Mammalian Genome 10, 168–175 (1999).

Mammalian Genome
Incorporating Mouse Genome

Models for prediction and recognition of eukaryotic promoters

Thomas Werner

GSF-National Research Center for Environment and Health, Institute of Mammalian Genetics, Ingolstädter Landstraβe 1, D-85764 Neuherberg, and Genomatix Software GmbH, Karlstraβe 55, D-80333 München, Germany

Received: 21 July 1998 / Accepted: 7 October 1998

Introduction

The enormous impact of various genome sequencing projects is bringing the importance of computer-assisted nucleotide sequence analysis to the attention of a constantly increasing number of scientists. Gene prediction is undoubtedly leading the list of important tasks in this context. It is also generally accepted that this task consists of the recognition of the exon/intron structure of the coding region as well as prediction of the corresponding promoter. However, there is much less agreement about the exact sequences that should be called a promoter.

In general, the promoter is an integral part of the gene and often makes sense only in the context of its own gene, especially if important parts of the regulation are determined outside of the promoter (for example, by an intron enhancer; Stamatoyannopoulos et al. 1997). The function of a promoter is to mediate and control initiation of transcription of that part of a gene that is located immediately downstream of the promoter (3′). This can be achieved either in an unregulated permanent manner (constitutive transcription) or in a highly regulated fashion by which transcription is subjected to the control of various extracellular and intracellular signals (regulated transcription). The DNA region required to fulfill this function can be determined by assays for promoter function in a heterologous context. Unfortunately, this simple scheme becomes blurred in the case of highly regulated promoters. Often complex regulation involves many more features than just the promoter; for example, enhancers, locus control regions (LCRs), and/or scaffold/matrix attachment regions (S/MARs, reviewed in Boulikas 1996). If any of these units, which are functionally completely different from promoters, happens to be located adjacent to the promoter, delineation of the promoter becomes difficult. This may be one of the reasons why promoter prediction programs almost exclusively focus on proximal promoter regions or even just on the core promoter. Therefore, I will refer to a promoter mainly as the region that is necessary to achieve transcriptional initiation, although this region may not be sufficient to determine the complete regulation of a gene.

What is the basic aim of computer-assisted promoter recognition? The most obvious answer is location of an important part of the regulatory region of a gene. However, promoter prediction can also be very useful in the context of gene prediction. The promoter by definition marks the beginning of the first exon of a gene, which is often difficult to predict, especially if the first exon is not translated or very short (sometimes even more than one promoter/first exon exists). The promoter regions also contain information complementary to the exons and introns because transcriptional regulation—which can play an important part in gene function—cannot be deduced from the predicted amino acid sequence. The promoter, once understood, may even yield first clues towards the function of a completely anonymous protein, for example, if the promoter is known to be tissue or cell specific. Prediction of the functionality of a promoter would also be welcome for gene therapy approaches to improve expression of newly created vector constructs.

Correspondence to: Neuherberg, Germany

Transcriptional promoters and their functions

All attempts to predict promoters necessarily include a model of how a promoter is organized or what the hallmarks of a promoter are. The strong as well as the weak points of different methods for promoter analysis/prediction are determined to a large extent by the accuracy of the underlying promoter model with respect to the biological organization and the quality of the training data used for modeling. Therefore, a description of the biological principles leading to the special structures of promoters will be necessary to understand the models. The functional setup of promoters is also intimately coupled with the basic events of transcriptional initiation. Therefore, I will briefly summarize the key events of transcriptional initiation before discussing the basic structure of a promoter. The focus throughout this review will be on polymerase II responsive promoters that are controlling the vast majority of cellular genes encoding proteins. Therefore, they are of special interest and, consequently, are the main targets for most of the programs available at present.

Key events in transcriptional initiation. Transcription can proceed only after a competent transcription complex consisting of RNA polymerase II (pol II) and several general transcription factors (GTFs) have been recruited to the promoter (Sauer and Tijan 1997). Usually, pol II is not capable of functionally interacting with a promoter and initiating transcription by itself and requires a host of cofactors to do so. This can be seen as a safeguard against unscheduled transcription, which would be disastrous for a cell. Basically, there are two different phases in transcriptional activation of a gene. The first step includes a variety of transcription factors (TFs) that bind to upstream promoter and enhancer sequences to form a multiprotein complex (Fig. 1a and b). In the second step, this complex directly or indirectly recruits a pol II complexed with some GTFs to the core promoter and the transcription start site (TSS) located within the core promoter. Subsequently, transcription is initiated by this initiation complex (more detailed in Fig. 1c), which itself is subject to regulatory influences of TFs.

Two types of cofactors are involved in transcriptional initiation, transcriptional accessory factors (TAFs) and general transcription factors (GTFs). The TAFs are involved in the TFIID complex, which binds to the TATA box via the TATA box binding protein (TBP). TFIID is involved in the transcription of most pol II promoters. Pol II also has an absolute requirement for at least two GTFs, called TFIIE and TFIIH, without which it cannot clear the promoter for elongation (Zawel and Reinberg 1995). RAP74, which is the large subunit of the GTF TFIIF, was recently shown to induce conformational changes in the pol II promoter complex,

