SECTION 1

Introductory principles

1

Genes and their expression

Dirk-Jan Kleinjan

The completion of the human genome project has heralded a new era in biology. Undoubtedly, knowledge of the genetic blueprint will expedite the search for genes responsible for specific medical disorders, simplify the search for mammalian homologues of crucial genes in other biological systems and assist in the prediction of the variety of gene products found in each cell. It can also assist in determining the small but potentially significant genetic variations between individuals. However, sequence information alone is of limited value without a description of the function and, importantly, of the regulation of the gene products. Our bodies consist of hundreds of different cell types, each designed to perform a specific role that contributes to the overall functioning of the organism. Every one of these cells contains the same 20 000 to 30 000 genes that we are estimated to possess. The remarkable diversity in cell specialization is achieved through the tightly controlled expression and regulation of a precise subset of these genes in each cell lineage. Further regulation of these gene products is required in the response of our cells to physiological and environmental cues. Most impressive perhaps is how a tightly controlled program of gene expression guides the development of a fertilised oocyte into a full-grown adult organism. The human genome has been called our genetic blueprint, but it is the process of gene expression that truly brings the genome to life. In this chapter we aim to provide a general overview of the physical appearance of genes and the mechanisms of their expression.

What is a gene?

The realization that certain traits are inherited from our ancestors must have been around for ages, but the study of these hereditary traits was first established by the Austrian monk Gregor Mendel. In his monastery in Brno, Czechoslovakia, he performed his famous experiments crossing pea plants and following a number of hereditary traits. He realised that many of these traits were under the control of two distinct factors, one coming from the male parent and one from the female. He also noted that the traits he studied were not linked and thus must have resided on separate hereditary units, now known as chromosomes, and that some appearances of a trait could be dominant over others. In the early 1900s, with the rediscovery of Mendel's work, the factors conveying hereditary traits were named ''genes'' by Wilhelm Johanssen. A large amount of research since then has led to our current understanding about what constitutes a gene and how genes work.

Genes can be defined in two different ways: the gene as a ''unit of inheritance'', or the gene as a physical entity with a fixed position on the chromosome that can be mapped in relation to other genes (the genomic locus). While the latter is the more traditional view of a gene the former view is more suited to our current understanding of the genomic architecture of genes. A gene gives rise to a phenotype through its ability to generate an RNA (ribonucleic acid) or protein product. Thus the

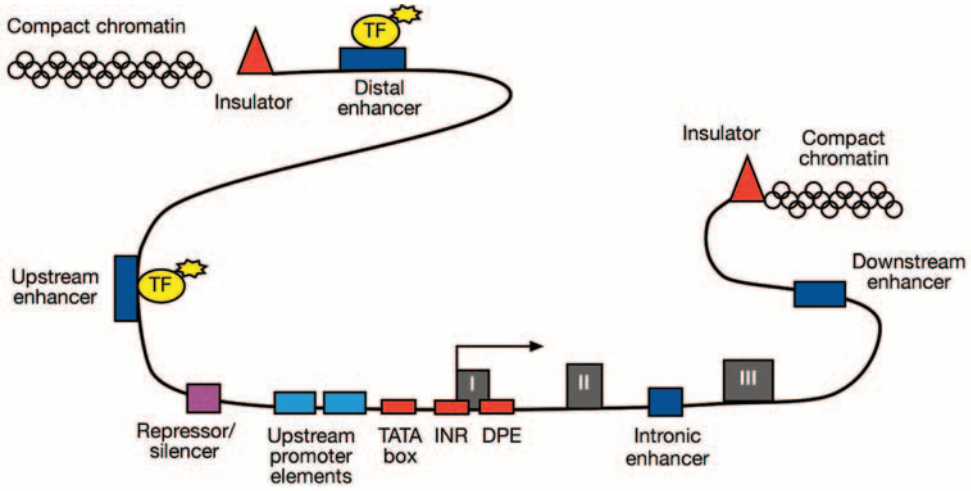

Figure 1.1 The chromosomal architecture of a (fictional) eukaryotic gene. Depicted here is a gene with three exons (grey boxes with roman numerals) flanked by a complex arrangement of *cis*-regulatory elements. The functions of the various elements are explained in the text.

