The Biology of Eukaryotic Promoter Prediction—a Review

Anders Gorm Pedersen[1†], Pierre Baldi[2], Yves Chauvin[2] and Søren Brunak[1]

[1]Center for Biological Sequence Analysis

Dept. of Biotechnology

The Technical University of Denmark

Building 208, DK-2800 Lyngby, Denmark

[2]Net-ID, Inc.

4225 Via Arbolada, suite 500

Los Angeles, CA 90042, USA

[†]Corresponding author. Tel.: (+45) 45 25 24 84, fax: (+45) 45 93 15 85,

email: gorm@cbs.dtu.dk

Abstract

Computational prediction of eukaryotic promoters from the nucleotide sequence is one of the most attractive problems in sequence analysis today, but it is also a very difficult one. Thus, current methods predict in the order of one promoter per kilobase in human DNA, while the average distance between functional promoters has been estimated to be in the range of 30-40 kilobases. Although it is conceivable that some of these predicted promoters correspond to cryptic initiation sites that are used *in vivo*, it is likely that most are false positives. This suggests that it is important to carefully reconsider the biological data that forms the basis of current algorithms, and we here present a review of data that may be useful in this regard. The review covers the following topics: (1) basal transcription and core promoters, (2) activated transcription and transcription factor binding sites, (3) CpG islands and DNA methylation, (4) chromosomal structure and nucleosome modification, and (5) chromosomal domains and domain boundaries. We discuss the possible lessons that may be learned, especially with respect to the wealth of information about epigenetic regulation of transcription that has been appearing in recent years.

Keywords: TATA-box, Initiator, nucleosomes, transcriptional initiation, genome analysis.

Introduction

Today, much attention in computational biology is focused on gene finding, *i.e.*, the prediction of gene location and gene products from experimentally uncharacterized DNA sequences. In this context, it is possible to use the prediction of promoter sequences and transcriptional start points as a "signal"—by knowing the position of a promoter one knows at least the approximate start of the transcript, thus delineating one end of the gene. This information is particularly helpful in connection with gene finding in DNA sequences from higher eukaryotes, where coding regions are present as small islands in a sea of non-coding DNA. However, the problem of predicting promoters is certainly also interesting in its own right. Thus, transcriptional initiation is the first step in gene expression, and generally constitutes the most important point of control. Through the elaborate mechanisms governing this process, specific genes can be turned on in a highly defined manner both spatially and temporally, as revealed for instance through the investigation of development in the fruit fly *Drosophila melanogaster* (Small *et al.*, 1991). It is by specifically turning on or off the transcription of sets of genes that cell types are determined in multicellular organisms. Transcriptional control also has direct implications for human health, since improper regulation of the transcription of genes involved in cell growth is one of the major causes of all forms of cancer.

Several different algorithms for the prediction of promoters, transcriptional start points, and transcription factor binding sites in eukaryotic DNA sequence now exist (Bucher *et al.*, 1996; Fickett & Hatzigeorgiou,

1997, Table 1). Although current algorithms perform much better than the earlier attempts, it is probably fair to say that performance is still far from satisfactory. Thus, the general picture is that when promoter prediction algorithms are used under conditions where they find a reasonable percentage of promoters, then the amount of falsely predicted promoters (false positives) is far too high. Existing methods predict in the order of one promoter per kilobase, while it is estimated that the human genome on average only contains one gene per 30-40 kilobases (Antequera & Bird, 1993). Promoter prediction algorithms have recently been thoroughly reviewed by Fickett and Hatzigeorgiou (1997; also see Bucher *et al*., 1996) with regard to both function and relative performance, and we will not attempt to repeat that work here but refer interested readers to these papers. Rather, we will review biological data that we consider to be relevant for computational biologists involved in construction of promoter finding algorithms, and also make suggestions for how this information may be put to use. We will emphasize the concept of epigenetic regulation, (*i.e.*, the "modulation of gene expression achieved by mechanisms superimposed upon that conferred by primary DNA sequence" (Gasser *et al.*, 1998)), since there is now a wealth of evidence that such mechanisms play an important role in all transcriptional control (*E.g.*, see (Simpson, 1991; Eden & Cedar, 1994; Paranjape *et al.*, 1994; Geyer, 1997; Gottesfeld & Forbes, 1997; Grunstein, 1997; Tsukiyama & Wu, 1997; Cavalli & Paro, 1998; Gasser *et al.*, 1998; Gregory & Hörz, 1998; Kellum & Elgin, 1998; Lu & Eissenberg, 1998)).

Basal transcription and core promoters

A core promoter is a binding site for RNA polymerase and general transcription factors

Eukaryotes have three different RNA-polymerases that are responsible for transcribing different subsets of genes: RNA-polymerase I transcribes genes encoding ribosomal RNA, RNA-polymerase II (which we will focus on in this review) transcribes genes encoding mRNA and certain small nuclear RNAs, while RNA-polymerase III transcribes genes encoding tRNAs and other small RNAs (Huet *et al.*, 1982; Breant *et al.*, 1983; Allison *et al.*, 1985).

RNA polymerase II (pol-II) consists of more than ten subunits, some of which are partly homologous to the α, β, and β' subunits of the RNA polymerase in *Escherichia coli*(Allison *et al.*, 1985). The eukaryotic pol-II enzyme is not in itself capable of specific transcriptional initiation *in vitro*, but needs to be supplemented with a set of so-called general transcription factors (GTFs) (Wasylyk, 1988; Zawel & Reinberg, 1993; Orphanides *et al.*, 1996). The most important of these factors are TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH. The GTFs are named after the order in which they were purified and discovered. Together, these factors, most of which consist of multiple subunits, contain approximately 30 polypeptides. Only when the polymerase and all the general transcription factors (with the possible exception of TFIIA) are assembled

on a piece of DNA transcription can commence. Although the assembly of this so-called pre-initiation complex was originally described to occur in an ordered stepwise manner (Orphanides *et al.*, 1996), newer experiments suggest that transcriptional initiation *in vivo* normally involves binding of a holoenzyme complex including pol-II and many or all of the GTFs in a single step (Greenblatt, 1997). The minimal pol-II promoter is typically defined as the set of sequences that is sufficient for assembly of such a pre-initiation complex, and for exactly specifying the point of transcriptional initiation *in vitro* (Fassler & Gussin, 1996, Figure 1). Transcription that is initiated by this minimal set of proteins is referred to as basal transcription.

More than one class of minimal promoters exist

Is it, then, this minimal set of sequences required for basal transcription that computational biologists should strive to recognize? One problem with this concept is that, in fact, there is not just one class of minimal promoters that have the above characteristics (Smale, 1994a). One important class of minimal (or core) promoters only consists of a TATA-box, which directs transcriptional initiation at a position about 30 bp downstream. The TATA-box has the consensus TATAAAA which appears to be conserved between most eukaryotes although several mismatches are allowed (Hahn *et al.*, 1989; Singer *et al.*, 1990; Wobbe & Struhl, 1990). It is bound by a subunit of TFIID known as TBP (the TATA binding protein) (Hernandez, 1993; Burley & Roeder, 1996). It has been found that TBP is present in the pre-initiation complexes with all three RNA polymerases (Hernandez, 1993; Burley & Roeder, 1996; Lee & Young, 1998) although new *in vitro* data suggest that under some specific circumstances it may be possible to initiate transcription without TBP (Wieczorek *et al.*, 1998), or with a TBP related factor (TRF) (Buratowski, 1997; Hansen *et al.*, 1997). In addition to TBP, TFIID also contains a number of TBP associated factors (TAFs) (Tanese & Tjian, 1993; Goodrich & Tjian, 1994; Tansey & Herr, 1997).

A second class of minimal promoters do not contain any TATA box (and are therefore referred to as TATA-less). In these promoters, the exact position of the transcriptional start point may instead be controlled by another basic element known as the initiator (Inr) (Smale, 1994a; Smale, 1997). The Inr is positioned so that it surrounds the transcription start point, and has the (rather loose) consensus PyPyAN[TA]PyPy, where Py is a pyrimidine (C or T), and N is any nucleotide (Bucher, 1990; Smale, 1994b). The first A is at the transcription start site, and the pyrimidine positioned just upstream of that is often cytosine. Promoters containing only an Inr are typically somewhat weaker than TATA-containing promoters (Smale, 1997). Interestingly, TBP has been shown also to participate in transcription initiated on TATA-less promoters (Smale, 1997). In addition to the two promoter-classes mentioned above, there are also promoters which have both TATA and Inr elements, and promoters that have neither (Smale, 1994a).

Another promoter element, which was recently discovered in both human and *Drosophila*, is present in

some TATA-less, Inr-containing promoters about 30 bp *downstream* of the transcriptional start point (Burke & Kadonaga, 1997). This element, which is known as the downstream promoter element (DPE), appears to be a downstream analog of the TATA box in that it assists the Inr in controlling precise transcriptional initiation.

Core promoters and promoter prediction

What does all this mean for promoter prediction? First of all, it is obvious that there is not one single type of core promoter. Instead, several combinations of at least three small elements are capable of directing assembly of the pre-initiation complex and sustaining basal transcription *in vitro*. Secondly, it appears that sequence downstream may also affect promoter function. This is important to keep in mind, since some promoter prediction methods focus on the upstream part of the promoter only. Thirdly, it is obvious that under any circumstance the elements described above cannot be the only determinants of promoter function. For instance, a sequence conforming perfectly to the Inr consensus will appear purely by chance approximately once every 512 bp in random sequence. Furthermore, in one study it was found that applying Buchers TATA box weight matrix (Bucher, 1990) to a set of mammalian non-promoter DNA sequences, resulted in an average of one predicted TATA box every 120 bp (Prestridge & Burks, 1993). Although this may in part be caused by the somewhat simplified description of a binding site that is implicit in a position weight matrix, it is nevertheless clear that there are far more perfect TATA-boxes in, for instance, the human genome than there are promoters. Since there is good evidence that such promiscuous transcriptional initiation does not take place *in vivo* (at least not to this degree), it is obvious that there are other important factors besides the core promoter elements. This is further supported by the fact that basal transcription (defined as transcription initiated by only pol-II and the GTFs on a minimal promoter) is practically non-existent *in vivo* (Paranjape *et al.*, 1994; Kornberg & Lorch, 1995; Fassler & Gussin, 1996; Gottesfeld & Forbes, 1997; Grunstein, 1997)—instead it is mainly an *in vitro* phenomenon that is useful for the analysis of the basal transcriptional machinery. But what is the reason for this—why are cryptic core promoters, which must exist in abundance in any genome, not utilized in the living cell? At least part of the explanation appears to rely on the three-dimensional structure of DNA in the nucleus, which has been found to have a generally repressing effect on transcription (Paranjape *et al.*, 1994; Kornberg & Lorch, 1995; Gottesfeld & Forbes, 1997; Grunstein, 1997). We will return to chromosomal structure and the implications for promoter activity and promoter prediction below. First, however, we will consider the phenomenon of transcriptional activation.

