**CCAAT-box binding protein NF-Y (CBF, CP1) recognizes the minor groove and distorts DNA**

Antonella Ronchi, Marianna Bellorini, Nicola Mongelli and Roberto Mantovani

Dipartimento di Genetica e Biologia dei Microrganismi, Università di Milano, Via Celoria 26, 20133 Milano, Italy

Received September 21, 1995; Accepted October 11, 1995

**ABSTRACT**

The CCAAT box is one of the most common promoter elements. The evolutionarily conserved heteromeric factor NF-Y binds this sequence with high affinity and specificity. By comparing the methylation interference patterns of different sites, performing electrophoretic mobility shift assays (EMSA) with IC-substituted oligonucleotides and competition experiments with the minor groove binder (MGB) drugs distamicin A, tallimustine and Hoechst 33258 we show that NF-Y factor NF-Y binds this sequence with high affinity and specificity. The CCAAT box is a widely distributed regulatory sequence to distort the double helix by angles of 62-82°, depending on the site used, and suggest that nucleotides flanking the CCAAT pentanucleotide influence the degree of bending.

**INTRODUCTION**

The CCAAT box is a widely distributed regulatory sequence present in several promoters and enhancers. A computer-assisted search on 503 eukaryotic promoters indicated that a high proportion of them (25/30%) harbour this pentanucleotide in the forward or reverse orientation, often at position -60/80 with respect to the start site (1). Unlike other elements composed of palindromic sequences, the CCAAT box does not seem to have a symmetry axis. Several proteins have been reported to bind this or related sequences (2,3). Among these NF-Y, originally identified as a factor recognizing the conserved Y box element in the mouse MHC Class II Ea promoter, has a peculiar and almost absolute requirement for these five nucleotides and a strong preference for additional flanks (2,4). Interestingly, the CCAAT consensus derived by Bucher (PuPuCCAAT/GA/G) fits well with the optimal NF-Y binding site, but not with the consensus of other transcription factors that bind CCAAT-related sequences, such as CTF/NF-1 and C/EBP (1).

NF-Y has been shown to be involved in a number of different systems, in addition to MHC Class II promoters: MSV (2), HSP70 (3), α-, β- and γ-globins (5-7), albumin (8), Thy-1 (9), α-collagen (10), β-actin (11), IL4 (12,13), Gp91phox (14), TSP-1 (15), FGF-4 (16), RSV (17) and TK (18). This CCAAT box binding protein has been identified and, in some cases, biochemically characterized in several laboratories and, consequently, it has been given different names. Because of the heteromeric nature of the DNA binding complex, of competition analysis with Ea Y box oligonucleotides and supershift experiments with anti-NF-Y monoclonal and polyclonal antibodies, CP1 (binding to the MLP, β and γ-globin CCAAT box; 3,7), CBF (to α-collagen; 10), α-CP1 (to α-globin; 5), EFL to RSV; 17) and CBPtk (to human TK; 18) are different acronyms for the same entity. Therefore it is possible to compare the binding sequences and methylation interference patterns of all these bona fide NF-Y binding sites (2,3,5,7,8,11,14-17,18). In all the above-mentioned promoters functional experiments indicate that the CCAAT box and, hence NF-Y, plays an important and sometimes essential role (9,11-16,20-25).

NF-Y is a ubiquitous heteromeric complex formed by at least three subunits, all necessary for DNA binding (3,5,10): NF-YA (HAP2/CBF-B) and NF-YB (HAP3/CBF-A), whose genes have been cloned in a number of species, including maize, *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*, sea urchin, *Kluyveromyces lactis* (27-34), and the recently discovered *Kluyveromyces lactis* (27-34), and the recently discovered HAP5/CBF-C (35,36). Protein alignments of all the subunits evidenced highly conserved domains (70% identity) in all species. Biochemical and genetic analysis dissected the protein–protein and DNA binding subdomains of HAP3 (34) and HAP2/NF-YA (37-39). The remarkable degree of conservation between yeast and man is reminiscent of factors belonging to the basal transcription machinery: TBP for example, the major determinant of Pol II positioning, has been shown to bind DNA via a conserved C-terminal domain that is responsible for making asymmetrical contacts exclusively with the minor groove and for significantly bending the double helix by an angle of 80° (40-43). DNA-dic-tated asymmetry and protein-mediated distortion of the DNA are believed to be the rate limiting step in building up a competent transcription complex (44).

