In Silico Biology 4, 0011 (2003); ©2003, Bioinformation Systems e.V.

Content analysis of the core promoter region of human genes

Vladimir B. Bajic*, Vidhu Choudhary and Chuan Koh Hock

Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore

* corresponding author
Email: bajicv@i2r.a-star.edu.sg

Edited by E. Wingender; received September 25, 2003; revised and accepted November 28, 2003; published January 17, 2004

Abstract

We analyzed an extended core promoter regions covering [-70,+60] segment relative to the transcription start site of human promoters contained in the Eukaryotic Promoter Database. The analysis was made by using the Match program ver. 1.9 with an optimized setting and the TRANSFAC Professional database ver. 7.2. This analysis revealed that the most common transcription factor binding site in the examined collection of core promoters appears to be initiator (characterized by GEN_INI), which is expected. The other less obvious sites found were Spz1, E2F-1, ZF5, and C/EBP. The 'cap' site was also in this most common group. Over-representation of these sites relative to the non-promoter background data ranged from 0.3167 to 32.1645. These sites were characterized by being present in more than 60% of promoter sequences. Interestingly, the TATA-box has been found in only 11.63% of all examined promoters. The study is complemented by separate analyses of promoter groups having different GC content. These additional analyses revealed that the most common promoter elements found also include AP-2, CdxA, Pax-2, SRY, STAT1 and STAT5A. It was also observed that a number of promoter elements show strong preference either for the GC-rich or the GC-poor core promoters.

Key words: core promoters, promoter elements, transcription factor binding sites, promoter models

Introduction

Analysis of promoter region is important for:

a. elucidation of the mechanisms for transcriptional activation of genes,
b. annotation of transcriptional regulatory elements, and
c. development of efficient promoter prediction programs.

While detection of promoter region is itself a difficult problem, accurate determination of the transcription start site (TSS) location is far more difficult. The region around human TSS has distinct nucleotide compositional properties as found by Majewski and Ott, 2002, most distinct being those just in the close neighborhood of TSS. We do not know if similar properties could be found in other species, but suspect that at least for mammals this could be the case. The narrow region around the TSS has been considered important in understanding promoter functionality [Butler and Kadonaga, 2002] and this region is termed 'core promoter'. Core promoters are also considered to control the main mechanisms of basal transcription [Smale and Kadonaga, 2003]. The core promoter region [Werner, 1999; Butler and Kadonaga, 2002; Chen and Hampsey, 2002; Kadonaga, 2002; Smale and Kadonaga, 2003] can be defined as the minimal continuous segment of DNA sufficient for accurate initiation (and direction) of transcription. However, the definition is ambiguous regarding the length of the region covered by the core promoter. In most cases promoter elements (PEs) considered important for the functionality of the core promoter region can spread over the segment [-60,+50] relative to TSS, and thus it makes sense to consider this segment as representing the core promoter. In this study, for the reasons to be explained later, we used a bit extended region of [-70,+60] relative to TSS. Different PEs are considered characteristic of the core promoter region including Inr, TATA, BRE and DPE elements [Werner, 1999; Butler and Kadonaga, 2002; Chen and Hampsey, 2002; Kadonaga, 2002; Smale and Kadonaga, 2003], although some other elements, such as DCE or MED-1 [Smale and Kadonaga, 2003] are also found functional.

The focus of this study to the core promoter region stems from our intention to find regularities of this region which later on may help in development of computer programs that attempt to pinpoint TSS. In this light and due to the biological significance of this promoter sub-region, our interest is to examine what computational techniques for in silico detection of PEs can detect in this region, hoping that results can help us elucidate some aspects of the statistical composition of the core promoters. Additionally, we expect that it may be possible to computationally detect some PE patterns that are at least statistically over-represented in the core promoters, thus potentially having an increased likelihood of being biologically active. Such PEs could be good candidates for additional wet lab experiments.

Our expectations of what we should find and in what proportion in the core promoters are modulated by the current knowledge and understanding of the core promoter functionality, as well as the experimental and computational methods available. As an example, several independent studies performed on the larger sets of promoters of different organisms [Ohler *et al.*, 2002; Suzuki *et al.*, 2001], suggested that some of the typical core PEs could be detected by computational tools but not at the levels previously expected. For example, Ohler *et al.*, 2002 found that the TATA-box pattern can be detected in 33% of 1941 *Drosophila* promoters, while Suzuki *et al.*, 2001, found that it is present in only 32% of 1031 human promoters; Inr element has been found in 69% of 1941 *Drosophila* promoters by Ohler *et al.*, 2002. Of course, different threshold setting of the computational tools used in these studies could provide a different picture.

This leads to a natural question of which PEs can be found by computational means in the core promoter region as being over-represented in this region. Such an approach partly reduces the influence of thresholds in the computational tools used as it emphasizes differences of the content of core promoters relative to the background non-promoter sequence set. Another question of interest is that of the most common PEs present in the core promoters and it directly depends on the threshold settings used in the analysis. There are several issues related to these problems. One is the accuracy of the TSS locations in the examined promoter sequences since the core promoter region is very narrow. The other is the accuracy of the current PE-detecting computational tools. These tools could miss many real sites for any reasonable level of sensitivity. The third issue is the diversity of the promoters used in the analysis. If the promoter set is small, generally it will be very biased, and the results will not be representative. Anyway, until the proper experimental evaluation is done we will not

be able to know the true situation.