functional genetic unit must encompass not only the DNA (deoxyribonucleic acid) that is transcribed into RNA, but also all of the surrounding DNA sequences that are involved in its transcription. Those regulatory sequences are called the *cis*-regulatory elements, and contain the binding sites for *trans*-acting transcription factors. *Cis*-regulatory elements can be grouped into different classes which will be discussed in more detail later. Recently it has become recognized that *cis*-regulatory elements can be located anywhere on the chromosomal segment surrounding the gene from next to the promoter to many hundreds of kilobases away, either upstream or downstream. Notably, they can also be found in introns of neighboring genes or in the intergenic region beyond the next gene. Crucially, the concept of a gene as a functional genetic unit allows genes to overlap physically yet remain isolated from one another if they bind different sets of transcription factors (Dillon, 2003). As more genes are characterized in greater detail, it is becoming clear that overlap of functional genetic units is a widespread phenomenon.

The transcriptome and the proteome

An enormous amount of knowledge has been gained about genes since they were first discovered, including the fact that at the DNA level most genes in eukaryotes are split, i.e. they contain exons and introns (Berget *et al.*, 1977; Chow *et al.*, 1977) (Figure 1.1). The introns are removed from the RNA intermediate during gene expression in a process called RNA splicing. The split nature of many genes allows the opportunity to create multiple different messages through various mechanisms collectively termed alternative splicing (Figure 1.2). A fully detailed image of a complex organism requires knowledge of all the proteins and RNAs produced from its genome. This is the goal of proteomics, the study of the complete protein sets of all organisms. Due to the existence of alternative splicing and alternative promoter usage in many genes the complement of RNAs and proteins of an organism far exceeds the total number of genes present in the genome. It has been estimated that at least 35% of all human genes show variably spliced products (Croft *et al.*, 2000). It is not uncommon to see genes

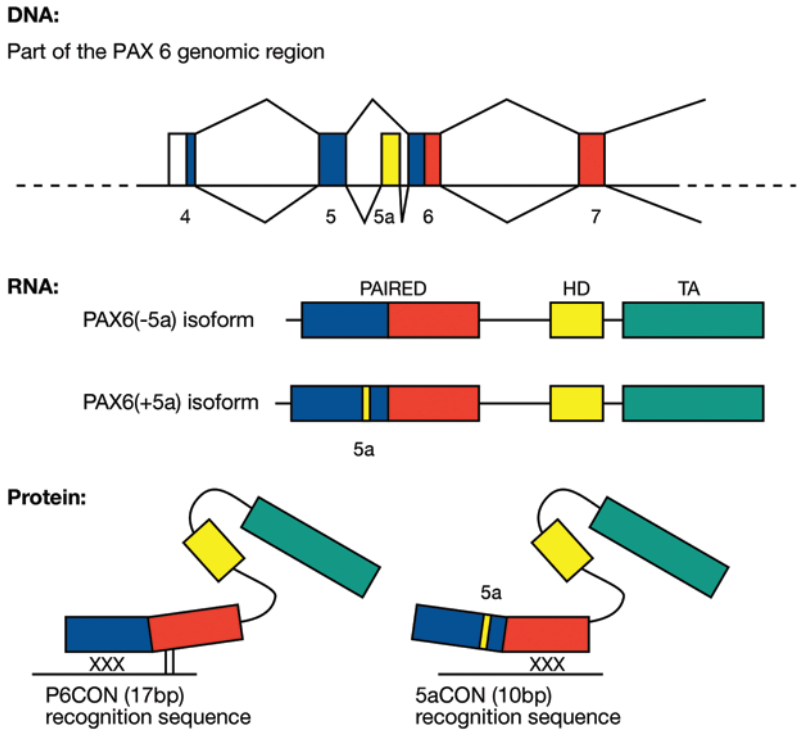

Figure 1.2 The impact of alternative splicing. As an example, part of the genomic region of the *PAX6* transcription factor gene, which has an alternative exon 5a, is shown. The inclusion or exclusion of this exon in the mRNA generates two distinct isoforms, PAX6(+5a) and PAX6(−5a). As a result of the inclusion of exon 5a an extra 14 amino acids are inserted into the paired box (PAIRED), one of its two DNA binding domains, the other being the homeobox domain (HD). The transactivation domain (TA) is also shown. This changes the conformation of the paired box causing it to bind to a different recognition sequence (5aCON) that is found in a different subset of target genes, compared with the –5a isoform recognition sequence (P6CON) (Epstein *et al.*, 1994).

