GENES

ADAPTED FROM

GENES & GENOMES: A CHANGING PERSPECTIVE BY M.SINGER & P. BERG

BLACKWELL SCIENTIFC PUBLICATIONS

Gene expression refers to process by which information encoded in DNA structure is read out in RNA and protein products. The expression of all cellular genes begins with the transcription of their nucleotide sequence into RNA. In that process, a region of one of two DNA strands is used as a template to direct the synthesis of an RNA chain by complementary base pairing. Genes that encode structural information for protein yield mRNA, others produce RNAs that are themselves parts of machinery needed to translate mRNA into proteins.

STRUCTURAL FEATURES OF EUKARYOTIC GENES

The structure and organisation of eukaryotic transcriptions units are considerably more complex than those of prokaryotes. Part of this complexity stems from the use of three discrete transcription systems. Each transcription system is considered in detail in subsequent sections, but it is useful to compare them briefly at this point.

Class I genes, which encode the 5,8S; 18S and 28S rRNAs, are transcribed by RNA polymerase I. All mRNAs and a variety of small nuclear RNAs (snRNAs) are derived from transcription of class II genes by RNA polymerase II. The tRNAs, 5S rRNA, and certain small cytoplasmic RNAs (scRNAs) are transcribed by RNA polymerase III from class III genes. Operationally, transcription by the three eukaryotic RNA polymerases can be distinguished by their relative sensitivities to α-amanitin, a poisonous bicylic octapeptide derived from the *Amanita* mushroom. RNA polymerae II is inactivated by a very low concentration (<0.1 μg/ml) of α-amanitin; a higher level (20 μg/ml) is needed to block RNA polymerase III transcription. RNA polymerase I remains active with even 200 μg/ml.

As anticipated, the three different RNA polymerases require different regulatory sequences in order to initiate transcription, and the typical localisation of regulatory sequences relative to the transcription start sites are distinctive for each enzyme. Transcription by RNA polymerase I and II depends in part on nucleotide sequence located within about 100 base pairs (bp) surrounding the 5´end of their respective transcription units. However, these polymerases almost invariably also require transcription factors that bind at specific sequences located nearby or frequently several kilobase pairs (kb) distant from the transcription start sites. Critical sequences required for transcription of most class III genes are located within the coding region, the sequences 5´ to the transcribed segment generally being dispensable or only marginally involved. The sequences that specify the 3´ termini of the respective functional RNA products are also distinctive for each of the RNA polymerase system. Prokaryotic and eukaryotic genes share the same logic in their basic design, but the differences in molecular detail are substantial. A revised definition of a eukaryotic gene may help summarise the essence of these differences. Admittedly, no single definition of a eukaryotic gene could be satisfactory to everyone or to every example. The one we have adopted emerges from the molecular structure of genes have a wide range of functions in diverse eukaryotic organisms. The definition takes account of differing locations and kinds of DNA sequence elements that influence gene expression. It also recognises, implicitly, that phenotypically observable mutations arise from changes in regulatory signals as well as coding sequences.

We define a gene as a combination of DNA segments that together comprise an expressible unit, a unit that results in a formation of a specific functional gene product that may be either an RNA molecule or a polypetide. The DNA segments that define the gene include the following:

1) the **transcription unit** refers to the contiguous stretch of DNA that encodes the sequence in the primary transcript; this include

a) the coding sequence of either the mature RNA or protein product,

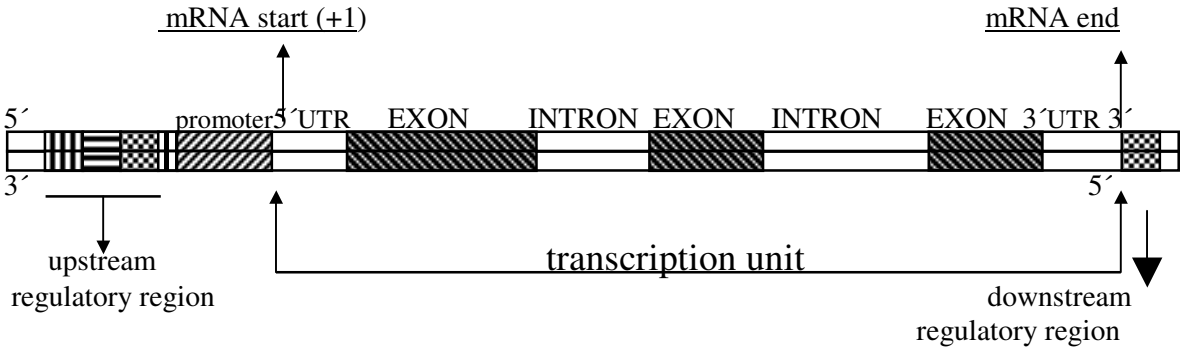
b) the introns,
c) the 5´ and 3´ **UTR**s (untranslated region) that appear in mature mRNAs as well as the spacer sequences that are removed during the processing of primary transcripts of RNA coding genes (class I and III);