In order to get a more clear picture of what can be found in the core promoter regions we used the Eukaryotic Promoter Database (EPD) [Praz *et al.*, 2002] to reduce the potential problem of the accuracy of TSS locations. To reduce further a potential bias which small promoter datasets can introduce, we have restricted our analysis to human promoters since these are from the most represented species in EPD (more than half of all entries in the current version of EPD). By utilizing currently the most sophisticated tools for finding PEs, we performed a detailed statistical analysis of the core promoter content for 1771 human promoters. To conduct this study we used the Match program [Kel *et al.*, 2003] ver. 1.9, and the collection of vertebrate matrix profiles contained in the TRANSFAC Professional database [Matys *et al.*, 2003] ver.7.2. This analysis has provided us with a very lucid picture of the computationally detected PEs in the extended core promoter region, their distributional properties and their overrepresentation w.r.t. non-promoter sequences. This analysis is complemented by the analyses of four different core promoter groups, where groups were made based on the GC content of the examined sequences. As such, the results could provide clues for computational annotation of human core promoters and potentially may help in designing more accurate TSS-finding programs.

Materials and methods

As the core promoter data we used the region of [-70,+60] relative to the annotated TSS location as given in the EPD. Since the accuracy of the TSS location in the EPD is within ±10 nt, we decided to extend the expected core promoter region of [60,+50] to [-70,+60] to accommodate for the possible imprecision of the annotated TSS. We also selected all human promoters which belong to the non-redundant group as defined in EPD.

As the non-promoter sequences we extracted by the FIE program [Chong *et al.*, 2003; Chong *et al.*, 2002] ver. 2, segments downstream of the most 5' estimated TSS location for 10568 genes, covering the region [+5001,+6000] relative to TSS.

We used the concept of over-representation of a particular pattern in one group of sequences w.r.t. another group of sequences. There exists no consensus on how the over-representation should be calculated. In our case we express this over-representation by a measure we call over-representation index (ORI) and we define it as

$$ORI(PE_i) = \frac{Density_{promoter}(PE_i)}{Density_{non-promoter}(PE_i)} \times \frac{Proportion_{promoter}}{Proportion_{non-promoter}},$$

$$ORI(PE_i) = \frac{\frac{Patt_p}{TotalLength_{promoter}} \times \frac{N_p}{N_{promoter}}}{\frac{Patt_{np}}{TotalLength_{non-promoter}} \times \frac{N_{np}}{N_{non-promoter}}},$$

where PE_i is the i-th PE pattern, $Patt_p$ is the number of patterns PE_i found in core promoter sequences, while $Patt_{np}$ is that one found in non-promoter sequences; $Density_{promoter}(PE_i)$ is the density at which this pattern is found in promoter sequences, $Density_{non\text{-}promoter}(PE_i)$ is the density at which this pattern is found in non-promoter sequences; N_p is the number of promoter sequences where

the pattern is found, N_{np} is the number of non-promoter sequences where the pattern is found, $TotalLength_{promoter}$ is the total length of promoter sequences and $TotalLength_{non\text{-}promoter}$ is the total length of non-promoter sequences. Also, $N_{promoter}$ and $N_{non\text{-}promoter}$ are the total number of promoter and non-promoter sequences, respectively. In our case we made a random sampling of background sequences selecting the sequences of the same length (130 bp) as is the length of core promoter sequences. From each background sequence of length 1000 bp we in this way selected 100 samples and included them in the background sequence set.

Statistical significance can be estimated from the frequencies of predictions. In the core promoter set the frequency of prediction equals $Density_{promoter}(PE_i)$. Similarly, for the background set the frequency of predictions equals $Density_{non\text{-}promoter}(PE_i)$. Ratio of these shows how more probable is to find a pattern PE_i in the core promoter set than in the background set. We, however, will use ORI since it takes into account not only the number of patterns found, but also the proportion of sequences in which the pattern is found.

The GC content of a sequence was determined as the ratio of the sum of G and C nucleotides over the total number of nucleotides in the sequence. We analyzed the GC ranges of [0, 0.4), [0.4, 0.5), [0.5, 0.6) and [0.6, 1] as in Kel-Margoulis *et al.*, 2003.

Other aspects of the tools and resources we used were presented earlier.

