BIOINFORMATICS DISCOVERY NOTE Vol. 21 no. 8 2005, pages 1295–1300
doi:10.1093/bioinformatics/bti172

Sequence analysis

Synergy of human Pol II core promoter elements revealed by statistical sequence analysis

Naum I. Gershenzon and Ilya P. Ioshikhes*

Department of Biomedical Informatics, The Ohio State University, 3184 Graves Hall, 333 W. 10th Avenue, Columbus, OH 43210, USA

Received on July 29, 2004; revised on November 3, 2004; accepted on November 20, 2004
Advance Access publication November 30, 2004

Downloaded from https://academic.oup.com/bioinformatics/article/21/8/1295/249581 by guest on 16 August 2022

ABSTRACT

Motivation: The subject of our paper is bioinformatics analysis of the distinguishing features of human promoter DNA sequences, in particular of synergetic combinations of core promoter elements therein. We suppose that specific scenarios of transcription initiation are essentially related to various particular implementations of the interaction of basal transcription machinery with promoter DNA, depending on the presence and mutual positioning of core promoter elements.

Results: In addition to the combinations of core promoter elements previously experimentally confirmed [TATA box and Initiator (Inr), Downstream Promoter Element (DPE) and Inr, and TFIIB recognition element (BRE) and TATA box] we propose other alternate synergetic combinations: BRE and Inr, BRE and DPE, and TATA and DPE with respective models. The suggestion is based on a high statistical significance of the alternate combinations in promoters, comparable with the significance of the known combinations. We also present arguments that the BRE element is statistically more important than previously thought, and suggest possible mechanisms of action of the core elements in the promoters with multiple transcription start sites.

Contact: ioschikhes-1@medctr.osu.edu

Supplementary information: Supplementary information is available at http://bmi.osu.edu/~ilya/synergy/Gershenzon_SuppMat-R.pdf

INTRODUCTION

Gene transcription is a multi-step, multi-level process involving many transcription factors. A fundamental step of the transcription initiation is an interaction of the basal transcription machinery [also named pre-initiation complex (PIC)] with a core promoter area of DNA spanning about ±40 bp around the transcription start site (TSS) (compare Smale and Kadonaga, 2003; Butler and Kadonaga, 2002; Zhang, 1998; Lewis *et al*., 2000). So far a few core-promoter elements have been found to be a target for the basal machinery. The most common elements are TATA box, Initiator (Inr), Downstream Promoter Element (DPE), and TFIIB recognition element (BRE) (Smale and Kadonaga, 2003).

There are a few general TFs (TFIIA, B, D, E, F and H) necessary for successful initiation of transcription (Roeder, 1996; Orphanides *et al*., 1996; Nikolov and Burley, 1997; Hampsey, 1998). Despite a diversity of scenarios TFIID always plays the central role in this process (Burley and Roeder, 1996; Burke and Kadonaga, 1997), acting in cooperation (synergy) with the core promoter elements and/or specific TFs (Nikolov and Burley, 1997; Hampsey, 1998; Lemon and Tjian, 2000). The TFIID consists of TATA Binding Protein (TBP) and at least 12 transcription associated factors (TAFs) (Green, 2000). In the TATA box-containing (TATA+) promoters, TBP binding starts the process of PIC formation. In the absence of the TATA box (TATA-less promoters), TAFs bind to DNA and/or to other TFs in order to involve TFIID (and TBP) in PIC (Burke and Kadonaga, 1997; Zenzie-Gregory *et al*., 1993; Martinez *et al*., 1995; Tsai and Sigler, 2000). Several combinations of the core-promoter elements were found to be synergistically advantageous for transcription initiation: TATA box and Inr (O'Shea-Greenfield and Smale, 1992; Emami *et al*., 1997), DPE and Inr (Burke and Kadonaga, 1997; Zhou and Chiang, 2001), and BRE and TATA box (Tsai and Sigler, 2000; Lagrange *et al*., 1998). In the present study, we discovered the high statistical significance of other combinations of the core-promoter elements suggesting existence of the additional synergetic combinations: BRE and Inr, TATA box and DPE, and BRE and DPE.

The statistics of the core elements for human promoters still remains obscure even for the most studied elements like TATA box and Inr. TATA-containing promoters were historically discovered first and the TATA box was thought to be the universal promoter element (Butler and Kadonaga, 2002). The TATA-less promoters were obtained several years later and their percentage in the total number of studied promoters has decreased steadily since: from 78% (Bucher, 1990) to 64% (Babenko *et al*., 1999) to 32% (Suzuki *et al*., 2001). There is also no consistency with Inr-containing promoters: 60% (Bucher, 1990) and 85% (Suzuki *et al*., 2001). The DPE was mainly studied in *Drosophila* (Kutach and Kadonaga, 2000). It was shown that DPE is conserved from *Drosophila* to human (Burke and Kadonaga, 1997); however, so far only one human gene with DPE has been experimentally studied (Zhou and Chiang, 2001). Few human genes with functional BRE have actually been investigated (Lagrange *et al*., 1998; Tsai and Sigler, 2000), so a general role of the BRE element was still under question (Smale and Kadonaga, 2003). In this paper, we give statistics of the aforementioned core-promoter elements and their synergetic combinations, both those previously described and those suggested herein. These statistics are based only on an examination of the presence of the element motifs defined by respective position weight matrices or consensus sequences. The actual functionality of each individual element in each individual gene is beyond the scope of this article.