The exact mechanisms of transcriptional activation by NF-Y are not known. Although the CCAAT box alone is not able to activate even if multimerized, it appears to increase the activity of neighbouring enhancer motifs (45,46). In the Ea promoter, which has no functional TATA box, NF-Y seems to play an important role in correctly positioning the start site, together with an initiator element (20,47). The strong -60/80 position preference (1) might suggest a structural role in promoter activation.

Because of the evolutionary conservation of the NF-Y subunits across species and the role that the CCAAT sequence plays in...
diverse promoter contexts in different organisms, we felt that it was important to know more about how NF-Y binds DNA and whether, upon binding, it is able to introduce distortions in the double helix.

**MATERIALS AND METHODS**

**Plasmids**

The starting plasmid for the circular permutation assays was pBend2, described by Aryia et al. (48). The Ea Y 39mer (2) was inserted by blunt end ligation between the Sail and Xbal sites of the poly linker. Oligonucleotides corresponding to the human HSP70 (CCCGGACCCCTGAGACGACATCGAGGCA), MSV (TGAACCAAGCCAATCAGTGCTGCT) and human e-globin (ACACAGGTCAGCCTTGACCAATTTTAAG) CCAAT boxes were blunt end cloned into the Xbal site.

**Electrophoretic mobility shift assay (EMSA)**

Binding reactions for NF-Y were performed according to Chodosh et al. (3) and Mantovani et al. (39), using recombinant NF-YA (20 ng), purified NF-YB/C (5 ng) and labelled oligonucleotides (10 000 c.p.m.) containing Ea, α-globin and the distal γ-globin (2,7) CCAAT boxes. EMSA with the API site (AGCTTTGACAGGCTAGGTAGCTG) CCAAT boxes were blunt end cloned into the XbaI site.

**Calculations of bending angles**

Location of the points of flexure and amplitudes of the bending angles were performed according to the method described in Thompson and Landy (49). Briefly, the mobilities of the NF-Y–DNA complexes were normalized to the mobilities of the corresponding free DNA fragments and bending angles calculated from the ratio between the fastest and the slowest migrating complexes in EMSA and then by linear interpolation between points obtained with A-tract DNA standards (49), according to the formula $\mu_M/\mu_E = \cos \omega/2$, where $\mu_M$ is the relative mobility of the complex exactly in the middle, $\mu_E$ is the relative mobility of the complex at the end of the fragment and $\omega$ is the angle of deviation. To determine bending centres the normalized mobility of each NF-Y–DNA complex was plotted as a function of the
Figure 2. Specificity of MGBs-DNA interactions. (A) Lack of EMSA inhibition by MGBs on the API site. A labelled API site corresponding to the collagenase TRE (TPA-responsive element) was incubated with 5 μg HeLa nuclear extract either in the absence (lane 1) or in the presence of increasing concentrations of three MGB drugs: distamycin A (lanes 2–5), tallimustine (lanes 6–9) and Hoechst 33258 (lanes 10–13). (B) A labelled Y box oligonucleotide was incubated with the indicated amounts of MGBs in the absence of NF-Y protein.

distance between the centre of the CCAAT sequence and the end of the DNA fragment. The bend was determined as the position at which the NF-Y–DNA complex was at a minimum.

RESULTS

Comparison of the methylation interference patterns of NF-Y on different sites

Because of its wide distribution, high affinity and specificity for CCAAT box sequences, several investigators identified NF-Y in biochemical characterizations of CCAAT box binding proteins involved in the regulation of different promoters. In many such studies methylation interference assays were performed to precisely pinpoint the nucleotides involved in protein–DNA interactions. This assay involves treatment of the target oligonucleotide with dimethylsulfate, which methylates Gs at the N7 position and As at the N3 position, separation of the free from the bound oligonucleotides by EMSA and identification of purine residues necessary for stable binding as missing bands in a sequencing gel. Such experiments not only give precise information about the sequence requirements of the protein, but also suggest whether the protein contacts the major or the minor groove (50): protection of Gs indicates that the protein is contacting the major groove, while a missing A is indicative of minor groove binding.