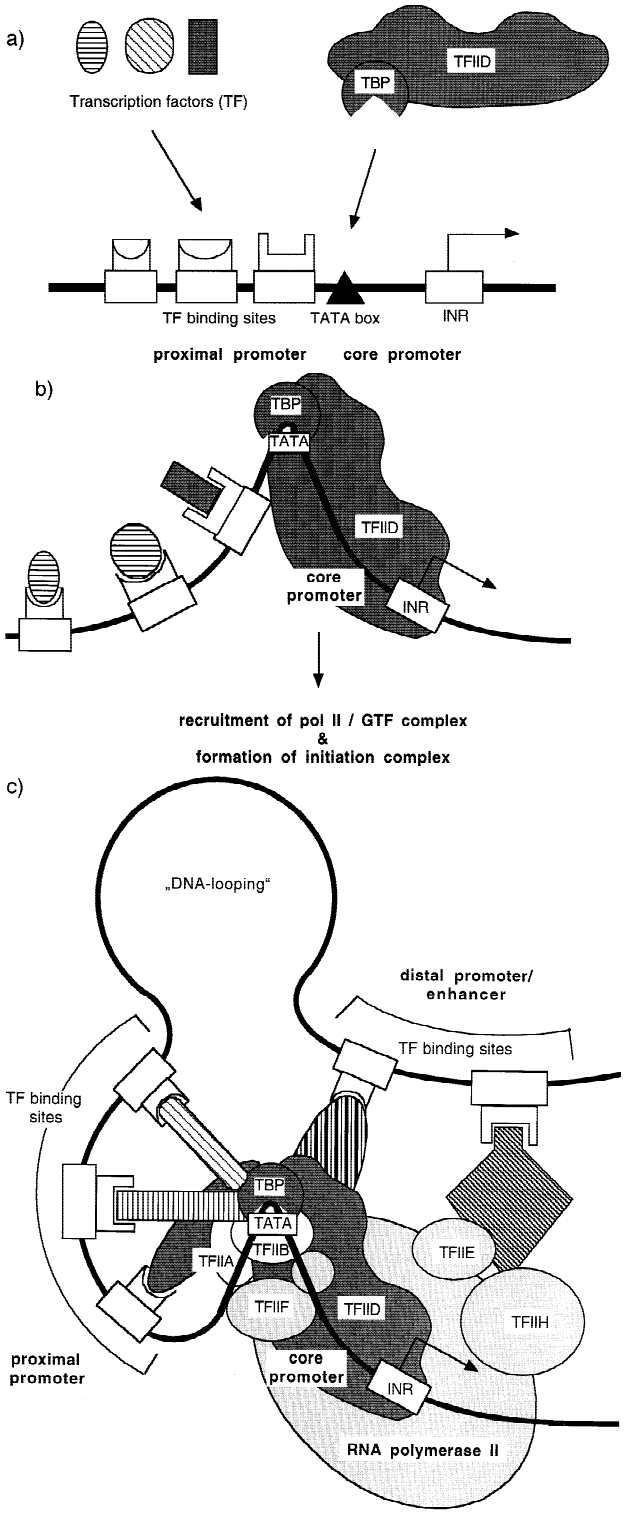

Fig. 1. Assembly of the activator/promoter complex on the proximal and core promoter region. **a)** Schematic representation of the proximal promoter with three specific transcription factor (TF) binding sites and the core promoter represented by the TATA box (black triangle) and the initiator region (INR). The transcription start site (TSS) is indicated by the angled arrow. **b)** Binding of the TFs and the TFIID complex (including the TATA box binding protein TBP). TBP binding induces a 90° bend in the promoter DNA. **c)** Subsequently the polymerase II/GTF complex is loaded to yield the complete initiation complex.

which most likely brings the polymerase into closer contact with the promoter region (Forget et al. 1997).

In addition to these GTFs, there is a multitude of transcriptional activator and repressor proteins (TFs) involved in transcriptional regulation. Only specific subsets of these factors are bound to promoters of individual genes. These proteins interact either directly with TAFs or form a ternary complex with a TAF. They may also contact one or more transcriptional mediators (for example, see Korzus et al. 1998; Montminy 1997) that do not bind DNA on their own but establish protein-protein links (Horvai et al. 1997). Figure 1 is a schematic representation of the cascade of events leading to the formation of the initiation complex prior to transcriptional initiation.

Once the complete complex including TFs, TAFs, GTFs, and pol II is assembled on the promoter, this is called the initiation complex, which is now competent to initiate RNA synthesis. I would like to call the minimum region of DNA allowing formation of a functional initiation complex the promoter. This can be reduced to a core promoter or may include one or several upstream and/or downstream elements. I am well aware that this is an arbitrary definition and that such a promoter can be determined exactly only by laboratory experimental research. Other regions directly upstream to this minimum promoter will be called upstream regulatory regions as long as they cannot be clearly identified as separate functional units (for example, enhancers). There is a gradual transition from upstream promoter elements to enhancer elements (position- and orientation-independent activator regions), which is not necessarily obvious from inspecting the sequence. Basically, the same is true for elements located downstream of the minimum promoter, which also may be downstream promoter elements as well as part of enhancers.

Basic structure of a polymerase II promoter. The structure of a pol II promoter can be seen as a composition of several regions with different functions. To start from inside out in terms of function, a promoter must contain a transcription start site (TSS), often located inside a so-called initiator region (INR), and one or several essential binding sites for GTFs, which may also be located downstream of the TSS in some cases (downstream elements). One of the most prominent GTF binding sites is the TATA box recognized by TBP, which is part of the TFIID complex (Sauer and Tijan 1997). The region of the promoter that is sufficient to determine the precise transcription start site will be referred to a the core promoter. The minimum promoter is the region that is capable of initiating basal transcription and may include a few more sites located close to the TATA box or the TSS. The region immediately 5′-adjacent to the minimum promoter constitutes the proximal promoter, which usually extends about 200–300 nucleotides upstream of the TSS. This region is responsible for the modulation of transcription, at least in part. The promoter CCAAT box is an example of a relatively common upstream TF binding site located in the proximal part of the promoter. This part may be extended further upstream (that is, in 5′-direction) by the distal promoter sequences. Figure 2 shows the general organization of a schematic pol II promoter.

The distal part of a promoter is also the most variable one with respect to composition as well as length. Binding sites for virtually all known transcription factors can be found in the distal regions of promoters, which may encompass anything from 100 nucleotides (as in the avian C-type LTR promoter; Ruddell 1995) to more than 2 kb, as was shown for some sea urchin promoters (Kirchhamer et al. 1996). There is no clear-cut defined 5′-boundary for promoters. The only difference between distal promoter and enhancer sequences is the positional and orientation independence of the enhancer function (Rippe et al. 1995; Thompson and McKnight 1992). This again is not obvious from the sequence analysis alone,

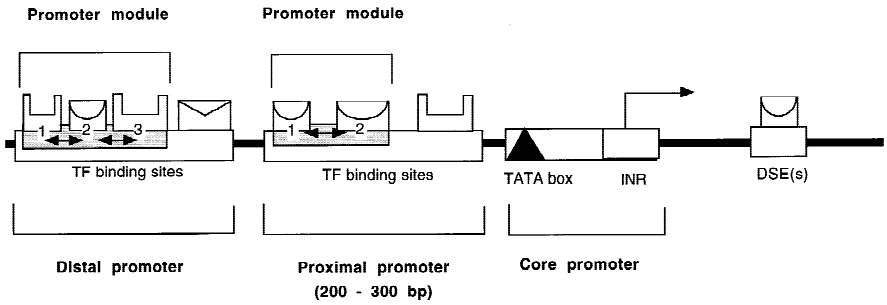

Fig. 2. Schematic structure of a polymerase II promoter. Boxes above the indicated promoter regions indicate TF binding sites. The transcription start site (TSS) is indicated by an angled arrow. Two examples of promoter modules are indicated by gray shading in the distal and the proximal promoter region. Numbers and arrows symbolize spatial and organizational restrictions within the modules.

because the internal organization of an enhancer is not fundamentally different from that of distal promoter sequences.

