Genome wide study of NF-Y type CCAAT boxes in unidirectional and bidirectional promoters in human and mouse

Antti Häkkinen, Shannon Healy, Howard T. Jacobs, Andre S. Ribeiro

www.elsevier.com/locate/yjtbi

Genome wide study of NF-Y type CCAAT boxes in unidirectional and bidirectional promoters in human and mouse

Antti Häkkinen, Shannon Healy, Howard T. Jacobs, Andre S. Ribeiro

PII: S0022-5193(11)00227-X
DOI: doi:10.1016/j.jtbi.2011.04.027
Reference: YJTBI 6458

To appear in: *Journal of Theoretical Biology*

Received date: 15 May 2010
Revised date: 5 April 2011
Accepted date: 23 April 2011

Cite this article as: Antti Häkkinen, Shannon Healy, Howard T. Jacobs and Andre S. Ribeiro, Genome wide study of NF-Y type CCAAT boxes in unidirectional and bidirectional promoters in human and mouse, *Journal of Theoretical Biology*, doi:10.1016/j.jtbi.2011.04.027

Genome wide study of NF-Y type CCAAT boxes in unidirectional and bidirectional promoters in human and mouse

Antti Häkkinen[a], Shannon Healy[a], Howard T. Jacobs[b], Andre S. Ribeiro[a,*]

[a] *Laboratory of Biosystem Dynamics, Computational Systems Biology Research Group, Dept. of Signal Processing, Tampere University of Technology, Finland*
[b] *Institute of Medical Technology and Tampere University Hospital, University of Tampere, Finland*

Abstract

A subset of CCAAT boxes are known binding sites for the transcription factor NF-Y. We characterize their number, mismatches to the consensus sequence, and locations in bidirectional and unidirectional promoter sequences in human and mouse. We confront the findings with an analytical null model of DNA sequences and find that NF-Y type CCAAT boxes play key, but distinct roles in the two types of promoters. They are found above chance in both, but in unidirectional only when having few mismatches. In bidirectional, the relative positions of multiple boxes differ from what is expected by chance, suggesting the need for contiguity. In agreement, when there are four boxes (four-box configurations), these have much lower number of mismatches than expected in bidirectional promoters alone. Positioning of the first box differs in the two types of promoters and the null model, and mismatches and positioning are found to be correlated. Finally, four-box configurations are conserved between human and mouse, supporting the relevance of the findings. We conclude that bidirectional and unidirectional promoters, while sharing some similarities, appear to possess distinct regulatory mechanisms at the sequence level.

Keywords:

*Corresponding author
Email addresses: antti.hakkinen@tut.fi (Antti Häkkinen), shannon.healy@tut.fi (Shannon Healy), howard.t.jacobs@uta.fi (Howard T. Jacobs), andre.ribeiro@tut.fi (Andre S. Ribeiro)

DNA sequences, Transcription Factor, Transcription Factor Binding Site

1. Background

In mammalian genomes many pairs of genes are co-regulated by bidirectional promoters (Engstrom et al., 2006). These pairs are organized in a head-to-head configuration with transcription start sites (TSS) less than 1 kb apart (Trinklein et al., 2004). The transcriptional activity of the genes in such pairs is, in general, co-regulated (Zanotto et al., 2007).

A recent study (Zanotto et al., 2008) characterized the bidirectional promoter of two mouse nuclear genes encoding components of the mitochondrial translational apparatus, mitoribosomal protein S12 (Mrps12) and mitochondrial seryl-tRNA ligase (Sarsm). Their TSSs are less than 200 bp apart, the intervening region containing four CCAAT boxes, which are transcription factor binding sites (TFBS) of the transcription factor (TF) NF-Y that facilitate efficient bidirectional transcription. These boxes play an identical role in the homologous human promoter. Altering the sequences of any of the boxes, e.g. by deletion, it is possible to change the relative rates of transcription of the two genes while not necessarily changing their combined transcriptional activity, or vice versa (Zanotto et al., 2009).

A similar array of CCAAT boxes interacting with NF-Y regulates the human SOX3 promoter (Krstic et al., 2007), which is unidirectional, suggesting that these boxes may also allow the transcriptional regulation of unidirectional promoters. The existence of multiple NF-Y type CCAAT boxes in promoters is common in cell cycle-regulated genes (Colter et al., 2005; Li et al., 1998; Muller et al., 2007). The boxes appear to recruit histone acetylase complexes such as p300 (Caretti et al., 2003; Salsi et al., 2003) and other TFs (Roder et al., 1999; Imbriano et al., 2005). Defects or malfunctioning of these CCAAT boxes have been linked to several diseases, including some forms of cancer (Farina et al., 1999; Pabst et al., 2001).

Previous studies suggest that NF-Y type CCAAT boxes are common regulators of bidirectional mammalian promoters (Lin et al., 2007; Zanotto et al., 2009). However, several questions remain regarding their role as regulators of gene expression. Are there any common patterns in the number, sequence, orientation or spacing of these boxes between different classes of promoters? Are these box arrays also present in unidirectional promoters? If so, is their function the same as in bidirectional promoters? What are the mismatch

distributions of these TFBSs in bidirectional and unidirectional promoters and are these mismatches informative of their functioning? Finally, are the sequences equally conserved in both type of promoters, and if not, why?

These questions can only be addressed using a proper "null model", since only by comparing the findings on numbers, positioning, and variability with a null model it is possible to determine whether these features are under selection, or are what is expected by chance.

Here we present the first genome-wide study of the occurrence of NF-Y type CCAAT boxes in bidirectional and unidirectional promoters combined with an exact null model to rigorously evaluate the statistical significance of the findings. Given the importance of NF-Y type CCAAT boxes in the regulation of many genes, it is of interest to analyze their presence at the genome-wide level. We perform this analysis in the promoter sequences of human and mouse, by searching and characterizing the number, the mismatches, and locations of these boxes in bidirectional and unidirectional promoters. We then confront the results with a null model here proposed. Our procedure can be used to search for any sequence with a predefined number of possible mismatches, and quantify precisely whether their number and location differ significantly from what is expected by chance.

2. Methods

2.1. Search of NF-Y type CCAAT boxes

The DNA sequences of the complete genome of human (Homo sapiens) and mouse (Mus musculus), were extracted from UCSC Genome Browser (`http://genome.ucsc.edu/`) database (December 1st, 2008). Promoter sequences were extracted according to the annotations of TSSs.

To extract bidirectional promoters we define a consecutive sequence of nucleotides (substring) of less than 1 kb (kilobase) in length as a bidirectional promoter if and only if there is a TSS downstream of the sequence in both strands and the substring is not a substring of another promoter determined by this criteria (Trinklein et al., 2004). Conversely, a substring of 1 kb in length is defined as a unidirectional promoter if and only if a TSS is downstream of the substring and one (or both) of the following conditions are met: i) the same strand has another TSS annotation within 1 kb, or, ii) the reverse complementary strand has no annotated TSS within 10 kb.

The consensus sequence for NF-Y binding CCAAT box was defined to be 5'-`YRRCCAATCA`-3' (Bi et al., 1997; Koessler et al., 2004) and also its re-

verse complement (i.e. 5'-TGATTGGYYR-3'), according to (Mantovani, 1998), as previously used in (Zanotto et al., 2009). In previous studies, the reverse orientation of the NF-Y binding CCAAT boxes was found to be slightly favored over the forward orientation (Mantovani, 1998). We note that this is only applies to unidirectional promoters, since in bidirectional promoters the orientation of the box is ambiguous. As certain mismatches in the sequences do not affect the binding significantly (Mantovani, 1998; Zanotto et al., 2009), we allow mismatches to some extent in all positions but the CCAAT box, as evidence suggest that this box is strictly required for NF-Y to bind (Dorn et al., 1987; van Huijsduijnen et al., 1987). In case of overlapping matches, only the first is recorded. Overlapping is possible and is more likely as the number of allowed mismatches increases. Such possibility is accounted for in the null model.

