Gene Structure

Jörg Bungert, PhD

Email: jbungert@ufl.edu
Phone 352-273-8098
Objectives

• Know the differences in promoter and gene structure between prokaryotes and eukaryotes.
• Know that some eukaryotic genes have alternative promoters and alternative exons.
• Understand the role of DNA methylation and insulator function in the imprinted expression of H19/IGF2.

Reading: Lodish 7th edition, chapter 6 (pp 225-232), chapter 6 (pp. 263-266).
Overview Gene Structure

- Prokaryotic Genes are intronless and are often organized in operons that encode for polycistronic RNAs encoding multiple proteins.
- Eukaryotic Genes are monocistronic and often split containing exons and introns, which are removed after transcription from the pre-mRNA.
- Prokaryotic genes are regulated by DNA elements located relatively close (within 200bp) to the genes or operons.
- Eukaryotic genes are often regulated by combination of DNA elements that are located close to the genes (promoters and upstream regulatory sequences) or located far away (enhancers and locus control regions).
- The prokaryotic genome is accessible whereas the eukaryotic genome is packed into chromatin.
Gene and Promoter Structure in Bacteria

(a) Prokaryotes

E. coli genome

trp operon

Start site for trp mRNA synthesis

Transcription

trp mRNA 5’ 3’

Start sites for protein synthesis

Translation

Proteins

Bacterial Promoter

TTGACA

TATAAT

extended -10 +1

UP-element

-35 -10 +1
Gene Structure in Yeast

(b) Eukaryotes

Yeast chromosomes

IV

TRP1

TRP2

V

TRP5

VII

TRP3

XI

Transcription and RNA processing

trp mRNAs

Translation

Proteins

3 2 5 4
Gene Structure in Higher Eukaryotes

β-Globin genomic DNA
Start site for RNA synthesis
Poly(A) site

Primary RNA transcript
3’ cleavage and addition of poly(A) tail

β-Globin mRNA

Exon
Intron
UTR
m7Gppp

(A)ₙ
Poly(A) tail
Intron excision, exon ligation
Gene Structure in Higher Eukaryotes

(a) Simple transcription unit

Gene

Control regions

Exon 1
Exon 2
Exon 3

Cap site

Poly(A) site

mRNA

5’

3’
Gene Structure in higher Eukaryotes

(b) Complex transcription units

Gene

Exon 1
Exon 2
Exon 3
Exon 4

mRNA₁ 5’

or

mRNA₂ 5’

3’

Gene

Exon 1
Exon 2
Exon 3

mRNA₁ 5’

or

mRNA₂ 5’

3’

Gene

Exon 1A
Exon 1B
Exon 2
Exon 3

mRNA₁ 5’

or

mRNA₂

3’
Alternative Exons in the Fibronectin Gene

Alternative Promoters in the GATA-1 Gene Locus

P1: Testis      P2: Hematopoietic Cells
RNA Pol I and III Gene and Promoter Structure

A. rRNA Repeats

B. Components in the IGS

Pre-rRNA promoter DNA

+1

Box A

Box B

Box C

tRNA Gene

5s RNA Gene
Prokaryotic and yeast genes are normally regulated by cis-elements that are located in relative close proximity (200 bp) to the gene. Higher eukaryotic cells often utilize DNA regulatory elements that can be located far away from the genes, either upstream or downstream, or even within introns of genes.
Organization of Genes within Chromosomes: Drosophila Hox Gene Cluster

The Hox genes encode for transcription factors that are involved in controlling functions required for body axis formation. During mammalian development, qualitative and quantitative differences in Hox gene expression leads to various vertebrate morphologies. It is very important that the Hox genes are transcribed at the right levels both temporally and spatially. Mutations in Hox genes lead to homeotic transformations, characterized by the transformation of one body part into a different one (for instance ultrabithorax in drosophila). The Hox genes are organized into clusters and the order of the genes reflect the order of expression along the body axis, meaning that the most 5’ genes of the clusters are expressed in posterior parts of the body.

Allison, Fundamental Molecular Biology (2011)
The ENCODE Project Consortium
(Encyclopedia of DNA Elements)
(see Nature 447(7146):799-816; 2007)

Biological complexity revealed by ENCODE.
(A) Representation of a typical genomic region portraying the complexity of transcripts in the genome. (Top) DNA sequence with annotated exons of genes (black rectangles) and novel TARs (hollow rectangles). (Bottom) The various transcripts that arise from the region from both the forward and reverse strands. (Dashed lines) Spliced-out introns. Conventional gene annotation would account for only a portion of the transcripts coming from the four genes in the region (indicated). Data from the ENCODE project reveal that many transcripts are present that span across multiple gene loci, some using distal 5' transcription start sites.