Activated transcription and regulatory sequences

Control of transcription by transcription factors binding to regulatory elements

It has been found that in order to sustain transcription *in vivo*, a core promoter needs additional short regulatory elements (Paranjape *et al.*, 1994; Kornberg & Lorch, 1995; Gottesfeld & Forbes, 1997; Grunstein, 1997). These elements are located at varying distances from the transcriptional start point. Thus, some regulatory elements (so-called proximal elements) are adjacent to the core promoter, while other elements can be positioned several kilobases upstream or downstream of the promoter (so-called enhancers). Both types of elements are binding sites for proteins (transcription factors) that increase the level of transcription from core promoters (Wasylyk, 1988; Johnson & McKnight, 1989; Mitchell & Tjian, 1989; McKnight & Yamamoto, 1992; Zawel & Reinberg, 1993; Fassler & Gussin, 1996). This phenomenon is referred to as activated transcription. Proteins that repress transcription by binding to similar DNA elements also exist. Unlike the small number of GTFs, there are several thousands of different transcription factors able to bind to regulatory elements. In fact, it has been estimated that factors involved in transcriptional regulation make up several percent of the proteins encoded in the vertebrate genome.

One very important way transcription factors achieve transcriptional activation is by recruitment of the basal transcriptional machinery to the promoter through protein-protein interactions, either directly or through adaptor proteins (Pugh & Tjian, 1990; Tanese & Tjian, 1993; Stargell & Struhl, 1996; Ptashne & Gann, 1997). In the case of binding sites located far from the promoter, it is believed that the protein-protein interactions involve looping out of the intervening DNA (Adhya, 1989; Matthews, 1992).

Regulatory regions, controlling the transcription of eukaryotic genes, typically contain several transcription factor binding sites strung out over a large region. Some of these individual binding sites are able to bind several different members of a family of transcription factors, or perhaps different dimeric complexes of related monomers. Which particular factor that binds to a given site, therefore, not only relies on the binding site, but also on what factors are available for binding in a given cell type at a given time. It is by the modular and combinatorial nature of transcriptional regulator regions that it is possible to precisely control the temporal and spatial expression patterns of the tens of thousands of genes present in higher eukaryotes (Schirm *et al.*, 1987; Dynan, 1989; Diamond *et al.*, 1990; Lamb & McKnight, 1991; McKnight & Yamamoto, 1992). Thus, any given gene will typically have its very own pattern of binding sites for transcriptional activators and repressors ensuring that the gene is only transcribed in the proper cell type(s) and at the proper time during development. Other genes are expressed only in response to extracellular stimuli such as for instance blood sugar level or viral infection, while still others are expressed more or less constitutively in most cell types. The latter class of genes include those encoding proteins involved in basal

metabolism, and are sometimes referred to as "house-keeping genes". Transcription factors themselves are, of course, also subject to similar transcriptional regulation, thereby forming transcriptional cascades and feed-back control loops. Striking and beautiful examples of the complexity of transcriptional regulation include, for instance, the *Drosophila even-skipped* gene (Small *et al.*, 1991; Jackle & Sauer, 1993), and the human β-globin gene (Evans *et al.*, 1990; Minie *et al.*, 1992; Crossley & Orkin, 1993; Higgs & Wood, 1993).

Transcription factor binding sites and promoter prediction

While this is all very nice and interesting from a biologists point of view, it seems to spell big trouble for promoter prediction. Not only are there thousands of transcriptional regulators, many of which have recognition sequences that are not yet characterized, but any given sequence element might be recognized by different factors in different cell types. Alternatively, a perfect consensus binding site near a promoter might never be bound because the corresponding factor is not present under the set of circumstances where the gene is transcribed. As in the case of core promoters, the fact that the regulatory elements are short and not completely conserved in sequence furthermore means that similar elements will be found purely by chance all over the genome. In accordance with this qualitative evaluation, statistical analysis also indicates that the density of (currently known) regulatory elements does not contain sufficient information to discriminate between promoters and non-promoters (Prestridge & Burks, 1993; Zhang, 1998a).

Although this may seem disheartening, it is important to remember that in the cell, after all, promoters *are* recognized correctly by the transcriptional apparatus. In some form the necessary information must therefore be present. But what, then, is the reason for the poor performance? Firstly, one fundamental reason may be that in most computational approaches, promoters are being searched for in single stranded sequence. The transcription apparatus, however, is designed to deal with chromatin, not single stranded DNA, as template. (This situation is very different from a search for, say, the correct exon-intron organization in a gene, where the biological object being processed by the splicing machinery is in fact single stranded pre-mRNA.) In a promoter prediction algorithm all this boils down to a proper way of representing the essential symmetries in double stranded versus single stranded sequence. It is possible that the use of strand-invariant encodings of DNA sequences may be helpful in this regard (Baldi *et al.*, 1998). One example where such symmetries are known to be important, is in connection with transcription factor binding sites that are functional in both orientations. As we suggested above, another possibility is that perhaps it is necessary to reconsider the nature of the questions we are asking. Is the problem in the form "discriminate between all promoters and all non-promoters" possibly too large for any single algorithm? Perhaps specific sub-classes of promoters need specific sub-algorithms that look for specific signals in order to be detected, and the

bigger problem can then only be solved by piecing together many such methods? The prediction of muscle-specific promoters is one step in this direction (Fickett, 1996a; Fickett, 1996b; Wasserman & Fickett, 1998). Of course, it is also conceivable that there is enough information in the binding sites, but that we simply have not yet figured out how to properly integrate the signals. Another possibility is that there is some element of the transcriptional regulation that cannot be directly deduced from the DNA sequence, but instead relies on mechanisms operating at other levels. In fact, it is well known that higher-level chromatin folding represses general transcription, and that unfolding of DNA through the action of histone acetyltransferases is important for transcriptional activation (Paranjape *et al.*, 1994; Kornberg & Lorch, 1995; Grunstein, 1997). We will return to possible ways of attacking this particular problem below.

Under all circumstances, it seems like a good idea to look for additional signals (besides transcription factor binding sites) that are correlated with the presence of transcription starts. Of course, the fact that a signal is helpful in determining whether a promoter is genuine, does not necessarily mean that the signal is involved in transcriptional regulation. Thus, the presence of a downstream coding region is helpful in identifying promoters in bacteria, but the coding region generally has no influence on the transcriptional process. Rather, it is the evolutionary history of bacteria that has resulted in very compact genomes where promoters are located in short intergenic regions. A feature that may be correlated with promoters in vertebrate genomes for similarly historical reasons, are the so-called CpG islands that we will discuss in the next section.

DNA methylation and CpG islands

Most CG dinucleotides in vertebrate genomes are methylated

In approximately 98% of the vertebrate genome the self-complimentary dinucleotide CpG is normally methylated at the 5 position on the cytosine ring (Bird, 1993; Bird *et al.*, 1995). (The p in CpG denotes the phosphodiester linkage). This is in contrast to the situation in non-vertebrate multicellular eukaryotes where methylation is either absent or confined to a small fraction of the genome. The vertebrate methylation pattern is established early in embryogenesis and is inherited by daughter cells after cell division. A specific cytosine methyltransferase is responsible for this by acting only on newly replicated CpG dinucleotides that are base-paired to an already methylated CpG (Holliday, 1993). Interestingly, CpG dinucleotides have been found to be present in vertebrate genomes much less frequently than would be expected from the mononucleotide frequencies. Specifically, the level of CpG is about 25% of that expected from base composition. This depletion is believed to be a result of accidental mutations by deamination of 5-methylcytosine to thymine. Since the product of this mutation (thymine) is indistinguishable from endogenous nucleotides it

cannot be recognized by DNA repair systems, and over evolutionary time CpG dinucleotides will therefore tend to mutate to TpG (Coulondre *et al.*, 1978; Bird, 1980; Jones *et al.*, 1992).

Methylated DNA is transcriptionally repressed

The functional importance of DNA methylation *in vivo* has been demonstrated by targeted disruption of the gene for cytosine methyltransferase in mice (Li *et al.*, 1992). Mutant mouse embryos display significantly lower levels of DNA methylation and die in mid-gestation. It is believed that this phenotype is related to the fact that DNA methylation in vertebrates has been found to have a repressing effect on transcriptional initiation, possibly mediated by the binding of a specific methyl-CpG binding protein (Boyes & Bird, 1992; Eden & Cedar, 1994). It is likely to be side-effects from the lack of methylation-mediated repression that cause mutant mouse embryos to die. Based on these observations it has been suggested that general DNA methylation may have evolved as a way of reducing background noise transcription, and that it made the subsequent development of the complex vertebrate lineage possible (Bird, 1993).

CpG islands are unmethylated regions that often overlap the 5' end of genes

In the context of promoter prediction, however, it is the unmethylated 2% of the genome that is of main interest. It has been found that vertebrate genomes contain CpG islands—regions about 1–2 kilobases in length where the dinucleotide CpG is present at the expected frequency and in unmethylated form (Gardiner-Garden & Frommer, 1987; Aïssani & Bernardi, 1991a; Antequera & Bird, 1993; Craig & Bickmore, 1994; Cross & Bird, 1995). Interestingly, the locations of these islands are almost always coincident with the 5' end of genes, often overlapping the first exon. Specifically, it has been estimated that about 56% of all human genes (*i.e.*, about 45,000) are associated with a CpG island (Antequera & Bird, 1993). In the same study it was estimated that the human genome contains approximately 22,000 house-keeping genes, all of which are associated with a CpG island, and about 58,000 tissue-restricted genes, of which approximately 40% (23,000) are associated with CpG islands (Antequera & Bird, 1993). It is not entirely clear how the demethylated status of these regions is maintained, but it has been shown that in some cases it is dependent on binding of the transcription factor Sp1 (Macleod *et al.*, 1994).

It has been suggested that CpG islands are in fact evolutionary remnants of the deamination event mentioned above (Antequera & Bird, 1993; Cross & Bird, 1995). According to this hypothesis, most promoters have somehow been kept methylation-free, and have therefore retained the original level of CpG dinucleotides. Hence, some promoters now stand out as obvious CpG islands compared to the surrounding regions of CpG-depleted DNA. If this is correct, then it appears that the methylation-free state has been maintained more strongly in house-keeping promoters than in tissue-restricted ditto.