2) the minimal sequences needed to initiate correct transcription (the promoter) and to create the proper 3´ terminus of mature RNA;

3) the sequence elements that regulate the rate of transcription initiation; this includes sequences responsible for inducibility and repression of transcription and cell, tissue and temporal specificity of transcription - these regions are so varied in their structure, position, and function as to defy a simple inclusive name - among them are **enhancers** and **silencers** , sequences that influence transcription initiation from a distance, irrespective of their orientation relative to the **transcription start site**.

Neither the DNA sequences that influence a gene´ s configuration within chromatin nor those that regulate its topology are included in our definition of the gene.

The essential features of a prototypical eukaryotic protein coding gene (class II) are depicted above:



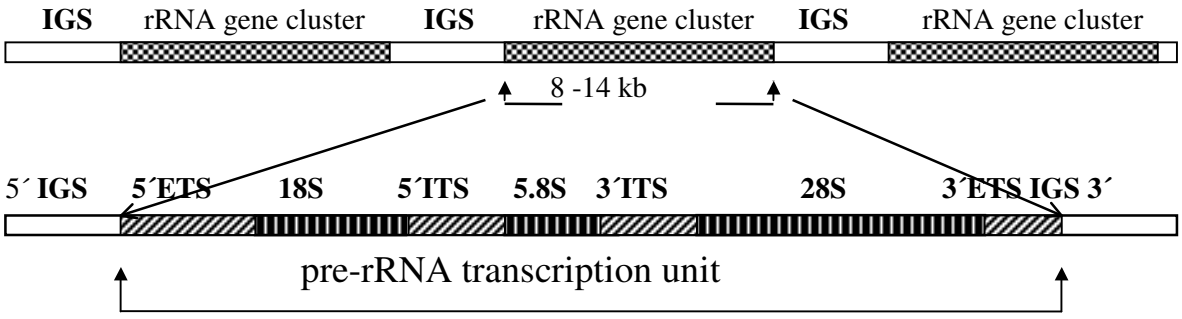
Structural features of a prototypic eukaryotic protein coding gene. The gene´s transcription unit is defined by the transcription start site (bp +1) and the region within which transcription terminates. The latter occurs beyond the end of the mature mRNA sequence, but only rarely is it a specific site. Sequences required for transcription initiation generally lie upstream within 100 bp of +1 but often extend for hundreds or even thousands of base pair. Besides a promoter and upstream regulatory sequences, many genes posses regulatory sequences within introns or beyond the 3´ end of mature mRNA. The regulatory sequences are shown interspersed with different symbols and different shades indicate possible arrangements of distinctive types of regulatory sequences.

In this representation, the beginning of the gene is shown at the left and the end of the gene at the right, consistent with the widespread convention for diagramming transcription from left to right. This means that the DNA strand with the same sequence as the RNA transcript, **the sense strand**, is 5`→ 3´ from left to right and is generally shown as the upper strand. The template for transcription, the **non-sense strand**, is 3´→ 5´ from left to right (lower strand). For convenience the sense strand is often the only one shown. The position of the first nucleotide in the transcript is designed **+1**, and those downstream of +1 (i.e., within the transcription unit) are given positive numbers (e.g. +2, +12, +34 ...). Nucleotides that are upstream of **+1** (nontranscribed sequences) are assigned negative numbers (-4, -12...). By contrast with almost invariable colinearity that exists between prokaryotic genes and respective RNAs they encode, many eukaryotic genes have mosaic structures. In this context, mosaicism refers to the interspersion of coding (**exon**) and non-coding (**intervening** or **intron**) sequence within the transcription unit. Introns occurs most frequently in genes that encode proteins and tRNAS and more rarely in rRNAs. With exceptions of some genes encoding the five

histones, α and β **interferons**, and several mammalians virus proteins, all known vertebrate protein coding genes contains introns. Introns vary in size, number and location from one gene to another. Nevertheless, the same genes in different species often have the same number of introns at analogous position, although the length of the introns and their nucleotide sequences may differ markedly. Generally, the number of introns per genes increases proportionally to the length of the protein coding sequence, and the exon sizes tend to be 300bp, on average. Overall, the total length of introns sequences exceeds the total length in exons, frequently by two to ten times, but occasionally by much more than ten times.