In addition to these features, specific DNA or RNA structural elements, like intrinsically curved DNA, direct or inverted repeat elements, may also influence the formation of the initiation complex.

Modular organization in promoter structures. The TF binding sites within a promoter (or the upstream regulatory sequences) do not show any obvious patterns with respect to location and orientation within the promoter sequences. Apparently, binding sites for any specific factor may occur almost anywhere in a promoter. For example, functional AP-1 binding sites can be located far upstream, as in the rat bone sialoprotein gene, where an AP-1 site located about 900 nucleotides upstream of the TSS suppresses expression (Yamauchi et al. 1996). An AP-1 site located close to the TSS plays an important role in the expression of Moloney murine leukemia virus (Sap et al. 1989). Functional AP-1 sites have also been described inside exon 1 (downstream of the TSS) of the propiomelanocortin gene (Boutillier et al. 1995), as well as within the first intron of the *fra-1* gene (Bergers et al. 1995). Similar examples can be found for many other TF binding sites, illustrating that transcription factor binding sites do not show any general correlation with specific promoter regions. To summarize this, TF binding sites can be found virtually everywhere in promoters, but not in every promoter. A closer look reveals that the particular function of an AP-1 binding site (for example, activating or repressing) often critically depends on the relative location and especially on the context of the binding site. For example, the AP-1 site in the above-mentioned rat bone sialoprotein gene overlaps with a set of glucocorticoid responsive element (GRE) half sites, and this overlap is crucial for the suppressive function.

As a consequence of context requirements, TF sites are often grouped together, and such functional groups have been described in many cases. A systematic attempt to collect synergistic or antagonistic pairs of TF binding sites is the COMPEL database (Kel et al. 1995; Heinemeyer et al. 1998). In many cases, a specific promoter function (for example, a tissue-specific silencer) will require more than two sites. Such promoter subunits are called modules, and I will refer to groups of TF binding sites that carry a specific function independent of the promoter context (and consequently are transferable) as *promoter modules.* A more detailed definition of promoter modules has been given recently by Arnone and Davidson (1997). The organization of binding sites (and probably also of other promoter elements) of a promoter module appears to be much more restricted than the apparent variety of TF sites and their distribution in the whole promoter suggests. Within a promoter module, both sequential order and distance can be crucial for function, indicating that these modules may be the critical determinants of a promoter rather than individual binding sites. Since promoters can contain several modules that may use overlapping sets of binding sites, the conserved context of a particular binding site is detectable only if the corresponding module can be identified. Below, I will illustrate this modular concept on a very well-known general promoter module, the core promoter. However, the basic principles of module organization are also true for most, if not all other promoter modules and are neither peculiar nor restricted to the core promoter.

Structure and function of the core promoter module. The core promoter module can functionally be defined by its capability to assemble the transcription initiation complex and orient it specifically towards the TSS of the promoter (Zawel and Reinberg 1995). This can be achieved by various combinations of about four distinguishable core promoter elements, which constitute in principle a core promoter module as shown on top in Fig. 3. This module includes the TATA box, the initiator region (INR), an upstream activating element, and a downstream element. However, this is also where the straightforward definition of a core promoter module ends, because not all four elements are always required.

Successful positioning of the initiation complex can be initiated on TATA box containing promoters by the TFIID complex, leading via TFIIB to the assembly of an initiation complex (Conaway and Conaway 1991). If an appropriate upstream TF binding site cooperates with the TATA box, no special initiator or downstream sequences might be required, which allows assembly of a functional core promoter module from just two of the four elements (Fig. 3a). This represents one type of a distinct core promoter that contains a TATA box, common among cellular genes in general.

However, as is known from a host of TATA-less promoters, the TATA box is by no means an essential element of a functional core promoter. An INR combined with a single upstream element has also been shown to be capable of specifically initiating transcription (Gilinger and Alwine 1993; Fig. 3b), and a remarkable array of four different upstream TF sites (SP1, AP1, ATF, or TEF1) was shown to confer inducibility by T-antigen to this very simple promoter. This is an example of a TATA-less distinct promoter that can be found in several genes from the hematopoietic lineage.

The third combination is called a composite promoter and consists of both a TATA box and an initiator. This combination can be found in several viral promoters (Fig. 3c), and it has been shown that an additional upstream TF binding site can influence whether the TATA box or the initiator element will be determining the promoter properties (Colgan and Manley 1995). The authors showed that upstream elements can significantly increase the efficiency of the INR in this combination, while especially SP-1 sites made the TATA box almost obsolete in their example. The combination of TATA box with an INR had the general effect of inducing resistance against the detrimental effects of a TFIIB mutant, which interfered with expression from TATA-only promoters. This is also an example of the more indirect effects of specific arrangements in promoters that may not be apparent unless special conditions occur.

The last group are so-called Null-promoters, which have neither a TATA box nor an initiator and rely exclusively on upstream and downstream elements (Novina and Roy 1996; Fig. 3d).

Basically, at least the four different core promoter types de-

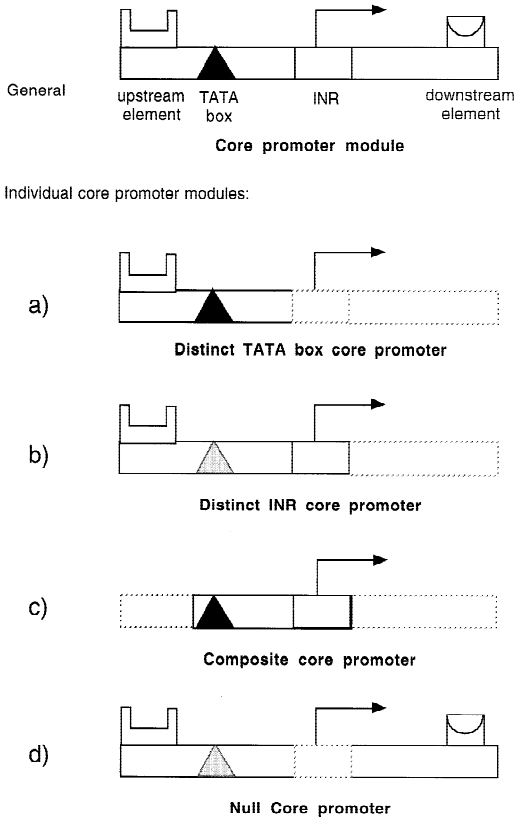

Fig. 3. Four possible arrangements for a functional core promoter module. On top, a schematic complete core promoter module is shown. **(a–d)** Stippled regions and sites within the individual core promoter modules are not recognizable by sequence analysis and/or not crucial for the particular core promoter function, which is determined by the elements shown as solid boxes.

tailed above have been identified so far; all represent valid combinations of core promoter sites (reviewed in Novina and Roy 1996; Fig. 3 a–d). If the combinations involving upstream and downstream elements are also considered, seven possible core promoter modules are possible (most of which can be actually found in genes).