2.2. Generalized null model

The significance of the findings is accessed by statistical hypothesis testing. Let Σ be a finite alphabet and let there be $\gamma \in \mathbb{N}$ random strings, denoted by $R_i^{\ell_i}$. Let the length of the ith string be iid (independent and identically distributed) $\ell_i \sim \mathcal{D}_\ell$, and let each character be iid $c_{ik} \sim \mathcal{D}_c$. The $\mathcal{D}_\ell$ and $\mathcal{D}_c$ are arbitrary univariate discrete distributions with support of $\ell \in \mathbb{N}$ and $c \in \Sigma$, respectively. To determine the probability that a random string $R_i^{\ell_i}$ contains one or more sequences $S_1, \ldots, S_k$ as a substring, say $\mathrm{P}[R_i^{\ell_i} \in L]$, we find the regular language L and build a deterministic finite state automaton (DFA) that recognizes that language.

Let $M = (Q, q_0, A, \Sigma, \delta)$ be the DFA, Q being the finite set of states, $q_0 \in Q$ the initial state, $A \subseteq Q$ the accepting states, Σ the alphabet, and $\delta(q, c) : Q \times \Sigma \mapsto Q$ the transition function. Let $\mathbf{P}$ be the stochastic matrix of size $|Q| \times |Q|$ corresponding to the DFA M, i.e. if $I(q) : Q \mapsto \{1, \ldots, |Q|\}$ is an arbitrary bijective mapping $\mathbf{P}$ is defined as follows:

$$\forall\ q, q' \in Q : [\mathbf{P}]_{I(q),I(q')} = \sum_{q'=\delta(q,c)} \mathrm{P}\left[c_{ik} = c\right] \ , \tag{1}$$

where $\mathrm{P}[c_{ik} = c]$ is the probability of a random character $c_{ik} \in \Sigma$ being c. The probability for the DFA M to transit from state i to state j through exactly k transitions is given by $[\mathbf{P}^k]_{i,j}$, thus:

$$\mathrm{P}[R_i^{\ell_i} \in L] = \mathbf{e}_{I(q_0)}{}^* \mathbf{P}^{\ell_i} \left(\sum_{q \in A} \mathbf{e}_{I(q)} \right) \ , \tag{2}$$

where $\mathbf{e}_i$ is the unit vector having the ith element equal to 1.

We construct the DFA M using the Aho-Corasick algorithm (Aho and Corasick, 1975), for which $|Q|$ is proportional to the sum of the length of the strings. To obtain the DFAs for multiple matches, we concatenate several of the DFAs. Next, e.g. given that k out of γ sequences met a certain condition, we use the two-tailed hypothesis testing to determine if the findings differ from the by-chance findings suggested by the null model. The p-value is:

$$p = 1 + F(\mathrm{E}[K] - |k - \mathrm{E}[K]|) - F(\mathrm{E}[K] + |k - \mathrm{E}[K]|) \ , \tag{3}$$

where $K \sim \mathcal{D}_K$ is a random variable representing the number of promoters meeting the criterion under the null hypothesis H_0, $\mathrm{E}[K]$ is its expected value, and $F(k) = \mathrm{P}[K \leq k]$ is the cumulative distribution function (CDF) of the distribution $\mathcal{D}_K$. It follows that the distribution is a binomial $\mathrm{Bin}(\gamma, \mathrm{P}[R^{\mathcal{D}_\ell} \in L])$, where:

$$\mathrm{P}[R^{\mathcal{D}_\ell} \in L] = \sum_{j=0}^{\infty} \mathrm{P}[\ell = j] \ \mathrm{P}[R_i^j \in L] \ , \tag{4}$$

is the probability mass of the mixture density of the probability mass functions (PMF) of different promoter lengths.

To determine the spatial probability that the nth match is at position k, we construct languages L^i such that L^i recognizes i (or more) matches to the consensus. Denoting a random string of length ℓ by R^ℓ, the probability is:

$$\mathrm{P}[\cdot] = \mathrm{P}[k \leq \ell] \left(\mathrm{P}[R^k \in L^n] - \mathrm{P}[R^{k-1} \in L^n] \right) \ , \tag{5}$$

for a given length distribution $\mathcal{D}_\ell$. The position k denotes the ending position of a match. To obtain the distance between $(n-1)$th and nth match, we modify the length distribution s.t. instead of $\mathcal{D}_\ell$, we use the distribution of the remaining lengths after exactly $(n-1)$ matches, $\mathcal{D}_\ell{}'$. Eq. 5 can be readily evaluated using the PMF obtained using the convolution:

$$f_\ell{}'(k') = \sum_{k=k'+1}^{\infty} f_\ell(k) \left(\mathrm{P}[R^{k-k'} \in L^{n-1}] - \mathrm{P}[R^{k-k'-1} \in L^{n-1}] \right) \ , \tag{6}$$

Several kinds of null hypotheses about DNA sequences have been used (Fitch, 1983). Typically, one or more sequences of non-related DNA are used, either as-is or randomly shuffled. Random shuffling guarantees the

independence between nucleotides but this is not strictly required, as our method can be extended for Markov chains of order up to $\min_i |S_i|$, and, with some modifications, even higher.

It is known that in many promoters in human and mouse, some dinucleotides are more common than others (Bajic et al., 2006). The different multinucleotide distributions can be imposed to the null model (e.g. using first- or second-order Markov chains). Instead, we use the alphabet $\Sigma = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$, and the experimentally observed PMFs for the nucleotide distribution $\mathcal{D}_c$ and the promoter length distribution $\mathcal{D}_\ell$. This is because it is unknown whether and to what extent multinucleotide distributions in mouse and human DNA sequences are due to physical-chemical constrains or due to evolutionary pressure. Since we aim to use a null model whose sequence is absent of any evolutionary pressure, we opted for not imposing any dependence in the distribution of nucleotides.

In that regard, for example, while many promoters have CpG islands and TATA boxes, many others do not (Bajic et al., 2006). This implies that they do not have to be necessarily present due to any physical-chemical constraints, and that selection can remove or add these motifs and alter the ratio of dinucleotides. Therefore, our null model should not assume a priori these distributions, if it is to be used to determine if a sequence has been selected in or out.

3. Results and Discussion

3.1. NF-Y type CCAAT box occurrence

First, we searched for NF-Y type CCAAT boxes in human and mouse promoters, as NF-Y (also called CBF) is currently considered the principal CCAAT-box-binding activator (Mantovani, 1998). The search for the consensus-binding sequence of NF-Y was made in both strands, allowing varying number of mismatches in specific nucleotides, as described. The findings, their corresponding expected probabilities given by the null model, and the hypothesis testing are shown in Table 1.

The occurrence of NF-Y type CCAAT boxes is similar in human and mouse although substantially different from their null models. In bidirectional promoters, the boxes were found above chance. This finding is in agreement with a study using chromatin-immunoprecipitation (ChIP) on 118 regions in the human genome that showed that CCAAT boxes, among a few

Table 1: Percentage of promoters meeting each criterion in human, mouse, and null model promoter sequences, and the results of hypothesis testing at significance level $\alpha = 0.01$. Percentages of findings above chance in dark gray background, below chance in light gray background and, if the null hypothesis H_0 is not rejected, the background is white. Criteria of type k-N-ε denote k or more NF-Y type CCAAT boxes, up to ε mismatches each.

