences, and silencers. The core promoter sequence consists of a TATA box that is approximately located at position -33 base pair (bp), a TFIIB (transcription factor II B) recognition element just upstream of the TATA box, an initiator sequence centred on the transcription start site, and a downstream promoter element.

The core promoter is recognized by the general transcription factors which associate with RNA polymerase II to form a preinitiation complex. RNA polymerase II is incapable of binding DNA by itself. The class II preinitiation complex contains RNA polymerase II and six general transcription factors: TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH. The first factor binding to the TATA box is TFIID which is a large complex composed of a TATA-binding protein (TBP) and more than eight TBP-associated factors (TAFs). The assembled preinitiation complex dictates the starting point and the direction of transcription, but conducts the transcription of all protein-encoding genes at a basic level. The gene-specific transcription factors are able to make physical contact with the initiation complex. Through

this physical contact the transcription factors exert their control over transcriptional activation and repression. The transcription factors which positively regulate transcription are called activators, whereas repressors have a negative effect on transcription. These factors finally arrange the transcription activity of RNA polymerase II.

1.2 CCAAT box

The transcriptional regulation of gene expression is mediated by the binding of regulatory factors to specific *cis*-acting DNA sequence elements. The CCAAT box is one of the most common *cis*-acting elements in eukaryotic RNA polymerase II promoters (Gelinas *et al.* 1985). A bioinformatic study in a set of 13,010 human genomic promoter sequences revealed that the CCAAT is the second most frequent consensus sequence (FitzGerald *et al.* 2004). In a statistical analysis of more than 500 eukaryote promoters, 30% of the promoters contain the CCAAT box, which is often positioned between 60 and 100 bp upstream of the transcription start site in the forward or reverse orientation (Bucher 1990). In higher eukaryotes, the pentanucleotide CCAAT is conserved and the flanking nucleotides and their positions before and after the CCAAT motif are necessary for transcriptional activities (Dorn *et al.* 1987, Mantovani 1998).

In fungi, the genes possessing CCAAT boxes are involved in tricarboxylic acid cycle and oxidative phosphorylation (Forsburg and Guarente 1989a), mitochondrial function (Buschlen *et al.* 2003), penicillin biosynthesis (Brakhage *et al.* 1999), nitrogen metabolism (Dang *et al.* 1996). In mammals, the CCAAT motifs have been found in promoter regions of all kinds of genes which are ubiquitous, tissue or developmental specific, either inducible or constitutive (Maity and de Crombrugghe 1998, Mantovani 1999, Matuoka and Chen 2002). Many CCAAT box elements have been experimentally defined (Testa *et al.* 2005) and it was striking that genes with these CCAAT boxes encode proteins with functions in DNA transcription, signal transduction, chromatin modifying, RNA processing, nuclear trafficking, cytoskeleton, metabolism, apoptosis, and cellular senescence.

The CCAAT motif is also present in plant genes. CCAAT boxes are identified in the promoters of the wheat histone-encoding genes (Ito *et al.* 1995, Taoka *et al.* 1998) and the soybean heat shock gene (Rieping and Schoffl 1992). The photosynthesis gene *AtpC* (Bezhani *et al.* 2001, Kusnetsov *et al.* 1999) and the genes for light-harvesting chlorophyll a/b-binding proteins (Lhcbs) (Carre and Kay 1995, Folta and Kaufman 1999, Kehoe *et al.*

1994) contain the CCAAT-motifs in promoter regions and belong to inducible genes with high significance for photosynthesis and energy delivery. The CCAAT box in the gene *FLOWERING LOCUS T* (*FT*) cooperates with other elements to regulate the gene expression (Blackman and Michaels 2010, Kumimoto *et al.* 2008, Tiwari *et al.* 2010). A consensus stress-responsive element among the endoplasmic reticulum stress-activated genes contains the CCAAT element (Liu and Howell 2010).

1.3 Transcription factor NF-Y

Transcription factors are sequence-specific DNA-binding proteins which are able to regulate the expression of genes at the transcriptional level. According to the specificity, transcription factors are grouped into two classes: general transcription factors and gene-specific transcription factors. The general transcription factors can attract the RNA polymerases to their respective promoters and support only a basal level of transcription. Gene-specific transcription factors can be activators or repressors of gene expression. Eukaryotic transcription factors contain at least two domains: a DNA-binding domain such as a zinc finger domain, a homeodomain, a basic leucine zipper (bZIP) domain, a basic helix-loop-helix (bHLH) domain, or a β-barrel; and a transcription-activation domain which can consist of an acidic, glutamine-rich, or proline-rich domain (Weaver 2008). According to the sequence and structure of the DNA-binding domain, each super transcription factor family has a specific mechanism of DNA binding (Luscombe *et al.* 2000).