The only apparent common denominator would be that there must be *at least one (or more)* core promoter elements within a certain region. This assumption is wrong: If there are at least two elements present both spacing and/or sequential order within the core promoter module is of utmost importance regardless of the presence or absence of individual elements (as a rule, there appear to be some exceptions). Moreover, many distinct promoters have requirements for specific upstream or downstream elements and will not function with any other TF. Moving around the initiator, the TATA box and to some extent also upstream elements can have profound effects on promoter functions. For example, insertion of just a few nucleotides between the TATA box and an upstream TF binding site (MyoD) in the desmin promoter cuts the expression levels by more than half (Li and Capatanaki 1994). Moreover, the promoter structure can affect later stages of gene expression such as splicing (Cramer et al. 1997). It was also shown for the rat beta-actin promoter that a few mutations around the transcription start site (that is, within the initiator) can render that gene subject to translational control (Biberman and Meyuhas 1997).

As a final note, even the concept of one general TATA box and one general INR is an oversimplification. There are several clearly distinguishable TATA boxes in different promoter classes (Frech et al. 1996, 1997a, 1998; Frech and Werner 1996a), and the same is true for the INR region, which also has several functionally distinct implementations as the glucocorticoid-responsive INR in the murine thymidine kinase gene (Rhee and Thompson 1996), the C/EBP binding INR in the hepatic growth factor gene promoter (Jiang and Zarnegar 1997), or the YY-1 binding INR (Usheva and Shenk 1996).

Most of the principles of variability and restrictions detailed above for the core promoter module are also true for other promoter modules. The bottom line is that the vast majority of alternative arrangements of the elements that can be seen in a particular promoter might not contribute to the function of the promoter. Module-induced restrictions are not necessarily obvious from the primary sequences. Figure 1c shows a schematic pol II promoter with the initiation complex assembled, which illustrates that it matters where a specific protein is bound to the DNA in order to allow proper assembly of the molecular jigsaw puzzle of the initiation complex. This is not immediately obvious from inspection of promoter sequences, because there exist several (but strictly limited) alternative solutions to the assembly problem. As complicated as Figure 1c may appear, it still ignores all aspects of chromatin rearrangements and nucleosomal positions, which also play an important role in transcriptional regulation. An example of the profound influence of these effects on promoter-protein complex assembly and function has been detailed for the ostecalcin promoter in a study by Stein and associates (1995). However, chromatin-related effects are not yet considered in any of the promoter prediction methods. Therefore, I cannot go into any more details here.

Problems, basic tools, and resources for promoter analysis

The determination of the full extent of the transcript (5′ and 3′ extremities of the primary RNA) and the location of promoter regions are still unreliable (see also Claverie 1997). What is the difference from the exon predictions, which are apparently much more reliable and are being regularly used already? One very important difference is lack of continuous sequence similarities in promoters, in contrast to the open reading frames present in coding exons, which allow successful application of alignment programs. Also, in most cases promoter sequences of homologous genes (for example, of human and mouse origin) are overall much more divergent than the corresponding coding regions. In some cases homology is still strong enough to allow identification of homologous promoters (Duret and Bucher 1997). In other cases similarity searches are no longer able to identify promoters of the same class (Frech et al. 1998). Another peculiarity is that many promoters are regulated by one or several signaling pathways. Specific requirements for responsiveness to diverse stimuli add to structural and organizational divergence of the promoters, vividly evident from the highly complex organization of some sea urchin promoters, which have been studied in great detail (Kirchhamer et al. 1996). Even promoters regulated in the same pathway can be completely different in sequence, because there is often more than one promoter structure to achieve similar expression patterns. For example, several completely distinct promoter organizations are found in muscle specifically expressed genes involving modules of different sets of transcription factor binding sites (reviewed in Firulli and Olson 1997).

On the other hand, many promoter features like TF binding sites can be found in a wide variety of promoters and do not convey any specificity or function of their own. This may also be the reason why such binding sites can be readily detected throughout the genome and not only in promoters. Nevertheless, transcription factor binding sites are crucial promoter elements, and methods to detect such sites have been developed and used for more than a decade now.

Table 1. Internet accessible methods to detect promoter elements (transcription factor binding sites).

Program	Availability	Comments
Matlnspector	http://www.gsf.de/cgi-bin/matsearch.pl http://genomatix.gsf.de/cgi-bin/matinspector/matinspector.pl	Matinspector matrix library (includes TRANSFAC[a] matrices)
SIGNAL SCAN	http://bimas.dcrt.nih.gov/molbio/signal/	IUPAC consensus library based on TFD[b]
MATRIX SEARCH	http://bimas.dcrt.nih.gov/molbio/matrixs/	IMD matrix library (TRANSFAC+TFD)
TFSearch	http://pdap1.trc.rwcp.or.jp/research/db/TFSEARCH.html	TRANSFAC* matrices
TESS	http://agave.humgen.upenn.edu/utess/tess/	TRANSFAC* matrices

[a] TRANSFAC: Heinemeyer et al. 1998.
[b] TFD: Gosh 1993.

Detection of transcription factor binding sites (TF sites). The majority of the known transcription factors recognize short DNA stretches of about 10–15 nucleotides in length which show different degrees of internal variation. This has been accounted for by the derivation of IUPAC consensus sequences, which indicate the predominant nucleotide or nucleotide combination at each position in a set of example sequences.