with a dozen or more different transcripts. There are also remarkable examples of hundreds or even thousands of functionally divergent mRNAs (messenger RNAs) and proteins being produced from a single gene. In the human genome such transcript-rich genes include the neurexins, N-cadherins and calcium-activated potassium channels (e.g. Rowen *et al.*, 2002). Thus the estimated 35 000 genes in the human genome could easily produce several hundred thousand proteins or more.

Variation in mRNA structure can be brought about in many different ways. Certain exons can be spliced in or skipped. Introns that are normally excised can be retained in the mRNA. Alternative 5′ or 3′ splice sites can be used to make exons shorter or longer. In addition to these changes in splicing, use of alternative promoters (and thus start sites) or alternative polyadenylation sites also allows production of multiple transcripts from the same gene. (Smith and Valcarcel, 2000). The effect which these alternative splice events can have on the structure of the resulting protein is similarly diverse. Functional domains can be added or left out of the encoded protein. Introduction of an early stop codon can result in a truncated protein or an unstable RNA. Short peptide sequences can be included in the protein that can have very specific

effects on the activity of the protein, e.g. they can change the binding specificity of transcription factors or the ligand binding of growth factor receptors. The inclusion of alternative exons can lead to a change in the subcellular localization, the phosphorylation potential or the ability to form protein–protein interactions. The *DSCAM* gene of *Drosophila* provides a particularly striking example of the number of proteins that can be generated from a single gene. This gene, isolated as an axon guidance receptor responsible for directing axon growth cones to their targets in the Bolwig organ of the fly, has 24 exons. However, 4 of these exons are encoded by arrays of potential alternative exons, used in a mutually exclusive manner, with exon 4 having 12 alternatives, exon 6 having 48 alternatives, exon 9 having 33 alternatives and exon 17 having another 2. Thus, if all of the possible combinations were used, the *DSCAM* gene would produce 38 016 different proteins (Schmucker *et al.*, 2000). This is obviously an extreme example, but it highlights the fact that gene number is not a reliable marker of the protein complexity of an organism. Additional functional variation comes from post-translational modification. Post-translational modifications are covalent processing events which change the properties of a protein by proteolytic cleavage or by addition of a modifying group to one or more amino acids (e.g. phosphorylation, glycosylation, acetylation, acylation and methylation). Far from being mere ''decorations,'' post-translational modification of a protein can finely tune the cellular functions of each protein and determine its activity state, localization, turnover, and interactions with other proteins.

Gene expression

The first definition of the gene as a functional unit followed from the discovery that individual genes are responsible for the production of specific proteins. The difference in chemical nature between the DNA of the gene and its protein product led to the concept that a gene codes for a protein. This in turn led to the discovery of the complex apparatus that allows the DNA sequence of a gene to generate an RNA intermediate which in turn is processed into the amino acid sequence of a protein. This sequence of events from DNA to RNA to protein has become known as the central dogma of molecular biology. Recent progress has revealed that many of the steps in the pathway from gene sequence to active protein are connected. To provide a framework for the large number of events required to generate a protein product we will follow a generalized pathway from gene to protein as follows.

The gene expression pathway usually starts with an initial signal, e.g. cell cycle progression, differentiation, hormonal stimulation. The signal is conveyed to the nucleus and leads to activation of specific transcription factors. These in turn bind to *cis*-regulatory elements, and, through interaction with other elements of the transcription machinery, promote access to the DNA (chromatin remodelling) and facilitate the recruitment of the RNA polymerase enzymes to the transcription initiation site at the core promoter. In eukaryotes there are three RNA polymerases (RNAPs; see also below). Here we will focus on the expression of genes transcribed by RNAPII, although many of the same basic principles apply to the other polymerases. Soon after RNAP II initiates transcription, the nascent RNA is modified at its 5′ end by the addition of a ''cap'' structure. This ^{7Me}G cap serves to protect the new RNA transcript from attack by nucleases and later acts as a binding site for proteins involved in nuclear export to the cytoplasm and in its translation (Proudfoot, 1997). After the ''initiation'' stage RNAP II starts to move 5′ to 3′ along the gene sequence to extend the RNA transcript in a process called ''elongation''. The elongation phase of transcription is subject to regulation by a family of elongation factors (Uptain *et al.*, 1997). The coding sequences (exons) of most genes are interrupted by long non-coding sequences (introns), which are removed by the process of mRNA splicing. When RNAP II reaches the end of a gene it stops transcribing