The proteins that recognize and bind to TATA box are named TBPs. For the GC box, the binding protein is designated the transcription factor SP1. The transcription factors which specifically bind to the CCAAT box are called heme activator proteins (HAP) in yeast, CCAAT binding factors (CBF) in rat, and nuclear factor Y (NF-Y) in mouse, human and *Arabidopsis* (Mantovani 1999, Siefers *et al.* 2009). Most of the information obtained about NF-Y is provided from studies in yeast and mammals. NF-Y is a heterotrimeric complex consisting of three subunits: NF-YA (CBF-B, HAP2), NF-YB (CBF-A, HAP3) and NF-YC (CBF-C, HAP5), which are all required for DNA-binding of NF-Y (Maity and de Crombrugghe 1998, Mantovani 1999, Matuoka and Yu Chen 1999). In yeast, an additional subunit HAP4 is recruited and required for transcription activation (Forsburg and Guarente 1989b).

NF-Y is found in many organisms. Proteins which bind to the CCAAT box element were first characterized in the yeast *Saccharomyces cerevisiae* (Guarente *et al.* 1984). In rat, CBF was purified from liver nuclear extracts as a nuclear factor that interacts with a CCAAT motif of the α2(1) collagen promoter (Maity *et al.* 1988). NF-Y was identified as a nuclear factor that interacts with a CCAAT motif of the major histocompatibility complex (MHC) class II promoter in mouse B-cell (Dorn *et al.* 1987). The presence and diversity of NF-YA, NF-YB and NF-YC were examined using complementation and computational approaches in *Arabidopsis* (Edwards *et al.* 1998). The *Triticum aestivum* NF-Y family was identified in the global DNA databases by computational analysis (Stephenson *et al.* 2007). A NF-Y family was also reported in rice (Thirumurugan et al. 2008), maize (Nelson *et al.* 2007), *Brachypodium distachyon* (Cao et al. 2011), *Brassica napus* (Albani and Robert 1995), sunflower (Fambrini *et al.* 2006), and tomato (Ben-Naim *et al.* 2006).

1.4 Structures and molecular functions of NF-Y

In mammals, the biochemical mechanism of the subunit association is well described. Firstly, NF-YB and NF-YC interact with each other in the cytoplasm to form a stable heterodimer mediated through the histone-fold motifs. Neither NF-YB nor NF-YC contains a nuclear localization signal. The dimerization is prerequisite for the nuclear localization. The NF-YB/NF-YC dimer is imported into the nucleus by an importin 13. Independently, NF-YA is translocated into the nucleus in an importin β-mediated way (Kahle *et al.* 2005). Then, the heterodimer offers a suitable position for NF-YA interaction. NF-YA is not able to associate with NF-YB or NF-YC individually (Sinha *et al.* 1995). The Electrophoretic Mobility Shift Assay (EMSA) with truncated NF-Y subunits indicates that the NF-Y complex contains one molecule of each subunit (Sinha *et al.* 1996). At last, the formed heterotrimer turns into a functional transcription complex, which can bind to DNA with sequence specificity and high affinity (Bi *et al.* 1997). As a result, the gene transcription is regulated by NF-Y in either a positive or a negative way (Ceribelli *et al.* 2008, Peng and Jahroudi 2002). The domain structures of NF-Y subunits are represented in Figure 1.

