

Molecular Basis of Transcription

A Focus Topic at BioScience2006, held at SECC Glasgow, U.K., 23–27 July 2006. Edited by S. Busby (Birmingham, U.K.), R. Weinzierl (Imperial College London, U.K.) and R. White (Glasgow, U.K.).

Perspectives on the RNA polymerase II core promoter

T. Juven-Gershon, J.-Y. Hsu and J.T. Kadonaga¹

Section of Molecular Biology, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0347, U.S.A.

Abstract

The RNA polymerase II core promoter is a critical yet often overlooked component in the transcription process. The core promoter is defined as the stretch of DNA, which encompasses the RNA start site and is typically approx. 40–50 nt in length, that directs the initiation of gene transcription. In the past, it has been generally presumed that core promoters are general in function and that transcription initiation occurs via a common shared mechanism. Recent studies have revealed, however, that there is considerable diversity in core promoter structure and function. There are a number of DNA elements that contribute to core promoter activity, and the specific properties of a given core promoter are dictated by the presence or absence of these core promoter motifs. The known core promoter elements include the TATA box, Inr (initiator), BRE^u {BRE [TFIIB (transcription factor for RNA polymerase II) recognition element] upstream of the TATA box} and BRE^d (BRE downstream of the TATA box), MTE (motif ten element), DCE (downstream core element) and DPE (downstream core promoter element). In this paper, we will provide some perspectives on current and future issues that pertain to the RNA polymerase II core promoter.

The core promoter is the gateway to transcription

Many biological phenomena are dependent on the proper control of transcription. In eukaryotes, there are four known nuclear RNA polymerases, and RNA polymerase II is devoted primarily to the transcription of protein-coding genes. One key step in this process is the decision whether or not to initiate transcription. Many complex events involving many factors lead to transcription initiation. Ultimately, however, the final decision whether or not to start gene transcription occurs at the core promoter. Hence, in some respects, the core promoter is the gateway of the transcription process (for recent reviews, see [1–4]).

Focused versus dispersed transcription initiation

The RNA polymerase II core promoter is usually defined to be the stretch of DNA that directs the initiation of transcription. This definition is perhaps deceptively simple. In practice, particularly in vertebrates, two very different strategies for transcription initiation have been observed. First, there is focused initiation, wherein transcription initiates at a single nucleotide or within a narrow region of several nucleotides. Focused initiation occurs with core promoters that contain sequence motifs such as the TATA box, Inr (initiator) and DPE (downstream core promoter element). Notably, these core promoter motifs are located at distinct positions relative to the +1 start site. Secondly, there is dispersed initiation, wherein transcription initiates at multiple weak start sites that are dispersed over a broad region of approx. 50–150 nt. Dispersed initiation occurs typically in CpG islands, which are GC-rich stretches of DNA (typically ~0.5–2 kb in length) that are often associated with house-keeping genes. It is possible that dispersed initiation is due to the presence of multiple weak core promoters. Although it is estimated that approximately one-third of human promoters contain CpG islands, the mechanism of dispersed transcription initiation remains to be clarified.

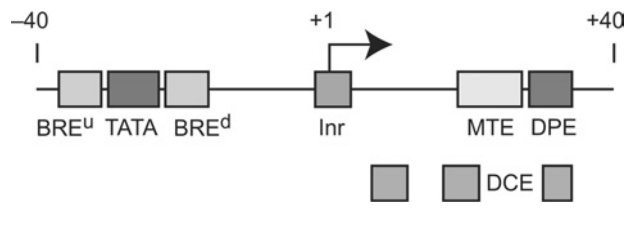
Key words: downstream core promoter element (DPE), initiator (Inr), motif ten element (MTE), RNA polymerase II, TATA box, transcription factor for RNA polymerase II (TFIIB).

Abbreviations used: TFIIB, transcription factor for RNA polymerase II; BRE, TFIIB recognition element; BRE^d, BRE downstream of the TATA box; BRE^u, BRE upstream of the TATA box; DCE, downstream core element; DPE, downstream core promoter element; GTF, general transcription factor; Inr, initiator; MTE, motif ten element; Sp1, specificity protein 1; TBP, TATA-box-binding protein; TAF, TBP-associated factor.