CpG islands and promoter prediction

The correlation between CpG islands and promoters may therefore be historical rather than functional, but it is nevertheless likely to be useful in connection with promoter prediction. In fact, this might be just the kind of global signal we mentioned above. There is currently no publicly available promoter finding software that utilize this correlation, but Junier, Krogh, and Bucher are currently developing such an algorithm that combines a search for CpG islands with an HMM-based detection of core promoter elements (Anders Krogh, personal communication). Furthermore, CpG island detection can be performed using a feature in the WebGene server (Milanesi & Rogozin, In press, Table 1) and is also available as a feature in the GRAIL gene finder, although it is not currently used as a signal for promoter finding in that method (Matis *et al.*, 1996; Uberbacher *et al.*, 1996, Table 1). All these methods define CpG islands according to Gardiner-Garden and Frommer (1987). According to this definition, a CpG island is a region that (1) is more than 200 bp long, (2) has more than 50% G+C (*i.e.*, $p_G + p_C > 0.5$), and (3) has a CpG dinucleotide frequency that is at least 0.6 of that expected on the basis of the nucleotide content of the region (*i.e.*, $p_{CpG} > 0.6 \times p_C \times p_G$). However, as mentioned above the phenomenon is only useful for analysis of vertebrate genomes, meaning that we will under all circumstances have to employ alternative methods in connection with the non-vertebrate multicellular eukaryotic model organisms (*e.g, Drosophila*, *Dictyostelium* and *C. elegans*) currently being sequenced.

It has been noted that, at least in warm-blooded vertebrates, there is a correlation between CpG islands and GC-rich isochores. (Isochores are long genomic segments with homogeneous base-composition, found in vertebrate genomes. They are divided into different families that are characterized by having different GC-levels (Bernardi, 1993; Bernardi, 1995)). However, the causality of this correlation is somewhat unclear (Aïssani & Bernardi, 1991a; Aïssani & Bernardi, 1991b; Cross *et al.*, 1991; Cross & Bird, 1995).

Chromosomal structure and transcriptional repression

DNA is packaged in the form of chromatin

In the nucleus of eukaryotic cells, DNA is packaged in the form of chromatin (Kornberg, 1977; McGhee & Felsenfeld, 1980; Widom, 1989). In a human cell, this compaction makes it possible to fit the approximately two meters of genomic DNA into a nucleus that is only a few micrometers in diameter. The fundamental repetitive unit of chromatin fibers is the nucleosome core particle which consists of approximately 146 bp of DNA wrapped around an octamer composed of two molecules each of the four core histones (H2A, H2B, H3, and H4) (Richmond *et al.*, 1984; Luger *et al.*, 1997). Higher-order structures are formed by folding of nucleosomal arrays and are stabilized by interaction with other nuclear proteins including perhaps the linker

histone H1 (Wolffe *et al.*, 1997; Ramakrishnan, 1997).

Chromatin represses transcription

This densely packed state limits the accessibility of the DNA for the basal transcriptional apparatus and has been found to inhibit transcriptional initiation *in vivo* (Paranjape *et al.*, 1994; Kornberg & Lorch, 1995; Gottesfeld & Forbes, 1997; Grunstein, 1997). Compared to naked DNA, chromatin is therefore in a state of transcriptional repression. This is presumably one reason why cryptic core promoters seem to be practically inactive in living cells, and is also important for the very tight regulation of gene expression *in vivo*. Thus, while activator proteins typically increase transcription from a naked DNA template around ten-fold, the activation seen *in vivo* can be a thousand-fold or more. Hence, derepression of transcription by partial unfolding of chromatin constitutes an important part of gene regulation (Tsukiyama & Wu, 1997; Davie, 1998; Mizzen & Allis, 1998; Turner, 1998; Workman & Kingston, 1998), and several transcription factors and transcriptional co-activators have been shown to work by disrupting or remodeling chromatin structure (Brownell & Allis, 1995; Brownell *et al.*, 1996; Mizzen *et al.*, 1996; Ogryzko *et al.*, 1996; Pazin & Kadonaga, 1997; Mizzen & Allis, 1998; Turner, 1998; Workman & Kingston, 1998). Besides the generally repressive effect of chromatin on transcription, there are also several known cases where precisely positioned nucleosomes are directly involved in transcriptional regulation (Richard-Foy & Hager, 1987; Simpson, 1991; Hayes & Wolffe, 1992; Lu *et al.*, 1994; Simpson *et al.*, 1994; Wolffe, 1994; Zhu & Thiele, 1996).

DNA structure and promoter prediction

The fact that chromosome structure is important for transcriptional regulation, suggests that the analysis of DNA structure and nucleosome positioning might be helpful in connection with promoter prediction. It is relevant that DNA three-dimensional structure, and consequently also nucleosome positioning *in vivo* has been found to be influenced by the exact nucleotide sequence (Klug *et al.*, 1979; Dickerson & Drew, 1981; Hagerman, 1984; Drew & Travers, 1985; Satchwell *et al.*, 1986; Richard-Foy & Hager, 1987; Calladine *et al.*, 1988; Bolshoy *et al.*, 1991; Simpson, 1991; Hunter, 1993; Goodsell & Dickerson, 1994; Lu *et al.*, 1994; Brukner *et al.*, 1995a; Bolshoy, 1995; Iyer & Struhl, 1995; Wolffe & Drew, 1995; Hunter, 1996; Ioshikhes *et al.*, 1996; Widom, 1996; Zhu & Thiele, 1996; Liu & Stein, 1997). This means that it is perhaps possible to capture essential features of even the epigenetic parts of transcriptional regulation through structural analysis of the DNA sequence. We suggest that the additional use of such signals is likely to improve the performance of promoter prediction algorithms (Benham, 1996; Karas *et al.*, 1996; Pedersen *et al.*, 1998; Baldi *et al.*, 1998).

As one example of how the use of second-order characteristics of the sequence (in this case DNA bendability) can be used to investigate promoters, we will briefly describe some recent results from our group (Pedersen *et al.*, 1998). By analyzing a large set of unrelated (*i.e.*, non-sequence similar) human pol-II promoter sequences using sequence-dependent models of DNA structure, we have recently found what appears to be a general structural feature that is present in a majority of the investigated promoter sequences (Figure 2). Specifically, computational analysis using three independent models of DNA flexibility (Satchwell *et al.*, 1986; Brukner *et al.*, 1990; Brukner *et al.*, 1995a; Brukner *et al.*, 1995b; Hassan & Calladine, 1996) shows that a set of promoters with low sequence similarity displays an average tendency for low bendability upstream of the TATA-box, and high bendability downstream of the transcriptional start point. Within the downstream region there are strong indications of periodic sequence and bendability patterns in phase with the DNA helical pitch. This periodic pattern is very similar to that known from X-ray structures of the nucleosome core particle and tabulations of preferred sequence locations on nucleosomes. These results therefore indicate that on average the DNA in the region downstream of the start point in a large set of unrelated promoters is able to assume a macroscopically curved structure (*e.g.*, to be wrapped around protein) very similar to that of DNA in a nucleosome. Since the length of the high bendability region is approximately the same as the length of DNA wrapped around a histone octamer, it is tempting to suggest that this is a signal for positioning nucleosomes right at the transcriptional start point. Positioning of nucleosomes near the transcriptional start point could be related to the tight regulation of gene expression that is often observed *in vivo*. We suggest the use of this structural profile as one extra signal for promoter finding, and are currently in the process of developing such an algorithm (Pedersen, Baldi, Chauvin, and Brunak, in preparation). Briefly our method is based on two sensors: one that detects the overall structural profile (the high-to-low bendability shift) and another that looks for a periodic sequence pattern.

It has been noted that a DNA-bendability profile averaged over all possible heptamers conforming to the very loose consensus sequence of the initiator element (PyPyA$_{+1}$N[TA]PyPy) displays a single distinct high-bendability peak at position $+1$. This is caused by the fact that all eight triplets described by the subconsensus PyA$_{+1}$N have high bendability (Pedersen *et al.*, 1998). It was therefore tentatively suggested that at least part of the sequence requirements for a functional Inr are of a structural nature. Based on this and similar observations, it is tempting to suggest that some of the sequence heterogeneity that is seen in transcription factor binding sites, may in reality represent a more conserved structural motif underneath (Lisser & Margalit, 1994; Karas *et al.*, 1996; Grove *et al.*, 1998; de Souza & Ornstein, 1998).

Chromosomal domains and domain boundaries

An additional level of chromosomal structure that may be relevant for promoter function is the somewhat controversial organization of eukaryotic chromosomes into very large loops through attachment to a proteinaceous matrix (Laemmli *et al.*, 1992; Cremer *et al.*, 1993; Saitoh & Laemmli, 1993; Vazquez *et al.*, 1993; Dillon & Grosveld, 1994; Bode *et al.*, 1995; Gardiner, 1995). The DNA sequences that bind to nuclear matrix *in vitro* (and thus define the base of these loops) are called Matrix or Scaffold Attachment Regions (MARs/SARs) (Laemmli *et al.*, 1992; Bode *et al.*, 1995). It has been suggested that DNA loops may correspond to units of gene regulation, *i.e.*, regions within which transcriptional enhancers or repressors are constrained to act (Bonifer *et al.*, 1991; Sippel *et al.*, 1993; Dillon & Grosveld, 1994; Karpen, 1994; Schübeler *et al.*, 1996). The concept of domains of transcriptional control is closely related to the phenomenon of position effect variegation (PEV), *i.e.*, the variable levels of expression that are observed when a transgene is inserted at different locations in the DNA of a cell (Fraser & Grosveld, 1998; Gasser *et al.*, 1998; Kellum & Elgin, 1998). One possible cause of such variable expression is the spreading of a tightly packed chromatin structure from adjacent chromosomal regions into the DNA of the transgene (Karpen, 1994). The cooperative spreading of multiprotein complexes that interact with nucleosomes is believed to be at the base of heterochromatin formation, and may take place in a manner reminiscent of the propagation of silencing complexes (involving Rap1, Sir3, and Sir4 proteins) at yeast telomeres (Hecht *et al.*, 1996). Another cause of position effects is the accidental insertion of the transgene adjacently to an enhancer (Kellum & Elgin, 1998). Interestingly, it has been found that elements that are present in flanking regions of some eukaryotic genes have the ability to counteract such position effects. *E.g.*, some elements have been found to prevent enhancers from inappropriately activating promoters of neighboring genes when inserted between the enhancer and the promoter (Geyer, 1997; Kellum & Elgin, 1998), while others seem to limit spreading of heterochromatin (Karpen, 1994; Gasser *et al.*, 1998; Mihaly *et al.*, 1998). It is an attractive hypothesis that the proposed physical domain boundaries also act as functional domain boundaries, and examples are indeed known in which MARs also have insulator function (Laemmli *et al.*, 1992). However, there are also examples where no such correlation is seen (Geyer, 1997). Although the correlation between physical and functional domain boundaries apparently is not complete, it is nevertheless of great potential interest for promoter finding purposes that there are sequence elements which delimit domains of commonly regulated gene expression. Thus it seems very likely that if such elements are generally present in the flanking regions of most genes, then the additional use of these global signals will be beneficial for the performance of promoter finding algorithms.