Despite the complexity and diversity of the biochemical interactions between the basal machinery and the core-promoter sequence,

*To whom correspondence should be addressed.

© The Author 2004. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oupjournals.org

these interactions essentially related to can be considered as one between the different parts of PIC and DNA through the core promoter elements (Smale and Kadonaga, 2003). The hypothesis behind our research is that specific scenarios of the transcription initiation are essentially various particular implementations of the general PIC–DNA interaction, depending on the presence and mutual positioning of different core promoter elements.

Based on statistical analysis we will examine the following particular questions in order to check this hypothesis:

(1) How many known human promoters follow known scenarios of the interaction of the basal machinery and DNA? In particular, the transcription of how many promoters is guided by the TATA box and/or by any of the known synergetic combinations?

(2) May statistical analysis suggest new scenarios?

(3) Do all known promoters contain at least one known core-promoter element at a position where it is able to function?

(4) Do the four known core-promoter elements play any role in transcription of genes with multiple start sites (MSS) promoters. In particular, is there any correlation between the multiple TSS positions and positions of the core-promoter elements?

(5) What is a relationship between core-promoter elements and CpG island?

DATA AND METHODS

A total of 1871 non-redundant human promoter sequences from the Eukaryotic Promoter Database (EPD) release 75 (http://www.epd.isb-sib.ch) and 8793 human promoters from the Database of Transcriptional Start Sites (DBTSS) (http://dbtss.hgc.jp/index.html) were used for statistical analyses as two separate datasets. We also constructed a small test set of 27 human promoters with MSS (see Supplementary Material 1). This set was utilized to analyze the statistics of core-promoter elements in MSS promoters. Each promoter was considered several times, one time for each known TSS, so the total number of sequences in this set is 107. The software package, Promoter Classifier (Gershenzon and Ioshikhes, 2005) (available at http://www.bmi.osu.edu/∼ilya/promoter_classifier/) was used for statistical analysis.

We exploit the idea that due to evolution, the motifs necessary for promoter regulation have been preserved in a promoter area and, therefore, their occurrence frequencies there are far from random. So the statistical analysis of averaged positional distribution of the element's occurrence frequency ($OF_i = n_i/N_s$, where n_i is the number of promoters containing a considered element centered at position i in N_s aligned promoter sequences) is the main method of our investigation. To find the element's occurrence frequency distribution we scan each promoter sequence at each position by respective weight matrix or motif consensus. We examine the presence of the core-promoter elements and relations between the elements in different subsets of human promoters. To implement this strategy we divided all three datasets into subsets (the respective subsets for EPD are available in the Sequence Supplementary Material). To extract a subset of promoter sequences containing the TATA box or Inr element at their functional positions, the positional weight matrices (PWM) with optimal cut-off values (Table 1) were applied (Bucher, 1990). We define the TATA or Inr element as being present at a certain position if the PWM score at this position exceeds the cut-off value, and define the element to be absent at this position otherwise. Since there are no matrices for DPE and BRE, we matched 5 out of 5 letters and 6 out of 7 for the DPE and BRE consensuses (Smale and Kadonaga, 2003), respectively.

We used the same parameters to extract subsets containing known synergetic combinations, yet the respective elements had to be placed at their experimentally defined synergetic distance from one another. The distances between the elements in the remaining combinations were chosen based on the positions of the respective elements in the known combinations.

Table 1. The parameters of core-promoter elements

Name	Consensus	Length/left	Cut-off	Window	dS EPD	dS DBTSS
TATA	TATAAA	15/3	0.79	−33 to −23	52.0	46.1
Inr	YYANWYY	8/2	0.814	−5 to +6	11.1	22.3
DPE	RGWYV	5/0	5 out of 5	+23 to +33	5.0	11.0
BRE	SSRCGCC	7/0	6 out of 7	−42 to −32	6.8	15.9

List of the core promoter elements (col. 1); consensus motif in an NC-IUB nomenclature (http://www.chem.qmul.ac.uk/iubmb/misc/naseq.html) (col. 2); the length of motif (at left) and the distance between center and 5′ end (at right) (col. 3); cut-off value for weight matrix (col. 4); applied windows for the center of motifs (col. 5); and statistical significance (dS) of the occurrence frequency of an element in the respective window in EPD and DBTSS databases (col. 6). All respective P-values are less than 0.0001 which is considered to be extremely statistically significant.