	Human		Mouse		Human (null)		Mouse (null)	
Criterion	Bi	Uni	Bi	Uni	Bi	Uni	Bi	Uni
γ	1718	37201	1447	31407	N/A	N/A	N/A	N/A
1-N-0	4.77	2.91	7.81	4.22	0.29	1.48	0.32	1.54
1-N-1	16.82	13.95	21.15	16.04	3.06	13.84	3.28	14.41
1-N-2	27.65	36.45	32.41	39.14	11.53	44.94	12.26	46.40
1-N-3	34.81	57.42	39.74	61.75	22.30	72.33	23.56	73.91
1-N-4	38.77	68.23	42.78	72.99	28.62	83.05	30.16	84.38
1-N-5	39.46	70.44	43.33	75.22	30.17	85.16	31.77	86.41
2-N-0	0.41	0.21	0.62	0.25	0.00	0.01	0.00	0.01
2-N-1	4.54	2.20	6.57	2.50	0.07	0.99	0.08	1.07
2-N-2	10.77	10.37	13.06	10.95	1.08	11.87	1.21	12.76
2-N-3	16.18	24.84	19.49	28.14	4.15	36.30	4.60	38.35
2-N-4	17.99	35.81	22.11	40.66	6.97	52.40	7.65	54.81
2-N-5	18.86	38.47	22.60	43.63	7.77	56.27	8.52	58.71
3-N-0	0.06	0.04	0	0.04	0.00	0.00	0.00	0.00
3-N-1	1.40	0.58	1.31	0.41	0.00	0.05	0.00	0.05
3-N-2	4.89	2.88	5.74	2.98	0.08	2.16	0.10	2.42
3-N-3	7.45	9.10	9.26	10.40	0.67	13.32	0.80	14.64
3-N-4	8.85	15.98	10.44	18.67	1.49	25.27	1.75	27.40
3-N-5	8.96	17.93	10.78	20.90	1.77	28.75	2.07	31.08
4-N-0	0	0.01	0	0.01	0.00	0.00	0.00	0.00
4-N-1	0.70	0.17	0.55	0.11	0.00	0.00	0.00	0.00
4-N-2	1.98	0.99	2.28	0.85	0.01	0.29	0.01	0.34
4-N-3	3.73	3.17	4.22	3.36	0.09	3.77	0.12	4.32
4-N-4	4.42	6.38	5.04	7.34	0.27	9.60	0.35	10.83
4-N-5	4.42	7.45	5.11	8.59	0.35	11.64	0.44	13.09

other small motifs, are over-represented in bidirectional promoters, in comparison to unidirectional ones (Lin et al., 2007).

Interestingly, our analysis of a much larger set of promoters further shows that in unidirectional promoters, the boxes are also found above chance, provided that they have a low number of mismatches from consensus. However, unlike in bidirectional, they were found below chance when a large number of mismatches from consensus were allowed. The striking differences between our findings in bidirectional and unidirectional promoters suggest that the boxes perform different functions in the two types of promoters.

We hypothesize that, in general but not exclusively, each box functions as an independent TFBS in unidirectional promoters, while in bidirectional there are usually arrays of multiple boxes, as in Mrps12/Sarsm in both mouse and human (Zanotto et al., 2007, 2009) which possesses four such boxes. NF-Y binding is likely to determine the ratio of expression of the two genes in such pairs, depending on how it interacts with the boxes, which depends on their number, mismatches, and positioning in the template. This was shown to be the case in Mrps12/Sarsm promoter, as deleting any of the boxes alters the ratio between the transcription rates of the two genes (Zanotto et al., 2009). Further, the deletion of boxes usually also affects the ratio by which each of the two genes is expressed, rather than their combined expression rate. Importantly, no single deletion blocked transcription completely (Zanotto et al., 2007, 2008). This suggests that one can expect several configurations of these boxes in promoters, instead of a single fixed configuration.

These observations are supported by our finding that the boxes in bidirectional promoters can sustain more mismatches from consensus. The occurrence, above chance, of boxes with many mismatches indicates that they maintain some functionality (otherwise they would not exist beyond what is expected by chance), and act co-operatively. On the other hand, since in unidirectional promoters boxes with high number of mismatches are "selected out" (i.e. are found below chance), the same is likely not to occur in unidirectional promoters. Interestingly, if boxes with many mismatches were totally non-functional in unidirectional promoters, their amount should be similar to what is predicted by the null model. Instead, they were found below chance. We suggest that such boxes with mismatches allow a TF, other than NF-Y, to bind, creating a co-operative array which would be pernicious in some way to the function of unidirectional promoters.

3.2. First box position

The boxes form co-operative arrays in some bidirectional promoters. Thus, we studied the spatial locations of the boxes in the two types of promoters, starting with the location of the first box found upstream of each TSS in human and mouse promoters.

Results for human and mouse are shown in Fig. 1, and are almost identical. The probability distributions for the locations of NF-Y type CCAAT boxes in bidirectional promoters are shown in the top plot of Fig. 1. There is a region adjacent to the 5' end of the TSS (from −50 to 0) that is less prone than in the null model to contain the first box, suggesting that, during evolution, the boxes have been selectively eliminated from this region; alternatively, some other class of sequence element is constrained to be located in this region. Finally, another possible explanation is that placing a CCAAT box in this region causes a change in the selection of the TSS as reported in (Kabe et al., 2005), since NF-Y plays a role in the recruitment of RNA polymerase II. NF-Y type CCAAT boxes were most common in the subsequent region, between −120 and −50 (particularly around position −85). Beyond that, several small peaks were found, e.g. around −200, −350 (in human only), and −580. Beyond −600, NF-Y type CCAAT box occurrence was the same as expected by chance, implying that such boxes located at this distance or greater from the TSS might not be functional.

As in the bidirectional case, in unidirectional promoters (bottom plot in Fig. 1) there was also a region where the first box was preferentially located, again around −85, with a comparatively barren region closer to the TSS. Beyond that and until approximately −800, the findings appear to match the predictions of the null model. Next, there is a larger peak centered on −930. Specific reasons for the existence of a peak in this location are unknown. It is noted that the y-axis scales on the upper and lower figures are not the same, and that the findings in unidirectional thus differ less from the null model than in the case of bidirectional promoters.

3.3. Subsequent box positions

We determined the relative positions of each subsequent NF-Y type CCAAT box (when present) up to the fourth box relative to the position of the first such element. The positions of these boxes for bidirectional promoters are presented in Fig. 2. The results for human and mouse are similar and show significant "non-random" positioning. In general, each subsequent NF-Y type CCAAT box appeared to be preferentially located at approximately −40 from

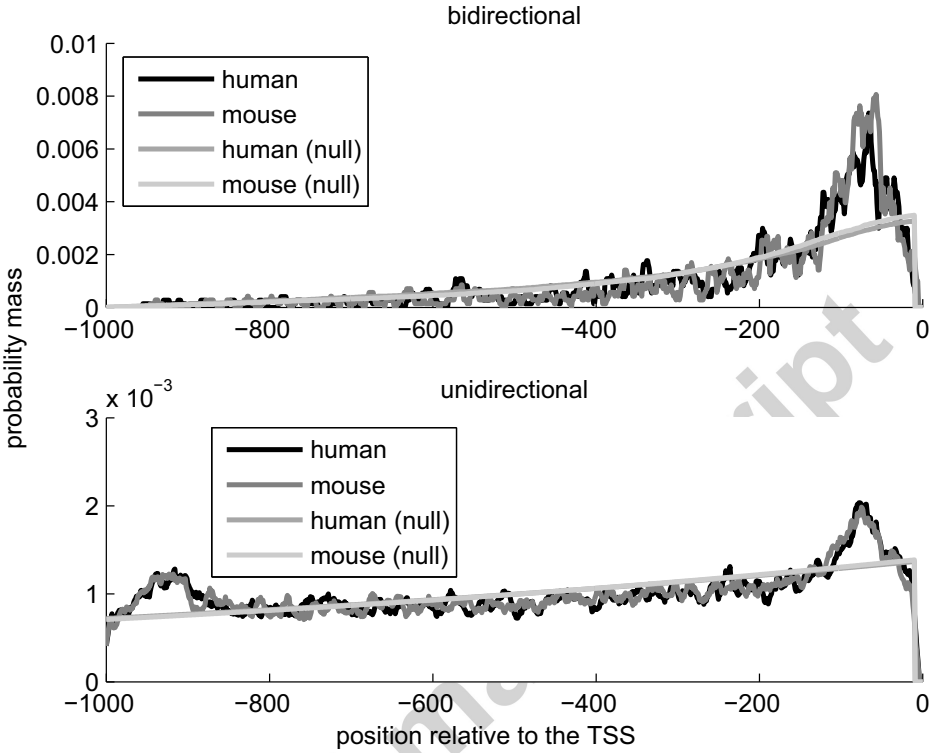

Figure 1: First box position relative to the TSS, up to 3 mismatches. Upper figure bidirectional promoters, lower figure unidirectional promoters. Experimental results were smoothed using moving average window of 10.

the previous box. The region immediately adjacent to the previous box was strongly disfavored. These features were found to be slightly more robust in mouse than in human. The −40 peak was also found in unidirectional promoters, but was far less pronounced peak. Thus, we conclude that, at least in bidirectional promoters, NF-Y type CCAAT boxes function in combination, in sets of two to four boxes, with very specific distances between them. Relative positioning is thus inferred to be a condition for functionality. It may be noted that the chance of finding two, three or four boxes arrayed with this precise spacing is increasingly small. Thus, the relative positioning of multiple boxes, when present, appears to be under stronger selection than the position of the first box.