The problem of computational MAR-detection has been taken up (Boulikas, 1995; Singh, 1997), but since there is currently very little experimentally verified data available, it is difficult to truly asses the

performance and usefulness of such methods. There is, however, no doubt that anyone interested in promoter prediction will be well advised to follow the developments in this field closely.

Discussion

The poor performance of current promoter finding algorithms is likely to indicate that these methods do not take into account enough relevant biological data. This does however not mean that improvement of such algorithms necessarily has to include explicit modeling of the biological reality. Indeed we believe there is much to be said for the inherent unbiasedness of purely data driven techniques. Rather, it means that it is very important to take biological knowledge into account when deciding (1) what to predict and (2) what data should be included when designing these methods.

The goal of promoter prediction—general *vs.* specific

There is, obviously, a conceptual difference between trying to recognize all eukaryotic promoters, and recognizing only those being active in a specific cell type or at a specific time during embryogenesis. In order to solve the first problem it is necessary to identify a set of features that are common between all promoters, and not present in the rest of the genome. Alternatively the promoter finding problem might be divided into several sub-problems, meaning that it may be necessary to construct specific algorithms to recognize specific classes of promoters. Some progress has been made along these lines in the work on predicting muscle-specific promoters (Fickett, 1996a; Fickett, 1996b; Wasserman & Fickett, 1998). It is also likely that prediction of promoters in single species, or perhaps groups of species, could be relevant since there without doubt are species specific promoter characteristics. It is not at all clear from the present results whether one approach is better than the other, and there is probably much to be learned from attempting to implement either solution. Here we mainly want to stress the importance of keeping the alternatives in mind.

Another fundamental question is whether promoter prediction algorithms should attempt to predict the exact transcriptional start point or the general region in which the promoter is likely to reside. We believe that it is probably beneficial to combine algorithms addressing these two problems, since the specific signals involved are apparently distinct to some degree (core promoter elements and structural features *vs.* enhancers, locus control regions, MARs, and other long-range and epigenetic signals).

In the case of exact start site prediction a related problem arises, namely how to evaluate the performance of predictions that are near but not at the start site. This is a problem both when testing the performance of existing methods, but also during development of new algorithms. It is not easy to give objective guidelines for when a prediction should be accepted. Most will probably agree that if a method consistently predicts

start sites to be within relatively few bp of the annotated sites then the algorithm is doing very well. This view also makes good biological sense: in many cases promoters do display more than one start. Furthermore, there are probably many cases where only one of several start sites is reported in available databases—a situation which necessarily must result in some degree of ambiguity in prediction methods constructed from these data.

Epigenetic control and long range interactions—more signals for promoter finding

Another fundamental question is whether there is in fact enough information present in the local DNA sequence to define promoters. Thus, statistical analysis of the density of transcription factor binding sites suggests that this alone is insufficient to unambiguously discriminate between promoters and non-promoters (Prestridge & Burks, 1993; Zhang, 1998a). From a biological point of view this may seem to be a strange thought—the vertebrate cell, after all, is capable of correctly controlling the expression of several tens of thousands of genes. However, such a situation could arise due to epigenetic control of transcription, *i.e.*, regulation by signals superimposed upon the primary DNA sequence. In accordance with this, it is known that chromatin folding and unfolding do indeed play a very important role in transcriptional control, as does DNA methylation (Boyes & Bird, 1992; Eden & Cedar, 1994; Paranjape *et al.*, 1994; Kornberg & Lorch, 1995; Gottesfeld & Forbes, 1997; Grunstein, 1997). The fact that transcriptional initiation can be dependent on long range interactions between factors bound at promoters and other factors bound at distant sites, further suggests that local DNA sequence alone may not be sufficient to define a promoter.

We believe that one way of approaching this problem is to include the use of more global signals in addition to the local signals mainly used by most current algorithms (TATA-box, Inr, CCAAT-box, *etc*.). Exactly what global signal(s) to use is an open question, but based upon the data reviewed here, two interesting new candidates emerge: (1) CpG islands, which in vertebrate genomes often are correlated with promoter position, and (2) chromosomal domain boundaries or domain insulators, which may in some cases delimit transcriptional units in genomic DNA. The use of several signals simultaneously might be implemented in a manner similarly to the "grammatical" approaches known from general gene finding methods (Borodovsky & McIninch, 1993; Krogh *et al.*, 1994; Matis *et al.*, 1996; Rosenblueth *et al.*, 1996; Uberbacher *et al.*, 1996; Burge & Karlin, 1997; Lukashin & Borodovsky, 1998). and it would then be natural to also include other feature detectors normally used in general gene finding (coding DNA, splice sites, poly-A signals, *etc*.).

In our experience, the simultaneous use of signals at both the local and global levels has a consistently beneficial effect on the performance of feature detectors. Examples include the prediction of splice sites, signal peptides, translation start sites, and glycosylation sites (Brunak *et al.*, 1991; Hansen *et al.*, 1995; Korning *et al.*, 1996; Nielsen *et al.*, 1996; Nielsen *et al.*, 1997; Pedersen & Nielsen, 1997; Tolstrup *et al.*,

1997; Hansen *et al.*, 1998). Presumably, the reason behind this behavior is that instead of having a fixed, absolute threshold for when a putative local signal is genuine, global signals are in effect able to modulate the cut-off for the local signal. Consider, for instance, the case of predicting signal peptide cleavage sites in protein sequences: obviously the likelihood that a putative cleavage site is functional, depends on whether it is adjacent to a signal peptide-like sequence (positioned upstream) or not. In this context, it is perhaps also interesting that some researchers have found that within any given promoter sequence, it is usually the strongest signal that is correct, regardless of the absolute strength (Hutchinson, 1996; Audic & Claverie, 1998).

Choice of data sets for construction of algorithms

Careful selection of the data set that is used for constructing and testing a promoter finding algorithm is of course of the utmost importance. Many existing methods use recognition of transcription factor binding sites, based on lists of sites present in databases such as TRANSFAC (Wingender *et al.*, 1998). Since this presumably is very similar to what the cell does, it is an intuitively pleasing approach and one that seems likely to succeed. However, some of the sites present in these databases may have only limited experimental evidence—a situation which could seriously affect the performance of algorithms that use them. It is always advisable to personally check the original references to sites of interest, although this will prove to be very hard work if more than a few sites are to be include in an analysis. There are undoubtedly also biologically based problems regarding the use of lists of transcription factor binding sites. Thus, it is well known that if a given transcription factor binds to sites present in a set of, *e.g.*, 50 different promoters, then the sequences of these different sites may display considerable diversity. The same will also be the case for the strength (binding energies) of the corresponding protein-DNA interactions. For instance, the protein may in some cases interact with other proteins that enable it to bind to very weak sites. Under all circumstances: if all 50 binding sites are included in a database and subsequently used to design a method for recognizing the site, then the information present in the weak binding sites can be said to be overemphasized, resulting perhaps in a method that predicts far too many false positives. It is an interesting thought that binding energies in the different protein-DNA interactions could be used to weight the different sequences when constructing recognition algorithms. Unfortunately, it is likely that the different conditions used in different experimental setups will make such comparisons practically useless. Furthermore, any truly successful method should of course be able to also predict the weaker sites, so under-estimating their importance may prove problematic as well. We have no simple answer to this problem.

An alternative to using lists of transcription factor binding elements, is to rely on annotated transcription start sites. The presence of an initiation site must necessarily imply that the surrounding sequence has

promoter activity. One advantage to this approach is that usually there is very little doubt about the validity of the experimental evidence for a transcriptional initiation site. The eukaryotic promoter database (EPD) is based on such experimentally determined transcription starts and is furthermore curated, meaning that all included start sites have been checked with the original literature (Perier *et al.*, 1998). A new web-interface for accessing this database seems very useful (Table 1). It is of course also possible to extract sequences directly from general nucleotide databases such as GenBank, EMBL, or the DNA Data Bank of Japan (Benson *et al.*, 1998; Stoesser *et al.*, 1998; Tateno *et al.*, 1998) by looking for feature keys which indicate that an initiation site is present (*e.g.*, "mRNA" and "prim_transcript"). In this way it is usually possible to collect larger data sets than when using EPD, but it is naturally also a tradeoff with respect to the quality of the data.

How much sequence to include is an open question, but based on biological data it at least seems that it is important to include sequence from both sides of the start point. The amount of sequence also depends on which prediction approach is chosen. Thus, it is obvious that the use of global signals such as CpG islands or MARs necessitates the inclusion of more flanking DNA.

Once a set of sequences has been extracted it may be relevant to reduce the redundancy or homology present within the data set. This is especially important when statistically based methods (including neural networks) are used, since statistics will otherwise be biased for the over represented sequences that are present in all databases. In this regard, we have developed an automatic redundancy reduction technique that is based on the use of pairwise alignments and a cut-off selected from the parameters of the extreme-value distribution which the resulting alignment scores follow (A. G. Pedersen, S. Brunak, and H. Nielsen, in prep.).

Perspectives: DNA structural analysis—more levels to one signal?

There are indications that DNA sequence-dependent three-dimensional structure may be important for transcriptional regulation both at the level of single binding sites and at the level of entire promoter regions (Lisser & Margalit, 1994; Karas *et al.*, 1996; Benham, 1996; Liu & Stein, 1997; Grove *et al.*, 1998; Pedersen *et al.*, 1998; de Souza & Ornstein, 1998). Thus, it is possible that sequence heterogeneity of the different binding sites of a transcription factor may to some degree reflect an underlying conserved DNA structure. Furthermore, it is also possible that the same is true for sequence surrounding the transcriptional start point. It may therefore be relevant to explicitly assist promoter finding algorithms in recognizing structural features of DNA sequences. One way of doing this is to encode DNA sequences in the form of DNA structural parameters for all overlapping triplets or dinucleotides (Karas *et al.*, 1996; Baldi *et al.*, 1998; Pedersen *et al.*, 1998). Preliminary investigations suggest that such structural encodings do indeed contain sufficient

information to recognize features that are normally considered to be present at the sequence level (Baldi *et al.*, 1998, N. Tolstrup, personal communication). Furthermore, it is an intriguing possibility that this may be a way of approaching "epigenetic" signals such as chromatin folding and nucleosome positioning from the sequence.

Acknowledgments

AGP and SB are supported by a grant from the Danish National Research Foundation. The work of PB and YC is in part supported by an NIH SBIR grant to Net-ID, Inc. We thank Lise Hoffmann for critical comments on this manuscript.

References

Adhya, S. (1989). Multipartite genetic control elements: communication by DNA looping. *Annu. Rev. Genet.*, **23**, 227–250.

Aïssani, B. & Bernardi, G. (1991a). CpG islands: features and distribution in the genomes of vertebrates. *Gene*, **106**, 173–183.

