Biochimica et Biophysica Acta 1839 (2014) 118–128

Contents lists available at ScienceDirect

Biochimica et Biophysica Acta

journal homepage: www.elsevier.com/locate/bbagrm

Review

Chromatin and DNA sequences in defining promoters for transcription initiation☆

Ferenc Müller [b,*], Làszlò Tora [a,c,**]

[a] Cellular Signaling and Nuclear Dynamics Program, Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC), UMR 7104 CNRS, UdS, INSERM U964, BP 10142, F-67404 Illkirch Cedex, CU de Strasbourg, France
[b] School of Clinical and Experimental Medicine, College of Medical and Dental Sciences, University of Birmingham, B15 2TT Edgbaston, Birmingham, UK
[c] School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, 637551, Singapore

ARTICLE INFO

Article history:
Received 26 June 2013
Received in revised form 11 November 2013
Accepted 11 November 2013
Available online 22 November 2013

Keywords:
Epigenetic mark
Nucleosome depleted region
CpG island
Core promoter
RNA polymerase II
Transcription initiation

ABSTRACT

One of the key events in eukaryotic gene regulation and consequent transcription is the assembly of general transcription factors and RNA polymerase II into a functional pre-initiation complex at core promoters. An emerging view of complexity arising from a variety of promoter associated DNA motifs, their binding factors and recent discoveries in characterising promoter associated chromatin properties brings an old question back into the limelight: how is a promoter defined? In addition to position-dependent DNA sequence motifs, accumulating evidence suggests that several parallel acting mechanisms are involved in orchestrating a pattern marked by the state of chromatin and general transcription factor binding in preparation for defining transcription start sites. In this review we attempt to summarise these promoter features and discuss the available evidence pointing at their interactions in defining transcription initiation in developmental contexts. This article is part of a Special Issue entitled: Chromatin and epigenetic regulation of animal development.

1. Introduction

One of the key events in eukaryotic gene regulation and consequent transcription is the assembly of general transcription factors (called GTFs or TFIIs) and RNA polymerase II (Pol II) into a pre-initiation complex (PIC) at core promoters. Initially GTFs were purified biochemically from human HeLa cell or rat liver extracts and defined in vitro as a set of factors essential for accurate transcription initiation at a strong TATA box- and initiator element (Inr)-containing viral core promoter [1,2]. The Pol II core promoter is defined as the stretch of DNA from where the transcription of a Pol II transcribed RNA is started [3]. GTFs include TFIID, which is composed of TATA-binding protein (TBP) and 13 TBP-associated factors (TAFs) [4–6]. According to the textbook view, the binding and recognition of core promoter sequences by the canonical TFIID is the first step that nucleates PIC formation. These early in vitro experiments led to generalised models of Pol II transcription initiation that were biased by the nature of the strong viral promoters, the naked DNA templates used and the cellular extracts, prepared from highly differentiated cells. Subsequent functional and genetic studies carried out in model systems, which reflect various stages of development in different model organisms, revealed the existence of alternative initiation complexes that have been suggested to replace canonical TFIID ([7–9] and refs therein). The diversity of the core Pol II promoter binding machinery has also been coupled with the characterisation of many different core promoter types and chromatin architectures at promoters [3,10]. The diversities in core promoter binding factors, core promoter elements and the associated chromatin signatures argue for a yet unappreciated dynamic regulatory step in transcription that is central to cellular homeostasis and thus to many developmental processes.

The complexity arising from a variety of promoter-associated DNA motifs their binding factors and associated chromatin properties brings an old question to the fore once again: how is a promoter defined in the cells of an organism? This question has not been satisfactorily answered despite 30 years of investigation. Accumulating evidence suggests that several parallel acting mechanisms are involved in orchestrating a pattern marked by the state of chromatin and general transcription factor binding in preparation for directing transcription at predefined sites. These mechanisms can be associated with at least 3 sets of core promoter features in vertebrates, which together provide an integrated platform for transcription initiation: i) DNA sequence motifs in promoters, such as the TATA-box, remain important factors in recruitment of initiation complex-forming transcription factors, as suggested

☆ This article is part of a Special Issue entitled: Chromatin and epigenetic regulation of animal development.