The concept of nucleotide weight matrix (NWM) descriptions has been developed in the 1980s to overcome shortcomings of IUPAC strings (for example, Staden 1984; Stormo and Hartzel 1989). However, weight matrices require predefined matrices, which are more complicated to construct than IUPAC (Nagase et al. 1998; Naitou et al. 1997) strings and do require specific software tools. This delayed widespread use of weight matrices for almost a decade since the methods were principally available. They remained mostly unused because only a few special matrices had been defined (Bucher 1990). In 1995 two (overlapping) matrix libraries for TF sites were compiled for the first time and became widely available almost simultaneously (Quandt et al. 1995; Chen et al. 1995). They still represent the only libraries of their kind and are being used in several tools, some of which are available through the world-wide-web (WWW, see Table 1). Available methods for TF binding site definition and detection have been surveyed recently (Frech et al. 1997b), and an extensive comparison of their capabilities was also published (Frech et al. 1997c). For convenience, Table 1 summarizes which methods are available in the internet, with emphasis on programs featuring a WWW-interface.

Other promoter elements. Besides TF sites, there exist a couple of other individual elements or sequence properties that are associated with promoter sequences. Among these are higher GpC content (GpC islands; Shago and Giguere 1996), secondary structure elements like the HIV-1 TAR region (for example, Bohjanen et al. 1997), cruciform DNA structures (for example, Wang et al. 1998), or simple direct repeats (for example, Bell et al. 1997). Three-dimensional structures like curved DNA (Kim et al. 1995) also influence promoter function. Most of these elements can be detected by computer-assisted sequence analysis (Chetouani et al., 1997; Schuster et al. 1997; Nakaya et al. 1995; Nielsen et al. 1995) but none of them is really promoter specific and can be found frequently outside of promoters. The secret of promoter function lies in the combination of several promoter elements that need to cooperate in transcriptional activation, which none of them can achieve alone. This also summarizes the main problem of promoter recognition. It is necessary to compile several individually weak signals into a composite signal which then indicates a potential promoter.

Promoter prediction tools

There are several ways in which promoter recognition tools can be categorized. I will focus on the main principles and intended usage of the programs rather than technical details, as this will also be the main interest of experimentally working scientists. Two fundamentally different approaches have been pursued so far in order to achieve promoter recognition. The majority of programs focus on *general promoter recognition,* which represents the first category. One group of programs in this category concentrates on recognition of core promoter properties and infers promoter location solely on that basis, while the other group consists of programs that take into account also the proximal promoter region of about 250–300 nucleotides upstream of the TSS. General recognition models are usually based on training sets derived from the Eukaryotic Promoter Database (EPD; Perier et al. 1998) and various sets of non-promoter sequences. The beauty of these approaches is their generality, which does not require any specific knowledge about a particular promoter to make a prediction. This can be a big advantage if nothing is known about the promoter to be predicted. The bad news is limited specificity, that is, a huge burden of false positive predictions in longer sequences.

The second category of tools aims at *specific promoter recognition,* relying on combinations of individual elements. Here again, one group focuses on small modules usually containing just two elements, while another approach includes several elements from the proximal promoter region. The beauty of this approach is excellent specificity, which is extremely helpful if only promoters of a certain class are of interest. The bad news here is limited applicability, that is, each promoter class needs to have a predefined model available before sequences can be analyzed for these promoters.

I will not discuss individual methods, since an excellent practical test involving the majority of available tools based on general promoter models has been carried out recently (Fickett and Hatzigeorgiou 1997). This study concluded that none of these methods is clearly superior to its peers.

Definition and detection of specific regulatory modules

The modular organization of promoter functions has long been recognized from experimental work. However, until recently few attempts to include this knowledge into computer models of promoters were made, probably owing to missing methods and compilations of individual elements. Definition of promoter modules is greatly facilitated by specialized databases like the COMPEL database (Heinemeyer et al. 1998; Kel et al. 1995) or the TRRD (Heinemayer et al. 1998; Kel et al. 1994). The number of publications describing experimental analysis of regulatory modules also increased dramatically just in the last two years, which is best demonstrated by the fact that the highly acclaimed journal *Molecular and Cellular Biology* introduced a new regular section called transcription control.

Many ongoing high-throughput studies of expression profiles in various organisms provide another milestone for the successful implementation of promoter modules (Nagase et al. 1998; Naitou et al. 1997). Access to expression data for a large collection of

Table 2. Available promoter/promoter regions prediction tools.

Program	Availability	Comments
Promoter prediction WWW-accessible		
FunSiteP	http://transfac.gbf.de/dbsearch/funsitep/fsp.html	includes proximal promoter
NNPP	http://www-hgc.lbl.gov/projects/promoter.html	core promoter
PromFD	http://beagle.colorado.edu/~chenq/Hypertexts/ PromFD.ht ml	includes proximal promoter
	chenq@beagle.colorado.edu	for further information
Promoter Scan	http://biosci.umn.edu/software/proscan/ promoterscan.htm	includes proximal promoter
TSSG/TSSW	http://dot.imgen.bcm.tmc.edu:9331/ gene-finder/gf.html	includes proximal promoter triplet & hexamer frequencies
	Download/inquire	
Audic/Claverie	audic@newton.cnrs-mrs.fr	Markov models
PromFind	ftp://iubio.bio.indiana.edu/molbio/ibmpc http://www.rabbithutch.com	hexamer frequencies for further information
XGRAIL	ftp://arthur.epm.ornl.gov/pub/xgrail	core promoter in gene context
Promoter module/ region recognition	***WWW-accessible***	
FastM	http://www.gsf.de/cgi-bin/fastm.pl http://genomatix.gsf.de/cgi-bin/fastm2/fastm.pl	module of 2 TF sites uses Matinspector library
TargetFinder	http://gcg.tigem.it/TargetFinder.html	module of 1 TF site combined with 1 annotated feature
	Download/inqure	
GenomeInspector	http://www.gsf.de/biodv/genomeinspector.html ftp://ariane.gsf.de/pub/unix/genomeinspector/	correlation analysis, e.g. 2 TF binding sites
Muscle-specific regions	ficketjw@molbio.sbphrd.com	contact for download
Xlandscape	ftp://beagle.colorado.edu/pub/Landscape/ xland.v1.tar.Z	word frequencies, not promoter specific

promoters allows selection of specific training sets which apparently share one or more functional modules manifested by a common expression pattern either with respect to spatial distribution or time course of transcription. Such preselected training sets will allow more wide-spread application of most of the approaches described below and should help to rapidly increase the number of descriptions of functional promoter modules.