Aïssani, B. & Bernardi, G. (1991b). CpG islands, genes, and isochores in the genomes of vertebrates. *Gene*, **106**, 185–195.

Allison, L. A., Moyle, M., Shales, M. & Ingles, C. J. (1985). Extensive homology among the largest subunits of eukaryotic and prokaryotic RNA polymerases. *Cell*, **42**, 599–610.

Antequera, F. & Bird, A. (1993). Number of CpG islands and genes in human and mouse. *Proc. Natl. Acad. Sci. USA*, **90**, 11995–11999.

Audic, S. & Claverie, J.-M. (1997). Detection of eukaryotic promoters using Markov transition matrices. *Computers Chem.*, **21**, 223–227.

Audic, S. & Claverie, J. M. (1998). Visualizing the competitive recognition of tata-boxes in vertebrate promoters. *Trends Genet.*, **14**, 10–11.

Baldi, P., Chauvin, Y., Brunak, S., Gorodkin, J. & Pedersen, A. G. (1998). Computational applications of DNA structural scales. *ISMB*, **6**.

Benham, C. J. (1996). Computation of DNA structural variability — a new predictor of DNA regulatory sequences. *CABIOS*, **12**, 375–381.

Benson, D. A., Boguski, M. S., Lipman, D. J., Ostell, J. & Ouellette, B. F. (1998). Genbank. *Nucleic Acids Res.*, **26**, 1–7.

Bernardi, G. (1993). The isochore organization of the human genome and its evolutionary history—a review. *Gene*, **135**, 57–66.

Bernardi, G. (1995). The human genome: organization and evolutionary history. *Annu. Rev. Genet.*, **29**, 445–476.

Bird, A. (1980). DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res.*, **8**, 1499–????

Bird, A. (1993). Functions for DNA methylation in vertebrates. *Cold Spring Harbor Symp. Quant. Biol.*, **58**, 281–285.

Bird, A., Tate, P., Nan, X., Campoy, J., Meehan, R., Cross, S., Tweedle, S., Charlton, J. & Macleod, D. (1995). Studies of DNA methylation in animals. *J. Cell Sci. Suppl.*, **19**, 37–39.

Bode, J., Schlake, T., Ríos-Ramírez, M., Mielke, C., Stengert, M., Kay, V. & Klehr-Wirth, D. (1995). Scaffold/matrix-attached regions: structural properties creating transcriptionally active loci. *Int. Rev. Cyt.*, **162A**, 389–454.

Bolshoy, A. (1995). CC dinucleotides contribute to the bending of DNA in chromatin. *Nature Struct. Biol.*, **2**, 447–448.

Bolshoy, A., McNamara, P., Harrington, R. E. & Trifonov, E. N. (1991). Curved DNA without A-A: experimental estimation of all 16 DNA wedge angles. *Proc. Natl. Acad. Sci. USA*, **88**, 2312–2316.

Bonifer, C., Hecht, A., Sauressig, H., Winter, D. M. & Sippel, A. E. (1991). Dynamic chromatin: the regulatory domain organization of eukaryotic gene loci. *J. Cell. Biochem.*, **47**, 99–108.

Borodovsky, M. & McIninch, J. (1993). Genemark: Parallel gene recognition for both dna strands. *Comput. Chem.*, **17**, 123–133.

Boulikas, T. (1995). Chromatin domains and prediction of MAR sequences. *Int. Rev. Cytol.*, **162A**, 279–388.

Boyes, J. & Bird, A. (1992). Repression of genes by DNA methylation depends on CpG density and promoter strength: evidence for involvement of a methyl-CpG binding protein. *EMBO J.*, **11**, 327–333.

Breant, B., Huet, J., Sentenac, A. & Fromageot, P. (1983). Analysis of yeast RNA polymerases with subunit-specific antibodies. *J. Biol. Chem.*, **258**, 11968–11973.

Brownell, J. E. & Allis, C. D. (1995). An activity gel assay detects a single catalytically active histone acetyltransferase subunit in *Tetrahymena* macronuclei. *Proc. Natl. Acad. Sci. USA*, **92**, 6364–6368.

Brownell, J. E., Zhou, J., Ranalli, T., Kobayashi, R., Edmondson, D. G., Roth, S. Y. & Allis, C. D. (1996). *Tetrahymena* histone acetyltransferase A: a homologue to yeast Gcn5p linking histone acetylation to gene activation. *Cell*, **84**, 843–851.

Brukner, I., Jurukovski, V. & Savic, A. (1990). Sequence-dependent structural variations of DNA revealed by DNase I. *Nucleic Acids Res.*, **18**, 891–894.

Brukner, I., Sanchez, R., Suck, D. & Pongor, S. (1995a). Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides. *EMBO J.*, **14**, 1812–1818.

Brukner, I., Sanchez, R., Suck, D. & Pongor, S. (1995b). Trinucleotide models for DNA bending propensity: comparison of models based on DNase I digestion and nucleosome positioning data. *J. Biomol. Struct. Dyn.*, **13**, 309–317.

Brunak, S., Engelbrecht, J. & Knudsen, S. (1991). Prediction of human mRNA donor and acceptor sites from the DNA sequences. *J. Mol. Biol.*, **220**, 49–65.

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

Bucher, P., Fickett, J. W. & Hatzigeorgiou, A. (1996). Computational analysis of transcriptional regulatory elements: a field in flux. *CABIOS*, **12**, 361–362.

Buratowski, S. (1997). Multiple TATA-binding factors come back into style. *Cell*, **91**, 13–15.

Burge, C. & Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, 78–94.

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

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

Calladine, C. R., Drew, H. R. & McCall, M. J. (1988). The intrinsic structure of DNA in solution. *J. Mol. Biol.*, **201**, 127–137.

Cavalli, G. & Paro, R. (1998). Chromo-domain proteins: linking chromatin structure to epigenetic regulation. *Curr. Opin. Cell Biol.*, **10**, 354–360.

Chen, Q. K., Hertz, G. Z. & Stormo, G. D. (1997). Promfd 1.0: a computer program that predicts eukaryotic pol II promoters using strings and IMD matrices. *CABIOS*, **13**, 29–35.

Chen, Q. K. & Stormo, G. Z. H. G. D. (1995). Matrix search 1.0: a computer program that scans dna sequences for transcriptional elements using a database of weight matrices. *Comput. Appl. Biosci.*, **11**, 563–566.

Coulondre, C., Miller, J. H., Farabough, P. J. & Gilbert, W. (1978). Molecular basis of base substitution hotspots in *Escherichia coli*. *Nature*, **274**, 775–780.

Craig, J. M. & Bickmore, W. A. (1994). The distribution of CpG islands in mammalian chromosomes. *Nature Genetics*, **7**, 376–382.

Cremer, T., Kurz, A., Zirbel, R., Dietzel, S., Rinke, B., Schröck, E., Speichel, M. R., Mathieu, U., Jauch, A., Emmerich, P., Schertan, H., Ried, T., Cremer, C. & Lichter, P. (1993). Role of chromosome territories in the functional compartmentalization of the cell nucleus. *Cold Spring Harbor Symp. Quant. Biol.*, **58**, 777–792.

Cross, S., Kovarik, P., Schmidtke, J. & Bird, A. (1991). Nonmethylated islands in fish genomes are GC-poor. *Nucleic Acids Res.*, **19**, 1469–1474.

Cross, S. H. & Bird, A. (1995). CpG islands and genes. *Curr. Opin. Genet. Dev.*, **5**, 309–314.

Crossley, M. & Orkin, S. H. (1993). Regulation of the beta-globin locus. *Curr. Opin. Genet. Dev.*, **3**, 232–237.

Davie, J. R. (1998). Covalent modification of histones: expression from chromatin templates. *Curr. Opin. Gen. Dev.*, **8**, 173–178.

de Souza, O. N. & Ornstein, R. L. (1998). Inherent DNA curvature and flexibility correlate with TATA box functionality. *Biopolymers*, **46**, 403–415.

Diamond, M. I., Miner, J. N., Yoshinaga, S. K. & Yamamoto, K. R. (1990). Transcription factor interactions: selectors of positive or negative regulation from a single DNA element. *Science*, **249**, 1266–1274.

Dickerson, R. E. & Drew, H. (1981). Structure of a B-DNA dodecamer. II. Influence of base sequence on helix structure. *J. Mol. Biol.*, **149**, 761–786.

Dillon, N. & Grosveld, F. (1994). Chromatin domains as potential units of eukaryotic gene function. *Curr. Opin. Genet. Dev.*, **4**, 260–264.

Drew, H. R. & Travers, A. A. (1985). DNA bending and its relation to nucleosome positioning. *J. Mol. Biol.*, **186**, 773–790.

Dynan, W. S. (1989). Modularity in promoters and enhancers. *Cell*, **58**, 1–4.

Eden, S. & Cedar, H. (1994). Role of DNA methylation in the regulation of transcription. *Curr. Opin. Genet. Dev.*, **4**, 255–259.

Evans, T., Felsenfeld, G. & Reitman, M. (1990). Control of globin gene transcription. *Annu. Rev. Cell Biol.*, **6**, 95–124.

Fassler, J. S. & Gussin, G. N. (1996). Promoters and basal transcription machinery in eubacteria and eukaryotes: concepts, definitions, and analogies. *Meth. Enz.*, **273**, 3–29.

Fickett, J. W. (1996a). Coordinate positioning of MEF2 and myogenin binding sites. *Gene*, **172**, GC19–GC32.

Fickett, J. W. (1996b). Quantitative discrimination of MEF2 sites. *Mol. Cell. Biol.*, **16**, 437–441.

Fickett, J. W. & Hatzigeorgiou, A. G. (1997). Eukaryotic promoter recognition. *Genome Res.*, **7**, 861–878.

Fraser, P. & Grosveld, F. (1998). Locus control regions, chromatin activation and transcription. *Curr. Opin. Cell Biol.*, **10**, 361–365.

Frech, K., Danescu-Mayer, J. & Werner, T. (1997). 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.

Gardiner, K. (1995). Human genome organization. *Curr. Opin. Genet. Dev.*, **5**, 315–322.

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

Gasser, S. M., Paro, R., Stewart, F. & Aasland, R. (1998). Epigenetic control of transcription. Introduction: the genetics of epigenetics. *Cell. Mol. Life Sci.*, **54**, 1–5.

Geyer, P. K. (1997). The role of insulator elements in defining domains of gene expression. *Curr. Opin Gen. Dev.*, **7**, 242–248.

Goodrich, J. A. & Tjian, R. (1994). TBP-TAF complexes: selectivity factors for eukaryotic trancription. *Curr. Opin. Cell Biol.*, **6**, 403–409.

Goodsell, D. S. & Dickerson, R. E. (1994). Bending and curvature calculations in B-DNA. *Nucleic Acids Res.*, **22**, 5497–5503.