To estimate the statistical significance of the occurrence frequency of an element or synergetic combination in the respective functional window, we calculated a parameter statistical significance, dS, measured in units of standard deviation $StD = \sqrt{N_{out}}$ $dS = (N_{in} - N_{out})/\sqrt{N_{out}}$, where N_{in} is the number of occurrences of an element or combination inside its functional window and N_{out} is the number of occurrences of that element or combination in the average interval of the same length outside the functional window. Since MSS promoters may contain core elements in several positions, to calculate statistical significance for the MSS dataset we use the value N_{out} [respectively recalculated $(N_{out}(\text{MSS}) = N_{out}(\text{EPD}) * 107/1871)$] from the EPD dataset.

In order to comprehend a correlation of the core elements and their combinations with CpG islands we divided the datasets to subsets with CpG island spanning TSS (CpG+) and without it (CpG-less). For implementation the commonly used parameters of CpG island (Gardiner-Garden and Frommer, 1987) were applied: (i) the length is over 200; (ii) over 50% of nucleotides are G or C; and (iii) the ratio of observed/expected CG dinucleotides ($N_{CG} * L/N_C * N_G$) exceeds 0.6. Here L is the length of the window considered; N_{CG} is the number of CG dinucleotides; and N_C and N_G are the number of nucleotides C and G, respectively, in that window. We scan over the sequences with window $L = 100$ starting from position −200 bp and ending at position 100 for the 5′ end of the window. The promoter is considered having a CpG island if the combined length of overlapping windows which satisfy criteria (ii) and (iii) exceeds 200. Since the 3′ ends of EPD sequences are defined up to +100 bp only, for them the window size L starting from the position +1 was shrunk accordingly.

RESULTS

To define an interval (window) for a functional position of a given element we considered the distribution of the element's occurrence frequency along the promoters. For both databases, we found the unambiguous maximums for the occurrence frequencies of the centers of the TATA and Inr elements at positions −28 and +1 respectively (see Figs 1 and 2 in Supplementary Material 2), which is consistent with the known functional positions of these elements. The occurrence frequency of the TATA box is essentially larger in the window (−33 to −23 bp) than in the surrounding area. We consider this window as functional for the TATA box. For the Inr element the functional window is (−5 to +6 bp) since the TSS position in EPD is defined with the accuracy ±5 bp (Cavin Périer *et al.*, 1998). Since DPE works in cooperation with Inr if positioned 27 bp

Downloaded from https://academic.oup.com/bioinformatics/article/21/8/1295/249581 by guest on 16 August 2022

Table 2. The distribution of elements in different subsets of promoters from the EPD database

	TATA		Inr		DPE		BRE	
	%	*N*	%	*N*	%	*N*	%	*N*
TATA+	**21.8**	**407**	61.9	252	24.1	101	11.8	48
TATA−			45.4	665	24.8	371	28.1	411
Inr+	27.5	252	**49.0**	**917**	25.8	242	21.1	193
Inr−	16.3	155			23.6	230	27.8	266
DPE+	21.4	101	51.3	242	**25.2**	**472**	27.3	129
DPE−	21.9	306	48.3	675			23.6	330
BRE+	10.5	48	42.1	193	28.1	129	**24.5**	**459**
BRE−	25.4	359	51.3	724	24.3	343		

The cells contain a percentage (%) and the absolute number (*N*) of promoters with an element at its functional position named in the header of a column from the subset of promoters named in the left column. For example, the values 61.9 and 252 in the cells (column Inr, line TATA+) are the percentage and absolute number of the promoters containing Inr element from the TATA+ subset of promoters. The boldface numbers indicate percentage and the absolute number of promoters containing an element named in the header of a column from the whole promoter database.

Table 3. The distribution of elements in different subsets of DBTSS database

	TATA		Inr		DPE		BRE	
	%	*N*	%	*N*	%	*N*	%	*N*
TATA+	**10.4**	**915**	57.9	532	23.4	214	13.8	126
TATA−			47.3	3725	24.8	1959	26.9	2120
Inr+	12.5	532	**48.4**	**4257**	24.9	1070	24.7	1052
Inr−	8.4	383			24.4	1103	26.3	1194
DPE+	9.8	214	49.2	1070	**24.7**	**2173**	25.5	553
DPE−	10.6	701	48.1	3187			25.6	1693
BRE+	5.6	126	46.8	1052	24.6	553	**25.5**	**2246**
BRE−	12.1	789	49.0	3205	24.7	1620		

The cells have the same meaning as for Table 2.