Table 1 lists the high affinity NF-Y binding sites for which methylation interference data are available. They recognize a single DNA binding factor, variously termed in different laboratories, whose identity with NF-Y has been established by competition with an Ea Y box oligonucleotide and/or with anti-NF-Y antibodies (5–18). The protected nucleotides in methylation interference patterns are indicated as asterisks and open circles represent partial protection. Several common features are evident.

The two purines at positions +1 and +2 are protected only if one (α- and γ-globin and Gp91) or two Gs (TSP1) are present. Adenines are never protected, whether single (UAS2UP1, RSV, α-globin, γ-globin, β-actin and Gp91 phox) or double (Ea, albumin and MLP).

The two Gs on the bottom strand at positions +3/+4 and the two As on the top strand at positions +5/+6 are always protected (with the exception of the +3 G in FGF-4 and the +5 A in Ea).

The A on the bottom strand at position +7 is completely (MLP, α-globin, γ-globin, β-actin, albumin, TSP1, Gp91 and FGF-4) or partially protected (Ea, UAS2UP1 and RSV).

Protection of Gs on the top strand at positions +8/+10 is evident in some sites (RSV, TSP1, Ea and albumin) but not in others, whereas an A at position +9 is usually protected (Ea, UAS2UP1, β-actin, albumin, Gp91 and MLP), the only exceptions being α-globin and FGF-4. On the bottom strand such positions are never protected, despite the presence of several Gs.

Overall these data suggest that: (i) the NF-Y binding site extends over one turn of the double helix (10–11 bp) and only in MLP, RSV and FGF-4 can partial protections outside the consensus core be scored; (ii) the NF-Y binding site can be tentatively separated into two halves. The left hand PuPuCC (+1/+4) would contact the protein through the major groove, since Gs, but not As, are protected on both the top and bottom strands. On the right hand (+5/+9) NF-Y would make essential contacts in the minor groove.
IC substitutions identify essential and non-essential nucleotides

To gain further information on the latter point we exploited the fact that hydrogen bonding on purine and pyrimidine rings in inosine-C pairs resemble G-C in the major groove and A-T in the minor groove (50) and therefore inosine-cytosine mutagenesis can give informations regarding major/minor groove contacts. This feature was successfully used to study the modality of TBP binding to the AdML TATA box (51,52). An oligonucleotide containing IC nucleotide pairs instead of ATs allowed normal interaction with TBP, an indication that only the minor groove was involved in protein–DNA interactions. Such findings were later fully confirmed by crystallographic studies. Similarly, TCF-1 and SRY were also shown to interact mainly with the minor groove (53). It should be noted that the TBP consensus is relatively loose (54-56) and no single nucleotide was shown to be absolutely essential. In the case of NF-Y the sequence-specific requirements are quite high, since saturation mutagenesis studies on a number of different sites clearly established the need for every single base pair of the CCAAT pentanucleotide (2,10,11,17,22) with the exception of the +7 T in two cases, where a C can be found instead (19,57). We introduced IC substitutions into each single nucleotide of the Ea CCAAT box oligonucleotide and performed competition EMSA analysis using a labelled Y box 22mer and purified NF-Y (2,39). Binding to an IC substitution mutant would indicate minor groove binding, while lack of binding would suggest either major groove binding or selective recognition in the minor groove of a specific determinant in the A/T pair.

Figure 1 shows the results of such competition experiments. Even at a very high molar excess (500-fold) no competition was essentially observed with -52IC, -55IC and -56IC mutant oligonucleotides, while -54IC and -57IC competed essentially as did the wild-type oligonucleotide. An intermediate situation was evident for the -53IC mutation, which competed with a 10-fold lower affinity. Evidence from the methylation interference experiments lends strong support to the idea that the two Cs at the +3/+4 positions are contacted in the major groove, so lack of competition of -52IC and, partially, of -53IC, confirm these data. Normal binding of -54IC strongly points to minor groove contacts, since other mutations at this position alter binding severely (2,17,22). -55IC and -56IC are more puzzling, because if the minor groove is indeed implicated, it has to harbour enough sequence-specific information to account for the almost absolute requirement for these two base pairs. Alternatively, in contradiction of the methylation interference data, this result could indicate that only the major groove is implicated. Finally, good competition of -57IC was not surprising, given the presence in high affinity binding sites of either a C (Ea), an A (γ-globin) or a G (α-globin, albumin).