1.4.1 NF-YA

The NF-Y subunits are conserved in eukaryotes (Li *et al.* 1992). The conserved region in NF-YA has been identified to consist of two domains: the N-terminal part required for subunit interaction, and the C-terminal part recognizing and binding to specific DNA sequences (Xing *et al.* 1993, Xing *et al.* 1994). The circular dichroism (CD) spectrum and the proline

substitution assay suggested that the two domains form α-helices (Mantovani *et al.* 1994, Xing *et al.* 1994). A spacer region exists between these two domains. The spacing and sequence of this spacer are variable from different organisms (Maity and de Crombrugghe 1992, Olesen and Guarente 1990). Outside the highly conserved area, the amino-terminal portion of NF-YA are very rich in glutamines (Q) (35%) and adjoined to a small serine/threonine (S/T)-rich segment at its carboxyl terminal (Maity *et al.* 1990). A Q-rich and S/T-rich segment has been identified to activate the transcription *in vivo* and *in vitro* (Coustry *et al.* 1995, Coustry *et al.* 1996, di Silvio *et al.* 1999). In yeast, HAP2 lacks the sequence corresponding to the Q-rich region; instead, there is a regulatory protein HAP4 which undertakes the responsibility of transcriptional activation (Forsburg and Guarente 1989b, Stebbins and Triezenberg 2004). However, for *Arabidopsis* NF-YA, the Q content is low in the corresponding region (Gusmaroli *et al.* 2002). The transcriptional activation assay indicates that *Arabidopsis* NF-YA do not contain a transcription activation domain (Park 2006).

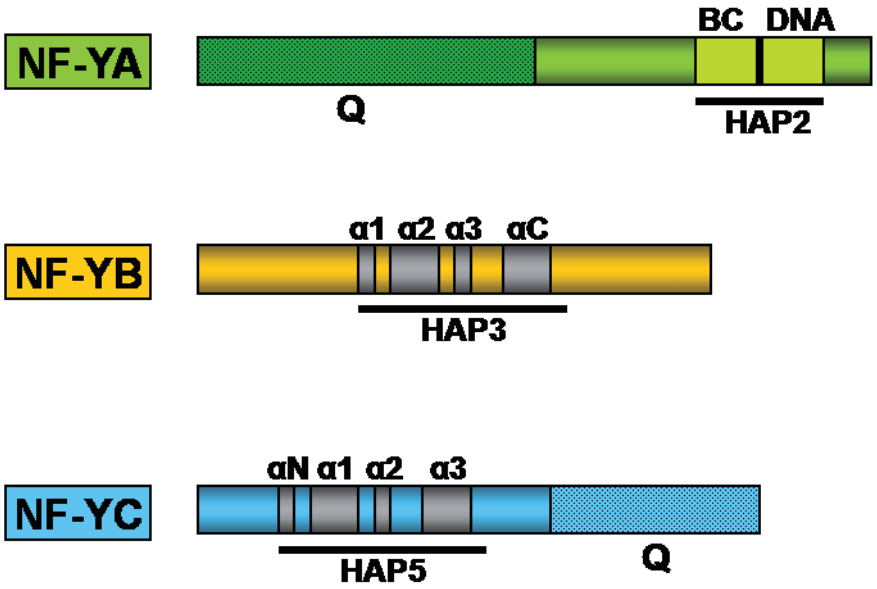

Figure 1. Schematic representation of NF-Y subunits in *H. sapiens* from Mantovani et al. (1999).

The conserved domains that are identified in the homologous yeast HAP2, HAP3 and HAP5 are underlined. The HAP2 domain of NF-YA is composed of NF-YB-NF-YC dimer binding domain (BC) and DNA binding domain (DNA). The N-terminal of NF-YA has a glutamine-rich transcriptional activation domain (Q). NF-YB and NF-YC have a highly conserved histone fold motif, which contains four α-helices (grey boxes). The C-terminal of NF-YC also has a transcriptional activation domain (Q).

1.4.2 NF-YB and NF-YC

The conserved regions of NF-YB and NF-YC show homology to the histone fold motif (HFM) in terms of structure and amino acid sequence (Baxevanis *et al.* 1995, Masiero *et al.* 2002, Romier *et al.* 2003, Siefers *et al.* 2009). The histone fold is composed of a short α-helix (α1) followed by a loop and β-strand segment (L1), a ling helix (α2), another short loop and β-strand (L2), and a short helix (α3). The high-resolution crystal structure of nucleosomes indicates that the HFMs interact with each other head to tail manner in the heterodimeric histones (Arents and Moudrianakis 1995, Luger *et al.* 1997). Sequence comparison reveals that NF-YB is closely related to the H2B histone and NF-YC is similar to H2A histone. A biochemical study provided the evidence that NF-YB, like H2B, contains a fourth helix, αC, and NF-YC contains the αN helix which is also present in H2A (Liberati *et al.* 1999).