Gottesfeld, J. M. & Forbes, D. J. (1997). Mitotic repression of the transcriptional machinery. *TIBS*, **22**, 197–202.

Greenblatt, J. (1997). RNA polymerase II holoenzyme and transcriptional regulation. *Curr. Opin. Cell Biol.*, **9**, 310–319.

Gregory, P. D. & Hörz, W. (1998). Life with nucleosomes: chromatin remodelling in gene regulation. *Curr. Opin. Cell. Biol.*, **10**, 339–345.

Grove, A., Galeone, A., Yu, E., Mayol, L. & Geiduschek, E. P. (1998). Affinity, stability and polarity of binding of the TATA binding protein governed by flexure of the TATA box. *J. Mol. Biol.*, **282**, 731–739.

Grunstein, M. (1997). Histone acetylation in chromatin structure and transcription. *Nature*, **389**, 349–352.

Hagerman, P. J. (1984). Evidence for the existence of stable curvature of DNA in solution. *Proc. Natl. Acad. Sci. USA*, **81**, 4632–4636.

Hahn, S., Buratowski, S., Sharp, P. A. & Guarente, L. (1989). Yeast TATA-binding protein TFIID binds to TATA elements with both consensus and nonconsensus sequences. *Proc. Natl. Acad. Sci. USA*, **86**, 5718–5722.

Hansen, J., Lund, O., Tolstrup, N., Gooley, A., Williams, K. & Brunak, S. (1998). NetOglyc: Prediction of mucin type O-glycosylation sites based on sequence context and surface accessibility. *Glycoconjugate Journal*, **15**, 115–130.

Hansen, J. E., Lund, O., Engelbrecht, J., Bohr, H., Nielsen, J. O., Hansen, J.-E. S. & Brunak, S. (1995). Prediction of O-glycosylation of mammalian proteins: specificity patterns of UDP-GalNAC: polypeptide *N*-acetylgalactosaminyltransferase. *Biochem J.*, **308**, 801–813.

Hansen, S. K., Takada, S., Jacobson, R. H., Lis, J. T. & Tjian, R. (1997). Transcription properties of a cell type-specific TATA-binding protein, TRF. *Cell*, **91**, 71–83.

Hassan, M. A. E. & Calladine, C. R. (1996). Propeller-twisting of base-pairs and the conformational mobility of dinucleotide steps in DNA. *J. Mol. Biol.*, **259**, 95–103.

Hayes, J. J. & Wolffe, A. P. (1992). The interaction of transcription factors with nucleosomal DNA. *BioEssays*, **14**, 597–603.

Hecht, A., Strahl-Bolsinger, S. & Grunstein, M. (1996). Spreading of transcriptional repressor sir3 from telomeric hetrochromatin. *Nature*, **383**, 92–96.

Hernandez, N. (1993). TBP, a universal eukaryotic transcription factor? *Genes Dev.*, **7**, 1291–1308.

Higgs, D. R. & Wood, W. G. (1993). Understanding erythroid differentiation. *Curr. Biol.*, **3**, 548–550.

Holliday, R. (1993). Epigenetic inheritance based on DNA methylation. *EXS*, **64**, 452–468.

Huet, J., Sentenac, A. & Fromageot, P. (1982). Spot-immunodetection of conserved determinants in eukaryotic RNA polymerases. Study with antibodies to yeast RNA polymerases subunits. *J. Biol. Chem.*, **257**, 2613–2618.

Hunter, C. A. (1993). Sequence-dependent DNA structure: the role of base stacking interactions. *J. Mol. Biol.*, **230**, 1025–1054.

Hunter, C. A. (1996). Sequence-dependent DNA structure. *Bioessays*, **18**, 157–162.

Hutchinson, G. B. (1996). The prediction of vertebrate promoter regions using differential hexamer frequency analysis. *CABIOS*, **12**, 391–398.

Ioshikhes, I., Bolshoy, A., Derenshteyn, K., Borodovsky, M. & Trifonov, E. N. (1996). Nucleosome DNA sequence pattern revealed by multiple alignment of experimentally mapped sequences. *J. Mol. Biol.*, **262**, 129–139.

Iyer, V. & Struhl, K. (1995). Poly (dA:dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure. *EMBO J.*, **14**, 2570–2579.

Jackle, H. & Sauer, F. (1993). Transcriptional cascades in *Drosophila*. *Curr. Opin. Cell Biol.*, **5**, 505–512.

Johnson, P. F. & McKnight, S. L. (1989). Eukaryotic transcription regulatory proteins. *Ann. Rev. Biochem.*, **58**, 799–839.

Jones, P. A., Rideout, W. M., Shen, J.-C., Spruck, C. H. & Tsai, Y. C. (1992). Methylation, mutation and cancer. *BioEssays*, **14**, 33–36.

Karas, H., Knüppel, R., Schultz, W., Sklenar, H. & WIngender, E. (1996). Combining structural analysis of DNA with search routines for the detection of transcription regulatory elements. *CABIOS*, **12**, 441–446.

Karpen, G. H. (1994). Position-effect variegation and the new biology of heterochromatin. *Curr. Opin. Genet. Dev.*, **4**, 281–291.

Kellum, R. & Elgin, S. C. R. (1998). Chromatin boundaries: punctuating the genome. *Curr. Biol.*, **8**, R521–R524.

Klug, A., Jack, A., Viswamitra, M. A., Kennard, O., Shakked, Z. & Steitz, T. A. (1979). A hypothesis on a specific sequence-dependent conformation of DNA and its relation to the binding of the *lac*-repressor protein. *J. Mol. Biol.*, **131**, 669–680.

Knudsen, S. (Submitted). Promoter1.0: a novel algorithm for the recognition of pol-II promoter sequences.

Kondrakhin, Y. V., Kel, A. E., Romashchenko, N. A. K. A. G. & Milanesi, L. (1995). Eukaryotic promoter recognition by binding sites for transcription factors. *Comput. Appl. Biosci.*, **11**, 477–488.

Kornberg, R. D. (1977). Structure of chromatin. *Annu. Rev. Biochem.*, **46**, 931–954.

Kornberg, R. G. & Lorch, Y. (1995). Interplay between chromatin structure and transcription. *Curr. Opin. Cell. Biol.*, **7**, 371–375.

Korning, P. G., Hebsgaard, S. M., Rouze, P. & Brunak, S. (1996). Cleaning the GenBank *Arabidopsis thaliana* data set. *Nucleic Acids Res.*, **24**, 316–320.

Krogh, A., Brown, M., Mian, I. S., Sjolander, K. & Haussler, D. (1994). Hidden Markov models in computational biology: Applications to protein modeling. *J. Mol. Biol.*, **235**, 1501–1531.

Laemmli, U. K., Käs, E., Poljak, L. & Adachi, Y. (1992). Scaffold-associated regioons: cis-acting determinants of chromatin structural loops and functional domains. *Curr. Opin. Gen. Dev.*, **2**, 275–285.

Lamb, P. & McKnight, S. L. (1991). Diversity and specificity in transcriptional regulation: the benefits of hetrotypic dimerization. *Trends Biochem. Sci.*, **16**, 417–422.

Lee, T. I. & Young, R. A. (1998). Regulation of gene expression by TBP associated proteins. *Genes Dev.*, **12**, 1398–1408.

Li, E., Bestor, T. H. & Jaenisch, R. (1992). Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell*, **69**, 915–926.

Lisser, S. & Margalit, H. (1994). Determination of common structural features in *Escherichia coli* promoters by computer analysis. *Eur J Biochem*, **223**, 823–830.

Liu, K. & Stein, A. (1997). DNA sequence encodes information for nucleosome array formation. *J. Mol. Biol.*, **270**, 559–573.

Lu, B. Y. & Eissenberg, J. C. (1998). Time out: developmental regulation of heterochromatic silencing in *Drosophila*. *Cell. Mol. Life Sci.*, **54**, 50–59.

Lu, Q., Wallrath, L. L. & Elgin, S. C. R. (1994). Nucleosome positioning and gene regulation. *J. Cell. Biochem.*, **55**, 83–92.

Luger, K., Maeder, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. (1997). Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*, **389**, 251–260.

Lukashin, A. & Borodovsky, M. (1998). GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.*, **26**, 1107–1115.

Macleod, D., Charlton, J., Mullins, J. & Bird, A. (1994). Sp1 sites in the mouse *aprt* gene promoter are required to prevent methylation of the CpG islands. *Genes Dev.*, **8**, 2282–2292.

Matis, S., Xu, Y., Shah, M., Guan, X., Einstein, J. R., Mural, R. & Uberbacher, E. C. (1996). Detection of rna polymerase ii promoters and polyadenylation sites in human dna sequence. *Comput. Chem.*, **20**, 135–140.

Matthews, K. S. (1992). DNA looping. *Microbiol. Rev.*, **56**, 123–136.

McGhee, J. D. & Felsenfeld, G. (1980). Nucleosome structure. *Annu. Rev. Biochem.*, **49**, 1115–1156.

McKnight, S. L. & Yamamoto, K. R. (1992). Transcriptional Regulation. Cold Spring Harbor Laboratory Press, Plainview, New York.

Mihaly, J., Hogga, I., Barges, S., Galloni, M., Mishra, R. K., Hagstrom, K., Müller, M., Schedl, P., Sipos, L., Gausz, J., Gyurkovics, H. & Karch, F. (1998). Chromatin domain boundaries in the bithorax complex. *Cell. Mol. Life Sci.*, **54**, 60–70.

Milanesi, L., Muselli, M. & Arrigo, P. (1996). Hamming clustering method for signals prediction in 5' and 3' regions of eukaryotic genes. *CABIOS*, **12**, 399–404.

Milanesi, L. & Rogozin, I. B. (In press). Prediction of human gene structure. In M.J.Bishop, (ed.) *Guide to Human Genome Computing*. Academic Press, Cambridge.

Minie, M., Clark, D., Trainor, C., Evans, T., Reitman, M., Hannon, R., Gould, H. & Felsenfeld, G. (1992). Developmental regulation of globin gene expression. *J. Cell Sci. Suppl*, **16**, 15–20.

Mitchell, P. J. & Tjian, R. (1989). Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science*, **245**, 371–378.

Mizzen, C. A. & Allis, C. D. (1998). Linking histone acetylation to transcriptional regulation. *Cell. Mol. Life Sci.*, **54**, 6–20.

Mizzen, C. A., Yang, X.-J., Kokubo, T., Brownell, J. E., Bannister, A. J., Owen-Hughes, T., Workman, J. L., Berger, S. L., Kouzarides, T., Nakatani, Y. & Allis, C. D. (1996). The $TAF_{II}250$ subunit of TFIID has histone acetyltransferase activity. *Cell*, **87**, 1261–1270.

Nielsen, H., Engelbrecht, J., Brunak, S. & von Heine, G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.*, **10**, 1–6.