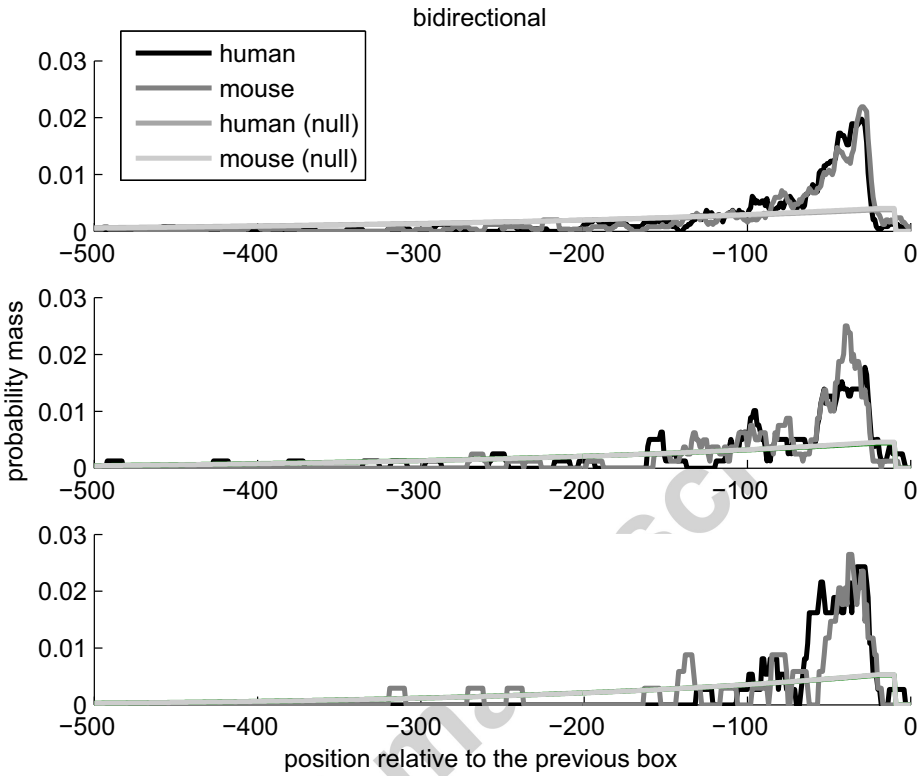

Figure 2: Second (top), third (middle), and fourth (bottom) box positions relative the previous box, up to 3 mismatches, in bidirectional promoters. Results from the searches in mouse and human sequences were smoothed using a moving average window of 10.

In unidirectional this is less frequent, as depicted by Fig. 3. Interestingly, in bidirectional, the strongest correlation in position is between the third and the fourth boxes, whereas in unidirectional the probability of precise spacing is similar between all boxes.

We found that in mouse, on average, the distance between the boxes is slightly smaller than in human, especially the relative positioning of boxes 2, 3 and 4. In general, the box positions approximately match the configuration in the Mrps12/Sarsm bidirectional promoter (Zanotto et al., 2009), with four consecutive, equidistant boxes (≈ 40 nucleotides apart), suggesting that NF-Y binding to a precisely spaced array of boxes might commonly regulate the expression of genes driven by bidirectional promoters.

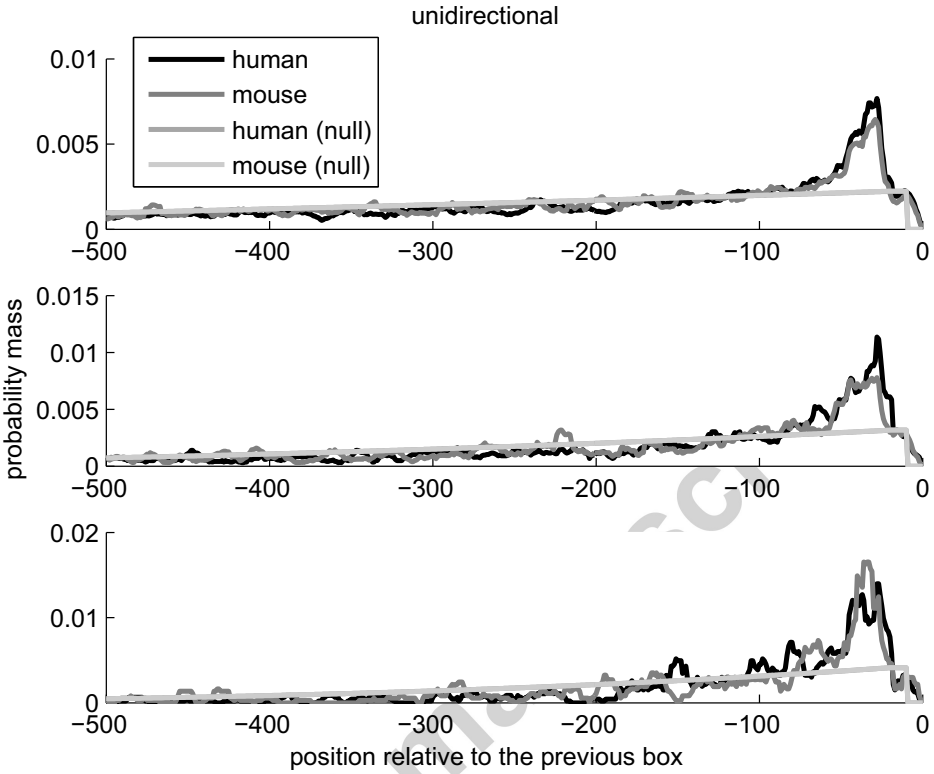

Figure 3: Second (top), third (middle), and fourth (bottom) box positions relative the previous box, up to 3 mismatches, in unidirectional promoters. Results from the searches in mouse and human sequences were smoothed using a moving average window of 10.

3.4. Four-box configurations

We analyzed promoters with exactly four NF-Y type CCAAT boxes, allowing up to 5 mismatches from consensus in each such element. Previous studies of the Mrps12/Sarsm bidirectional promoter (Zanotto et al., 2009) showed that NF-Y binds to a specific spatial configuration of four consecutive boxes, approximately 40 nucleotides apart from each other. However, measurement of the transcriptional activity in promoters with boxes deleted showed that most single and double deletions do not completely hamper expression (Zanotto et al., 2007, 2008, 2009), although they can change the ratio of expression levels of the two genes of the pair. Thus, it is likely that multiple variations of the box configuration (in both number and position

of boxes) in the Mrps12/Sarsm promoter might still be under regulation by NF-Y.

The locations of boxes in bidirectional promoters with four boxes are shown in Fig. 4. The relative positions, in general, differ significantly from what would be expected by chance. In general, they are contained within small regions as expected given the preferential distance between consecutive boxes (Fig. 2). As an example, 50% of the sets of four boxes in human and 38% in mouse are within a range of 200 nucleotides or less, significantly above from what would be expected by chance (Table 2).