STRUCTURE AND EXPRESSION OF CLASS I GENES

The transcription of class I genes by RNA polymerase I accounts for almost half transcriptional activity of most eukaryotic cells. The sole product of this transcriptional process is a precursor of ribosomal RNA (pre-rRNA) that is processed by sequencial cleavages to the mature 5.8S; 18S and 28S rRNA species. The fourth rRNA species - 5S - is encoded by a class III gene. The number of genes encoding rRNA ranges between a hundred and several thousand, depending upon eukaryotic species. They are located in one or few specific chromosomes (human chromosome 13p; 14p; 15p 21p and 22p) at the morphologically distinctive region called **nucleolar organisers**. During interphase, these regions are incorporated into **nucleoli** , structures which rRNAs are actively transcribed, pre-rRNAs are processed, and ribosomes are assembled. The rRNA genes of virtually all eukaryotic organisms are arranged in long " head- to-tail" repeats. In all species, each rRNA transcription unit encodes 18S, 5.8S and 28S rRNAs arranged in that order starting from 5´ end. The three rRNA coding sequences are both flanked and separated by transcriber spacer segments. These are called, respectively, the **external transcribed spacer** (5´ and 3´**ETSs**) and **internal transcribed spacer** (5´ and 3´**ITSs**). The region between tandemly repeated transcription units ranges between a few kb to nearly 30 kb in length and was originally referred to as the nontranscribed spacer (NTS). Because it is now evident that the so-called nontranscribed spacers are transcribed and that transcriptions in this region are a relevant if not an essential feature of the regulation of rDNA expression, it is more appropriate to use the term **intergenic spacer (IGS)** for this region**.** During transcription, the nascent pre-rRNA continuously associates with the corresponding ribosomal proteins and becomes methylated at specific base and ribose residues. After completion of the transcript, nucleolar endonucleases cleave the pre-rRNA in the nucleoprotein complex, first at the 5´ end of the 5.8S rRNA sequence and subsequently at the 5´ end of the 18S rRNA and the 3´ end of the 28S rRNA, respectively. Subsequent endonucleolytic cleavages produce the three mature sized rRNAs. The RNA segments corresponding to the transcribed spacers are rapidly destroyed and do not accumulate.



Arrangement of eukaryotic rRNA gene cluster. Each rRNA gene cluster about 8 to 14 bp in length, depending on the specie, is separated from another by intergenic spacer (IGS) which is highly variable in length but relatively constant within species. Each rRNA gene cluster constitute a single transcription unit consisting of the 5´ external transcribed spacer (5´ETS) 18S rRNA coding sequence, 5´internal transcribed spacer (5´ITS) 5.8 S rRNA coding sequence, 3´ITS, 28S rRNA coding sequence and 3´ETS.

STRUCTURE AND EXPRESSION OF CLASS II GENE

Class II genes are transcribed in the nucleus by RNA polymerase II. RNA polymerase II, like the others RNA polymerases, requires an array of additional proteins to form a functional transcriptional complex. Many are DNA binding proteins that recognise one or more of the sequence elements that together constitute the gene´s promoter. Some transcription factors appear to function through protein-protein interactions with others factors. By their interaction with the different DNA sequence motifs and with each other, the various transcription factors form complex proteins assemblies that regulate the ability of RNA polymerase II initiate transcription. Most of the complexes act positively to increase transcription initiation, but some are known to behave negatively and thereby act as repressor. In sense, RNA polymerase II provides the transcribing machine, and the interactions of the factors with the regulatory signals in the DNA, and with each other, determine where, when and how fast the machine operate.