(''termination''), the newly synthesized RNA is cleaved off (''cleavage'') and a polyadenosine tail is added to the 3′ end of the transcript ('polyadenylation') (Proudfoot, 1997).

As transcription occurs in the nucleus and translation in the cytoplasm (though some sort of translation proofreading is thought to occur in the nucleus, as part of the ''nonsense-mediated decay'' process, see below), the next phase is the transport of the transcript to the cytoplasm through pores in the nuclear membrane. This process is mediated by factors that bind the mRNA in the nucleus and direct it into the cytoplasm through interaction with proteins that line the nuclear pores (Reed and Hurt, 2002). Translation of mRNA takes place on large ribonucleoprotein complexes called ribosomes. It starts with the localization of the start codon by translation initiation factors and subunits of the ribosome and once again involves elongation and termination phases (Dever, 2002). Finally the nascent polypeptide chain undergoes folding, in some cases assisted by chaperone proteins, and often post-translational modification to generate the active protein.

The process of nonsense-mediated mRNA decay (NMD) is increasingly recognized as an important eukaryotic mRNA surveillance mechanism that detects and degrades mRNAs with premature termination codons (PTC+ mRNAs), thus preempting translation of potentially dominant-negative, carboxyl-terminal truncated proteins (Maquat, 2004). It has been known for more than a decade that nonsense and frameshift mutations which induce premature termination codons can destabilize mRNA transcripts in vivo. In mammals, a termination codon is recognized as premature if it lies more than about 50 nucleotides upstream of the final intron position, triggering a series of interactions that leads to the decapping and degradation of the mRNA. Although still controversial, it has been suggested that for some genes regulated alternative splicing is used to generate PTC+ mRNA isoforms as a means to downregulate protein expression, as these alternative mRNA isoforms are degraded by NMD rather than translated to make protein. This system has been termed regulated unproductive splicing and translation (RUST) (Neu-Yilik *et al.*, 2004; Sureau *et al.*, 2001; Lamba *et al.*, 2003).

Transcriptional regulation

As follows clearly from the previous section, the expression of a gene can be regulated at several stages in the process from DNA to protein product: at the level of transcription; RNA stability and export; and at the level of translation or post-translational modification or folding. However, for most genes transcriptional regulation is the main stage at which control of expression takes place. In this section we take a more detailed look at the issues involved in RNAPII transcription.

Promoters and the general transcription machinery

Gene expression is activated when transcription factors bind to their cognate recognition motifs in gene promoters, in interaction with factors bound at *cis*-regulatory sequences such as enhancers, to form a complex that recruits the transcription machinery to a gene. A typical core promoter encompasses 50–100 basepairs surrounding the transcription start site and forms the site where the pre-initiation complex, containing RNAPII, the general transcription factors (GTFs) and coactivators, assemble. The promoter thus positions the start site as well as the direction of transcription. The core promoter alone is generally inactive in vivo, although it may support low or basal levels of transcription in vitro. Activators greatly stimulate transcription levels and the effect is called activated transcription.