The first attempts in this direction have to be credited to Claverie and Sauvaget, who published a method detecting modules of two distinct elements in a preset distance and orientation in 1985 and demonstrated that this concept can be useful for detection of heat-shock promoters (Claverie and Sauvaget 1985). This already includes the biological restrictions seen in promoter modules, which may be one reason why even simple module models can be very successful.

Claverie and Sauvaget had not much choice but to encode the sequences directly into their search patterns, since no compilations were available at that time. A more recent approach based on the same philosophy, FastM, was derived from the program ModelGenerator (Frech et al., 1997) and takes advantage of the existence of NWM libraries. It can be accessed via a WWW-interface and allows straightforward generation of any two TF binding site modules by simple selection from the MatInspector Library (Quandt et al. 1995). This now allows definition and detection of synergistic TF binding site pairs. FastM models can successfully identify promoters sharing such composite elements, but are not promoter specific. Composite elements can also be located in enhancers or similar structures.

Fickett also employed the idea of a two-TF binding site module to successfully detect a subclass of muscle-specific regulatory sequences governed by a combination of MEF2 and MyoD (Fickett 1996). However, this was also a very specific approach, and no general tool resulted from that work, but the MEF2/MyoD model can be successfully used with FastM. Wasserman and Fickett (1998) published a modeling approach based on clustering of a preselected set of NMW (defined in this study) correlated with muscle-specific gene expression, with which they were able to detect about 25% of muscle-specific regulatory regions in *genomic* sequences and more than 60% in their *test* set. They describe their method as a regulatory module detection. However, their results suggest that they probably detect a collection of different modules with respect to the definition given in this review. This is a very interesting approach, which has potential for further development and demonstrates that modules and their combinations are very suitable to describe functionally similar groups of promoter. The authors will make the non-commercial software used in their approach available on request (contact J. Fickett).

TargetFinder is another web-based search tool that combines a MatInspector search for TF sites with context information taken from the sequence annotations. TargetFinder relies on predefined annotations from GenBank or EMBL. The advantage is that TargetFinder basically also follows the same philosophy as the two module-detecting programs discussed above, but allows inclusion of features that have been annotated by experimental work for which no search algorithm exists. A collection of programs potentially useful for target gene detection has been reviewed recently (Lavorgna et al. 1998).

A method focusing on the detection of correlated elements also not specific for promoter analysis is GenomeInspector (Quandt et al. 1996a), which can use potential TF binding sites based on MatInspector searches (Quandt et al. 1995). The program has been shown to be able to locate organizational features within promoters (Quandt et al. 1996b). This tool should be useful to determine modules which then can be used to define search criteria for FastM or TargetFinder (see Table 2 for available programs).

So far only one system attempting specific definition and recognition of whole promoters, including proximal and probably some distal promoter regions, has been published and applied to several examples. The ModelGenerator system carries the clustering of promoters to the degree of functionally related promoters (defining promoter classes) and allows development of complex models recognizing promoters of a specific class with extremely high specificity (Frech et al. 1996, 1997, 1998). However, generation of these models also requires a highly preselected training set of promoters from the particular class, and the development of the final model is a lengthy process. ModelInspector is the corresponding search machine and uses predefined models to scan sequences. Provided a library exists, usage of ModelInspector is as simple as using the popular TF binding site search programs Signal Scan or MatInspector.

The unprecedented specificity of ModelGenerator models (about 1 match in 2.5 million base pairs for an action promoter model with no false negatives in mammalian sequences; Frech et al. 1998) is a consequence of the close resemblance of the computer model to the modular promoter organization. However, this close match is both the greatest advantage as well as the greatest obstacle of this method. The models are so specific that they do not recognize any other promoters but members of their own class (no overall sequence similarity required). This is good news if the sequence to be analyzed contains a promoter from a promoter class for which such a model is available. In this case, not only is a match reported, but simultaneously a lot about the functionality and inner structure of the promoter is revealed by the class assignment. However, the method cannot be applied if a predefined model does not exist for the promoter of interest. General promoter prediction is absolutely out of range for this approach until a sufficiently large library of promoter models is available, which is reminiscent of the situation with weight matrices about 10 years ago.

Conclusions

As the human and mouse genome sequencing projects enters the production mode, the fully automated annotation of megabase-long anonymous genomic sequences is the next big challenge in bioinformatics. There are two major demands to be satisfied with respect to promoter analysis, which none of the above-mentioned methods presently is able to deliver. First, there are lots of anonymous sequences to be analyzed; this necessitates the broad range of promoter recognition that general prediction tools offer. However, there are also mega- and soon gigabases of sequences to be analyzed, which necessitates the specificity reached by the ModelGenerator models in order to yield useful results.

As the practical tests of Fickett and Hatzigeorgiou (1997) clearly demonstrated, all general promoter prediction tools exhibit specificities and sensitivities in the same order of magnitude regardless of the particular algorithm. This is not surprising, because these methods by definition ignore the modular design of promoters, which is an essential part of biological specificity. Therefore, I assume that the general promoter prediction is already about as good as it gets, and I do not expect more than moderate increases in specificities. However, a meaningful analysis of really large sequences like whole chromosomes requires specificities about 2 to 3 orders of magnitude higher than that of the general methods. This is what specific promoter models have been shown to be principally capable of, suggesting a possible solution to the problem. Unfortunately, although considerable efforts are under way to generate the necessary promoter library, this will take some time and, as in case of the matrix libraries, will be able to cover only part of the promoter classes present.

Therefore, although such models might represent a gold standard with regard to specificity, existing and improved general promoter recognition programs will remain indispensable workhorses for systematic analysis of genomic sequences. Accepting that their specificities probably cannot be dramatically enhanced for the principal reasons detailed above, it will be necessary to divide sequences into shorter regions with which those methods can successfully cope.

The combination of general promoter prediction with exon/intron predictions also holds great potential for the improvement of promoter recognition and might be a way to escape the intrinsic lower specificity of general promoter prediction. Several groups have already recognized this, and projects are under way to implement such combinatorial approaches. I can also envisage synergistic combinations between specific promoter modeling and general prediction tools, whereby the specific models (or parts of them) could be employed to filter the original output of the general prediction programs.

I have no doubt that continuous improvements of general promoter prediction and various combinatorial approaches, together with growing libraries of highly specific models, will narrow the gap between the accuracy of promoter prediction and exon/intron prediction tools. After all, given the enormous amount of anonymous sequences to be expected and the available experimental capacities for functional analysis, I cannot see any alternative to computer-assisted evaluation of genomic sequences prior to experimental analysis.