Analysis of the extended core promoter region as a single group

We analyzed the core promoter region of human promoters contained in the EPD with the aim at finding a) the most common putative PEs present in these regions as revealed by the application of currently the most sophisticated computational tools for PE-finding, and b) the most over-represented patterns in core promoters. There have been 1771 promoter sequences which belonged to the non-redundant group of human promoters, where non-redundancy was as defined in EPD. We also used ~10.6 Mb of non-promoter sequences as the background against which the promoter content has been evaluated. In the analysis, all vertebrate matrix profiles contained in the TRANSFAC database ver.7.2 were used in order not to miss any of the potential PEs which have matrix model in TRANSFAC. The setting for the Match program was selected to be the minimum of the sum of false positive and false negative predictions as described in Kel *et al.*, (2003). The found PEs in the core promoters have been compared to those found in the background sequence data and the number of promoters and non-promoters in which they were detected has been determined, as well as their over-representation in the promoter sequence group. See details of this comparison in Materials and Methods section. The results are given in Table 1a (sorted according to frequency in core promoters) and Table 1b (sorted according to over-representation index (ORI)) for all PEs which are contained in more than 10% of the core promoters. Positional distributions of all PEs that have been found in at least 60% of the core promoters have been also analyzed and the results are depicted in Figures 1-4. These figures are obtained by calculating all PE sites (which are found present in more than 60% of the promoters is [-70,+60] domain) contained in windows of length 25 bp, and shifted along DNA in steps of 5 bp.

Table 1a: PEs found in more than 60% of the core promoters. The on-line version of this table contains all PEs which are found in at least 10% of the core promoters.

Strand	Name of TFBS	Total # of promoters in which TFBS is	% of promoters in which TFBS is	Total # of TFBSs	Over-representation

		found	found	found	index (ORI)
-1	GEN_INI	1397	78.88	7191	0.5247
+1	Spz1	1353	76.40	4279	6.1791
+1	GEN_INI	1351	76.28	6412	0.4898
+1	E2F-1	1315	74.25	3370	20.2050
+1	ZF5	1253	70.75	3864	32.1645
-1	E2F-1	1237	69.85	2698	17.4793
-1	ZF5	1193	67.36	3533	27.4143
+1	C/EBP	1138	64.26	2344	0.3167
-1	Spz1	1098	62.00	2966	3.6005
-1	Cap	1098	62.00	1846	0.5212
+1	Cap	1097	61.94	1759	0.4962

One may notice that the over-representation index (ORI) (for the definition of ORI see Materials and Methods section) for the PE patterns presented in Table 1a ranges from 0.3167 to 32.1645 for the PEs present in at least 60% of the core promoters. We observe that some of the most common PEs, such as GEN_INI, C/EBP and the 'cap' site, are in fact under-represented in core-promoters as opposed to the background non-promoter sequences, while the other group containing Spz1, E2F-1 and ZF5 patterns are over-represented.

Table 1b: Top 10 PEs based on the over-representation index (ORI) value. The on-line version of this table contains all PEs which are found in at least 10% of the core promoters.

Strand	Name of TFBS	Total # of promoters in which TFBS is found	% of promoters in which TFBS is found	Total # of TFBSs found	Over-representation index (ORI)
+1	E2F"	1019	57.54	2184	84.3371
-1	E2F"	927	52.34	1837	66.4509
-1	CREB"	251	14.17	405	46.2131
+1	CREB"	234	13.21	392	43.6498
+1	ETF"	671	37.89	1250	42.2698
+1	Elk-1"	376	21.23	553	40.1891
-1	Elk-1"	393	22.19	579	34.0581
+1	ZF5"	1253	70.75	3864	32.1645
-1	ZF5"	1193	67.36	3533	27.4143
-1	LBP-1"	299	16.88	335	20.3961

We observe from Table 1b that patterns which are most over-represented in the core promoter group are usually not the most common. In the top 10 PEs with the highest ORI, only E2F and ZF5 appear in more than 50% of core promoters, while each of the other patterns such as CREB, ETF, Elk-1 and LBP-1 are present in less than 40% of core promoters.

The search for PEs was two-stranded. We have opted to use a two-stranded search for PEs since there is no reason to assume absolute symmetry in the distribution of PEs on the two DNA strands, although many of the PEs are not directional. For example, TATA-box is uni-directional. This approach should enable us to observe potential differences in strand-specific distributions of some PEs, which can further contribute to our goals. This is why we can find that some sites have similar distribution on the two DNA strands, while some others do not.

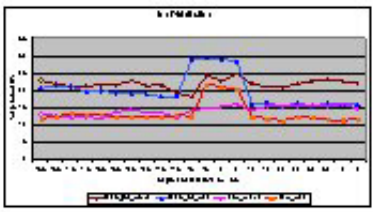

Figure 1: There is a distinct preferred region of Inr on the plus strand covering the segment [-8,+7]. Distributions of GEN_INI and 'cap' sites are highly correlated on the plus strand, as well as on the minus strand.

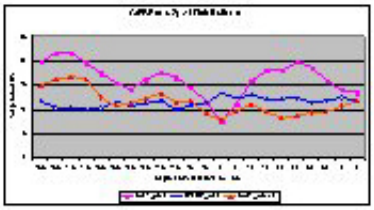

Figure 2: Distributions of C/EBP and Spz1 elements. Spz1 on the plus strand has a preferred region away from TSS.

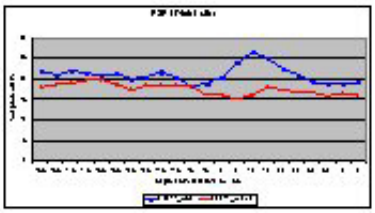

Figure 3: Distribution of E2F-1 binding site on the plus strand shows preference in the downstream region around [+7,+22].