In unidirectional promoters (figure not shown), the higher variability of distances between the boxes suggests that their relative position does not follow a specific pattern. This implies that, in general, they are independent TFBSs, which are nevertheless likely to be functional as they were found to occur above chance.

We measured the mean μ and variance σ^2 of distances between consecutive boxes and the correlation ρ between their positions (Table 2). In bidirectional promoters the locations of the boxes was highly correlated, above 0.9. The only exception was between the first and second boxes in human, which are weakly correlated. One possible explanation is that, in human, three-box configurations might be more common than four-box ones, compared with mouse, with any additional (fourth) box functionally superfluous. However, it is noted that such proportions would change slightly if one had considered independently cases with five or more boxes. Such promoters might have four "active" boxes and "superfluous" ones.

Interestingly, the correlation between the box positions in unidirectional promoters is slightly higher than in the null model (but much lower than in bidirectional promoters). One possibility is that some apparently unidirectional promoters are in fact bidirectional, with the "other" direction used to transcribe e.g. a short non-coding RNA that has not yet been mapped.

Generally, the findings suggest that, in unidirectional promoters, NF-Y type CCAAT boxes do not function in the same manner as in bidirectional promoters since, in the unidirectional case, there is much less conservation across promoters and correlation between the positions of different boxes in each promoter.

We studied the degree of conservation of the four-box configurations between human and mouse. Although there is some evidence suggesting that certain dinucleotides might mutate at faster rates than others, it is unknown if such higher rates result in faster changes on these sites in long time scales,

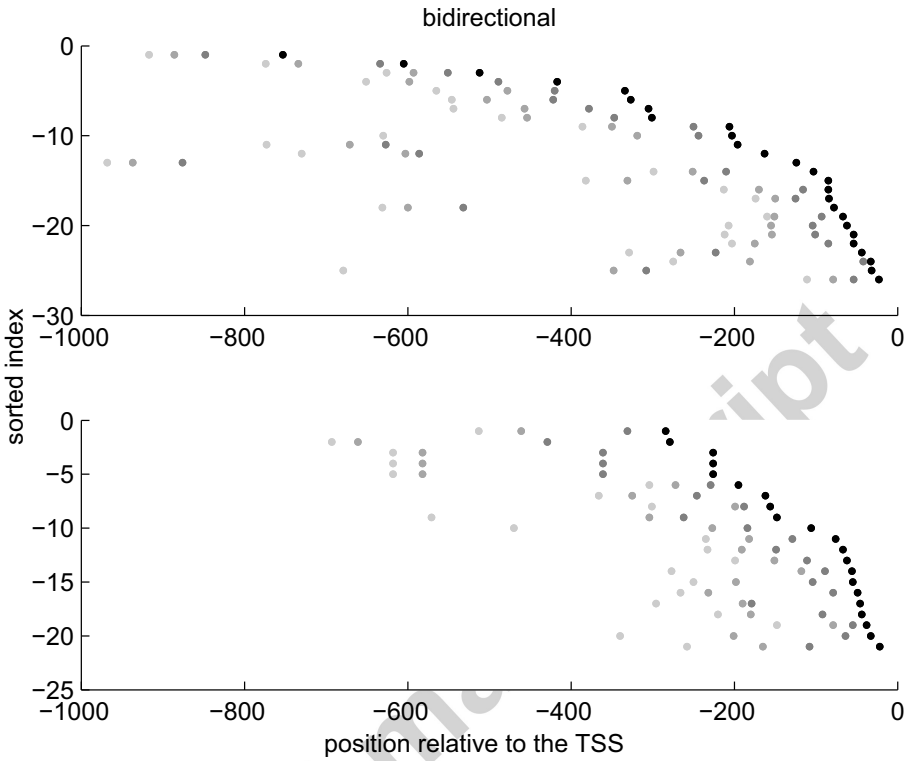

Figure 4: All four-box configurations found in bidirectional promoters, sorted by the position of the first box. Human (upper plot) and mouse (lower plot). First box in black and the next boxes are in increasingly lighter grey.

as that depends on if the changes are conserved or not. Thus, rates of mutation are not considered when computing significance of findings. The significance of findings is made by confronting the conservation rate of the boxes with the mean conservation rate of all nucleotides of the promoters.

We first located each human (mouse) bidirectional promoter and their neighboring genes. Then, using Ensembl (`http://www.ensembl.org/`), we obtained the protein sequences produced by the sense and antisense genes, and mapped them to the best matching mouse (human) protein sequences using BLAST (Altschul et al., 1990), thus obtaining the corresponding genes. The intervening region between the two genes was selected for further analysis. We found that 18 out of 26 cases in human and 14 out of 21 cases in

Table 2: Spatial features of four-box configurations in unidirectional and bidirectional promoters along with the null predictions. Null prediction was obtained using Monte-Carlo simulations rather than analytically. X_i denote random variables of the position of ith box and $\mathrm{P}[\cdot] = \mathrm{P}[X_4 - X_1 \leq 200]$ the fraction of four-box promoters where boxes separated by 200 bp or less.

	Human		Mouse		Human (null)		Mouse (null)	
Measure	Bi	Uni	Bi	Uni	Bi	Uni	Bi	Uni
$\mathrm{P}[\cdot]$	0.50	0.10	0.38	0.08	0.06	0.03	0.03	0.02
ρ_{X_1,X_2}	0.67	0.70	0.95	0.72	0.41	0.61	0.66	0.60
ρ_{X_2,X_3}	0.99	0.72	0.95	0.78	0.77	0.66	0.73	0.66
ρ_{X_3,X_4}	0.95	0.71	0.92	0.67	0.74	0.60	0.68	0.63
$\mu_{X_2-X_1}$	140.5	180.3	74.0	181.8	191.8	199.1	173.7	202.5
$\mu_{X_3-X_2}$	65.1	187.1	93.9	178.8	150.3	204.1	150.1	201.5
$\mu_{X_4-X_3}$	72.2	175.5	81.5	190.1	151.2	200.8	184.0	193.1
$\sigma_{X_2-X_1}$	181.0	170.1	43.3	158.5	150.6	160.8	121.7	164.1
$\sigma_{X_3-X_2}$	30.4	174.1	75.4	149.8	116.2	161.9	118.2	160.8
$\sigma_{X_4-X_3}$	78.3	165.5	68.2	175.1	130.9	160.3	135.2	152.9

mouse (i.e. 69.2% and 66.7%, respectively) could be mapped to the other genome.

Next, for the cases that could be mapped, we obtained the optimal alignment between nucleotides of the pairs of promoters using the Smith-Waterman algorithm (Smith and Waterman, 1981), and from there, the differences between the sequences (fraction of deletions, insertions and conserved nucleotides). Results are shown in Table 3.

We performed an exact multinomial test to obtain a p-value to test if the number of deletions, insertions and conservations in the boxes is similar to that of all nucleotides in the promoter or differs significantly, i.e. to test whether the boxes were more strongly conserved than average in cases i) and ii). The small p-values indicate that the differences between fractions of deleted, inserted and conserved nucleotides between nucleotides that form the boxes and all other nucleotides are significant (significance level $\alpha = 0.01$). From the comparison, in both human and mouse, NF-Y type CCAAT boxes are strongly conserved.

Table 3: Fraction of nucleotides and CCAAT boxes deleted, inserted, and conserved (%) in the aligned subregions when mapping i) the 18 human bidirectional promoters containing exactly four NF-Y type CCAAT boxes to mouse bidirectional promoters and ii) the 14 mouse promoters to human promoters, and the p-values for the distributions to match.

		Total	Deleted	Inserted	Conserved	P-value
i	Nucleotides	18460	30.36	30.96	38.67	7.0×10^{-3}
	Boxes	117	28.21	28.21	43.59	
ii	Nucleotides	9536	29.12	31.39	39.49	9.3×10^{-3}
	Boxes	85	25.88	29.41	44.71	

3.5. Number of mismatches

We investigated the distribution of the number of mismatches. First, we studied these distributions as a function of the position relative to the 5' end of the TSS. In Fig. 5 are the results for bidirectional and in Fig. 6 for unidirectional promoters.