(B) Representation of the various regulatory sequences identified for a target gene. For Gene 1 we show all the component transcripts, including many novel isoforms, in addition to all the sequences identified to regulate Gene 1 (gray circles). We observe that some of the enhancer sequences are actually promoters for novel splice isoforms. Additionally, some of the regulatory sequences for Gene 1 might actually be closer to another gene, and the target would be misidentified if chosen purely based on proximity.

Fr. Genome Research 17:669; 2007
Example of a complex locus: The **SNURF-SNRPN** Locus

The SNURF-SNRPN locus represents a genomic region that is imprinted, meaning that the genes located in this locus are expressed from only one of the two alleles. The SNURF/SNRPN genes are expressed from the chromosome inherited from the father (paternal), whereas the UBE3A gene is expressed only on the chromosome inherited from the mother (maternal). Imprinting is regulated by epigenetic mechanisms involving methylation of CpGs, which are often associated with a repressed chromatin configuration.
The H19/IGF2 gene locus is another example for an imprinted gene locus. The H19 gene is expressed from the maternal allele while the IGF2 gene is expressed from the paternal allele. Gene expression is regulated by an enhancer element located downstream of the H19 gene and an imprinting control region (ICR) located between the H19 gene and the IGF2 gene. The ICR functions as an insulator (enhancer blocker) in the maternal allele thus preventing the enhancer from activating the IGF2 gene. The insulator function is mediated by the protein CTCF. On the paternal allele the ICR is methylated at CpG dinucleotides, which prevents binding of CTCF. The ICR no longer functions as an insulator on the paternal allele and the enhancer activates IGF2 gene expression. Methylation of the ICR represses H19 gene expression.
The β-globin genes encode subunits of Hemoglobin and are almost exclusively expressed in erythroid cells. Different genes in the β-globin gene locus are expressed in a developmental-stage specific manner.
Methylation of DNA in Higher Eukaryotes at CpG sites

DNA methylation occurs at cytosine within the sequence ‘CG’

Catalyzed by DNA methyltransferases

Found in mammals, plants, bacteria, & insects

~1% of all cytosines are methylated in mammals

DNA methylation patterns are precisely replicated with the DNA sequence

DNA methylation is usually associated with transcriptionally repressed genes

When unmethylated DNA becomes methylated, any transcribed genes associated with that region are usually shut off
Fundamental Difference in the Organization of Prokaryotic versus Eukaryotic Genomes

Prokaryotic DNA is circular and accessible

Eukaryotic DNA is linear and packed in chromatin; accessibility is restricted and regulated

DNA binding proteins
Functional Nuclear Architecture: CT-IC Model

a) Chromosome territory (CT) with loops containing active genes (red) folding into the interchromatin compartment (IC).

b) Separate arm domains of chromosomes (short and long arms, p and q; here red and green) as well as centromere (asterisk). Inset shows active genes (white) located away from centromere and repressed genes (black associated with centromere).

c) CTs have variable chromatin densities (dark brown, dense; yellow, less dense)

d) Late replicating, gene poor chromatin domains (red) are located at the nuclear periphery or associated with the lamina of the nucleolus. Early replicating gene rich domains (green) are in the nuclear plasma.

e) Active genes (white dots) are at the surface of convoluted chromatin fibers. Inactive genes are located towards the interior.

f) Protein complexes (orange dots) in the IC (green) for transcription, splicing, DNA-repair and DNA replication.

g) IC (green) expanding between ~1Mb CT domains. Inset: Finest branches of IC and ~100kb chromatin domains. Active genes (white), at surface.

Two ~1 Mb chromatin domains or “subcompartments” linked by a chromatin fiber (a). Each ~1Mb domain is built up as a rosette of looped ~100kb chromatin fibers. Loops are held together by loop base spring (CT anchor proteins). Three-dimensional models of the internal ultrastructure of a~ 1Mb chromatin domain is shown in c) and d). The nucleosome chain is compacted into a 30nm fiber, which becomes unfolded in small regions exposing individual nucleosomes (white dots in b). The arrow points to a red sphere, with a 30nm diameter representing transcription factor complexes. Nucleosomal model of 100Kb subdomain is shown revealing random-walk (zig-zag) nucleosome chains, one domain is shown in open configuration (yellow).
Eukaryotic DNA is anchored at a chromosome scaffold Matrix Attachment Region (MAR)
Chromosome Scaffold

Loops of DNA

Protein scaffold