Regulation of Simplex and Complex Transcription

Introduction:

Many eukaryotes are estimated to have 20,000–25,000 genes. Some of these are expressed (transcribed) in all cells all of the time, while others are expressed as cells enter a particular pathway of differentiation or as conditions in and around the cells change. In the early 1980s, transcription researchers primarily explored DNA–protein interactions in vitro. Research focused on the purification of sequences pecific DNA-binding proteins by affinity chromatography, analysis of the transcriptional activity of promoters by reporter gene assays, in vitro transcription assays that allowed the fractionation of the general transcription machinery, and assays such as electrophoretic gel mobility shift assays (EMSA) and DNase I foot printing for analysis of cis-acting DNA elements with trans-acting factors.

By the late 1980s, many sequence-specific DNA-binding proteins had been identified, purified, and their genes cloned. Upon further study, it became clear that in addition to DNA–protein interactions, protein–protein interactions were of critical importance for regulating gene transcription. This insight was followed closely by the realization in the early 1990s that chromatin structure, nuclear architecture, and cellular compartmentalization must also be taken into account. Sections within this chapter will cover protein-coding gene regulatory elements, transcription factors and their DNA-binding motifs, the general transcription machinery and the mechanism of RNA polymerase II transcription, transcriptional coactivators and corepressors, including chromatin modification and remodeling complexes, and signal-mediated nuclear import and export of proteins involved in regulating gene transcription.

Overview of transcriptional regulation:

The most important and widely used strategy for regulating gene expression is altering the rate of transcription of a gene. However, the control of gene expression can be exerted at many other levels, including processing of the RNA transcript, transport of RNA to the cytoplasm, translation of mRNA, and mRNA and protein stability. These additional levels of control are discussed in Chapters 13 and 14. There are also instances where genes are selectively amplified during development and, as a consequence, there is an increase in the amount of RNA transcript synthesized. The ribosomal RNA genes of Xenopus are an example of this form of gene regulation.

The regulation of transcription of protein-coding genes by RNA polymerase II (RNA pol II) will be highlighted. RNA pol II is located in the nucleoplasm and is responsible for transcription of the vast majority of genes including those encoding mRNA, small nucleolar RNAs (snoRNAs), some small nuclear RNAs (snRNAs), and microRNAs. Gene transcription is a remarkably complex process. The synthesis of tens of thousands of different eukaryotic mRNAs is carried out by RNA pol II. During the process of transcription, RNA pol II associates transiently not only with the template DNA but with many different proteins, including general transcription factors. The initiation step alone involves the assembly of dozens of factors to form a preinitiation complex. Transcription is mediated by the collective action of sequence-specific DNA-binding transcription factors along with the core RNA pol II transcriptional machinery, an assortment of coregulators that bridge the DNA-binding factors to the transcriptional machinery, a number of chromatin remodeling factors that mobilize nucleosomes, and a variety of enzymes that catalyze covalent modification of histones and other proteins. Not surprisingly, the transcription literature is replete with a sometimes-bewildering array of acronyms such as TBP, CBP, HDAC, LSD1, and SWI/SNF, to name a few.

There are two other important eukaryotic polymerases – RNA polymerase I and RNA polymerase III. RNA polymerase I resides in the nucleolus and is responsible for synthesis of the large ribosomal RNA precursor. RNA polymerase III is also located in the nucleoplasm and is responsible for synthesis of transfer RNA (tRNA), 5S ribosomal RNA (rRNA), and some snRNAs. Plants have a fourth nuclear polymerase, named RNA polymerase IV, which is an RNA silencing-specific polymerase that mediates synthesis of small interfering RNAs (siRNAs) involved in heterochromatin formation.

Protein-coding gene regulatory elements:

Expression of protein-coding genes is mediated in part by a network of thousands of sequence-specific DNA-binding proteins called transcription factors. Transcription factors interpret the information present in gene promoters and other regulatory elements, and transmit the appropriate response to the RNA pol II transcriptional machinery. Information content at the genetic level is expanded by the great variety of regulatory DNA sequences and the complexity and diversity of the multiprotein complexes that regulate gene expression. Many different genes and many different types of cells in an organism share the same transcription factors. What turns on a particular gene in a particular cell is the unique combination of regulatory elements and the transcription factors that bind them.