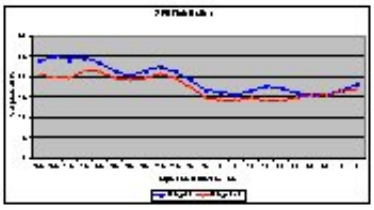

Figure 4: ZF5 binding sites show strong correlation in distributions over plus and minus strands.

It is interesting to observe that distributions of the found initiator sequences, characterized by patterns of GEN_INI and 'cap' site, are highly correlated on the plus strand, as well as on the minus strand. Distribution of the ZF5 pattern on the plus and minus strands is also highly correlated, implying that this pattern is not very directional. Local over-representation is prominent only for GEN_INI and 'cap' sites. For GEN_INI the preferred region is found on the plus strand around [-8,+7] regions about where we expect these sites. The region of over-representation for the 'cap' site is a bit shorter covering [-3,+7] segment.

The E2F-1 pattern on the plus strand has preferred region around [+7,+22]. ZF5 has as the preferred

region on both strands the [-70,-13] segment, while on the plus strand Szp1 is under-represented around position 5, and has preferred regions around [-70,-43] and position 29. Similarly, Spz1 on the minus strand has preferred region around [-70,-43] and it is highly correlated in distribution with Spz1 on the plus strand over the region [-70,12].

Analysis of core promoters as a function of the GC content

In order to analyze the effect of the GC content of the core promoter sequences on the found PE patterns, we have formed several groups of sequences on this basis as summarized in Table 2. Then for each pair of groups corresponding to the different GC content we have conducted the similar analysis as presented in the previous section.

Table 2: Different sequence groups based on the GC content.

GC content as a proportion of nucleotide content	Number of core promoters	Number of background sequences
[0, 0.4)	26	3236
[0.4, 0.5)	118	4565
[0.5, 0.6)	334	2195
[0.6, 1.0]	1293	572
Total number of sequences	1771	10568

Table 2 shows that the most present core promoters in our promoter dataset are those with the high GC content. This is in line with Majewski and Ott, 2002, and also supports Kel-Margoulis *et al.*, 2003, regarding the elevated GC content of promoter segments [-300,+50]. Since the most GC-rich group contains ~73% (1293 out of 1771) of all core promoters, it is obvious that the composition of promoters in terms of the found PE patterns, as given in the previous section, is biased towards the patterns from the promoters with the high GC content. Dividing data according to the GC content will help us to see the effects of the GC composition in the analyzed sequences. The most common PE patterns in different promoter groups are given in Tables 3-6. In these tables we show the top 10 PE patterns found in each GC group of sequences sorted according to frequency (Tables a) and according to the value of ORI (Tables b). The on-linev versions of these tables contain all PE patterns found in at least 10% of core promoters.

Tables 3-6:

Table 3a: Ten most common PEs found in the co range [0, 0.4). The extended version of found in at least 10% of this group of c core promoters in this group is 26.

Strand	Name of TFBS	Total # of promoters in which TFBS is found	% of in wh
-1	CdxA"	26	
+1	CdxA"	26	
+1	GEN_INI"	25	
+1	C/EBP"	25	
-1	GEN_INI"	24	
+1	cap"	23	
-1	STAT1"	22	
-1	C/EBP"	22	
-1	SRY"	22	
+1	C/EBPalpha"	21	

As can be seen, in the most GC-poor group (GC range [0, 0.4)), 10 most common PEs are CdxA, GEN_INI, C/EBP, 'cap' site, STAT1, SRY and C/EBPalpha. Their ORI ranges from 0.8052 to 1.6675. For the GC range of [0.4, 0.5), the 10 most common patterns are C/EBP, CdxA, GEN_INI, 'cap' site, STAT5A and Pax-2 with ORI in the range from 0.7611 to 1.5327. In the GC range of [0.5, 0.6) the 10 most common PE patterns found are GEN_INI, C/EBP, 'cap' site, STAT1, Spz1, STAT5A and Pax-2 having ORI in the range from 0.6897 to 1.9354. Finally, in the most GC-rich group with the GC content in the range [0.6, 1.0], we found that the 10 most common PE patterns are E2F-1, Spz1, ZF5, GEN_INI, AP-2 and E2F. Their ORI ranges from 0.6708 to 8.6349.

Regarding the PE patterns which are more over-represented in different core promoter groups, we find that in the most GC-poor group ORI for top 10 PEs ranges from 10.1173 to 248.826, but the most common of these patterns (HSF2 on plus strand) appears to be common for only 26.92 core promoters. In the core promoters with the GC content in the range of [0.4, 0.5), top 10 ORI range from 3.8785 to 15.6735. It is interesting to note that the most common of these PEs is the TATA-box element on plus strand and it appears in 38.14% of these core promoters. For the next core promoter group with the GC content of [0.5, 0.6), top 10 ORI range from 8.7451 to 30.1101. The most common of these 10 patterns is c-Ets-1(p54) on plus strand which appears in 26.05% of core promoters of this group. Finally, in the most GC-rich core promoter group, top 10 ORI range from 4.2953 to 9.8841. Here, the most common of these patterns is E2F on plus strand which appears common 69.91% of core promoters of this group.

