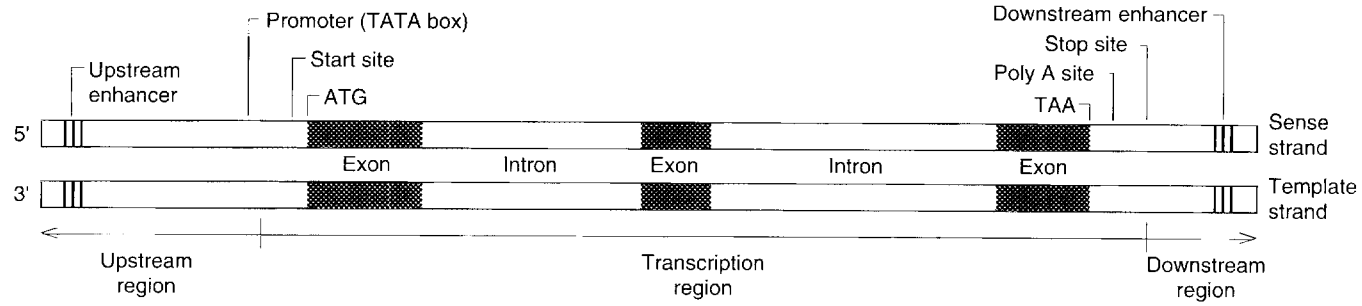
THE ANATOMY OF A EUKARYOTIC PROTEIN GENE



Each eukaryotic gene is placed in one of three classes according to which of the three eukaryotic RNA polymerases is involved in its transcription. The genes for RNAs are transcribed by RNA polymerases I and III. The genes for proteins, the class first brought to mind by the word "gene" and the class focused on here, are transcribed by RNA polymerase II (*pol* II).

Shown above are the components of a prototypic protein gene. By convention the sense strand of the gene, the strand with the sequence of DNA bases corresponding to the sequence of RNA bases in the primary RNA transcript, is depicted with its 5′-to-3′ direction coincident with the left-to-right direction. (Often only the sense strand of a gene is displayed.) The left-to-right direction thus coincides with the direction in which the template strand is transcribed. The terms "upstream" and "downstream" describe the location of one feature of a gene relative to that of another. Their meanings in that context are based on regarding transcription as a directional process analogous to the flow of water in a stream.

The start site is the location of the first deoxyribonucleotide in the template strand that happens to be transcribed. It defines the beginning of the transcription region of the gene. Note that the start site lies upstream of the DNA codon (ATG) corresponding to the RNA codon (AUG) that signals the start of translation of the transcribed RNA. The transcription region ends at some nonspecific deoxyribonucleotide between 500 and 2000 base pairs downstream of the poly A site. Within the poly A site are sequences that, when transcribed, signal the location at which the primary RNA transcript is cleaved and equipped with a "tail" composed of a succession of ribonucleotides containing the base A. (The poly A tail is thought to aid the transport of messenger RNA from the nucleus of a cell to the cytoplasm.) Note that the poly A site lies downstream of the DNA codon (here TAA) corresponding to one of the RNA codons (UAA) that signals the end of translation of the transcribed RNA.

Within the transcription region are exons and introns. Exons tend to be about 300 base pairs long; each is a succession of codons uninterrupted by stop codons. Introns, on the other hand, are not uninterruped successions of codons, and the RNA segments transcribed from introns are spliced out of the primary RNA transcript before translation. A few protein genes contain no introns (the human α–interferon gene is an example), most contain at least one, and some contain a large number (the human thyroglobulin gene contains about forty). Generally the amount of DNA composing the introns of a protein gene is far greater than the amount composing its exons.

Close upstream of the start site is a promoter sequence, where *pol* II binds and initiates transcription. A common promoter sequence in eukaryotic genes is the so-called TATA box, which has the consensus sequence 5′-TATAAA and is located at a variable short distance (about 30 base pairs) upstream of the start site.

The region upstream of the promoter and, less frequently, the downstream region or the transcription region itself contain sequences that control the rate of initiation of transcription. Although expression of a protein gene is regulated at a number of stages in the pathway from gene to protein, control of replication initiation is the dominant regulatory mechanism. (Primary among the other regulatory mechanisms is control of splicing.) The regulated expression of a gene (the when, where, and degree of expression) is the key to phenotypic differences between the various cells of a multicellular organism and also between organisms that possess similar genotypes.

Initiation of transcription is controlled mainly by DNA sequences (*cis* elements) and by certain proteins, many but not all of which are sequence-specific DNA-binding proteins (*trans*-acting transcription factors). Thus both temporal and cellular specificities of transcription control are governed by the availability of the different *trans*-acting transcription factors. Interactions of transcription factors with *cis* elements and with each other lead to formation of complex protein assemblies that control the ability of *pol* II to initiate transcription. Most of the complexes enhance transcription initiation, but some act as repressors. Enhancers and repressors can be located as far as 10,000 base pairs away from the transcription region.

Class I and class III genes differ from protein genes not only in their anatomies but also in the promoters, *cis* elements, and *trans*-acting factors involved in their transcription.