Nielsen, H., Engelbrecht, J., von Heijne, G. & Brunak, S. (1996). Defining a similarity threshold for a functional protein sequence pattern: the signal peptide cleavage site. *Proteins*, **26**, 165–177.

Ogryzko, V., Schiltz, R. L., Russanova, V., Howard, B. H. & Nakatani, Y. (1996). The transcriptional coactivators p300 and CBP are histone acetyltransferases. *Cell*, **87**, 953–959.

Orphanides, G., Lagrange, T. & Reinberg, D. (1996). The general transcription factors of RNA polymerase II. *Genes Dev.*, **10**, 2657–2683.

Paranjape, S. M., Kamakaka, R. T. & Kadonaga, J. T. (1994). Role of chromatin structure in the regulation of transcription by RNA polymerase II. *Annu. Rev. Biochem.*, **63**, 265–297.

Pazin, M. J. & Kadonaga, J. T. (1997). SWI2/SNF2 and related proteins: ATP-driven motors that disrupt protein-DNA interactions? *Cell*, **88**, 737–740.

Pedersen, A. G., Baldi, P., Chauvin, Y. & Brunak, S. (1998). DNA structure in human RNA polymerase II promoters. *J. Mol. Biol.*, **281**, 663–673.

Pedersen, A. G. & Nielsen, H. (1997). Neural network based prediction of translation initiation sites in eukaryotes: perspectives for EST and genome analysis. *ISMB*, **5**, 226–233.

Perier, R. C., Junier, T. & Bucher, P. (1998). The eukaryotic promoter database epd. *Nucleic Acids Res.*, **26**, 353–357.

Prestridge, D. S. (1995). Prediction of pol ii promoter sequences using transcription factor binding sites. *J. Mol. Biol.*, **249**, 923–932.

Prestridge, D. S. (1997). Signal scan: A computer program that scans dna sequences for eukaryotic transcriptional elements. *CABIOS*, **7**, 203–206.

Prestridge, D. S. & Burks, C. (1993). The density of transcriptional elements in promoter and non-promoter sequences. *Hum Mol Genet*, **2**, 1449–53.

Ptashne, M. & Gann, A. (1997). Transcriptional activation by recruitment. *Nature*, **386**, 569–577.

Pugh, B. F. & Tjian, R. (1990). Mechanism of transcriptional activation by Sp1: evidence for coactivators. *Cell*, **61**, 1187–1197.

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. *Nucl. Acids Res.*, **23**, 4878–4884.

Ramakrishnan, V. (1997). Histone H1 and chromatin higher-order structure. *Crit. Rev. Eukaryot. Gene Expr.*, **7**, 215–230.

Reese, M. G., Harris, N. L. & Eeckman, F. H. (1996). Large scale sequencing specific neural networks for promoter and splice site recognition. In Hunter, L. & Klein, T. E., (eds.) *Biocomputing: Proceedings of the 1996 Pacific Symposium*. World Scientific Publishing Co., Singapore.

Richard-Foy, H. & Hager, G. L. (1987). Sequence specific positioning of nucleosomes over the steroid-inducible MMTV promoter. *EMBO J.*, **6**, 2321–2328.

Richmond, T. J., Finch, J. T., Rushton, B., Rhodes, D. & Klug, A. (1984). Structure of the nucleosome core particle at 7 Å resolution. *Nature*, **311**, 532–537.

Rosenblueth, D. A., Thieffry, D., Huerta, A. M., Salgado, H. & Colladio-Vides, J. (1996). Syntactic recognition of regulatory regions in *Escherichia coli*. *CABIOS*, **12**, 415–422.

Saitoh, Y. & Laemmli, U. K. (1993). From the chromosomal loops and the scaffold to the classic bands of metaphase chromosomes. *Cold Spring Harbor Symp. Quant. Biol.*, **58**, 755–765.

Satchwell, S. C., Drew, H. R. & Travers, A. A. (1986). Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.*, **191**, 659–675.

Schirm, S., Jiricny, J. & Schaffner, W. (1987). The SV40 enhancer can be dissected into multiple segments, each with a different cell type specificity. *Genes Dev.*, **1**, 65–74.

Schübeler, D., Mielke, C., Maas, K. & Bode, J. (1996). Scaffold/matrix attached regions act upon transcription in a context-dependent manner. *Biochem.*, **35**, 11160–11169.

Schug, J. & Overton, G. C. (1997). Tess: Transcription element search software on the www. In *Technical Report CBIL-TR-1997-1001-v0.0, of the Computational Biology and Informatics Laboratory, School of Medicine, University of Pennsylvania*.

Simpson, R. T. (1991). Nucleosome positioning: occurrence, mechanisms, and functional consequences. *Prog. Nucleic Acids Res. Mol. Biol.*, **40**, 143–184.

Simpson, R. T., Roth, S. Y., Morse, R. H., Patterson, H. G., Cooper, J. P., Murphy, M., Kladde, M. P. & Shimizu, M. (1994). Nucleosome positioning and transcription. *Cold Spring Harbor Symp. Quant. Biol.*, **58**, 237–245.

Singer, V. L., Wobbe, C. R. & Struhl, K. (1990). A wide variety of DNA sequences can functionally replace a yeast TATA element for transcriptional activation. *Genes Dev.*, **4**, 636–645.

Singh, G. B. (1997). Mathematical model to predict regions of chromatin attachment to the nuclear matrix. *Nucleic Acids Res.*, **25**, 1419–1425.

Sippel, A. E., Schäfer, G., Faust, N., Hecht, A. & Bonifer, C. (1993). Chromatin domains constitute regulatory units for the control of eukaryotic genes. *Cold Spring Harbor Symp. Quant. Biol.*, **58**, 37–44.

Smale, S. T. (1994a). Core promoter architecture for eukaryotic protein-coding genes. In Conaway, R. C. & Conaway, J. W., (eds.) *Transcription: Mechanisms and regulation*. Raven Press, New York, NY, pp. 63–81.

Smale, S. T. (1994b). DNA sequence requirements for transcriptional initiator activity in mammalian cells. *Mol. Cell. Biol.*, **14**, 116–127.

Smale, S. T. (1997). Transcription initiation from TATA-less promoters within eukaryotic protein-coding genes. *Bioch. Biophys. Acta*, **1351**, 73–88.

Small, S., Kraut, R., Hoey, T., Warrior, R. & Levine, M. (1991). Transcriptional regulation of a pair-rule stripe in *Drosophila*. *Genes Dev.*, **5**, 827–839.

Solovyev, V. & Salamov, A. (1997). The gene-finder computer tools for analysis of human and model organisms genome sequences. In *Proceedings, Fifth International Conference on Intelligent Systems for Molecular Biology (ISMB-97)*. pp. 294–302.

Stargell, L. A. & Struhl, K. (1996). Mechanisms of transcriptional activation in vivo: two steps forward. *Trends Genet.*, **12**, 311–315.

Stoesser, G., Moseley, M. A., Sleep, J., McGowran, M., Garcia-Pastor, M. & Sterk, P. (1998). The embl nucleotide sequence database. *Nucleic Acids Res.*, **26**, 8–15.

Tanese, N. & Tjian, R. (1993). Coactivators and TAFs: a new class of eukaryotic transcription factors that connect activators to the basal machinery. *Cold Spring Harbor Symp. Quant. Biol.*, **58**, 179–185.

Tansey, W. P. & Herr, W. (1997). TAFs: guilt by association. *Cell*, **88**, 729–732.

Tateno, Y., Fukami-Kobayashi, K., Miyazaki, S., Sugawara, H. & Gojobori, T. (1998). Dna data bank of japan at work on genome sequence data. *Nucleic Acids Res.*, **26**, 16–20.

Tolstrup, N., Rouzé, P. & Brunak, S. (1997). A branch point consensus from *Arabidopsis* found by non-circular analysis allows for better prediction of acceptor sites. *Nucleic Acids Res.*, **25**, 3159–3163.

Tsukiyama, T. & Wu, C. (1997). Chromatin remodeling and transcription. *Curr. Opin. Gen. Dev.*, **7**, 182–191.

Turner, B. M. (1998). Histone acetylation as an epigenetic determinant of long-term transcriptional competence. *Cell. Mol. Life Sci.*, **54**, 21–31.

Uberbacher, E. C., Xu, Y. & Mural, R. J. (1996). Discovering and understanding genes in human DNA sequence using GRAIL. *Meth. Enz.*, **266**, 259–281.

Vazquez, J., Farkas, G., Gaszner, M., Udvardy, A., Muller, M., Hagstrom, K., Gyurkovics, H., Sipos, L., Gausz, J., Galloni, M., Hogga, I., Karch, F. & Schedl, P. (1993). Genetic and molecular analysis of chromatin domains. *Cold Spring Harbor Symp. Quant. Biol.*, **58**, 45–54.

Wasserman, W. W. & Fickett, J. W. (1998). Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.*, **24**, 167–181.

Wasylyk, B. (1988). Transcription elements and factors of RNA polymerase B promoters of higher eukaryotes. *CRC Crit Rev Biochem*, **23**, 77–120.

Widom, J. (1989). Toward a unified model of chromatin folding. *Annu. Rev. Biophys. Biophys. Chem.*, **18**, 365–395.

Widom, J. (1996). Short-range order in two eukaryotic genomes: relation to chromosome structure. *J. Mol. Biol.*, **259**, 579–588.

Wieczorek, E., Brand, M., Jacq, X. & Tora, L. (1998). Function of TAF(II)-containing complex without TBP in transcription by RNA polymerase II. *Nature*, **393**, 187–191.

Wingender, T. H. E., Hermjakob, I. R. H., Kel, A. E., Kel, O. V., Ignatieva, E. V., Ananko, E. A., Podkolodnaya, O. A., Kolpakov, F. A. & Kolchanov, N. L. P. N. A. (1998). Databases on transcriptional regulation: Transfac, trrd, and compel. *Nucl. Acids Res.*, **26**, 364–370.

Wobbe, C. R. & Struhl, K. (1990). Yeast and human TATA-binding proteins have nearly identical DNA sequence requirements for transcription *in vitro*. *Mol. Cell. Biol.*, **10**, 3859–3867.

Wolffe, A. P. (1994). Nucleosome positioning and modification: chromatin structures that potentiate transcription. *TIBS*, **19**, 240–244.

Wolffe, A. P. & Drew, H. R. (1995). DNA structure: implications for chromatin structure and function. In Elgin, S. C. R., (ed.) *Chromatin structure and gene expression*. IRL Press, Oxford, pp. 27–48.

Wolffe, A. P., Khochbin, S. & Dimitrov, S. (1997). What do linker histones do in chromatin? *BioEssays*, **19**, 249–255.

Workman, J. L. & Kingston, R. E. (1998). Alteration of nucleosome structure as a mechanism of transcriptional regulation. *Annu. Rev. Biochem.*, **67**, 545–579.