downstream from it (Burke and Kadonaga, 1997), we applied the window (28 − 5)–(28 + 5) bp for the DPE. BRE is shifted from the TATA box in the 5′ direction to the distance equal to the BRE length plus the 2 bp between the center of TATA element and its first 'T' (Tsai and Sigler, 2000). So the functional window for BRE is (−33 − 7 − 2 to −23 − 7 − 2 bp). Note that the occurrence frequencies of all four core-promoter elements at their functional windows are essentially larger than in the rest of the promoter area. Indeed, the statistical significance ranges from 5.0 (11.0) *StD* for DPE up to 52.0 (46.1) *StD* for the TATA box (Table 1, the last two columns). Hereafter the first number refers to EPD and the following number (in parentheses) refers to DBTSS. While the high levels of the statistical significance for the TATA box and Inr elements are not surprise, the high statistical significance of the DPE and BRE elements at their expected functional positions have never been revealed before for human genes.

Tables 2 and 3 represent percentages of the core-promoter elements in different EPD and DBTSS subsets. According to these data, half of the promoters, 49.0% (48.4%), have the Inr element at a functional position, only 21.8% (10.4%) have TATA box, 24.6% (24.6%) contain DPE, and 24.5% (25.5%) have BRE.

As we see, the percentage of the TATA+ promoters is much lower than even the minimal previous estimate (32%, Suzuki *et al*., 2001). Comparison of an absolute number of the TATA+ promoters (Tables 2 and 3) with those expected from the Suzuki's estimate for the EPD and DBTSS datasets (599 and 2814 sequences, respectively) gives a difference of 7.8 (35.8) *StD* below the estimate. The TATA+ promoters have a larger probability of having an Inr element than TATA-less promoters. Indeed, 61.9% (57.9%) of TATA+ promoters have Inr compared with 45.4% (47.3%) of TATA-less promoters. The presence of DPE is virtually irrelevant to the presence of the TATA box or Inr elements (Tables 2 and 3), in contrast to the *Drosophila* (Kutach and Kadonaga, 2000). The BRE-containing promoters 'prefer' to be TATA-less promoters: 28.1% (26.9%) of TATA-less promoters contain BRE versus 11.8% (13.8%) of TATA+ promoters. The majority of the promoters, 77.3% (74.3%), have at least one of four core-promoter elements at its functional position and 41.8% (44.1%) have only one element including TATA – 5.5% (2.9%), Inr – 20.1% (23.0%), DPE – 6.6% (8.4%), and BRE – 9.6% (9.8%). The list of promoters from EPD with no core-promoter elements at a functional position may be found in the Sequence Supplementary Material.

Table 4 shows the distances between the elements, percentages, actual numbers and statistical significance of promoters having combinations of the elements in the entire EPD and DBTSS datasets. We also calculated the percentages of all combinations (the last sub-column in columns 4 and 5) with the distances being the same as in column 3 plus one in both directions. The results clearly indicate the high statistical significance of the occurrence frequencies of all considered combinations in both promoter databases (the third subcolumn in columns 4 and 5) (see also Figs 1–6 of Supplementary Material 3). The data is consistent between two databases. The widening of the range of the distances between elements increases the percentages of the promoters containing respective combinations, preserving the ratios between them.

The presence of a CpG island essentially affects the promoter contents. The distributions of elements and their synergetic combinations for the CpG+ and CpG-less subset of promoters are presented in Table 5. As expected, the percentage of TATA+ promoters in the CpG-less subset is much higher than in CpG+ (Table 5, first line). However, still 13.3% (6.9%) of promoters with CpG island have a TATA box. The percentage of Inr+ promoters in the CpG-less subset is also higher than in CpG+ (second line). The presence of DPE is slightly more probable in the absence of CpG islands (third line). Note that statistical significances of occurrence frequency of the TATA box, Inr and DPE elements are high for both CpG+ and CpG-less subsets for both databases. Thus, our statistics do not confirm the widely held opinion that 'CpG islands usually lack consensus or near-consensus TATA boxes, DPE elements, or Inr elements' (Smale and Kadonaga, 2003). The BRE is the only element whose presence is much more probable in the CpG+ promoters [30.9% (33.4%) in CpG+ versus 9.7% (7.7%) in CpG-less]. The statistical significance of BRE is high for the CpG+ subset and non-substantial for the CpG-less subset (line 4). All six combinations of elements (with the exception of combinations with BRE in CpG-less subset) have high level of statistical significance in both subsets for both databases (lines 5–10).