Acknowledgments. I thank Rudolf Balling, Ruth Brack-Werner, Kornelie Frech, Kerstin Quandt, and Ralf Schneider for critically reading the manuscript. I also thank Christian Mirschberger for his help in preparation of the manuscript. This work was supported by EU grant BI04-CT95-0226 (TRADAT).

References

Arnone MI, Davidson EH (1997) The hardwiring of development: organization and function of genomic regulatory systems. Development 124, 1851–1864

Bell PJL, Higgins VJ, Dawes IW, Bissinger PH (1997) Tandemly repeated 147 bp elements cause structural and functional variation in divergent MAL promoters of *Saccharomyces cerevisiae.* Yeast 13, 1135–1144

Bergers G, Graninger P, Braselmann S, Wrighton C, Busslinger M (1995) Transcriptional activation of the fra-1 gene by AP-1 is mediated by regulatory sequences in the first intron. Mol Cell Biol 15, 3748–3758

Biberman Y, Meyuhas O (1997) Substitution of just five nucleotides at and around the transcription start site of rat beta-actin promoter is sufficient to render the resulting transcript a subject for translational control. FEBS Lett 405, 333–336

Bohjanen PR, Liu Y, GarciaBlanco MA (1997) TAR RNA decoys inhibit Tat-activated HIV-1 transcription after preinitiation complex formation. Nucleic Acids Res 25, 4481–4486

Boulikas T (1996) Common structural features of replication origins in all life forms. J Cell Biochem 60, 297–316

Boutillier AL, Monnier D, Lorang D, Lundblad JR, Roberts JL, et al. (1995) Corticotropin-releasing hormone stimulates propiomelanocortin transcription by cFos-dependent and -independent pathways: characterization of an AP1 site in exon 1. Mol Endocrinol 9, 745–755

Bucher P (1990) Weight matrix description of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. J Mol Biol 212, 563–578

Chen QK, Hertz GZ, Stormo GD (1995) MATRIX SEARCH 1.0: A computer program that scans DNA sequences for transcriptional elements using a database of weight matrices. Comp Appl Biosci 11, 563–566

Chetouani F, Monestie P, Thebault P, Gaspin C, Michot B (1997) ESSA: an integrated and interactive computer tool for analysing RNA secondary structure. Nucleic Acids Res 25, 3514–3522

Claverie JM (1997) Computational methods for the identification of genes in vertebrate genomic sequences. Hum Mol Genet 6, 1735–1744

Claverie J-M, Sauvaget I (1985) Assessing the biological significance of primary structure consensus patterns unsing sequence databanks. I. Heat-shock and glucocorticoid control elements in eukaryotic promoters. Comp Appl Biosci 2, 95–104

Colgan J, Manley JL (1995) Cooperation between core promoter elements influences transcriptional activity in vivo. Proc Natl Acad Sci USA 92, 1955–1959

Conaway JW, Conaway RC (1991) Initiation of eukaryotic messenger RNA synthesis. J Biol Chem 266, 17721–17724

Cramer P, Pesce CG, Baralle FE, Kornblihtt AR (1997) Functional association between promoter structure and transcript alternative splicing. Proc Natl Acad Sci USA 94, 11456–11460

Duret L., Bucher P (1997) Searching for regulatory elements in human noncoding sequences. Curr Opin Struct Biol 7, 399–406

Fickett JW (1996) Coordinate positioning of MEF2 and myogenin binding sites (reprinted from Gene-Combis, 172, GC19–GC32, 1996). Gene 172, GC19–GC32

Fickett JW, Hatzigeorgiou AC (1997) Eukaryotic promoter recognition. Genome Res 7, 861–878

Firulli AB, Olson EN (1997) Modular regulation of muscle gene transcription: a mechanism for muscle cell diversity. Trends Genet 13, 364–369

Forget D, Robert F, Grondin G, Burton ZF, Greenblatt J et al. (1997)

RAP74 induces promoter contacts by RNA polymerase II upstream and downstream of a DNA bend centered on the TATA box. Proc Natl Acad Sci USA 94, 7150–7155

Frech K, Werner T (1996) Specific modelling of regulatory units in DNA sequences. In *Pacific Symposium on Biocomputing '97*, RB Altman, AK Dunker, L Hunter, TE Klein, eds. (World Scientific), pp 151–162

Frech K, Brack-Werner R, Werner T (1996) Common modular structure of lentivirus LTRs. Virology 224, 256–267

Frech K, Danescu-Mayer J, Werner T (1997a) A novel method to develop highly specific models for regulatory units detects a new LTR in GenBank which contains a functional promoter. J Mol Biol 270, 674–687

Frech K, Quandt K, Werner T (1997b) Finding protein-binding sites in DNA sequences: the next generation. Trends Biochem Sci 22, 103–104

Frech K, Quandt K, Werner T (1997c) Software for the analysis of DNA sequence elements of transcription. Comp Appl Biosci 13, 89–97

Frech K, Quandt K, Werner T (1998) Muscle actin genes: a first step towards computational classification of tissue specific promoters. In Silico Biol 1, 0005

Ghosh D (1993) Status of the transcription factors database (TFD). Nucleic Acids Res 21, 3117–3118

Gilinger G, Alwine JC (1993) Transcriptional activation by simian virus-40 large T-antigen—requirements for simple promoter structures containing either TATA or initiator elements with variable upstream factor binding sites. J Virol 67, 6682–6688

Heinemeyer T, Wingender E, Reuter I, Hermjakob H, Kel AE et al. (1998) Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL. Nucleic Acids Res 26, 362–367

Horvai AE, Xu L, Korzus E, Brard G, Kalafus D, et al. (1997) Nuclear integration of JAK/STAT and Ras/AP-1 signaling by CBP and p300. Proc Natl Acad Sci USA 94, 1074–1079

Jiang JG, Zarnegar R (1997) A novel transcriptional regulatory region within the core promoter of the hepatocyte growth factor gene is responsible for its inducibility by cytokines via the C/EBP family of transcription factors. Mol Cell Biol 17, 5758–5770

Kel OV, Romachenko AG, Kel AE, Naumochkin AN, Kolchanov NA (1994) Structure of data representation in TRRD—database of transcription regulatory regions on eukaryotic genomes. Proceedings of the 28th Annual Hawaii International Conference on System Sciences [HICSS]; Biotechnology computing. (Los Alamitos, Calif.: IEE Computer Society Press), pp 42–51