Zawel, L. & Reinberg, D. R. (1993). Initiation of transcription by RNA polymerase II: a multi-step process. *Prog. Nucl. Acids Res. Mol. Biol.*, **44**, 67–108.

Zhang, M. Q. (1998a). A discrimination study of human core-promoters. In Altman, R., Donker, A. K., Hunter, L. & Klein, T. E., (eds.) *Proc. Pacific Symp. Biocomputing 1998*. World Scientific, Singapore, pp. 240–251.

Zhang, M. Q. (1998b). Identification of human gene core-promoters *in silico*. *Genome Res.*, **8**, 319–326.

Zhu, Z. & Thiele, D. J. (1996). A specialized nucleosome modulates transcription factor access to a *C. glabrata* metal responsive promoter. *Cell*, **87**, 459–470.

Table 1: Servers and software for promoter finding

Detection of pol-II promoters	
[1]Audic/Claverie	Send request to audic@newton.cnrs-mrs.fr
[2]CorePromoter	http://sciclio.cshl.org/genefinder/CPROMOTER/
[3]FunSiteP	http://transfac.gbf.de/dbsearch/funsitep/fsp.html
[4]ModelGenerator/ModelInspector	http://www.gsf.de/biodv/modelinspector.html
[5]PPNN	http://www-hgc.lbl.gov/projects/promoter.html
[6]PromFD 1.0	FTP to beagle.colorado.edu, directory: pub, file: promFD.tar
[7]PromFind	http://www.rabbithutch.com/
[8]Promoter 1.0	http://www.cbs.dtu.dk/services/promoter-1.0/
[9]Promoter Scan	http://biosci.umn.edu/software/proscan/promoterscan.htm
	http://bimas.dcrt.nih.gov/molbio/proscan/
[10]TSSG/TSSW	http://dot.imgen.bcm.tmc.edu:9331/gene-finder/gf.html
Detection of transcription factor binding sites	
[11]MatInd/MatInspector/FastM	http://www.gsf.de/biodv/matinspector.html
	http://www.gsf.de/biodv/fastm.html
[12]MATRIX SEARCH 1.0	Send request to chenq@boulder.colorado.edu
[13]PatSearch 1.1	http://transfac.gbf-braunschweig.de/cgi-bin/patSearch/patsearch.pl
[14]Signal Scan	http://bimas.dcrt.nih.gov/molbio/signal/
[15]TESS	http://www.cbil.upenn.edu/tess/
[16]TFSEARCH	http://pdap1.trc.rwcp.or.jp/research/db/TFSEARCH.html
General genefinders with pol-II detection and other feature detectors (MARs, CpG-islands)	
[17]GENSCAN	http://CCR-081.mit.edu/GENSCAN.html
[18]GRAIL	http://compbio.ornl.gov/Grail-1.3/
[19]MAR-Finder	http://www.ncgr.org/MarFinder/
[20]WebGene	http://itba.mi.cnr.it/webgene/

[1](Audic & Claverie, 1997), [2](Zhang, 1998b), [3](Kondrakhin *et al.*, 1995), [4](Frech *et al.*, 1997), [5](Reese *et al.*, 1996), [6](Chen *et al.*, 1997), [7](Hutchinson, 1996), [8](Knudsen, Submitted), [9](Prestridge, 1995), [10](Solovyev & Salamov, 1997), [11](Quandt *et al.*, 1995), [12](Chen & Stormo, 1995), [13](Wingender *et al.*, 1998), [14](Prestridge, 1997), [15](Schug & Overton, 1997), [16]Has not been published, [17](Burge & Karlin, 1997), [18](Matis *et al.*, 1996; Uberbacher *et al.*, 1996), [19](Singh, 1997), [20](Milanesi *et al.*, 1996; Milanesi & Rogozin, In press).

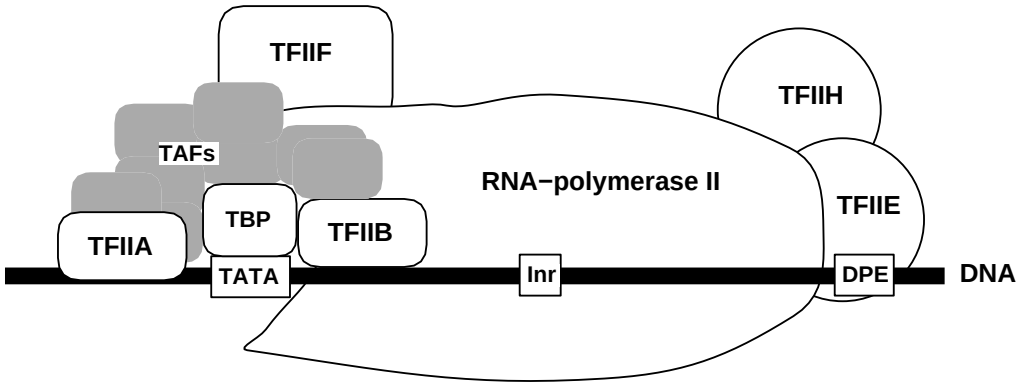
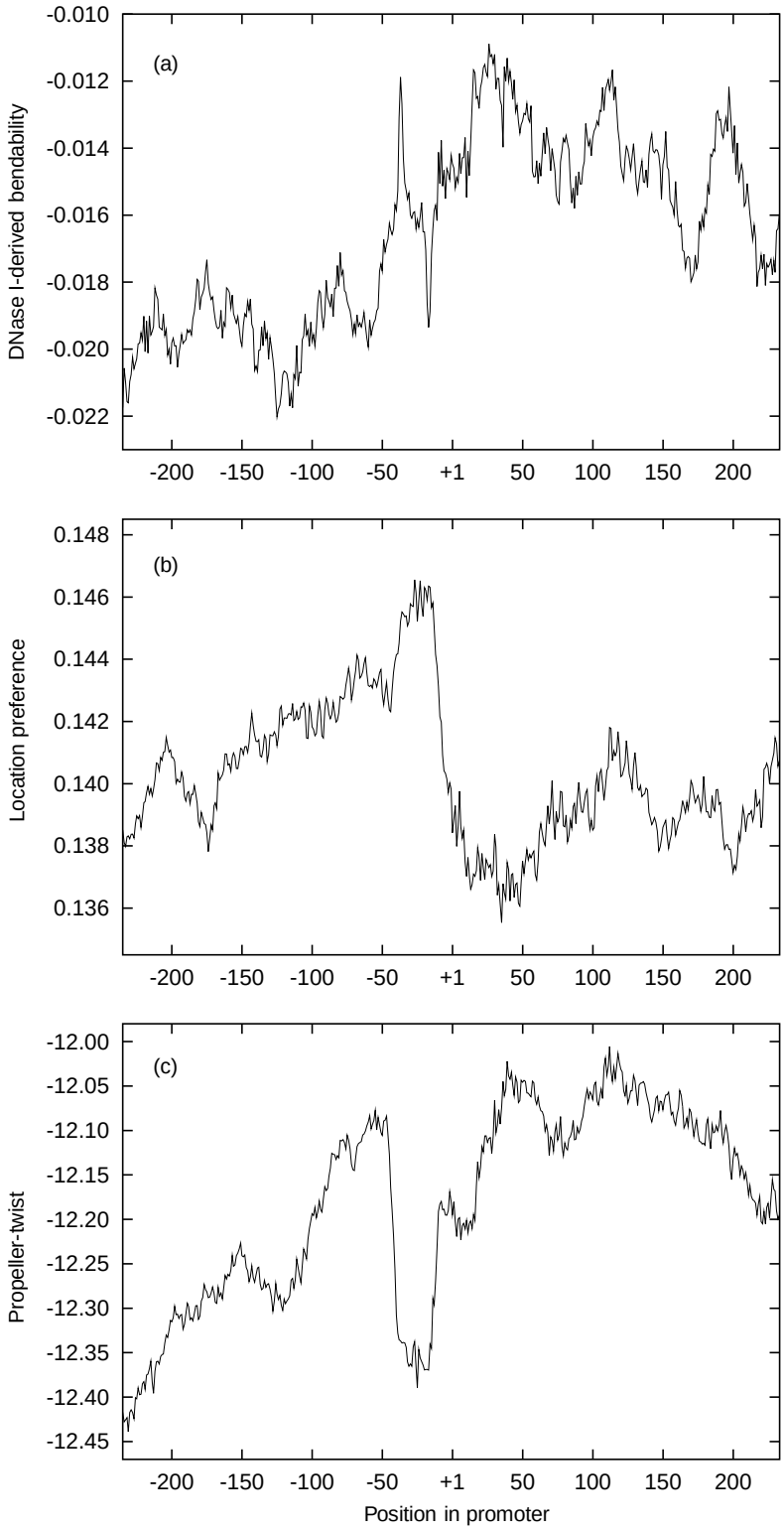


Figure 1: Core-promoter complexed with RNA-polymerase II and general transcription factors. Shown core promoter elements are the TATA-box (TATA, usually around -30), the initiator (Inr, around the start point), and the downstream promoter element (DPE, around $+30$). The DPE is present in some TATA-less, Inr-containing promoters (Burke & Kadonaga, 1997).

Figure 2: (Next page): Average bendability profiles of the human promoter sequences. Position $+1$ corresponds to the transcriptional start point. **(a)**, A non-redundant set of 624 human promoters was aligned using a hidden Markov model, and the average bendability for each position in the promoter calculated using DNase I-derived bendability parameters (Brukner *et al.*, 1995a). Larger values correspond to higher bendability (or propensity for major groove compressibility). The two peaks around position -30 are caused by TATA-box containing promoters. The profile has been smoothed by calculating a running average with a window of size 20. **(b)**, Average flexibility profile calculated from the aligned promoters using a trinucleotide model based on preferred sequence location on nucleosomes (Satchwell *et al.*, 1986). Lower values correspond to more flexible sequences which have less preference for being positioned specifically. **(c)**, Average flexibility profile based on propeller twist values from X-ray crystallography of DNA oligomers (Hassan & Calladine, 1996). Higher (less negative) propeller-twist corresponds to higher flexibility. Profiles **(b)** and **(c)** have been smoothed by calculating a running average with a window of size 30. Note that all three profiles show a tendency for higher flexibility downstream of the transcriptional start point. From (Pedersen *et al.*, 1998).

Figure 2


(a)
DNase I-derived bendability
-0.010
-0.012
-0.014
-0.016
-0.018
-0.020
-0.022
(b)
Location preference
0.148
0.146
0.144
0.142
0.140
0.138
0.136
(c)
Propeller-twist
-12.00
-12.05
-12.10
-12.15
-12.20
-12.25
-12.30
-12.35
-12.40
-12.45
-200 -150 -100 -50 +1 50 100 150 200
Position in promoter