* Correspondence to: F. Mueller, School of Clinical and Experimental Medicine, College of Medical and Dental Sciences, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK. Tel.: +44 12141 42895.
** Correspondence to: L. Tora, Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC), UMR 7104 CNRS, UdS, INSERM U964, BP 10142, F-67404 Illkirch Cedex, CU de Strasbourg, France. Tel.: +33 3 88 653444.
E-mail addresses: f.mueller@bham.ac.uk (F. Müller), laszlo@igbmc.fr (L. Tora).

1874-9399/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.bbagrm.2013.11.003

by the original textbook view; ii) a second promoter-associated property is a characteristic base composition at transcriptional start site regions, which may promote nucleosome-free regions or support nucleosome depletion by chromatin remodelling; and iii) thirdly, the enrichment of CG dinucleotides in many promoters appears to provide templates for epigenetic mechanisms involved in defining promoter regions such as DNA (hypo)methylation and deposition of promoter associated histone modifications, at least in mammals. In this review we attempt to summarise these promoter features and discuss the available evidence supporting their interactions in various developmental and other ontogenic contexts.

2. Positionally constrained sequence motifs involved in transcription initiation

2.1. Core promoter elements

Due to our still limited understanding of the exact molecular recognition mechanisms that lead to Pol II transcription initiation in vivo, prediction of eukaryotic Pol II promoters by sequence analysis remained uncertain and inaccurate. Pol II core promoters are recognised as modular in structure and many of them contain specific consensus sequence elements which are positionally constrained in relation to the transcription start site (TSS) [10–13]. The advancement of high-throughput-sequencing-based TSS mapping technologies, such as CAGE, Gro-cap, RAMPAGE and Cap seq allowed genome wide (GW) mapping of the transcription starting positions of mRNAs on a variety of eukaryotic genomes and thus, the very accurate identification of transcription start sites (TSSs) ([14–20] and refs therein). These GW mapping methods fostered generalised models to aid in understanding of eukaryotic Pol II transcription regulation [14]. Core promoters typically span between −50 and +50 relative to the +1 TSS. The modular structure of core promoters consists of multiple short sequence elements that can be dispersed or overlapping and surrounding the TSS. Their functions are mostly context-dependent ([12] and refs therein) and their organisation is often gene or pathway specific (i.e. see the ribosomal protein gene and translation-associated snoRNA gene promoter-associated TCT initiator in the next paragraph in this section). Although core promoters direct initiation of transcription, they exhibit different transcriptional initiation properties. Depending on the number of sites from where Pol II initiates transcription, core promoters can be divided in two main categories: “single” or “narrow” TSS-containing promoters, where Pol II transcription initiates at a single, or narrow cluster of base pairs, and “broad” TSS promoters, in which transcription is initiated at several seemingly independent sites within a window of about 50–100 bps [21,22]. Interestingly, genes with restricted spatial expression patterns often corresponding to tissue-specific structural genes tend to have a single TSS-containing promoter, while genes with ubiquitous spatial expression often have a broad TSS promoters [14,16] (see also below). Despite the fact that many eukaryotic core promoters contain specific sequence motifs, i.e. a TATA box and/or an Inr, there seemed to be no universal motifs, which would allow the unambiguous definition of a core promoter in a given eukaryotic genome. This view has recently been challenged by the Pugh lab by analysing the organisation of the of human TFIIB-containing initiation complexes genome-wide [23]. They reported the widespread existence of a human core promoter consensus ($SSRCGCCTATAWAWRNRTDKKKK(N)_{13}YYA_{+1}NWYY$) that could be built from 4 previously described consensus motifs: the upstream TFIIB responsive element (BRE_u; SSRCGCC) [24], the TATA box consensus (TATAWAWR) [25], the downstream BRE (BRE_d; RTDKKKK) [26] and the Inr ($YYA_{+1}NWYY$) [27], where the tolerance for mismatches in these elements would be 2–3–2–1, respectively [23]. Furthermore, this study detected about 8000 PICs at expressed coding gene promoters, and about 160,000 PICs genome-wide, but did not define whether the detected TFIIB-containing PICs were at single TSS-, or broad TSS-containing promoters. Note however, that TFIIB-recognition elements (BRE_u and BRE_d) can have positive or negative effects on transcription in a promoter context-dependent manner [28].