The pre-initiation complex that assembles at the core promoter consists of two classes of factors: (1) the GTFs including RNAPII, TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH (Orphanides *et al.*, 1996) and (2) the coactivators and corepressors

that mediate the response to regulatory signals (Myer and Young, 1998). In mammalian cells those coactivator complexes are heterogeneous and sometimes purify as a separate entity or as part of a larger RNAPII holoenzyme. The first step in the assembly of the pre-initiation complex at the promoter is the recognition and binding of the promoter by TFIID. TFIID is a multisubunit protein containing the TATA binding protein (TBP) and 10 or more TBP-associated factors (TAF_{II}s). A number of sequence motifs have been identified that are typically found in core promoters and are the recognition sites for TFIID binding: (1) the TATA box, usually found 25–30 BP upstream of the transcription start site and recognized by TBP, (2) the initiator element, (INR) overlapping the start site, (3) the downstream promoter element or DPE, located approximately 30 BP downstream of the start, (4) the TFIIB recognition element, found just upstream of the TATA box in a number of promoters (Figure 1.1). Most transcriptionally regulated genes have at least one of the above motifs in their promoter(s). However, a separate class of promoter, which is often associated with ubiquitously expressed ''housekeeping genes'', appears to lack these motifs but instead is characterized by a high G/C content and multiple binding sites for the ubiquitous transcription factor Sp1 (Smale, 2001; Smale and Kadonaga, 2003).

RNA polymerases

In eukaryotes nuclear transcription is carried out by three RNA polymerases, I, II and III, which can be distinguished by their subunit composition, drug sensitivity and nuclear localization. Each polymerase is specific to a particular class of target genes. RNAP I is localized in the nucleoli, where multiple enzymes simultaneously transcribe each of the many active 45S rRNA genes required to maintain ribosome numbers as cells proliferate. RNAPs II and III are both localized in the nucleoplasm. RNAP II is responsible for the transcription of protein-encoding mRNA as well as snRNAs and a growing number of other non-coding RNAs. RNAP III transcribes genes encoding other small structural RNAs, including tRNAs and 5S RNA. Each of the polymerases has its own set of associated GTFs.

RNAP II is an evolutionarily conserved protein composed of two major, specific subunits, RPB1 and RPB2, in conjunction with 10 smaller subunits. RPB1 contains an unusual carboxy-terminal domain (CTD), composed in mammals of 52 repeats of a heptapeptide sequence. Cycles of phosphorylation and dephosphorylation of the CTD play a pivotal role in mediating its function as a nucleating center for factors required for transcription as well as cotranscriptional events such as RNA capping, splicing and polyadenylation. Elongating RNAP II is phosphorylated at the Ser2 residues of the CTD repeats.

The manner in which the transcription machinery is assembled at the core promoter remains somewhat unclear. Initial observations seemed to suggest a stepwise assembly of the various factors at the promoter, starting with binding of TFIID to the TATA box. However, more recent research has focussed on recruitment of a single large complex called the holoenzyme. The latter view would certainly simplify matters, as the holoenzyme provides a single target through which activators bound to an enhancer or promoter can recruit the general transcription machinery (Myer and Young, 1998).

Cis-regulatory elements

Gene expression is controlled through promoter sequences located immediately upstream of the transcriptional start site of a gene, in interaction with additional regulatory DNA sequences that can be found around or within the gene itself. The sequences located in the region immediately upstream of the core promoter are usually rich in binding sites for a subgroup of ubiquitous, sequence-specific transcription factors including Sp1 and CTF/NF-I (CCAAT binding factor). These immediate upstream sequences are usually termed the regulatory promoter, while sequences found

at a greater distance are called *cis*-regulatory elements. Together with the transcribed regions of genes, the promoters and *cis*-regulatory elements form the working parts of the genome. It has been estimated that around 5% of the human genome is under evolutionary constraint, and hence may be assumed to contribute to the fitness of the organism in some way. However, less than a third of this functional DNA comprises coding regions, while the rest is made up of different classes of regulatory elements such as promoters, enhancers and silencers (which control gene expression) and locus control regions, insulators and matrix attachment regions (which mediate chromatin organization). There is, as yet, no clear understanding of how exactly promoters interact with the various *cis*-regulatory elements.