We found that 83 from 107 MSS promoters (i.e. 76.9%) contain at least one core-promoter element in the functional position relative to the TSS. This percentage is practically the same as for all

Downloaded from https://academic.oup.com/bioinformatics/article/21/8/1295/249581 by guest on 16 August 2022

Table 4. The statistical parameters of combinations of core elements

1. Combination	2. Position	3. Distance	4. EPD %	*N*	d*S*	±1	5. DBTSS %	*N*	d*S*	±1
TATA_Inr	−33 to −23	26–30	9.4	175	50.5	11.1	4.3	376	47.9	5.2
Inr_DPE	−5 to +6	27	2.1	40	6.7	5.1	1.9	165	12.2	4.6
BRE_TATA	−42 to −32	9	0.86	16	19.1	1.4	0.47	41	24.4	0.83
BRE_Inr	−42 to −32	35–39	5.5	103	7.0	7.6	6.7	592	19.8	8.8
BRE_DPE	−42 to −32	62–66	4.3	77	8.8	5.3	3.6	319	14.5	4.7
TATA_DPE	−33 to −23	53–57	3.2	60	25.5	3.8	1.3	118	19.9	1.8

Combination name (col. 1); position of the center of the first element of the combination in bp (col. 2); synergetic distance between the centers of the elements in bp (col. 3); the percentage (%), the absolute number (*N*) and statistical significance of occurrence frequency of promoters having this combination at respective positions with distance as in col 3 (cols 4 and 5 for EPD and DBTSS, respectively), the fourth sub-columns in columns 4 and 5 contain the percentage of promoters having this combination at distance as in col. 3 with a shift ±1 bp versus distance in col. 3. All respective *P*-values are less than 0.0001 which is considered to be extremely statistically significant.

Table 5. The percentage (%), absolute number (*N*) and statistical significance (d*S*) of elements and their synergetic combinations in CpG+ and CpG-less promoters calculated for EPD and DBTSS promoter databases (respective *P*-values are less than 0.0001 if $dS \geq 3.8StD$)

	CpG+						CpG-less					
	EPD			DBTSS			EPD			DBTSS		
	%	*N*	d*S*	%	*N*	d*S*	%	*N*	d*S*	%	*N*	d*S*
TATA	13.3	175	28.7	6.9	419	28.1	41.6	232	45.9	18.4	496	37.2
Inr	43.8	575	7.6	45.2	2752	18.2	61.3	342	8.5	55.8	1505	13.0
DPE	23.8	312	3.1	24.3	1480	8.9	28.7	160	4.3	25.7	693	6.5
BRE	30.9	405	6.7	33.4	2038	16.4	9.7	54	1.4	7.7	208	1.2
TATA_Inr	4.9	64	24.1	2.7	164	29.2	19.9	111	48.2	7.9	212	38.5
Inr_DPE	2.4	31	7.2	1.9	117	11.7	1.6	9	1.6	1.8	48	4.8
BRE_TATA	1.1	14	18.2	0.56	34	22.4	0.36	2	6.1	0.26	7	9.8
BRE_Inr	7.2	94	7.5	8.5	521	19.5	1.6	9	0.4	2.6	71	4.8
BRE _DPE	5.2	68	8.6	4.7	286	14.5	1.6	9	2.4	1.2	33	2.9
TATA_DPE	2.0	26	13.9	0.80	49	10.2	6.1	34	23.2	2.6	69	18.4

promoters from the both datasets. The statistical significance of the presence of any one of the four elements in the functional position is comparatively high for a relatively small dataset: $dS = 3.5StD$, P-value $= 0.0005$. Remarkably, the portion of MSS promoters containing BRE (29.6%) is larger than on average in the EPD/DBTSS datasets. Since the d*S* value is roughly proportional to the $\sqrt{}$ of the number of sequences, one may expect respective decrease of a statistical significance of every particular element on a small dataset. For example, one would expect the statistical significance of every element in the MSS promoters to be approximately in 4.2 [$\sqrt{(1871/107)}$] times lower than for EPD database (Table 1). BRE is the only element whose statistical significance exceeds the expectation, reaching $dS = 3.4StD$, $P = 0.0007$ in the MSS TATA-less promoters. Thus the presence of the BRE element in the CpG+ and MSS promoters is comparable with the presence of the TATA box in the CpG-less promoters.

DISCUSSION

The TATA box at position −26 to −30 and the Inr element around TSS enable the successful transcription initiation (O'Shea-Greenfield and Smale, 1992). In this scenario, the TFIID presumably binds to DNA through both the TATA box and Inr elements (see schematic representation on Fig. 1A). Only 9.4% (4.3%) of the promoters (Table 4) contain TATA and Inr at the synergetic distance. How does the transcription machinery work in the rest of the promoters?

The DPE element was found to be a target for TFIID and, in some cases, leads the transcription initiation in cooperation with the Inr element (Burke and Kadonaga, 1997) (Fig. 1B). Some of the TAFs [$TAF_{II}55$ was found to be one of them (Zhou and Chiang, 2001)] bind to the DPE motif and attract TFIID to DNA.

