Transcription 2:5, 201-206; September/October 2011; © 2011 Landes Bioscience

The unexpected traits associated with core promoter elements

Rivka Dikstein

Department of Biological Chemistry; The Weizmann Institute of Science; Rehovot, Israel

Key words: core promoter, TATA-box, TATA-less promoter, gene length, transcription elongation

The core promoter of eukaryotic coding and non-coding genes that are transcribed by RNA polymerase II (RNAP II) is composed of DNA elements surrounding the transcription start site. These elements serve as the docking site of the basal transcription machinery and have an important role in determining the position and directing the rate of transcription initiation. This review summarizes the current knowledge about core promoter elements and focuses on several unexpected links between core promoter structure and certain gene features. These include the association between the presence or absence of a TATA-box and gene length, gene structure, gene function, evolution rate and transcription elongation.

Introduction

Diversity in rates of gene expression is essential for basic cell functions and is controlled through several intricate mechanisms. Major contributors to gene expression rates are DNA cis-regulatory elements that vary between promoters of individual genes. Two types of DNA regulatory sequences control transcription of protein-encoding and non-coding genes in eukaryotes. The first type is gene specific enhancer elements that serve as the binding sites of transcription regulatory factors and can be divided into two classes: those that function independently of their position relative to the transcription start site (TSS) and those that can activate transcription only when located proximal to the TSS. The second type is the core promoter, which consists of sequence elements that surround the TSS. These elements serve as a docking site for general transcription factors (GTFs) and RNA polymerase II that assemble into a pre-initiation complex (PIC).[1,2] The core promoter has a crucial role in transcription as it serves as the acceptor site for the effects exerted by enhancer-bound transcription factors; it contributes to the overall transcription level and it determines the site of transcription initiation. Thus, the information encoded in the core promoter ensures proper regulation of gene expression. This review summarizes the current knowledge about core promoter types and their mechanism of action and focuses on several traits associated with the core promoter, some of which run beyond transcription initiation.

*Correspondence to: Rivka Dikstein; Email: rivka.dikstein@weizmann.ac.il
Submitted: 06/27/11; Revised: 07/20/11; Accepted: 07/21/11
DOI: 10.4161/trns.2.5.17271

Core Promoter Elements

The first core promoter element to be described in eukaryotes was the TATA-box.[3] The TATA-box is a highly conserved element, strictly located between -35 to -25 relative to the TSS (designated +1) in most eukaryotes. The TATA-box, which was once thought to be a universal element, is present in a smaller fraction of RNAP II genes than initially estimated (**Table 1**): between 20–46% in yeast (depending on the definition of the TATA-box sequence),[4,5] ~30% of Drosophila genes[6] and up to 35% of human genes.[7-10] The major core promoter-binding factor is TFIID, a large complex consisting of the TATA-binding protein (TBP) and 13 associated factors called TAFs. The TATA-box is directly recognized and bound by TBP, while the TAFs interact with sequences upstream and downstream to the TATA-box.[11-19] In certain promoters, the TATA-box cooperates with one or more elements to direct efficient transcription initiation (**Fig. 1**). For example, two TFIIB recognition elements (BRE), which are located either upstream (BREu) or downstream (BREd) of the TATA-box[20,21] function only together with the TATA-box. Similarly, the contribution to promoter strength of the TAF1 recognition element DCE that is located downstream relative to the TSS is also dependent on the presence of a TATA-box.[22] The presence or absence of a TATA-box in core promoters has been linked in yeast[4,5,23-25] and humans[26] to two pathways of pre-initiation complex assembly, one being TFIID dependent (weak TATA or TATA-less) and the other TFIID independent and SAGA dependent.