Protein-coding sequences make up only a small fraction of a typical multicellular eukaryotic genome. For example, they account for less than 2% of the human genome. The typical eukaryotic protein-coding gene consists of a number of distinct transcriptional regulatory elements that are located immediately 5′ of the transcription start site (termed +1). The regulatory regions of unicellular eukaryotes such as yeast are usually only composed of short sequences located adjacent to the core promoter. In contrast, the regulatory regions in multicellular eukaryotes are scattered over an average distance of 10 kb of genomic DNA with the transcribed DNA sequence only accounting for just 2 or 3 kb. Genes range in size from very small, such as a histone gene that is only 500 nt long with no introns, to very large. The largest known human gene encodes the protein dystrophin, which is missing or nonfunctional in the disease muscular dystrophy. The transcribed sequence is 2.5 million nucleotides in length, including 79 introns. It takes over 16 hours to produce a single transcript, of which more than 99% is removed during splicing to generate a mature mRNA.

Gene regulatory elements are specific cis-acting DNA sequences that are recognized by trans-acting transcription factors Cis-regulatory elements in multicellular eukaryotes can be classified into two broad categories based on how close they are to the start of transcription: promoter elements and long-range regulatory elements. In comparing the regulatory region of a particular gene with another in multicellular eukaryotes, there will be variation in whether a particular element is present or absent, the number of distinct elements, their orientation relative to the transcriptional start site, and the distance between.

Structure and function of promoter elements:

The "gene promoter" is loosely defined as the collection of cis-regulatory elements that are required for initiation of transcription or that increase the frequency of initiation only when positioned near the transcriptional start site. The gene promoter region includes the core promoter and proximal promoter elements. Proximal promoter elements are also sometimes designated as "upstream promoter elements" or "upstream regulatory elements."

Proximal promoter elements:

The regulation of TFIID binding to the core promoter element in yeast depends on an upstream activating sequence (UAS) located within a few hundred base pairs of the promoter. The vast majority of yeast genes contain a single UAS, which is usually composed of two or three closely linked binding sites for one or two different transcription factors. In contrast, a typical multicellular eukaryote gene is likely to contain several proximal promoter elements. Promoter proximal elements are located just 5′ of the core promoter and are usually within 70–200 bp upstream of the start of transcription. Recognition sites for transcription factors tend to be located in clusters. Examples include the CAAT box and the GC box. The CAAT box is a binding site for the CAAT-binding protein (CBF) and the CAAT/enhancer-binding protein (C/EBP). The GC box is a binding site for the transcription factor Sp1. Sp1 was initially identified as one of three components required for the transcription of SV40 early and late promoters. Promoter proximal elements increase the frequency of initiation of transcription, but only when positioned near the transcriptional start site. The transcription factors that bind promoter proximal elements do not always directly activate or repress transcription. Instead, they might serve as "tethering elements" that recruit long-range regulatory elements, such as enhancers, to the core promoter.

Structure and function of long-range regulatory elements:

Protein-coding genes of multicellular eukaryotes typically contain additional regulatory DNA sequences that can work over distances of 100 kb or more from the gene promoter. These long-range regulatory elements are instrumental in mediating the complex patterns of gene expression in different cells types during development. Such long-range regulation is not generally observed in yeast, although a few genes have regulatory sequences located further upstream than the UAS (e.g. silencers of the mating-type locus). The function of many long-range regulatory elements was confirmed by their effect on gene expression in transgenic animals. These elements tend to protect transgenes from the negative o positive influences exerted by chromatin at the site of integration. Long-range regulatory elements in multicellular eukaryotes include enhancers and silencers, insulators, locus control regions (LCRs), and matrix attachment regions (MARs).

Enhancers and silencers:

A typical protein-coding gene is likely to contain several enhancers which act at a distance. These elements are usually 700–1000 bp or more away from the start of transcription. The hallmark of enhancers is that, unlike promoter elements, they can be downstream, upstream, or within an intron, and can function in either orientation relative to the promoter. A typical enhancer is around 500 bp in length and contains in the order of 10 binding sites for several different transcription factors. Each enhancer is responsible for a subset of the total gene expression pattern. Enhancers increase gene promoter activity either in all tissues or in a regulated manner (i.e. they can be tissue-specific or developmental stage-specific). Similar elements that repress gene activity are called silencers.