TFIIB plays a central role in TSS selection as well as in the PIC assembly connecting TFIID and Pol II (Hawkes and Roberts, 1999; Fairley *et al*., 2002). In the presence of BRE, TFIIB binds to DNA immediately upstream of the TATA box and to TFIID to direct transcription (Tsai and Sigler, 2000; Lagrange *et al*., 1998) or to repress it (Evans *et al*., 2001) (Fig. 1C).

Note the common features of the aforementioned combinations: (1) all of them involve TFIID, and TBP binds to DNA regardless of the presence/absence of TATA box; (2) TFIID covers the TSS area; (3) the distance from the TSS to the edge of the complex is approximately the same (∼30–40 bp). Combinations BRE_Inr, BRE_DPE and TATA_DPE also satisfy these requirements. These combinations are presented in a number of promoters comparable with the three previous combinations with comparable statistical significance (Table 4). They may therefore be also considered as possible synergetic combinations of core-promoter elements (Fig. 1D–F).

Downloaded from https://academic.oup.com/bioinformatics/article/21/8/1295/249581 by guest on 16 August 2022

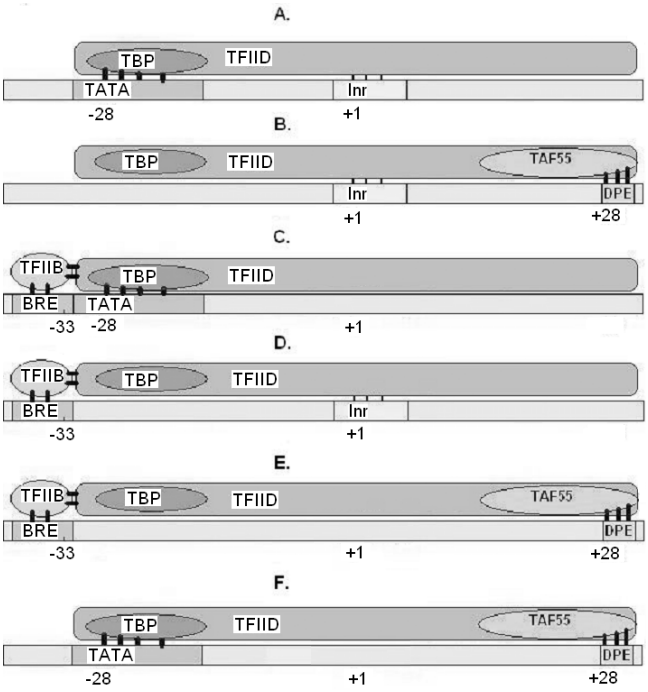

Fig. 1. Illustration of different possible scenarios of interaction between general TFs (TFIID and TFIIB) and core-promoter elements. The lower bar on each picture represents promoter area of DNA. TSS is placed at position +1. The bold black lines indicate interaction between the TFs or between the TFs and binding elements.

The following arguments show the possibility of synergy between the BRE and Inr elements. An essential part of the TATA-less promoters [28.1% (26.9%)] contains BRE (Tables 2 and 3). There is experimental evidence that TFIIB may recognize BRE directly, not necessarily through interaction with TBP (Lagrange *et al*., 1998). As in the TATA+ promoters, TFIIB can bind to the BRE motif of DNA and to some TAFs of the TFIID complex (Fig. 1D), attracting TFIID to DNA. It was found that non-sequence-specific bound TBP (i.e. bound not to the TATA box element) is also active in assembling PIC (Coleman and Pugh, 1995); so as in the Inr_DPE promoters, TBP could bind to the DNA upstream of TSS. The interaction of TFIID and Inr may create a stable complex as in the TATA_Inr case. The percentage of combination BRE_Inr is comparable with the TATA_Inr combination and the statistical significance of the former combination is high: 9.6 (8.8) *StD* (Table 4). This observation partially supports the statement that 'IIB–BRE interaction will play a role, possibly a dominant role, in preinitiation complex assembly and transcription initiation at TATA-less promoters' (Lagrange *et al*., 1998).

The same arguments also work for the BRE_DPE combination. Indeed, the subset TATA-less_Inr-less contains much more BRE [31.2% (27.5%)] than the subset TATA+Inr+ [12.6% (14.1%)]. The statistical significance of this combination is also high: 8.0 (19.3) *StD* (Table 4). In this case TFIIB binds to both the TFIID and the BRE motif, and TFIID through TAFs binds to the DPE motif (Fig. 1E).

Finally, the combination TATA_DPE [Fig. 1F, statistical significance 25.5 (19.9) *StD*] also may work in the framework described above: TFIID binds to the TATA box through TBP, and another part of TFIID binds to DPE. Of course, in such promoters the TATA box may be strong enough to start transcription (at least *in vitro*) alone (Burke and Kadonaga, 1997). Hypothetically, *in vivo*, when many subtle factors are essential for transcription regulation, the TATA and DPE elements placed at their functional positions could work synergistically.