In order to see how is the detection of some of the PEs influenced by the the GC content of core

promoters, we analyzed a collection of PEs made up of those which appear in at least 60% of all core promoters and extend it with the E2F patterns. The results are presented in Table 7 and graphically depicted in Figure 5. We can notice based on Figure 5 very clear groupings strongly dependent on the GC content of the core promoters. Obviously, some patterns are more likely to be found in the GC-poor group and vice versa.

Table 7: Distribution of selected PEs as a function of the GC content of core promoters.

PE pattern	G+C≤0.4% of core promoters where pattern is found	0.4≤G+C<0.5% of core promoters where pattern is found	0.5≤G+C<0.6% of core promoters where pattern is found	0.6≤G+C% of core promoters where pattern is found
cap_minus	76.92	73.73	70.36	58.47
E2F_plus	0	7.630	31.74	69.91
GEN_INI_minus	92.31	90.68	85.33	75.87
ZF5_plus	0	20.34	47.01	82.91
GEN_INI_plus	96.15	92.37	85.93	71.93
E2F-1_plus	15.38	24.58	58.08	84.15
cap_plus	88.46	84.75	73.95	56.23
C/EBP_plus	96.15	87.29	81.74	57.00
Spz1_plus	11.54	34.75	68.56	83.53
Spz1_minus	38.46	33.90	50.60	67.98
C/EBP_minus	84.62	94.92	71.86	46.09
E2F_minus	0	6.780	25.75	64.42
ZF5_minus	0	21.19	41.32	79.66
E2F-1_minus	11.54	21.19	55.69	79.12

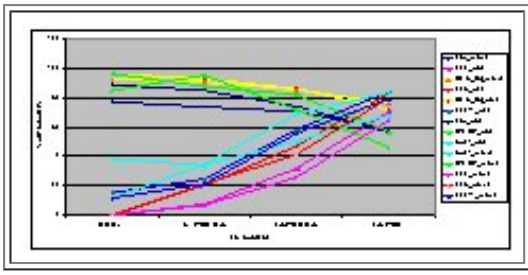

Figure 5: Proportion of the core promoters which contain specific PE depicted as a function of the core promoter's GC content.

We were also interested to see how the detection of the TATA-box is influenced by the GC content of the core promoter sequences. Table 8 summarizes the results. We observe that it is much more likely to detect TATA-box in the GC-poorer groups.

Table 8: TATA-box detection (on plus strand) in the core promoters as a function of the GC content.

GC content of core promoters	ORI	% of the core promoters with the TATA-box
0.0≤GC<0.4	4.2235	42.31
0.4≤GC<0.5	15.6735	38.14
0.5≤GC<0.6	30.1101	17.96
0.6≤GC≤1	37.0466	6.96

Discussion

With the exception of the study of Ohler *et al.*, 2002, for *Drosophila*, there has been no systematic study of the content of the core promoter region to reveal what the current computational tools single out as the relevant sequence motifs. Contrary to this study, the study of Ohler *et al.*, 2002 was not based on the predefined motif models. Our study is the first one for the human core promoters which employs all patterns of the known binding sites of vertebrates described by the matrix profile models contained in the latest release of the TRANSFAC Professional database, ver. 7.2. The results rely on the ability of the TRANSFAC database associated Match program to detect PE sites at one of its optimized setting.

In comparison with the previous studies, we have to point out that our results are dependent on the tools we have used for the analysis, the thresholds used for detection of the presence of PEs, as well as the promoter sequences and the background non-promoter sequences. Since other related studies [Ohler *et al.*, 2002; Suzuki *et al.*, 2001; Kel-Margoulis *et al.*, 2003] have been conducted under different conditions, no direct comparison with them is possible and thus, no conflict exists with regards to these studies.

We have adopted strand-specific search for PEs for the reasons we mentioned earlier. The difference that we have observed in the strand-specific distributions of some binding sites, such as E2F-1, could be result of many causes, for example the quality of TFBS models and thresholds used. They can also be statistical artifacts caused by the combination of sequence properties and the above mentioned potential causes. We cannot comment more except to notice that such computationally found differences suggest a necessity for an additional experimental study to elucidate the actual nature of these sites which show very different strand distributions. For example, for several of PEs we will show in Figs. 6-11 the change of the proportion of core promoters having a particular PE as a function of GC content.

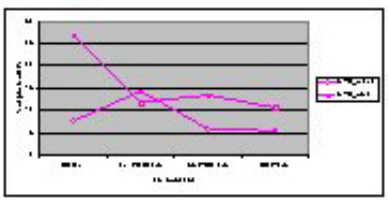

Figure 6: Change of the proportion of core promoters which contain MYB pattern in different GC content groups.

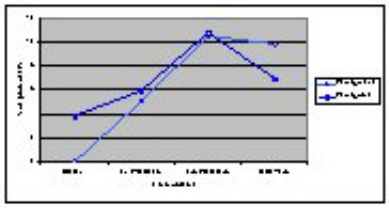

Figure 7: Change of the proportion of core promoters which contain PEA3 patterns in different GC content groups.