In both types of promoters, the mismatch distributions differ significantly from what is expected by chance. However, the spatial location also seems relevant. Both in unidirectional and bidirectional promoters there were fewer mismatches in the regions where the boxes are preferred, especially in unidirectional (Fig. 6). In bidirectional this was also true but the result is less robust as the sample size is smaller, especially in the regions far from the TSS. Nevertheless, it appears that the regions where the boxes are found above expectation is also where the number of mismatches from consensus is lower, consistent with selection for functionality.

In bidirectional promoters, the distributions differ up to 5 mismatches. However, in unidirectional promoters the distributions for more than 3 mismatches are very similar to the null prediction. This supports our hypothesis, that due to the boxes functioning in combination in bidirectional promoters, more mismatches are allowed. In unidirectional, it appears that for more than 3 mismatches, the boxes are non-functional, as they do not function in co-operation with other boxes.

We analyzed the number of mismatches within four-box configurations. Assuming that, in many cases, the boxes work in combination in bidirectional promoters, but not in unidirectional promoters, we hypothesized that it would be unlikely to find promoters with four boxes with the same mis-

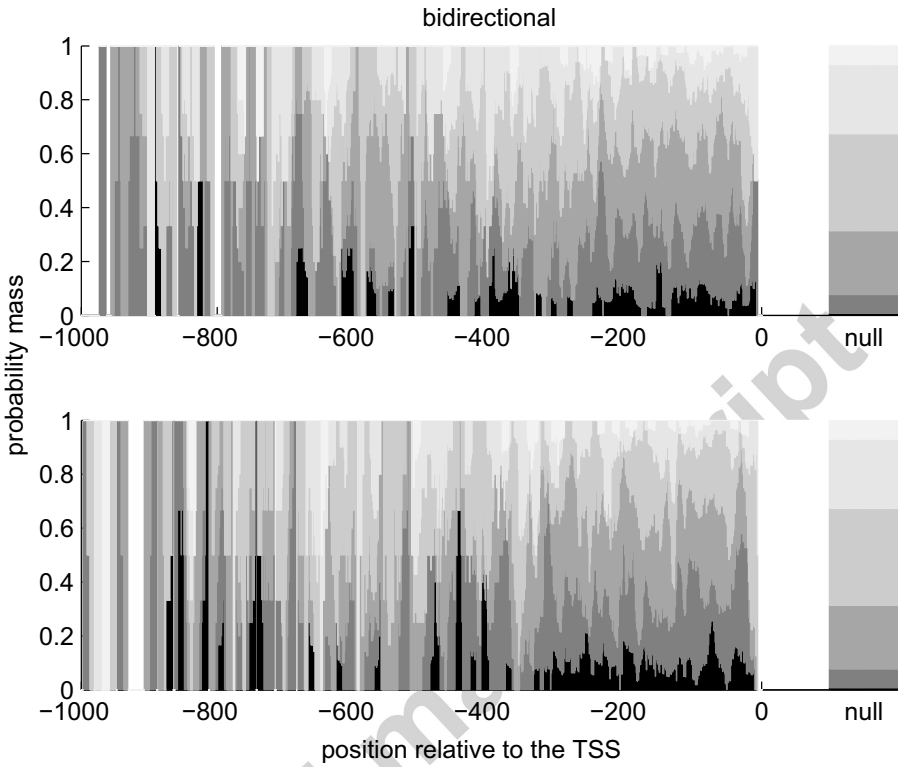

Figure 5: Distribution of mismatches in bidirectional promoters in each position. Upper figure for human, and lower for mouse. Results were smoothed using moving average window of 10. Bar on the right is the null prediction. Black: no mismatches, lightest gray: 5 mismatches.

match distributions in the two types of promoters. Since in unidirectional promoters the boxes are proposed to function as independent TFBSs, we expected results similar to the null model. Conversely, we hypothesized that the mismatch distribution in NF-Y type CCAAT boxes in bidirectional promoters with four boxes would be far from what is expected by chance at the single box level, as we expect the mismatches to be evenly distributed by the four boxes, which is not what the null model predicts. The results of this test are shown in Fig. 7 and confirm this prediction.

The results support the hypothesis. In bidirectional promoters the number of sets of four boxes where all boxes have 3 or less mismatches was

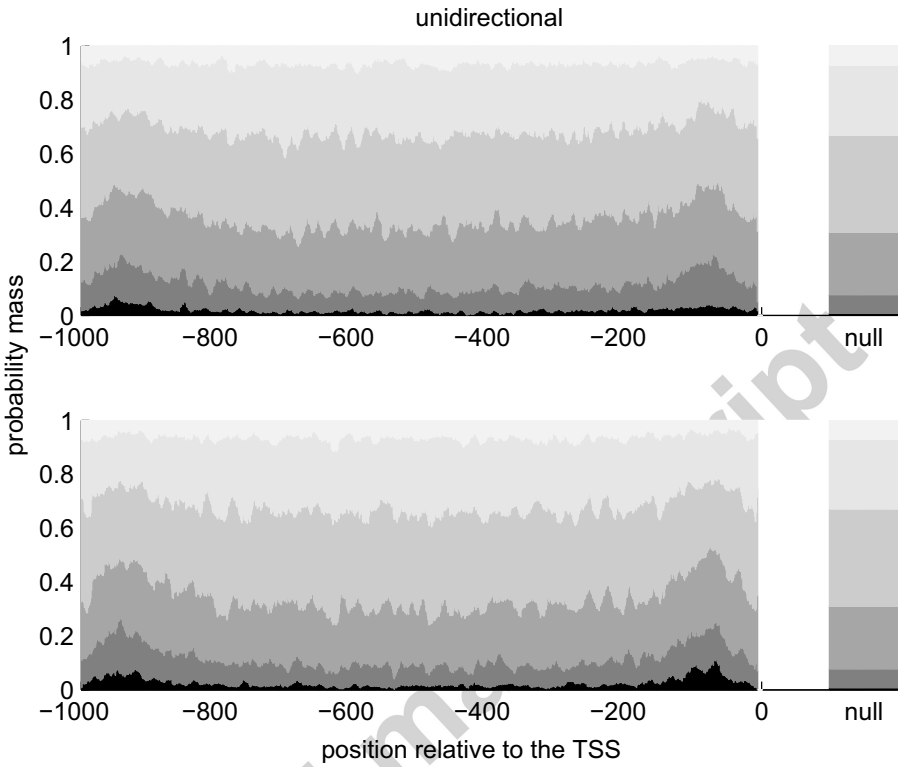

Figure 6: Distribution of mismatches in unidirectional promoters in each position. Upper figure for human, and lower for mouse. Results were smoothed using moving average window of 10. Bar on the right is the null prediction. Black: no mismatches, lightest gray: 5 mismatches.

significantly above chance or, equivalently, the cases where most boxes have a large number of mismatches was significantly below chance, whereas in unidirectional no clear difference from the null model was found.