As we have already mentioned the majority of promoters have at least one core-promoter element at a functional position. These elements can work as an anchor for the basal machinery. In many cases, the presence of a synergetic combination of two elements, which is much stronger than a single element, dictates the position of TSS. In other cases, the position of one core element plus the position of binding sites of non-general transcription factors, like Sp1, which interacts with both DNA and PIC, define the position of TSS (Liao *et al*., 1994). In any case, most likely the presence of any core element is beneficial for transcription initiation. Usually the promoters with strong synergetic combinations have SSS. If there are no such combinations, as in TATA-less_Inr-less promoters, the presence of core elements could possibly initiate multiple weak TSSs in a so-called initiation window of MSS promoters (Lin *et al*., 2001). This is consistent with the suggestion that the TSS positions in MSS promoters are defined in part by the positions of the core elements. (See Supplemental Material 4 for example of an MSS sequence with several core-promoter elements at functional positions.)

In order to minimize the possible database biases we used two different promoter databases, EPD and DBTSS. Comparisons show that both databases give, in general, consistent results (Tables 1–5). The only visible difference is between TATA + promoters: 21.8% for EPD and 10.4% for DBTSS. This discrepancy may be explained by the difference in database creation and fivefold difference in volume. The EPD database is a collection of experimentally defined promoters (Cavin Périer *et al*., 1998). The DBTSS promoters were identified by the mRNA start sites determined by a large-scale sequencing of the cDNA libraries constructed by the 'oligo-capping' method (Suzuki *et al*., 2001). So the percentage of the TATA+ promoters in the EPD database is higher since the TATA-containing promoters are more accessible for experimental analysis by standard start-site mapping techniques and hence were discovered. We have already mentioned that the maximal occurrence frequency of the Inr element is placed at position +1. The frequencies at the positions of the nearest neighbors (−1, +1) bp are approximately the same as an average occurrence frequency. This pattern is true for both the CpG+ and CpG-less subsets. This means that for both databases in the majority of (at least) Inr+ promoters the TSS position was defined with precise accuracy.

The most important conclusions of this study are: (1) The portion of the TATA+ promoters is just from 10 to 20% of all known human promoters. (2) The statistical significances of the occurrence frequency of the DPE and BRE elements at their experimentally defined functional positions are high, indicating that considerable amount of human genes use these elements for the transcription. (3) The combinations of the core-promoter elements such as BRE and Inr, TATA and DPE, and BRE and DPE are statistically as important as known synergetic combinations such as TATA_Inr, Inr_DPE and TATA_BRE suggesting that the former combinations may also work synergistically. (4) The high percentage and statistical significance of MSS promoters having core-promoter elements at functional positions suggests that those elements define the position of TSS in MSS promoters. (5) The high percentage and statistical significance

Downloaded from https://academic.oup.com/bioinformatics/article/21/8/1295/249581 by guest on 16 August 2022

of BRE, especially in CpG+ and MSS promoters, suggests that this element may be functional in many promoters including TATA-less promoters. (6) Approximately one-fourth of all promoters do not have any of the four core-promoter elements suggesting the existence of other yet undiscovered core elements.

ACKNOWLEDGEMENTS

We are grateful to L.F. Johnson, J. Kadonaga and M.Q. Zhang for their useful comments on this work.

REFERENCES

Babenko,V.N., Kosarev,P.S., Vishnevsky,O.V., Levitsky,V.G., Basin,V.V. and Frolov,A.S. (1999) Investigating extended regulatory regions of genomic DNA sequences. *Bioinformatics*, **15**, 644–653.

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.

Burke,T.W. and Kadonaga,J.T. (1997) The downstream core promoter element, DPE, is conserved from *Drosophila* to humans and is recognized by $TAF_{II}60$ of *Drosophila Genes Dev.*, **11**, 3020–3031.

Burley,S.K. and Roeder,R.G. (1996) Biochemistry and structural biology of transcription factor IID (TFIID). *Annu. Rev. Biochem.*, **65**, 769–799.

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.

Cavin Périer,R., Junier,T. and Bucher,P. (1998) The Eukaryotic Promoter Database EPD. *Nucleic Acids Res.*, **26**, 353–357.

Coleman,R.A. and Pugh,B.F. (1995) Evidence for functional binding and stable sliding of the TATA binding protein on nonspecific DNA. *J. Biol. Chem.*, **270**, 13850–13859.

Emami,K.H., Jain,A. and Smale,S.T. (1997) Mechanism of synergy between TATA and initiator: synergistic binding of TFIID following a putative TFIIA-induced isomerization. *Genes Dev.*, **11**, 3007–3019.