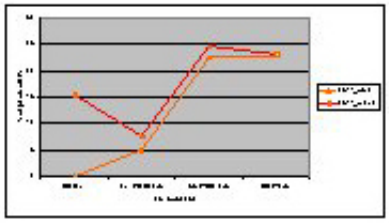

Figure 8: Change of the proportion of core promoters which contain Elk-1 patterns in different GC content groups.

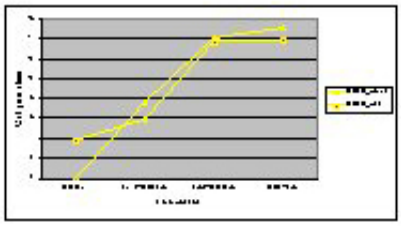

Figure 9: Change of the proportion of core promoters which contain CREB patterns in different GC content groups.

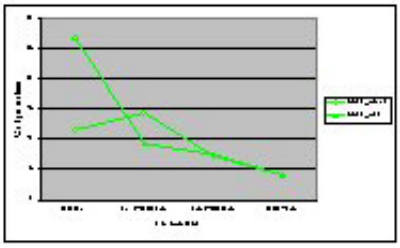

Figure 10: Change of the proportion of core promoters which contain HSF2 patterns in different GC content groups.

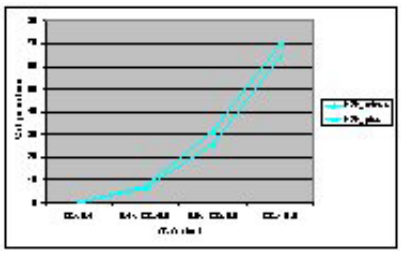

Figure 11: Change of the proportion of core promoters which contain E2F patterns in different GC content groups.

As can be noticed, for PEs shown in Figs. 6-10 the change in proportion of core promoters containing such elements on plus strand and minus strand are not quite similar, while for the PE pattern in Fig. 11, the behavior on the plus and minus strand is very similar. This support our approach to look for PE patterns as stand-dependent.

Our study reveals some interesting findings. To our surprise, the only core promoter element which has been detected in a significant proportion of the core promoters is Inr with its variants covered by the GEN_INI elements and the Bucher's [Bucher, 1990] 'cap' site. Some of the most expected patterns for core promoters have not been detected in a significant proportion of promoters with the tools and threshold settings used. For example, the TATA-box element was found in only 11.63% of all core promoter sequences on the plus strand. Three possible reasons could be given for this result:

1. the previous estimates of the presence of TATA-box elements are not good;
2. models of TATA-box elements used are not good;
3. optimized thresholds for detection of TATA-box elements used are not good.

We are not in position to determine the actual causes of this unexpected result for TATA-box patterns. Although the conditions for detection of PEs in this study were different from previous studies and thus the direct comparisons are not possible, we notice on a qualitative level that our result suggests a significant decline from the previous computational estimates of ~32% [Suzuki *et al.*, 2001] for the TATA-box. This is especially important as we counted here the whole [-70,+60] region and did not restrict consideration to the preferred region for the TATA-box which is centered around -30.

On the other hand, BRE and DPE elements do not have matrix models in the TRANSFAC database release that we used, and hence it is not possible to comment on their presence or absence.

Thus, are the results that we obtained very unusual? In fact, yes. While it is expected that a general initiation sequence (GEN_INI), and thus the 'cap' site [Bucher, 1990], are to be present around the TSS location, the suggested significant presence of the other found PEs are somehow strange.

E2F-1, the main member of the E2F family of transcription factors, is involved in numerous processes in control of cell-cycle and apoptosis, but has also other suggested roles such as in tumorigenesis [La

Thangue, 2003]. We found that in the core promoters the E2F-1 patterns are with the over-representation index of 20.205 (plus strand) and 17.4793 (minus strand) in [-70,+60]). C/EBP is the CCAAT/enhancer-binding protein which controls cell cycle progression [Cho and Kim, 2003]. Its over-representation index in core promoters for patterns on the plus strand is 0.3167 implying that this pattern is in fact under-represented in core promoters. The binding site for Spz1 gene (bHLH-Zip gene) is found in cDNA of mouse testis and epididymis [Hsu *et al.*, 2001]. Its over-representation index for patterns on the plus strand is 6.1791. ZF5 is a transcription factor with five zinc finger motifs which expresses in various tissues [Obata *et al.*, 1999]. We found it on the plus strand in 70.75% of core promoters with the over-representation index of 32.1645.

These finding suggest that it would make sense to experimentally examine the roles of these 'unexpected' PEs in the human core promoters. It may also be true that the 'conventional' perception of the important PE sites in the core promoter region should be enriched and expanded more than previously thought, or it may be that the human promoters in the dataset we used in this analysis is far too biased. There is also a possibility that the currently used computational tools are not sufficiently sophisticated. Different settings for the Match program, as well as the quality of profile matrices in TRANSFAC, have influence on the overall outcome of the computational analysis. We, however, have opted for the use of 'minimum sum of false positive and false negative predictions' since these provide balanced predictions which do not favor either over-optimistic predictions where the most of the predicted sites are false positives, or over-pessimistic predictions where many real sites will be missed which also cause conceptual problems in generalizing results of the analysis. Having in mind that the tools for finding PEs in DNA are gradually improving, we have full confidence that analysis of this type will contribute to our understanding of the role of different PEs in the transcription initiation process.