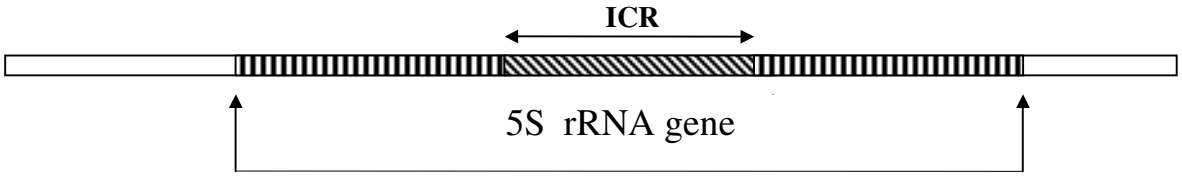
Class II genes encode all cytoplasmic mRNA and a variety of small nuclear RNAs. All transcripts of class II gene have characteristic modification (caps) at their 5´ ends (7methyl guanosine in mRNA). The methylated guanosine in the cap is linked to the first nucleotide of the transcript by a triphosphate bridge between their respective 5´ position. Almost all mRNA have characteristic polyadenylate (poly A) tails beginning varying distances beyond the end of their coding sequences the poly stretch is not encoded in the genes from which the mRNA are transcribed but is added in a separate postranscriptional reaction. Irrespective of whether there are specific terminations signals, transcription most frequently passes through polyadenylation sites. Consequently, the 3´ polyadenylated ends must be created by endonucleolytic scission followed by polimeration of adenylate residues onto the 3´ hidroxyl group created at the cleavage site. Two relatively closely spaced sequence are needed to specify the cleavage-polyadenylation site. One is the virtually invariant **AATAAA** sequence located 10 to 30 bp upstream of a **CA** dinucleotide, within or near the site at which cleavage and polyadenylation frequently occur. The nuclear mRNAs of most vertebrates have poly A tails between 200 and 300 nucleotides long. By contrast many histone mRNAs lack poly A tails. These modification are probably made cotranscriptionally or posttranscriptionally, but their physiological function is unknown. Except in rare instances mRNAs are derived from primary transcripts whose length and sequence are collinear with the DNA from which they are transcribed. Aside from capping and cleavage at 3´ end preliminary to polyadenilation, the principal processing step needed for the formation of functional mRNA is splicing.

The coding sequences of eukaryotic genes (**exons**) are frequently interrupted by noncoding stretches of DNA (**introns**). The frequency of split genes varies greatly among eukaryotic species. They are most prevalent in plants, animals and the viruses that infect them. All introns are transcribed as part of precursor RNAs and subsequently removed by a cleavage-ligation process called **splicing.** The introns in nuclear protein coding genes vary in size from 100 bp to well over 10 kb. Introns from corresponding genes in vertebrates species can be as dissimilar from each other in length and sequence as two introns from unrelated genes. The most distintive common features associated with introns are sequences at their 5´ (upstream or donor) and 3´ (dowstream or acceptor) borders. The

first two nucleotides at the 5´ end of the intron in RNA are virtually always **GU** , the next four are not invariant, although the sequence $^{A}_{G}$AGU appears to be consensus. The 3´ terminus of intron invariably ends with **AG**. The occurrence of consensus splice site sequence does not always predict that a intron can be spliced. Sometimes either or both of these splice site sequences occur within exons and introns at positions where splicing normally does not take place. However, such cryptic splice sites do function under certain circumstances. (e.g. when the authentic sites are altered or missing). Sometimes existing splice sites are not used. For example, retroviral genomes are derived from transcripts of the proviral DNA without splicing, but production of mRNA that encode certain viral proteins requires splicing. In this instance, the same transcripts may be spliced or remain unspliced.

STRUCTURE AND EXPRESSION OF CLASS III GENES

The RNAs transcribed from class III genes do not encode proteins. Instead, their products are small RNAs that are involved in protein synthesis (tRNAs and 5S rRNA), in intracellular protein transport (7SL RNA) and posttranscriptional processing (U6 RNA), beside others whose physiological roles are unknown. Class III genes fall in two groups. In one, which encodes the 5S rRNA, the primary transcript has the same length as the mature functional RNA; therefore, processing of the transcript is not required. The second group encodes tRNAs whose mature form are produced by removing nucleotides from both the 5´ and the 3´ ends of primary transcript by specific endonucleases. The 3´ terminal **- CCA -** sequence, which is required for the function of all tRNAs, is not encoded in the gene and is added by a special enzyme.. Because some tRNA genes contain an intron, splicing is required to produce their mature tRNA products. The signal governing transcription initiation and termination of various kind of class III genes are highly conserved among eukaryotic organisms. A distinctive feature of the sequence that control transcription initiation of many class III genes is their location within a the coding region . These **internal control region (ICR)** vary somewhat from one class III gene to another.



Structure of 5S rRNA gene´s. The internal control region (ICR) is composed of an A-box, an intermediate segment, and a C-box.