The initiator (INR) is a metazoan conserved element, strictly located around the TSS, with a consensus of YYANWYY. The INR can be weakly bound by RNAP II itself[27] and more strongly by a complex consisting of TFIIB, TFIID, TFIIF and RNAP II.[27,28] Within TFIID the subunits that form direct and specific contacts with the INR are TAF1 and TAF2.[12,29,30] The INR can function alone, together with the TATA-box or in conjunction with two specific core promoter elements, the DPE and the MTE (**Fig. 1**). The DPE and the MTE are mostly found in Drosophila promoters and both have a strict downstream location at +28 and +18, respectively, relative to the TSS, and both are recognized by TFIID through the TAF6 and TAF9 subunits.[31-34] Following computational analysis of a large number of mammalian TSSs, the INR consensus in mammals was recently suggested to be composed of only YR, where R corresponds to the TSS.[35,36] On the other hand, another study has

Table 1. Prevalence of the TATA box in various species

Species	Sequence	Fraction (%)	Ref.
Saccharomyces cerevisiae	TATAWAWR	~20	3
Saccharomyces cerevisiae	TATAWA	45.8	4
Drosophila	TATAAA (up to one mismatch)	29.3	5
Homo sapiens	TATAWA (up to one mismatch)	8.3	8
Homo sapiens	TATAWA (two mismatches)	27	8

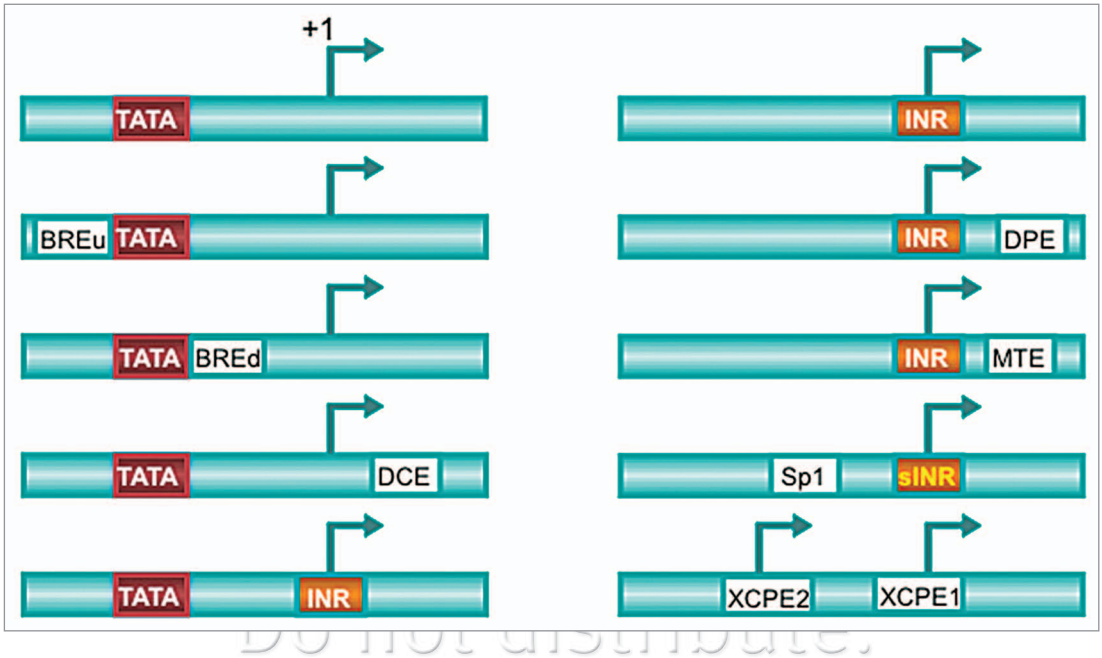

Figure 1. A scheme of several core promoters with a major TSS(s) which are governed by specific functional combinations of core promoter motifs.

identified, also by computational analysis a version of the INR, called "strict Initiator" (sINR), that is much less divergent than the INR as, unlike the INR, its core sequence is very strict and is flanked by additional conserved sequences, not shared by the INR.[37] sINR is specifically enriched in TATA-less promoters and functions in cooperation with a nearby Sp1 site (**Fig. 1**). Interestingly, while sINR can substitute for a canonical INR, it cannot be replaced by an INR, indicating that the small sequence variations are functionally very important.[37] Another element with an INR positional bias is the pyrimidine-rich TOP element present in many protein biogenesis genes.[38,39] The TOP element was recently reported to be active in Drosophila as well.[40]