Enhancers and repressors

Enhancers are stretches of DNA, commonly spanning a few hundred base pairs that are rich in binding sites for transcription factors, and which have a (usually positive) effect on the level of gene transcription. Most enhancers are tissue- or cell-type specific: in cells with sufficient levels of cognate binding factors *cis*-regulatory elements are often exposed as sites that are hypersensitive to DNaseI digestion. This supposedly reflects a local rearrangement in nucleosome positioning and/or local chromatin topology. During differentiation, hypersensitive site formation at promoters and enhancers usually precedes transcription. Transcriptional activators that bind to the *cis*-regulatory elements of a gene are modular proteins with distinct domains, including ones for DNA binding and transcriptional activation (''transactivation''). The DNA binding domain targets the activator to a specific sequence in the enhancer, while the transactivation domain interacts with the general transcription machinery to recruit it to the promoter. Efficient binding of transcription factors to an enhancer often requires cooperative combinatorial interaction with other activators having recognition binding sites nearby in the *cis*-regulatory element. With such a combinatorial system many layers of control can be achieved with a relatively small number of proteins and without the requirement that all genes be expressed in the same way. It also provides the plasticity required by metazoans to respond to developmental and environmental cues, and it effectively integrates many different signaling pathways to provide a complex regulatory network based on a finite number of transcription factors. Nevertheless, setting up and maintaining a tightly controlled program of gene expression requires a big input from our genetic resource, which is reflected in the fact that more than 5% of our genes are predicted to encode transcription factors (Tupler *et al.*, 2001).

Mechanisms of repression are generally less well understood than activation mechanisms, mainly because they are more difficult to study. Repression can occur in several ways: (1) through inactivation of an activator by post-transcriptional modification, dimerization or the blocking of its recognition site, (2) through inhibition of the formation of a pre-initiation complex, (3) mediated through a specific *cis*-regulatory repressor element and its DNA binding protein(s).

Locus control regions

In general, locus control regions (LCRs) share many features with enhancers, in that they coincide with tissue-specific hypersensitive sites, bind typical transcription factors and confer high levels of gene expression on their gene(s). However, LCRs subsume the function of enhancers along with a more dominant ''chromatin opening'' activity, i.e. they modulate transcription by influencing chromatin structure through an extended region in which they induce and maintain an enhanced accessibility to transcription factors. This activity is dominant such that it can override any negative effects from neighbouring regions. The defining characteristic of an LCR is its ability to drive copy-number-dependent, position-independent expression of a linked gene in transgenic assays, even

when the transgene has integrated (randomly) in a region of highly repressive centromeric heterochromatin (Fraser and Grosveld, 1998).

Boundary elements/insulators

Cis-regulatory control regions such as enhancers and LCRs can regulate gene expression over large distances, in some cases several hundreds of kilobases away (Lettice *et al.*, 2002). However, where necessary, mechanisms must have evolved to prevent the unwanted activation of adjacent gene loci. Mechanisms affecting how the genome manages to set up independent expression domains often invoke the use of insulators or boundary elements. These are *cis*-elements that are required at the borders of gene domains and thought to prevent the inappropriate effects of distal enhancers and/or encroaching heterochromatin. Elements that fit this profile have been identified and have been shown to function in assays as transcriptionally neutral DNA elements that can block or insulate the action of enhancers when located in between the enhancer and promoter. Similarly they can also block the influence of negative effects, such as mediated by silencers or by spreading of heterochromatin-like repression when flanking a reporter gene in certain assays. Examples of well-studied insulators are the *Drosophila* gypsy and scs/scs elements, and in vertebrates the IGF2/H19 DMR (differentially methylated region) and HS4 of the chicken β-globin locus (Bell *et al.*, 2001). All vertebrate insulators that have been analyzed so far require the binding of a protein called CTCF for its function.

Matrix attachment sites

Matrix or scaffold attachment sites (MAR/SARs), are DNA sequences isolated as fragments that remain attached to nuclear structures after stringent extraction with high salt or detergent. They are usually A/T rich and are thought to be the sequences where DNA attaches to the nuclear matrix, thus forming the looped structures of the chromosome that were once thought to demarcate separate gene domains. In some cases, MARs have been shown to coincide with transcriptional enhancers and insulators, however, it remains to be established whether this is coincidental or if MARs have a real function in transcriptional regulation (Hart and Laemmli, 1998).