Separation of promoters and background sequences to distinct GC content groups, has reveals very clear groupings of putative PEs in the core promoters. Based on Table 7 and Figure 5 we see that Spz1, ZF5, E2F and E2F-1 patterns are detected in considerably smaller proportion in the GC-poor core promoters than in the GC-rich ones. On the other hand, we notice that GEN_INI, 'cap' site and C/EBP patterns are not much influenced much by the GC content, although, generally, the proportion of the core promoters in which these patterns are found slightly decreases as the GC content of the core promoters increases. The least influences seems to be GEN_INI, while the most influenced of these three patterns appears to be C/EBP. In the most GC-rich group of core promoters, Spz1, ZF5, E2F and E2F-1 patterns appear to be detected in a higher proportion of promoters than the 'cap' site and C/EBP patterns.

Compared to the analysis when we did not consider core promoters separated into different groups based on the GC content, we observe that the PEs obtained by selecting the top 10 most common patterns in each of these promoter groups now contain CdxA, SRY, STAT1, STAT5A, Pax-2 and AP-2 patterns. For example, STAT1 and STAT5A are known to act as signal transducers and activators of transcription [Akira, 1999; Takeda and Akira, 2000]. But the other four, AP-2, CdxA, Pax-2 and SRY, are difficult to relate to the initiation of transcription process and thus their role has to be analyzed with caution. Namely, AP-2 transcription factor is implied in progression of malignant melanoma [Bar-Eli, 1999], while AP-2 adaptor complex is related to clathrin-coated vesicles budding from the plasma membrane [Hirst and Robinson, 1998]; CdxA transcription factor is a homeodomain protein found in chicken [Margalit *et al.*, 1993]; Pax-2 transcription factor is a member of the Pax family known to be involved in embryogenesis [Underhill, 2000]; SRY is the testis-determining factor, but its function is poorly understood [Harley *et al.*, 2003].

We observe (Table 8) that the core promoters with detected TATA-box patterns on the plus strand decrease with the GC content of core promoters from 42.31% to 6.96%, while at the same time ORI increases from 4.2235 to 37.0466. This suggests that the TATA-box patterns found in more GC-rich groups are statistically more significant.

Recently, Kel-Margoulis *et al.*, 2003, analyzed the region [-300, +50] relative to the TSS for a large group of human genes. Although we cannot make direct comparison of their and our results due to different conditions under which the analyses have been conducted (they used a different criterion for the Match program setting which allows only for 50% of the known sites to be detected and a different way to determine over-representation), we can still make a comparison at a qualitative level. At first glimpse, our results seem to contradict the findings of Kel-Margoulis *et al.*, 2003. They found that E2F patterns were under-represented in the [-300, +50] region. They also found that the C/EBP and TATA-box patterns were under-represented in the group of promoters with the GC content < 0.4, while they were over-represented in the other promoter groups. Our findings imply that E2F (and E2F-1) patterns appear in on plus strand in 69.91% (84.15%) of the core promoters in the most GC-rich group with ORI of 8.6349 (3.8112). We also found that C/EBP patterns are among the most common core promoter groups with the GC content < 0.6. They, however, are under-represented based on ORI values in groups with GC content $\geq$ 0.5. Also, in the most GC-poor group of core promoters C/EBP on minus strand is under-represented, while that on the plus strand is with ORI of 1.0048. In the group with GC content of [0.4, 0.5), C/EBP on the plus strand is under-represented based on ORI, while on the C/EBP on the minus strand is with ORI of 1.0725. In the group of core promoters with the GC content $\geq$ 0.6 the proportion of the promoters containing this pattern on the plus strand is 57%, while those having this pattern on the minus strand is 46.09%. For the TATA-box patterns we found that it is under-represented (in terms of proportion of core promoter sequences which contain them) in all four groups. However, these patterns were over-represented in terms of ORI value. Explanation for these findings is that essentially we have examined a region [-70,+60] different from the one studied in Kel-Margoulis *et al.*, 2003, and thus the properties we found are specific to this region only. In a way, properties of the core promoters are 'diluted' when considered over the longer region such as in [-300, +50]. We have used a specific background sequence set, different from those utilized by Kel-Margoulis *et al.*, 2003. Furthermore, we opted to use exclusively the EPD promoter data due to their reliability w.r.t. the TSS location, while in Kel-Margoulis *et al.*, 2003, the reference TSS locations have been based on the RefSeq data which result in less reliable TSS location estimate as discussed in Chong *et al.*, 2002; 2003. All these clarify why there are different qualitative conclusions about the content of the core promoter region as opposed to the [-300, +50] promoter region from Kel-Margoulis *et al.*, 2003. We can conclude that our analysis complements the one of Kel-Margoulis *et al.*, 2003.