Evans,R., Fairley,J.A. and Roberts,S.G. (2001) Activator-mediated disruption of sequence-specific DNA contacts by the general transcription factor TFIIB. *Genes Dev.*, **5**, 2945–2949.

Fairley,J.A., Evans,R., Hawkes,N.A. and Roberts,S.G. (2002) Core promoter-dependent TFIIB conformation and a role for TFIIB conformation in transcription start site selection. *Mol. Cell Biol.*, **22**, 6697–6705.

Gardiner-Garden,M. and Frommer,M. (1987) CpG islands in vertebrate genomes. *J. Mol. Biol.*, **196**, 261–282.

Gershenzon,N. and Ioshikhes,I. (2005) Promoter Classifier: software package for promoter database analysis. *Appl. Bioinformatics*, **4** (in press).

Green,M.R. (2000) TBP-associated factors (TAF_{II}s): multiple, selective transcriptional mediators in common complexes. *Trends Biochem. Sci.*, **25**, 59–63.

Hampsey,M. (1998) Molecular genetics of the RNA polymerase II general transcriptional machinery. *Microbiol. Mol. Biol. Rev.*, **62**, 465–503.

Hawkes,N.A. and Roberts,S.G.E. (1999) The role of human TFIIB in transcription start site selection *in vitro* and *in vivo. J. Biol. Chem.*, **274**, 14337–14343.

Kutach,A.K. and Kadonaga,J.T. (2000) The downstream promoter element DPE appears to be as widely used as the TATA box in Drosophila core promoters. *Mol. Cell. Biol.*, **20**, 4754–4764.

Lagrange,T., Kapanidis,A.N., Tang,H., Reinberg,D. and Ebright,R.H. (1998) New core promoter element in RNA polymerase II-dependent transcription: Sequence-specific DNA binding by transcription factor IIB. *Genes Dev.*, **12**, 34–44.

Lemon,B. and Tjian,R. (2000) Orchestrated response: A symphony of transcription factors for gene control. *Genes Dev.*, **14**, 2551–2569.

Lewis,BA., Kim,T.K. and Orkin,S.H. (2000) A downstream element in the human beta-globin promoter: evidence of extended sequence-specific transcription factor IID contacts. *Proc. Natl Acad. Sci. USA*, **97**, 7172–7177.

Liao,W.-C., Geng,Y. and Johnson,L.F. (1994) In vitro transcription of the TATAA-less mouse thymidylate synthase promoter: multiple transcription start points and evidence for bidirectionality. *Gene*, **146**, 183–189.

Lin,Y., Ince,T.A. and Scotto,K.W. (2001) Optimization of a versatile in vitro transcription assay for the expression of multiple start site TATA-less promoters. *Biochemistry*, **40**, 12959–12966.

Martinez,E., Zhou,Q., L'Etoile,N.D., Oelgeschlager,T., Berk,A.J. and Roeder,R.G. (1995) Core promoter-specific function of a mutant transcription factor TFIID defective in TATA-Box binding. *Proc. Natl Acad. Sci. USA*, **92**, 11864–11868.

Nikolov,D.B. and Burley,S.K. (1997) RNA polymerase II transcription initiation: a structural view. *Proc. Natl Acad. Sci. USA*, **94**, 15–22.

Orphanides,G., Lagrange,T. and Reinberg,D. (1996) The general transcription factors of RNA polymerase II. *Genes Dev.*, **10**, 2657–2662.

O'Shea-Greenfield,A. and Smale,S.T. (1992) Roles of TATA and initiator elements in determining the start site location and direction of RNA polymerase II transcription. *J. Biol. Chem.*, **267**, 1391–1402.

Roeder,R.G. (1996) The role of general initiation factors in transcription by RNA polymerase II. *Trends Biochem. Sci.*, **21**, 327–335.

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. *et al.* (2001) Identification and characterization of the potential promoter regions of 1031 kinds of human genes. *Genome Res.*, **11**, 677–684.

Tsai,F.T.F. and Sigler,P.B. (2000) Structural basis of preinitiation complex assembly on human Pol II promoters. *EMBO J.*, **19**, 25–36.

Zenzie-Gregory,B., Khachi,A., Garraway,I.P. and Smale,S.T. (1993) Mechanism of initiator-mediated transcription: evidence for a functional interaction between the TATA-binding protein and DNA in the absence of a specific recognition sequence. *Mol. Cell Biol.*, **13**, 3841–3849.

Zhang,M.Q. (1998) A discrimination study of human core-promoters. *Pac. Symp. Biocomput., 1998*, 240–251.

Zhou,T. and Chiang,C.-M. (2001) The intronless and TATA-less human $TAF_{II}55$ gene contains a functional initiator and a downstream promoter element. *J. Biol. Chem.*, **276**, 25503–25511.

Downloaded from https://academic.oup.com/bioinformatics/article/21/8/1295/249581 by guest on 16 August 2022