¹To whom correspondence should be addressed (email jkadonaga@ucsd.edu).

Figure 1 | Some core promoter motifs for RNA polymerase II

It is important to note that each of these motifs is present in only a subset of core promoters. DCE is a tripartite element. It is likely that there are other core promoter motifs that remain to be discovered. This Figure is roughly to scale.



In the remainder of this paper, we will describe studies of focused transcription initiation, which has been investigated more thoroughly than dispersed initiation. In simpler organisms, such as *Drosophila* or yeast, transcription appears to occur primarily via focused initiation. In the future, however, it will be very important to understand the factors and mechanisms that are employed in dispersed transcription initiation.

General/basal transcription factors

By itself, purified RNA polymerase II is incapable of promoter recognition and accurate transcription initiation. The polymerase requires additional auxiliary factors, which are commonly known as the GTFs (general transcription factors) or 'basal transcription factors', to initiate transcription (for a review, see [5]). The GTFs that mediate TATA-dependent transcription have been purified and identified. These factors are termed TFIIA (transcription factor for RNA polymerase II A), TFIIB, TFIID, TFIIE, TFIIF and TFIIH. TFIID is a multisubunit protein complex that contains TBP (TATA-box-binding protein) as well as over ten TAFs (TBP-associated factors). TBP as well as some of the TAFs are involved in core promoter recognition. TFIIB also interacts with core promoter motifs known as BREs (TFIIB recognition elements). It is important to note that the 'general' transcription factors do not function generally at all core promoters. For instance, the TATA-based GTFs will not mediate transcription from DPE-dependent core promoters [6]. It is likely that other 'general'/basal transcription factors exist.

Core promoter elements

A typical core promoter encompasses the RNA start site and extends approx. 35 nt upstream or downstream of the +1 site. Thus core promoters are usually approx. 40–50 nt in length, although they are sometimes as long as 70–80 nt. Core promoter activity is conferred by the presence of sequence motifs that recruit GTFs such as TFIID and TFIIB. These core promoter motifs are located at specific positions relative to the +1 transcription start site, as depicted in Figure 1.

The first-identified and best known core promoter element is the TATA box, which was discovered in the course of

sequencing the histone genes in *Drosophila* [7]. The Inr motif was identified by comparison of sequences that encompass the transcription start site [8] and then functionally defined as a discrete core promoter element [9]. The DPE was discovered through the analysis of the binding of purified TFIID to TATA-less genes [10,11]. The MTE (motif ten element) was found by a combination of computational and functional studies [12,13]. In addition, the DCE (downstream core element) was identified in the analysis of the β -globin gene [14,15]. The TATA, Inr, MTE, DPE and DCE are all recognition sites of the multisubunit TFIID complex. Thus TFIID is a key factor involved in core promoter binding. There are also two BREs, BRE^u and BRE^d (BREs that are upstream and downstream of the TATA box respectively) [16,17]. As noted above, the BRE motifs interact with the TFIIB basal transcription factor.

Some properties of core promoter motifs are as follows. The TATA box promotes transcription either by itself or in conjunction with other core promoter elements, such as the Inr. Thus the TATA box alone can confer core promoter activity. The addition of an Inr to a TATA box can sometimes focus transcription initiation to a site (or sites) within the Inr. The Inr encompasses the transcription start site and appears to be the most commonly occurring core promoter element. The Inr can weakly direct transcription by itself, but is usually found with other core promoter elements. The DPE acts co-operatively with the Inr. Likewise, the MTE also acts co-operatively with the Inr. Neither the DPE nor the MTE exhibits core promoter activity in the absence of an Inr. Most DPE- and MTE-containing core promoters lack a TATA box, and vice versa. There are, however, core promoters that contain both DPE and TATA motifs. In addition, synergy has been observed between the MTE and the TATA box as well as between the MTE and the DPE. The BRE^u and BRE^d motifs are extensions of the TATA box. Depending on the context, the BRE motifs have been found to have either a positive or a negative effect on transcriptional activity [16–18]. It is also interesting to note that the Inr can function with sequence-specific activators, such as Sp1 (specificity protein 1) [19]. Given that CpG islands usually contain multiple Sp1-binding sites, it is possible that dispersed transcription initiation in TATA-less CpG islands could occur via the Sp1 sites acting in conjunction with Inr or weak Inr-like sequences.