Conclusions

We performed a detailed analysis of the computationally discovered PE patterns in the core promoters of 1771 human genes. In addition to the expected initiator sequence, several other PE patterns have been found which are not expected to be present in a high proportion of the core promoters. Further experimental analysis could elucidate the role of these PE patterns in the core promoter region.

References

- Akira, S. (1999). Functional roles of STAT family proteins: lessons from knockout mice. Stem Cells **17**, 138-146.
- Bar-Eli, M. (1999). Role of AP-2 in tumor growth and metastasis of human melanoma. Cancer Metastasis Rev. **18**, 377-385.

- 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.
- Butler, J. E. and Kadonaga, J. T. (2002). The RNA polymerase II core promoter: a key component in the regulation of gene expression. Genes Dev. **16**, 2583-2592.
- Chen, B. S. and Hampsey M. (2002). Transcription activation: unveiling the essential nature of TFIID. Curr. Biol. **12**, R620-R622.
- Chong, A., Zhang, G. and Bajic, V. B. (2002). Information and sequence extraction around the 5'-end and translation initiation site of human genes. In Silico Biol. **2**, 0041.
- Chong, A., Zhang, G. and Bajic, V. B. (2003). FIE2: A program for the extraction of genomic DNA sequences around the start and translation initiation site of human genes. Nucleic Acids Res. **31**, 3546-3553.
- Cho, M. K. and Kim, S. G. (2003). Hepatocyte growth factor activates CCAAT enhancer binding protein and cell replication via PI3-kinase pathway. Hepatology **37**, 686-695.
- Harley, V. R., Clarkson, M. J. and Argentaro, A. (2003). The molecular action and regulation of the testis-determining factors, SRY (sex-determining region on the Y chromosome) and SOX9 [SRY-related high-mobility group (HMG) box 9]. Endocr. Rev. **24**, 466-487.
- Hirst, J. and Robinson, M. S. (1998). Clathrin and adaptors. Biochim. Biophys. Acta **1404**, 173-193.
- Hsu, S. H., Shyu, H. W., Hsieh-Li, H. M. and Li, H. (2001). Spz1, a novel bHLH-Zip protein, is specifically expressed in testis. Mech. Dev. **100**, 177-187.
- Kadonaga, J. T. (2002). The DPE, a core promoter element for transcription by RNA polymerase II. Exp. Mol. Med. **34**, 259-264.
- Kel, A. E., Gössling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O. V. and Wingender, E. (2003). MATCH: A tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Res. **31**, 3576-3579.
- Kel-Margoulis, O. V., Tchekmenev, D., Kel, A. E., Goessling, E., Hornischer, K., Lewicki-Potapov, B. and Wingender, E. (2003). Composition-sensitive analysis of the human genome for regulatory signals. In Silico Biol. **3**, 0013.
- La Thangue, N. B. (2003). The yin and yang of E2F-1: balancing life and death. Nat Cell Biol. **5**, 587-589.
- Margalit, Y., Yarus, S., Shapira, E., Gruenbaum, Y. and Fainsod, A. (1993). Isolation and characterization of target sequences of the chicken CdxA homeobox gene. Nucleic Acids Res. **21**, 4915-4922.
- Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A. E., Kel-Margoulis, O. V., Kloos, D. U., Land, S., Lewicki-Potapov, B., Michael, H., Munch, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele, S. and Wingender, E. (2003). TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res. **31**, 374-378.
- Majewski, J. and Ott, J. (2002). Distribution and characterization of regulatory elements in the human genome. Genome Res. **12**, 1827-1836.
- Obata, T., Yanagidani, A., Yokoro, K., Numoto, M. and Yamamoto, S. (1999). Analysis of the consensus binding sequence and the DNA-binding domain of ZF5. Biochem. Biophys. Res. Commun. **255**, 528-534.

- Ohler, U., Liao, G. C., Niemann, H. and Rubin, G. M. (2002). Computational analysis of core promoters in the Drosophila genome. Genome Biol. **3**, RESEARCH0087.
- Praz, V., Perier, R., Bonnard, C. and Bucher, P. (2002). The Eukaryotic Promoter Database, EPD: new entry types and links to gene expression data. Nucleic Acids Res. **30**, 322-324.
- Smale, S. T. and Kadonaga, J. T. (2003). The RNA polymerase II core promoter. Annu. Rev. Biochem. **72**, 449-479.
- Suzuki, Y., Tsunoda, T., Sese, J., Taira, H., Mizushima-Sugano, J., Hata, H., Ota, T., Isogai, T., Tanaka, T., Nakamura, Y., Suyama, A., Sakaki, Y., Morishita, S., Okubo, K. and Sugano, S. (2001). Identification and characterization of the potential promoter regions of 1031 kinds of human genes. Genome Res. **11**, 677-684.
- Takeda, K. and Akira, S. (2000). STAT family of transcription factors in cytokine-mediated biological responses. Cytokine Growth Factor Rev. **11**, 199-207.
- Underhill, D. A. (2000). Genetic and biochemical diversity in the Pax gene family. Biochem. Cell Biol. **78**, 629-38.
- Werner, T. (1999). Models for prediction and recognition of eukaryotic promoters. Mamm. Genome **10**, 168-175.