A current view of enhancer action

To explain how regulatory elements relay information to their target promoters through nuclear space, three models have been proposed: looping; tracking; and linking. The looping model predicts that an enhancer/LCR with its bound transcription factors loops through nucleoplasmic space to contact the promoter where it recruits or activates the general transcription machinery. Initial contact is supposed to occur through random collision while affinity between bound proteins will determine the duration of the interactions. In contrast, in the tracking (or scanning) model transcription factors assemble on the DNA at the enhancer and then move along the DNA fiber until they encounter their cognate promoter. At first view this model explains more easily how insulators located between enhancer and promoter can block the influence of enhancers on transcription. In the linking model, transcription factors bind at a distant enhancer, from where the signal is propagated via a growing chain of proteins along the DNA towards the promoter.

Recently, two novel techniques, 3C-technology (Tolhuis *et al.*, 2002) and RNA-TRAP (Carter *et al.*, 2002) have provided some evidence for a looping model in the regulation of the multigeneβ-globin locus. In these studies, based on the relative levels of cross-linking between various sites within the globin locus in erythroid cells, a spatial clustering of the *cis*-regulatory elements (including the active gene promoters, LCR and other DNAse hypersensitive (HS) sites) was found, with the intervening DNA and the inactive genes in the locus looping out. In brain tissue where the β-globin cluster is not

expressed, the DNA appeared to adopt a relatively straight conformation. These observations have led to the proposal of an active chromatin hub (ACH), a 3-D structure created by clustering of the relevant control elements and bound factors to create a nuclear environment amenable to gene expression. The tissue-specific formation of an ACH would create a mini ''transcription factory'', a local high concentration of transcription factors for the promoter to interact with. It remains to be seen whether ACH formation is a general phenomenon, but it is an attractive model that can explain the existence of distinct, autonomously controlled expression patterns from overlapping gene domains (de Laat and Grosveld, 2003).

Transcriptional regulation and chromatin remodeling

Chromatin structure

While DNA binding proteins and their interactions with the basic synthetic machinery drive transcription, it is now clear that the efficiency and the precision of this process are strongly influenced by higher nuclear organization. The DNA in our cells is packaged in a highly organized and compact nucleoprotein structure known as chromatin (Figure 1.3). This enables the very long strands of DNA to be packaged in a compact configuration in the nucleus. The basic organizational unit of chromatin is the nucleosome, which consists of 146 bp of DNA wrapped almost twice around a protein core, the histone octamer, containing two copies each of four histone proteins: H2A, H2B, H3 and H4 (Luger, 2003). Histones are small, positively charged proteins which are very highly conserved among eukaryotes. The structure created by the DNA wrapped around the nucleosomes is known as the 10 nm fibre, also referred to as the so-called ''beads on a string'' structure. The linker histones H1 and H5 can be found on the DNA in between the beads and assist in further compaction to create less well-defined levels of higher order chromatin folding (e.g. 30 nm fibre). In addition to histones, several other abundant proteins are commonly associated with chromatin, including various HMG proteins and HP1 (specifically at heterochromatin). Visually ''compact'' chromatin such as found at the centromeres is called heterochromatin. Silenced genes are thought to adopt a comparable compact and relatively inaccessible chromatin structure. Expressed genes tend to reside in what is called euchromatin, where genes and their control elements are more accessible to transcriptional activators by virtue of an open structure. Many aspects of chromatin structure are based on interactions between nucleosomal histones and DNA, neighbouring nucleosomes and the non-histone chromatin binding proteins. Most of these interactions involve the N-terminal tails of the core histones, which protrude from the compact nucleosome core and are among the most highly conserved sequences in eukaryotes. Post-translation modifications of the N-termini, in particular of histones H3 and H4, modulates their interaction potential and hence influences the folding and functional state of the chromatin fibre. Three types of modification are known to occur on histone tails: acetylation, phosphorylation and methylation (Spotswood and Turner, 2002).

Chromatin modification and transcription

To activate gene expression, transcriptional activator proteins must bind to and decompact repressive chromatin to induce transcription. To do so they frequently require the cooperation of the diverse family of transcriptional coactivator proteins, as mentioned earlier. The role of these coactivator protein complexes was initially obscure until it was found that many of them carry subunits that have one of two activities: (1) histone acetyl transferase (HAT) activity, or (2) adenosine triphosphate (ATP)-dependent chromatin remodeling activity.