3.6. Specificity of the findings to NF-Y versus other TFBSs

We also performed additional analysis to study to what extent the findings presented are specific to NF-Y, compared to other similar TFBSs. While allowing for mismatches is necessary to determine the possible binding sites of NF-Y (Zanotto et al., 2009), it may also allow the nucleotide sequences identified as binding sites for NF-Y to also be binding sites for another TF, in the case where the two TFBS are similar. Such an eventuality becomes

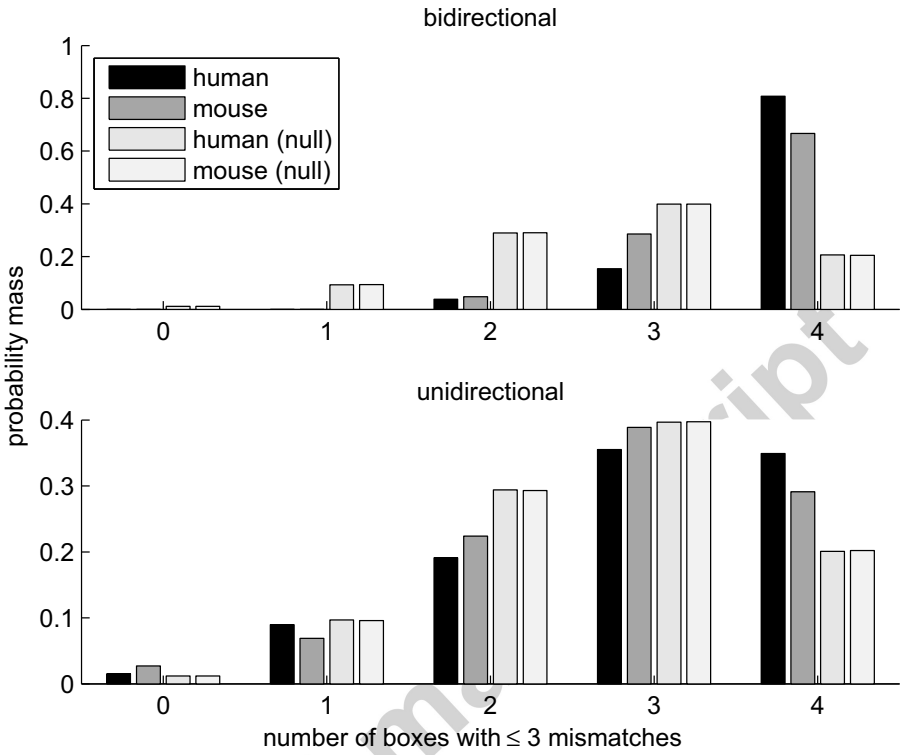

Figure 7: Distribution of number of boxes with up to 3 mismatches in bidirectional (top) and unidirectional (bottom) promoters with four boxes.

more likely, the higher the number of mismatches allowed in the search. Due to this, it is necessary to estimate the probability of this eventuality.

We verified that only a small fraction of our findings are potential binding sites of other TFs. This was done by comparing the findings in human and mouse genomes using the NF-Y binding site consensus with overlapping findings of two other TFBSs, namely the binding sites of transcription factors C/EBP and CTF/NF-I. Their consensus sequences were defined to be 5'-`RTTGCGYAAY`-3' for C/EBP (Osada et al., 1996), and 5'-`TTGGCNNNNNGCCAA`-3' for CTF/NF-I (Roulet et al., 2000).

The results in Table 4 indicate that regardless of the number of mismatches allowed to the NF-Y consensus sequence, there is only a small probability that an overlapping match with low number of mismatches to

a consensus sequence of other type is present. Namely, for any class of the NF-Y findings, there is less than 2% chance for an overlapping finding of the other TFBSs with no more than 1 mismatch, and less than 10% chance for no more than 2 mismatches. Thus, we expect that very few of the findings reported by a match to the NF-Y consensus defined, are functional TFBSs of TFs other than NF-Y.

Table 4: Distributions of the minimum number of mismatches to any overlapping C/EBP or CTF/NF-I consensus sequences, in percentage, for findings with various number of mismatches in NF-Y consensus. N-ε denotes that up to ε mismatches are allowed to the NF-Y consensus sequence.

Criterion	Mismatches to C/EBP consensus						
	0	1	2	3	4	5	6
N-0	0	0.21	7.52	41.11	47.03	4.13	0
N-1	0	0.42	7.56	39.59	47.44	4.98	0.01
N-2	0.01	0.51	7.68	37.88	48.34	5.57	0.00
N-3	0.00	0.59	7.44	35.84	49.48	6.64	0.01
N-4	0.00	0.48	6.76	34.99	50.42	7.33	0.01
N-5	0.00	0.46	6.50	34.37	51.03	7.63	0.01

Criterion	Mismatches to CTF/NF-I consensus							
	0	1	2	3	4	5	6	7
N-0	0.32	0.46	9.19	24.19	46.49	18.56	0.78	0
N-1	0.55	0.60	6.98	24.85	44.95	20.98	1.08	0.01
N-2	0.47	0.68	5.96	23.15	44.43	23.80	1.52	0.00
N-3	0.44	0.57	4.93	21.33	44.55	26.47	1.71	0.00
N-4	0.42	0.48	4.40	19.95	44.68	28.16	1.90	0.01
N-5	0.42	0.46	4.26	19.66	44.60	28.69	1.92	0.01

3.7. Comparison with alternative TFBS mapping techniques

Instead of the consensus-with-mismatches method used here, it is also possible to use other methods to locate putative TFBSs. Position weight matrix (PWM) based techniques (e.g. Staden, 1984) are one such alternative, allowing fuzzy or probabilistic matching, at the expense of higher computational cost.

Table 5: Intervals of PWM thresholds for which the results of the consensus-with-mismatch findings lie within the 90% confidence interval. N-ε denotes that up to ε mismatches are allowed to the NF-Y consensus sequence.

Criterion	PWM threshold range
N-0	$[9.9381, 10.0121]$
N-1	$[7.7843, 7.8172]$
N-2	$[5.3332, 5.3931]$
N-3	$[2.5715, 2.6928]$
N-4	$[-0.5868, -0.3383]$
N-5	$(-\infty, -2.6060]$

To find if such a matrix is well approximated by our choice of consensus sequence, we searched the promoters also with a PWM constructed from the findings reported in (Bi et al., 1997). The PWM was constructed such that the PWM score corresponds to the log-odds of the respective sequence over the background nucleotide distribution present in each set of promoters. For each possible value of threshold we then obtained the 90% confidence interval for the match count, and tested if one of our findings with consensus-with-mismatch technique lies within that interval. This was done by using the Beta-Binomial distribution. The set of ranges of thresholds that the consensus-with-mismatch findings correspond to are presented in Table 5. We found that for each number of mismatches, there is a range of PWM thresholds for which the results of the two methods do not differ significanly.

4. Conclusions

From a genome wide survey of promoter sequences in human and mouse we found that the amount of NF-Y type CCAAT boxes in both unidirectional and bidirectional promoters differ significantly from what is expected by chance, indicating that they play a key role. Further, since the boxes occupy specific locations in bidirectional, more than in unidirectional promoters, their performance is likely to depend on distance to the TSS.

Strikingly, in bidirectional promoters there is a strong correlation between the locations of boxes when multiple boxes exist. This correlation is far less pronounced in unidirectional promoters although still slightly higher than if they were randomly located. Further, in bidirectional promoters, many

boxes have an arbitrary number of mismatches, while in unidirectional, only boxes with a low number of mismatches are found significantly above chance.

The findings support the hypothesis that in bidirectional promoters the boxes function as a co-operative array, while in unidirectional they are independent TFBSs (whose position is under selection but not relative to one another). This suggests that these boxes function inherently differently in unidirectional versus bidirectional promoters, and that in the latter case the binding requires multiple boxes. This would explain why the number of mismatches from consensus, per box, is evenly distributed and higher in bidirectional promoters, since due to co-operation, mismatches on one box affect less the total binding strength between TF and TFBS, than if only one box was required for the binding.

The above conclusions are similar in human and mouse, suggesting similar functions in both, in agreement with the strong conservation observed in sets of four boxes in bidirectional promoters in human and mouse. The results on the location, number and mismatch tolerance in bidirectional promoters, its similarity in mouse and human, and the strikingly different results in unidirectional promoters, suggest that these two types of promoters are likely to be subject to distinct regulatory mechanisms.

The variability in positioning, mismatches and number of boxes in bidirectional promoters with multiple boxes, in both human and mouse, suggests that this regulatory mechanism has plasticity and multi-functionality, i.e. it allows different pairs of genes to differ in total and relative expression of the two genes and might facilitate interactions with other TFs, as suggested in (Dorn et al., 1987).