There are no universal core promoter elements. Rather, each of these motifs is found only in a subset of core promoters. For instance, although the TATA box is well known, recent studies suggest that it is present in only approx. 10–15% of human core promoters [20–22]. In addition, it is likely that many other core promoter motifs will be discovered in the future.

What are the distinguishing features of core promoter elements?

It is reasonable to ask what is the difference between a core promoter element and any other DNA sequence that affects

transcriptional activity. Some of the following guidelines might be useful to determine whether a DNA sequence motif is a core promoter element. First, it is important to identify and to characterize any putative core promoter motif in several different core promoters, that is, it is necessary to show that the sequence is indeed a 'motif' (i.e. a recurring sequence) rather than an isolated sequence. Secondly, the DNA sequence should be important for transcriptional activity. This property could be tested simply by mutational analysis. It is useful, in particular, to make a series of mutations that 'scan' across the putative core promoter element. In this manner, it can be determined whether the motif is a discrete element. That is, mutations inside the motif should affect transcription, whereas mutations that flank the motif should be neutral (unless a flanking sequence included another core promoter motif). Thirdly, if the DNA motif is a recognition site for a basal/general transcription factor, such as TFIID or TFIIB, then it is likely to be a core promoter element. In addition, it would be important to demonstrate, such as by mutational analysis, that binding of the basal/general factor to the motif affects core promoter activity. Fourthly, core promoter motifs tend to function optimally at a specific position relative to the transcription start site (or to the Inr motif). For instance, there is a strict spacing requirement for the positioning of the MTE and DPE motifs relative to the Inr. An increase or decrease of only a single nucleotide in the Inr-MTE or Inr-DPE spacing causes a severalfold decrease in core promoter activity [13,23]. The TATA box also exhibits a distinct yet slightly less rigid preference with regard to its location relative to the start site. (Specifically, the upstream 'T' in the TATA box is most commonly located at position -31 relative to the 'A' in the Inr consensus sequence.) Thus, if the potential core promoter motif exhibits a preferred location relative to the RNA start site, then it would seem more likely that the motif is involved in the basal transcription process. Fifthly, it is also useful to see if the putative core promoter motif can increase transcriptional activity when added to a separate core promoter. For example, the DPE was found to be able to restore activity that was lost upon mutation of the TATA box [10]. In addition, the MTE can restore activity that is lost upon mutation of the DPE or the TATA box [13]. These experiments show a gain of function upon addition of the motif, and complement the loss of function that is seen upon mutation of the motif. Lastly, it is important to note that these guidelines may not necessarily apply to all core promoter elements. Rather, these issues are considerations that could be tested in the study of new, putative core promoter motifs.

The core promoter is a regulatory element

The diversity in core promoter structure is mirrored by diversity in core promoter function. Different core promoters exhibit different requirements for basal/general transcription factors [6,24]. In addition, some transcriptional enhancers function specifically with core promoters that contain DPE or TATA motifs [25,26]. These experiments reveal that the core

promoter is not only important for the specification of the transcription start site but is also a regulatory element. This issue has been discussed extensively in other recent reviews [1-4]. For the purposes of this paper, it would be useful for the reader to appreciate the diversity in core promoter structure and activity, and to design and to interpret experiments with that knowledge. For instance, the study of transcriptional enhancers typically involves the use of standard reporter genes with a TATA-dependent core promoter. Such reporter genes would not, however, detect the presence or activity of a DPE-specific enhancer. Hence, in the study of transcriptional enhancers, it would be best to use their cognate core promoters.