Effects of minor groove binders on NF-Y binding

To discriminate between the two possibilities emerging from the competition experiments with the -55IC and -56IC substitution mutants we decided to use small drugs known to bind specifically in the minor groove (minor groove binders or MGBs) in EMSA inhibition experiments. Distamycin A is a sequence-specific DNA
over two orders of magnitude and implying a remarkable degree of selectivity for these drugs. It will be interesting to test the role of such compounds in functional CCAAT-dependant assays.

IC substitution experiments prove that despite minor groove binding, the +6 A and +7 T retain remarkable sequence specificity. Other minor groove binding factors, like TBP and HMG, are far less susceptible to single nucleotide changes (53–56), in fact, it has been shown that the entire TATA box can be mutated into IC sequences and still bind TBP normally (51,52). This is certainly not possible for NF-Y, which has a high degree of discriminatory power not only in the major groove, but for minor groove constituents as well.

Both methylation interference and competition with MGBs suggest that the NF-Y binding site can be divided in two halves: the 5'-part (PuPuCC) is in contact with the major groove, while the 3'-part (AATC/AG/G) interacts with the minor groove. NF-Y seems to lie on one side, contacting at least one full turn of the double helix. The apparent asymmetry of the CCAAT box probably mirrors asymmetric contact points on the NF-Y heteromeric protein complex. Perhaps the best known example of an asymmetric binding site is represented by the TATA box: TBP however, thanks to two direct repeats present in the C-terminal conserved domain, contains a remarkable symmetry in the protein surfaces presented for association with other polypeptides (40–43). The polarity comes from the non-conserved N-terminal region. No repeat is found in the DNA binding domains of the three NF-Y subunits contacting DNA. We can conclude that asymmetry is the rule for this transcription factor, both at the DNA and at the protein levels. This finding could imply important functional consequences. The CCAAT box, in fact, is present in both the forward (globins, albumin and β-actin) and reverse orientations (MHC Class II), statistically equally well represented (1). Moreover, the distance between the element and the binding sites of several other factors in different promoters is important, including in MHC Class II promoters. In some cases NF-Y seems to favour the binding of transcription factors to nearby sequences: RFX to the Dra X box (69) and C-EBP to the albumin D box (70).

The polarity of NF-Y binding could then be essential for functional consequences. The CCAAT box, in fact, is present in both the forward (globins, albumin and β-actin) and reverse orientations (MHC Class II), statistically equally well represented (1). Moreover, the distance between this element and the binding sites of several other factors in different promoters is important, including in MHC Class II promoters. In some cases NF-Y seems to favour the binding of transcription factors to nearby sequences: RFX to the Dra X box (69) and C-EBP to the albumin D box (70).

The polarity of NF-Y binding could then be essential for functional consequences. The CCAAT box, in fact, is present in both the forward (globins, albumin and β-actin) and reverse orientations (MHC Class II), statistically equally well represented (1). Moreover, the distance between this element and the binding sites of several other factors in different promoters is important, including in MHC Class II promoters. In some cases NF-Y seems to favour the binding of transcription factors to nearby sequences: RFX to the Dra X box (69) and C-EBP to the albumin D box (70).

The polarity of NF-Y binding could then be essential for functional consequences. The CCAAT box, in fact, is present in both the forward (globins, albumin and β-actin) and reverse orientations (MHC Class II), statistically equally well represented (1). Moreover, the distance between this element and the binding sites of several other factors in different promoters is important, including in MHC Class II promoters. In some cases NF-Y seems to favour the binding of transcription factors to nearby sequences: RFX to the Dra X box (69) and C-EBP to the albumin D box (70).

ACKNOWLEDGEMENTS

We would like to thank S.Ottolenghi for critically reviewing the manuscript and M.Bianchi for the gift of materials and helpful discussions. This work was supported by a Pharmacia/FCE grant to RM and Progetto Finalizzato ACRO and AIRC grants to S.Ottolenghi. AR is supported by Fondazione Italiana L.Giambrone per la Guarigione dalla Talassemia. MB is supported by a Milano Ricerche fellowship.

REFERENCES