The mutational analysis demonstrated that the HFMs are necessary for the subunit interaction and DNA binding (Kim *et al.* 1996, Sinha *et al.* 1996). In NF-YB, the entire histone fold is required for association with NF-YC. The α2, L2 and a part of α3 of the HFM of NF-YC are needed to form a stable NF-YB-NF-YC heterodimer. Three separate NF-YA-binding domains locate in the NF-YB-NF-YC heterodimer: one is within the α2 of NF-YB; the other two are in the α1, L1 and part of α3 of NF-YC. The interaction domains overlap in NF-YB and NF-YC. The α1, L1 and part of α2 in NF-YB and the portion of α1 in NF-YC are essential for DNA binding. NF-YC is similar to NF-YB, but has an additional C-terminal Q-rich region which is competent to activate the transcription (Coustry *et al.* 1996, di Silvio *et al.* 1999). In yeast, HAP5 contains no Q-rich region and the corresponding regions of NF-YC are particularly rich in prolines and hydrophobic residues in *Arabidopsis*. Unlike NF-YA and NF-YC, NF-YB seems to miss a domain with an activating potential (Serra *et al.* 1998).

1.4.3 NF-Y and chromatin

The chromatin is composed of the genomic DNA, linker histones and the nucleosomes which consists of heterooctamers containing a histone pair of the isoforms H2A, H2B, H3 and H4 (Luger *et al.* 1997). The x-ray crystallography analysis revealed that NF-YB and NF-YC interact through histone fold motifs in a head-to-tail way (Romier *et al.* 2003). The NF-YB-NF-YC dimer, similarly to H2A-H2B, binds to H3-H4 tetramer but not to H2A-H2B during the nucleosome formation (Caretti *et al.* 1999). Even after the completion of the nucleosome assembly, NF-Y is still able to interact with the genomic DNA assembled in the nuleosome. NF-YA is responsible to search their specific target, the CCAAT sequence. Once bound to the DNA, NF-Y preferably interacts with the minor groove and distorts the double helix

strands (Ronchi *et al.* 1995). The association of NF-Y and DNA results in the disruption of the nucleosomal structure (Coustry *et al.* 2001). The transcription initiation site is exposed by the distortion and the general transcription machinery binds to the DNA to permit the transcription starting.

Moreover, chromatin structure is modified by NF-Y via post-translational regulations of histones. The highly basic N-terminus of histone proteins surrounds the negatively charged DNA strands. This interaction contributes to a tight binding between DNA and the nucleosome forming the heterochromatin. For gene expression, acetylation, phosphorylation, and methylation of histones result in euchromatin state that allows the DNA release from histones (Jenuwein and Allis 2001). NF-Y is found to be associated with histone acetyltransferases: for example the human GCN5 with NF-YB/C dimer (Currie 1998), p300 with NF-YB (Li et al. 1998), and P/CAF with NF-YA (Jin and Scotto 1998). The genes targeted by NF-Y either contain positive histone marks for transcription or have negative histone marks correlating with transcriptional repression (Ceribelli *et al.* 2008).

1.5 Biological functions of NF-Y subunits in *Arabidopsis*

In *Arabidopsis*, about 5.9 % of the genome contains genes encoding transcription factors and half of the genes encode plant-specific transcription factors (Riechmann *et al.* 2000). In consistency to four representative databases, *Arabidopsis* transcription factors are in total classified in 72 families including the CCAAT transcription factor family (Mitsuda and Ohme-Takagi 2009).

In yeast and mammals, each of the three NF-Y subunits is encoded by a single gene. In contrast, the plant NF-Y subunits are encoded by large gene families. *Arabidopsis* NF-YA, NF-YB and NF-YC subunit families have ten, 13 and 13 members (Table 1), respectively (Siefers *et al.* 2009). Ten OsHAP2, 11 OsHAP3 and 7 OsHAP5 subunits are encoded in the rice genome (Thirumurugan *et al.* 2008). Similar quantities of NF-Y subunit families have been identified in wheat and Brachypodium (Cao *et al.* 2011, Stephenson *et al.* 2007). Although several plant NF-Y subunit members form functional trimeric complexes with mammalian or yeast NF-Y subunits (Kumimoto *et al.* 2008, Park 2006, Yazawa and Kamada 2007), no complete plant NF-Y complex has been reported. It is suggested that the individual plant NF-Y subunits play specific and essential roles during plant development and environmental adaptation.