At present it is not well understood how and where PICs are assembled on “broad” TSS-containing core promoters and how Pol II initiates transcription from seemingly random, independent sites within a broad window. At a “broad” TSS-containing core promoter several scenarios are conceivable: i) canonical or non-canonical TFIID complexes bind specifically to several sites at the broad promoter region and together with TFIIB determine multiple positions for each PIC formation that could also vary from cell to cell, ii) PICs assemble always at the same position, but the diffuseness in start site selection on these promoters occurs downstream of PIC assembly, or iii) the size of the nucleosome free regions would be longer on these broad TSS-containing promoters than on sharp TSS-containing promotes allowing several less specific TFIID/PIC positioning and the consequent multiple starts of transcription (see below).

As several excellent recent reviews have described the characterisation and sequence composition of all the identified core promoter elements [12–14,29] such as the TATA box [30–32], the Inr [27], the DPE (downstream promoter element) [33], the MTE (motif ten element) [34], the Motif 1 and 6 elements [35], the BRE_u and BRE_d [24,26], the DRE (DNA replication-related-element) [35,36], the DCE (downstream core element) [37], the XCPE (X core promoter element) [38], the TCE (translational control element) [39], and the PB (the pause button) [40] (see also Fig. 1 in [14]), here we focus on just one and briefly describe the TCT initiator element, which demonstrates a specialised core promoter type adapted to a specialised biological process.

The canonical transcription initiation region is characterised either by a full Inr motif (YYA(+1)NWYY) [41], or by YR(+1) consensus [21] in the TSS of polyadenylated transcripts of mammals. In contrast, a conserved but divergent sequence motif was described in a number of vertebrate genes including all of the ribosomal proteins [42,43]. One of the main features of the promoters of this gene class was a cytosine +1 at the dominant TSS surrounded by a tract of 4 to 13 pyrimidines, occasionally interrupted by one or two guanosine residues. This motif was termed as polypyrimidine initiator [42]. Later this initiator (also named as TCT) was further analysed functionally in *Drosophila* [44–46]. The TCT motif coincides with and replaces the canonical initiator at the TSS in nearly all *Drosophila* ribosomal protein genes, as well as in at least 48 human ribosomal protein genes [42,46]. In addition, the core promoters of some genes encoding translation initiation and elongation factors also contain a TCT element. In the TCT motif, which spans from −2 to +6 relative to the +1 transcription start site, pyrimidine nucleotides encompass the C+1 start site. This is a distinct type of initiation from that of the canonical, where Pol II has a very strong preference for A/G +1 start sites. Furthermore, in vitro foot-printing experiments demonstrated that the TCT element is not recognised by the canonical TFIID [46]. Thus, it appears that the TCT motif-containing promoters recruit a specialised PIC distinct from TFIID to achieve high levels of expression of ribosomal protein genes as well as to coordinate the relative amounts of corresponding translation initiation and elongation factors. This example demonstrates how a divergent and specialised sequence motif together with a yet unknown initiation complex coevolved for a specialised transcriptional function.

2.2. Differential usage of core promoter elements

Alternative core promoter usage in a developmental stage-specific manner is a means for the utilisation of alternative transcription initiation mechanisms for individual genes active during development [8,13]. For example, transcripts expressed in the oocyte in *Drosophila* are deposited and maternally inherited in the embryo and have been shown to be transcribed from alternative promoters, selectively utilised in development [47]. Core promoters of genes expressing maternally inherited transcripts showed differences in motif composition when compared to zygotically active promoters. For example, DRE, Ohler 1 and 6 elements are enriched