Even though a substantial fraction of promoters lack both TATA-box and INR there is only a limited number of characterized core elements that function independently of a TATA-box or INR. Two such elements called, XCPE1 and XCPE2, were identified in the hepatitis B virus X gene, each one directing a distinct TSS.[41,42] XCPE1 has a consensus sequence of DSGYGGRASM and is located from -8 to +2 relative to the TSS. It is present in ~1% of human core promoters and acts only in conjunction with other sequence-specific activators such as the NRF1, NF-1 and Sp1.[42] XCPE2 has a VCYCRTTRCMY consensus and drives transcription that is independent of TAFs but dependent on TBP and the mediator.[41] A third element is MED-1 found in TATA-less promoter with unclustered, multiple start sites.[43]

Several bioinformatics studies found that most of the mammalian TATA-less promoters are associated with CpG islands.[44-47] Another poorly investigated phenomenon characteristic to a large number of TATA-less and INR-less promoters is transcription initiation from multiple sites, as opposed to the single major site that is characteristic of promoters driven by a TATA-box or INR.[36] It is very likely that in such promoters the mechanism of transcription initiation is quite different. With a strict TSS site the PIC associates with the promoter through a specific site (TATA, INR or other), while with dispersed TSSs the PIC may not be associated with one specific element. One possibility is that the general machinery is recruited by a transcription factor bound to a proximal promoter element. In the absence of a direct docking site, RNAP II is more flexible and can initiate transcription at favorable nucleotides in the vicinity of the element. Another possibility is the presence of a number of docking sites on the same promoter to which the general machinery can weakly bind and direct transcription initiation. The two possibilities are not necessarily mutually exclusive.

Association of Core Promoter with Gene Length and the Relationship with Expression

The core promoter has been recently reported to be linked to structural features of genes. A statistical analysis of more than 14,000 human genes that were classified into groups according to their core promoter type revealed a remarkable observation. Genes with a TATA-box are, on average, 3-fold shorter than TATA-less genes. Furthermore, within the TATA containing genes, length is inversely correlated to the strength of the TATA-box, with one mismatch from the TATA-box consensus being associated with a more than 2-fold increase in gene length over canonical TATA genes. Differences in gene length are primarily due to the size and number of introns.[9]

Analysis of gene expression data of genes in the different core-promoter groups revealed the expected correlation between the strength of the TATA-box and expression levels,[9] confirming previous gene-specific studies.[69,74,75] In general an increase in gene length also correlates with reduced levels of expression, but the impact of gene length varies according to the core promoter. The inverse correlation of gene length with expression was found to be the highest for genes with a TATA-box and lowest for TATA-less genes. Having a TATA-box in the core promoter seems beneficial for expression of short genes, while its advantage diminishes with longer genes. We have therefore proposed that substantial variation in gene expression levels can be achieved through different combinations of TATA promoters with varying intron length. On the other hand, a TATA-less promoter ensures similar levels of expression regardless of gene length.

The sensitivity of TATA-box genes to increased gene length may be related to the bursty nature of transcription initiation directed by the TATA-box element.[69,76] Theoretical calculations predicted that bursty transcription initiation would be highly sensitive to elongation interruptions[77] because transcription initiation bursts create localized pools of RNAP II molecules over the gene. When elongation is interrupted, the leading RNAP II pauses or stalls, thereby increasing the risk that the following RNAP II molecules will collide and destroy the burst. Clearly, the chances of RNAP II to encounter an obstacle during elongation increase in proportion to gene length, which may affect initiation bursts more frequently. In contrast, in non-bursty transcription, the distance between each RNAP II molecule allows sufficient time for RNAP II to clear the block so risk of RNAP II collision is lower.

The TATA-Box and Evolution

The presence or absence of a TATA-box has been also linked to the rate at which genes are evolved. Examination of the transcriptional responses of four closely related yeast species to a variety of environmental stresses revealed that genes containing a TATA-box show an increase in interspecies variability in expression. This enhanced expression divergence of TATA-containing genes was confirmed in all eukaryotes.[78] Another study that examined the effects of naturally occurring mutations on gene expression in yeast also found that the sensitivity of gene expression to mutations is higher for genes with a TATA-box.[79] Thus, the transcription initiation mechanism associated with the TATA-box may facilitate evolution in gene expression.