Conclusions

The core promoter is a fascinating component of the transcription process, and has a diversity of structure and function that was largely unanticipated. There are new factors and mechanisms involved in basal transcription that need to be identified. In addition, the basis for the core promoter specificity of enhancers remains to be elucidated. It is also likely that additional core promoter motifs have yet to be discovered. There will, in the future, be many interesting and important avenues of research to be explored in the study of the RNA polymerase II core promoter.

The study of basal RNA polymerase II transcription in the laboratory of J.T.K. is supported by a grant from the National Institutes of Health (Bethesda, MD, U.S.A.; GM41249).

References

- Smale, S.T. (2001) *Genes Dev.* **15**, 2503-2508
- Butler, J.E.F. and Kadonaga, J.T. (2002) *Genes Dev.* **16**, 2583-2592
- Smale, S.T. and Kadonaga, J.T. (2003) *Annu. Rev. Biochem.* **72**, 449-479
- Kadonaga, J.T. (2002) *Exp. Mol. Med.* **34**, 259-264
- Orphanides, G., Lagrange, T. and Reinberg, D. (1996) *Genes Dev.* **10**, 2657-2683
- Lewis, B.A., Sims, 3rd, R.J., Lane, W.S. and Reinberg, D. (2005) *Mol. Cell* **18**, 471-481
- Goldberg, M.L. (1979) Ph.D. Thesis, Stanford University, Stanford, CA, U.S.A.
- Corden, J., Wasylyk, B., Buchwalder, A., Sassone-Corsi, P., Keding, C. and Chambon, P. (1980) *Science* **209**, 1405-1414
- Smale, S.T. and Baltimore, D. (1989) *Cell* **57**, 103-113
- Burke, T.W. and Kadonaga, J.T. (1996) *Genes Dev.* **10**, 711-724
- Burke, T.W. and Kadonaga, J.T. (1997) *Genes Dev.* **11**, 3020-3031
- Ohler, U., Liao, G., Niemann, H. and Rubin, G.M. (2002) *Genome Biology* **3**, research0087.1-0087.12
- Lim, C.Y., Santoso, B., Boulay, T., Dong, E., Ohler, U. and Kadonaga, J.T. (2004) *Genes Dev.* **18**, 1606-1617
- Lewis, B.A., Kim, T.-K. and Orkin, S.H. (2000) *Proc. Natl. Acad. Sci. U.S.A.* **97**, 7172-7177
- Lee, D.H., Gershenson, N., Gupta, M., Ioshikhes, I.P., Reinberg, D. and Lewis, B.A. (2005) *Mol. Cell. Biol.* **25**, 9674-9686
- Lagrange, T., Kapanidis, A.N., Tang, H., Reinberg, D. and Ebright, R.H. (1998) *Genes Dev.* **12**, 34-44
- Deng, W. and Roberts, S.G.E. (2005) *Genes Dev.* **19**, 2418-2423
- Evans, R., Fairley, J.A. and Roberts, S.G.E. (2001) *Genes Dev.* **15**, 2945-2949
- Smale, S.T., Schmidt, M.C., Berk, A.J. and Baltimore, D. (1990) *Proc. Natl. Acad. Sci. U.S.A.* **87**, 4509-4513

- 20 Bajic, V.B., Choudhary, V. and Hock, C.K. (2004) *In Silico Biol.* **4**, 109–125
- 21 Gershenzon, N.I. and Ioshikhes, I.P. (2005) *Bioinformatics* **21**, 1295–1300
- 22 Kim, T.H., Barrera, L.O., Zheng, M., Qu, C., Singer, M.A., Richmond, T.A., Wu, Y., Green, R.D. and Ren, B.A (2005) *Nature* **436**, 876–880
- 23 Kutach, A.K. and Kadonaga, J.T. (2000) *Mol. Cell. Biol.* **20**, 4754–4764
- 24 Willy, P.J., Kobayashi, R. and Kadonaga, J.T. (2000) *Science* **290**, 982–984
- 25 Ohtsuki, S., Levine, M. and Cai, H.N. (1998) *Genes Dev.* **12**, 547–556
- 26 Butler, J.E.F. and Kadonaga, J.T. (2001) *Genes Dev.* **15**, 2515–2519
-

Received 21 July 2006