In the methods, we mentioned that our null model should be free from any constrain in the ordering of the existing nucleotides, as there is no known physical constrain that would impede a certain sequence. The only imposed condition to our null model was the ratio of each nucleotide, which was set to be identical to those in human and mouse. It is of importance to note that we did not necessarily had to impose this as well. In fact, most likely, these ratios were also imposed by selection, as it selected for certain sequences more than others, instead of being due to physical-chemical constraints. For that reason, we also confronted the sequences of human and mouse with a null model where all nucleotides exist in equal ratios. The results do not differ, qualitatively, from those here presented.

Finally, the conservation of sets of four boxes between mouse and human and the overall number of boxes in both genomes confirms their widespread

use. This strengthens the conclusions of previous studies (Kabe et al., 2005) about the critical role of NF-Y as a regulator of transcription of many bidirectional promoters (Mantovani, 1998), and of CCAAT boxes as binding sites for several other TFs (Dorn et al., 1987).

5. Acknowledgements

We thank the Academy of Finland, TEKES, the Sigrid Juselius Foundation, and Tampere Univ. Hospital Medical Research Fund.

References

Aho, A.V., Corasick, M.J., 1975. Efficient string matching: An aid to bibliographic search. Commun. ACM 18, 333–340.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. A basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Bajic, V.B., Tan, S.L., Christoffels, A., Schönbach, C., Lipovich, L., Yang, L., Hofmann, O., Kruger, A., Hide, W., Kai, C., Kawai, J., Hume, D.A., Carninci, P., Hayashizaki, Y., 2006. Mice and men: Their promoter properties. PLoS Genet. 2, e54.

Bi, W., Wu, L., Coustry, F., de Crombrugghe, B., Maity, S.N., 1997. Dna binding specificity of the ccaat-binding factor cbf/nf-y. J. Biol. Chem. 272, 26562–26572.

Caretti, G., Salsi, V., Vecchi, C., Imbriano, C., Mantovani, R., 2003. Dynamic recruitment of nf-y and histone acetyltransferases on cell-cycle promoters. J. Biol. Chem. 278, 30435–30440.

Colter, D.C., Piera-Velazquez, S., Hawkins, D.F., Whitecavage, M.K., Jimenez, S.A., Stokes, D.G., 2005. Regulation of the human sox9 promoter by the ccaat-binding factor. Matrix Biol. 24, 185–197.

Dorn, A., Bollekens, J., Staub, A., Benoist, C., Mathis, D., 1987. A multiplicity of ccaat box-binding proteins. Cell 50, 863–872.

Engstrom, P.G., Suzuki, H., Ninomiya, N., Akalin, A., Sessa, L., Lavorgna, G., Brozzi, A., Luzi, L., Tan, S.L., Yang, L., Kunarso, G., Ng, E.L., Batalov, S., Wahlestedt, C., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., Wells, C., Bajic, V.B., Orlando, V., Reid, J.F., Lenhard, B., Lipovich, L., 2006. Complex loci in human and mouse genomes. PLoS Genet. 2, e47.

Farina, A., Manni, I., Fontemaggi, G., Tiainen, M., Cenciarelli, C., Bellorini, M., Mantovani, R., Sacchi, A., Piaggio, G., 1999. Down-regulation of cyclin b1 gene transcription in terminally differentiated skeletal muscle cells is associated with loss of functional ccaat-binding nf-y complex. Oncogene 18, 2818–28127.

Fitch, W.M., 1983. Random sequences. J. Mol. Biol. 163, 171–176.

van Huijsduijnen, R.A.M., Bollekens, J., Dom, A., Benois, C., Mathis, D., 1987. Properties of a ccaat box-binding protein. Nucleic Acids Res. 15, 7265–7282.

Imbriano, C., Gurtner, A., Cocchiarella, F., Agostino, S.D., Basile, V., Gostissa, M., Dobbelstein, M., Sal, G.D., Piaggio, G., Mantovani, R., 2005. Direct p53 transcriptional repression: in vivo analysis of ccaat-containing g2/m promoters. Mol. Cell. Biol. 25, 3737–3751.

Kabe, Y., Yamada, J., Uga, H., Yamaguchi, Y., Wada, T., 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.

Koessler, H., Kahle, J., Bode, C., Doenecke, D., Albig, W., 2004. Human replication-dependent histone h3 genes are activated by a tandemly arranged pair of two ccaat boxes. Biochem. J. 384, 317–326.

Krstic, A., Mojsin, M., Stevanovic, M., 2007. Regulation of sox3 gene expression is driven by multiple nf-y binding elements. Arch. Biochem. Biophys. 467, 163–173.

Li, Q., Herrler, M., Landsberger, N., Kaludov, N., Ogryzko, V., Nakatania, Y., Wolffe, A.P., 1998. Xenopus nf-y pre-sets chromatin to potentiate p300 and acetylation-responsive transcription from the xenopus hsp70 promoter in vivo. EMBO J. 17, 6300–6315.

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

Mantovani, R., 1998. A survey of 178 nf-y binding ccaat boxes. Nucleic Acids Res. 26, 1135–1143.

Muller, G.A., Heissig, F., Engeland, K., 2007. Chimpanzee, orangutan, mouse, and human cell cycle promoters exempt ccaat boxes and chr elements from interspecies differences. Mol. Biol. Evol. 24, 814–826.

Osada, S., Yamamoto, H., Nishihara, T., Imagawa, M., 1996. Dna binding specificity of the ccaat/enhancer-binding protein transcription factor family. J. Biol. Chem. 271, 3891–3896.

Pabst, T., Mueller, B., Zhang, P., Radomska, H., Narravula, S., Schnittger, S., Behre, G., Hiddemann, W., G., D., Tenen, 2001. Dominant-negative mutations of cebpa, encoding ccaat/enhancer binding protein-alpha (c/ebpalpha), in acutemyeloid leukemia. Nat. Genet. 27, 263–270.

Roder, K., Wolf, S., Sickinger, S., Schweizer, M., 1999. Fire3 in the promoter of the rat fatty acid synthase (fas) gene binds the ubiquitous transcription factors cbf and usf but does not mediate an insulin response in a rat hepatoma cell line. Eur. J. Biochem. 260, 743–751.

Roulet, E., Bucher, P., Schneider, R., Wingender, E., Dusserre, Y., Werner, T., Mermod, N., 2000. Experimental analysis and computer prediction of ctf/nfi transcription factor dna binding sites. J. Mol. Biol. 297, 833–848.

Salsi, V., Caretti, G., Wasner, M., Reinhard, W., Haugwitz, U., Engeland, K., Mantovani, R., 2003. Interactions between p300 and multiple nf-y trimers govern cyclin b2 promoter function. J. Biol. Chem. 278, 6642–6650.

Smith, T.F., Waterman, M.S., 1981. Identification of common molecular subsequences. J. Mol. Biol. 147, 195–197.

Staden, R., 1984. Computer methods to locate signals in nucleic acid sequences. Nucleic Acids Res. 12, 505–519.

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

Zanotto, E., Häkkinen, A., Teku, G., Shen, B., Ribeiro, A.S., Jacobs, H.T., 2009. Nf-y influences directionality of transcription from the bidirectional mrps12/sarsm promoter in both mouse and human cells. BBA Gene Regul. Mech. 1789, 432–442.

Zanotto, E., Lehtonen, V., Jacobs, H.T., 2008. Modulation of mrps12/sarsm promoter activity in response to mitochondrial stress. Biochim. Biophys. Acta 1783, 2352–2362.

Zanotto, E., Shah, Z.H., 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.

Highlights for "Genome wide study of NF-Y type CCAAT boxes in unidirectional and bidirectional promoters in human and mouse"

. Bidirectional and unidirectional promoters differ in regulatory mechanisms at the sequence level

. NF-Y type CCAAT boxes play key but distinct roles in Bidirectional and unidirectional promoters

. In bidirectional promoters, NF-Y type CCAAT boxes are often contiguous

. Several NF-Y type CCAAT four-box configurations are conserved between human and mouse