Conclusions and Perspectives

Although the core promoter is a key element in gene transcription, we still know very little about its structure and function in most of the TATA-less genes. It is likely that there are several mechanisms, yet to be discovered, by which distinct core promoter elements cooperate with distal enhancer elements to increase the rate of transcription. Additional unresolved issues that are of interest include: the identity of the trans-acting factors that bind to the different core promoter elements; the chromatin features associated with the different core promoter elements; the interplay between new core promoter elements and the succeeding stages of gene expression; the association of core promoter elements, other than the TATA-box, and gene features directly and indirectly related to transcription. Undoubtedly, a more systematic analysis of this basic component of transcriptional control is required in order to increase our ability to read the regulatory information encoded by the genome.

Acknowledgments

I would like to thank Dr. Sandra Moshonov for critical reading and editing the manuscript and the referees of this paper for their helpful suggestions. This work was supported by grants from the Israel Science Foundation, Israel Cancer Research Fund and The Pearl Welinsky Merlo Foundation Scientific Research Progress Fund. R.D. is the incumbent of the Ruth and Leonard Simon Chair of Cancer Research.

References

1. Juven-Gershon T, Hsu JY, Kadonaga JT. Perspectives on the RNA polymerase II core promoter. Biochem Soc Trans 2006; 34:1047-50.
2. Smale ST. Core promoters: active contributors to combinatorial gene regulation. Genes Dev 2001; 15:2503-8.
3. Lifton RP, Goldberg ML, Karp RW, Hogness DS. The organization of the histone genes in *Drosophila melanogaster*: functional and evolutionary implications. Cold Spring Harb Symp Quant Biol 1978; 42:1047-51.
4. Basehoar AD, Zanton SJ, Pugh BF. Identification and distinct regulation of yeast TATA box-containing genes. Cell 2004; 116:699-709.
5. Mencia M, Moqtaderi Z, Geisberg JV, Kuras L, Struhl K. Activator-specific recruitment of TFIID and regulation of ribosomal protein genes in yeast. Mol Cell 2002; 9:823-33.
6. Ohler U, Liao GC, Niemann H, Rubin GM. Computational analysis of core promoters in the Drosophila genome. Genome Biol 2002; 3:87.
7. Gershenzon NI, Ioshikhes IP. Synergy of human Pol II core promoter elements revealed by statistical sequence analysis. Bioinformatics 2005; 21:1295-300.
8. Kim TH, Barrera LO, Zheng M, Qu C, Singer MA, Richmond TA, et al. A high-resolution map of active promoters in the human genome. Nature 2005; 436:876-80.
9. Moshonov S, Elfakess R, Golan-Mashiach M, Sinvani H, Dikstein R. Links between core promoter and basic gene features influence gene expression. BMC Genomics 2008; 9:92.
10. Yang C, Bolotin E, Jiang T, Sladek FM, Martinez E. Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. Gene 2007; 389:52-65.
11. Gazit K, Moshonov S, Elfakess R, Sharon M, Mengus G, Davidson I, et al. TAF4/4b-TAF12 displays unique mode of DNA binding and is required for core promoter function of subset of genes. J Biol Chem 2009.

72. Dong D, Shao X, Deng N, Zhang Z. Gene expression variations are predictive for stochastic noise. Nucleic Acids Res 2011; 39:403-13.
73. Raser JM, O'Shea EK. Control of stochasticity in eukaryotic gene expression. Science 2004; 304:1811-4.
74. Hoopes BC, LeBlanc JF, Hawley DK. Contributions of the TATA box sequence to rate-limiting steps in transcription initiation by RNA polymerase II. J Mol Biol 1998; 277:1015-31.
75. Wobbe CR, Struhl K. Yeast and human TATA-binding proteins have nearly identical DNA sequence requirements for transcription in vitro. Mol Cell Biol 1990; 10:3859-67.
76. Yean D, Gralla J. Transcription reinitiation rate: a special role for the TATA-box. Mol Cell Biol 1997; 17:3809-16.
77. Dobrzynski M, Bruggeman FJ. Elongation dynamics shape bursty transcription and translation. Proc Natl Acad Sci USA 2009; 106:2583-8.
78. Tirosh I, Weinberger A, Carmi M, Barkai N. A genetic signature of interspecies variations in gene expression. Nat Genet 2006; 38:830-4.
79. Landry CR, Lemos B, Rifkin SA, Dickinson WJ, Hartl DL. Genetic properties influencing the evolvability of gene expression. Science 2007; 317:118-21.