Kel OV, Romaschenko AG, Kel AE, Wingender E, Kolchanov NA (1995) A compilation of composite regulatory elements affecting gene transcription in vertebrates. Nucleic Acids Res 23, 4097–4103

Kim J, Klooster S, Shapiro DJ (1995) Intrinsically bent DNA in a eukaryotic transcription factor recognition sequence potentiates transcription activation. J Biol Chem 270, 1282–1288

Kirchhamer CV, Yuh C-H, Davidson EH (1996) Modular cis-regulatory organization of developmentally expressed genes: two genes transcribed territorially in the sea urchin embryo, and additional examples. Proc Natl Acad Sci USA 93, 9322–9328

Korzus E, Torchia J, Rose DW, Xu L, Kurokawa R, et al. (1998) Transcription factor-specific requirements for coactivators and their acetyltransferase functions. Science 279, 703–707

Kraus RJ, Murray EE, Wiley SR, Zink NM, Loritz K, et al. (1996) Experimentally determined weight matrix definitions of the initiator and TBP binding site elements of promoters. Nucleic Acids Res 24, 1531–1539

Lavorgna G, Wagner A, Bonicelli E, Werner T (1998) Detection of potential target genes in silico?, Trends Genet 14, 375–376

Li H, Capatanaki Y (1994) An E box in the desmin promoter cooperates with the E-box and MEF-2 sites of a distal enhancer to direct muscle-specific transcription. EMBO J 13, 3580–3589

Montminy M (1997) Transcriptional activation—Something new to hang your HAT on. Nature 387, 654–655

Nagase T, Ishikawa K, Miyajima N, Tanaka A, Kotani H, et al. (1998) Prediction of the coding sequences of unidentified human genes. IX. The complete sequences of 100 new cDNA clones from brain which can code for large proteins in vitro. DNA Res 5, 31–39

Naitou M, Hagiwara H, Hanaoka F, Eki T, Murakami Y (1997) Expression profiles of transcripts from 126 open reading frames in the entire chromosome VI of *Saccharomyces cerevisiae* by systematic Northern analyses. Yeast 13, 1275–1290

Nakaya A, Yamamoto K, Yonezawa A (1995) RNA secondary structure prediction using highly parallel computers. Comp Appl Biosci 11, 685–692

Nielsen DA, Novoradovsky A, Goldman D (1995) SSCP primer design based on single-strand DNA structure predicted by a DNA folding program. Nucleic Acids Res 23, 2287–2291

Novina CD, Roy AL (1996) Core promoters and transcriptional control. Trends Genet 12, 351–355

Perier RC, Junier T, Bucher P (1998) The eukaryotic promoter database EPD. Nucleic Acids Res 26, 353–357

Quandt K, Frech K, Karas H, Wingender E, Werner T (1995) Matind and Matinspector: New fast and versatile tools for detection of consensus matches in nucleotide sequence data. Nucleic Acids Res 23, 4878–4884

Quandt K, Grote K, Werner T (1996a) GenomeInspector: a new approach to detect correlation patterns of elements on genomic sequences. Comp Appl Biosci 12, 405–413

Quandt K, Grote K, Werner T (1996b) GenomeInspector: Basic software tools for analysis of spatial correlations between genomic structures within megabase sequences. Genomics 33, 301–304

Rhee K, Thompson EA (1996) Glucocorticoid regulation of a transcription factor that binds an initiator-like element in the murine thymidine kinase (Tk-1) promoter. Mol Endocrinol 10, 1536–1548

Rippe K, Vonhippel PH, Langowski J (1995) Action at a distance: DNA-looping and initiation of transcription. Trends Biochem Sci 20, 500–506

Ruddell A (1995) Transcriptional regulatory elements of the avian retroviral long terminal repeat. Virology 206, 1–7

Sap J, Muñoz A, Schmitt J, Stunnenberg H, Vennström B (1989) Repression of transcription mediated at a thyroid homone response element by the v-erb-A oncogene product. Nature 340, 242–244

Sauer F, Tjian R (1997) Mechanisms of transcriptional activation: differences and similarities between yeast, Drosophila, and man. Curr Opin Genet Dev 7, 176–181

Schuster P, Stadler PF, Renner A (1997) RNA structures and folding: from conventional to new issues in structure predictions. Curr Opin Struct Biol 7, 229–235

Shago M, Giguere V (1996) Isolation of a novel retinoic acid-responsive gene by selection of genomic fragments derived from CpG-island-enriched DNA. Mol Cell Biol 16, 4337–4348

Staden R (1984) Computer methods to locate signals in nucleic acid sequences. Nucleic Acids Res 12, 505–519

Stamatoyannopoulos JA, Clegg CH, Li Q (1997) Sheltering of gamma-globin expression from position effects requires both an upstream locus control region and a regulatory element 3′ to the (A)gamma-globin gene. Mol Cell Biol 17, 240–247

Stein GS, Vanwijnen AJ, Stein J, Lian JB, Montecino M (1995) Contributions of nuclear architecture to transcriptional control. Int Rev Cytol 162A, 251–278

Stormo GD, Hartzell III GW (1989) Identifying protein-binding sites from unaligned DNA fragments. Proc Natl Acad Sci USA 86, 1183–1187

Thompson CC, McKnight SL (1992) Anatomy of an enhancer. Trends Genet 8, 232–236

Usheva A, Shenk T (1996) YY1 transcriptional initiator: protein interactions and association with a DNA site containing unpaired strands. Proc Natl Acad Sci USA 93, 13571–13576

Wang WD, Chi TH, Xue YT, Zhou S, Kuo A, et al. (1998) Architectural DNA binding by a high-mobility-group/kinesin-like subunit in mammalian SWI/SNF-related complexes. Proc Natl Acad Sci USA 95, 492–498

Wassermann WW, Fickett JW (1998) Identification of regulatory regions which confer muscle-specific gene expression. J Mol Biol 278, 167–181

Yamauchi M, Ogata Y, Kim RH, Li JJ, Freedman LP et al. (1996) AP-1 regulation of the rat bone sialoprotein gene transcription is mediated through a TPA response element within a glucocorticoid response unit in the gene promoter. Matrix Biol 15, 119–130

Zawel L, Reinberg D (1995) Common themes in assembly and function of eukaryotic transcription complexes. Annu Rev Biochem 64, 533–561