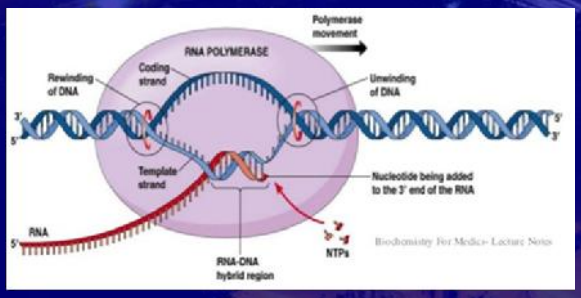
http://smtom.lecture.ub.ac.id/
Password:

https://syukur16tom.wordpress.com/
Password:

Lecture 12.
PROTEIN SSYNTHESIS: TRANSCRIPTION



http://image.slidesharecdn.com/transcription-140308034234-phpapp02/95/dna-transcription-part1-27-638.jpg

LEARNING OUTCOMES

Students, after mastering materials of the present lecture, should be able

1. to explain genetic information and its relation to plant growth and development.
2. to explain genome structure, annotation and organization.
3. to explain transcription of genetic information including RNA polymerase, transcriptional machinery, promoters, and control of transcription factors.

LECTURE OUTLINE

I. INTRODUCTION
1. Definition
2. Genome Size

II. GENOME ORGANIZATION
1. Basic Structure
2. Genome Annotation
3. Genome Organization

III. TRANSCRIPTION
1. Definition
2. RNA Polymerase
3. Transcriptional Machinery
4. Promoters
5. Control of Transcription Factors

I. INTRODUCTION

1. Definition

- The essence of every characteristic of a species, from its morphology to timing of senescence, resides in information (genes) stored in its DNA (genome).
  - A genome is an organism's complete set of DNA, including all of its genes.
  - Each genome contains all of the information needed to build and maintain that organism. In humans, a copy of the entire genome—more than 3 billion DNA base pairs—is contained in all cells that have a nucleus.
- Protein synthesis is the main process by which genomes control characteristics of species.
- Protein synthesis is one of the most fundamental biological processes by which individual cells build their specific proteins.

- Nuclear DNA contain most of an organism's genes, but plastids and mitochondria also contain DNA which encodes some of the gene products required for organelle function and reproduction.
- Traits arising from organellar genes can be clearly recognized by their uniparental inheritance patterns. It is based on the fact that in most species mitochondria and chloroplasts are inherited only through the maternal contribution to the seed.
  - Several traits encoded by organellar genes are in commercial use in plant crops, such as cytoplasmic male sterility in sorghum (Sorghum sp.) and maize (Zea mays).
- In the 1950s, it was found that pieces of genetic material could move from place to place in the genome. These mobile genetic elements were then found to be ubiquitous components of genomes.

2. Genome Size

- Complete genome sequences for a large number of microorganisms and complete, or nearly complete, sequences for numerous higher eukaryotes including many plant species have been obtained through large-scale, relatively inexpensive DNA sequencing.
- The first complete plant genome sequence, that of Arabidopsis thaliana, was published in 2000.
- The size of the nuclear genome varies widely among organisms, with haploid DNA content (C value) ranging from 10^7 to 10^{11} base pairs (bp).
- The human genome is in the middle of this range, at 3×10^9 bp. It has been assumed that the apparent complexity of an organism correlates roughly with genome size, and genomes of humans>most insects >fungi, but this correlation is not universal.

- For example, some amphibians have genomes almost 50 times larger than that of humans, and cartilaginous fish generally have larger genomes than bony fish.
- Plant genomes are represented throughout the size range, with one of the smallest known plant genomes belonging to Arabidopsis thaliana, and one of the largest to a member of the lily family, Fritillaria assyriaca (Fig. 9.2).
- In the mid-20th century, the term C-value paradox was coined to describe the lack of a direct relationship between genome size and organismal complexity.
- With the sequence of many genomes in hand, the sequence features of different classes of DNA that contribute to the paradox are known.

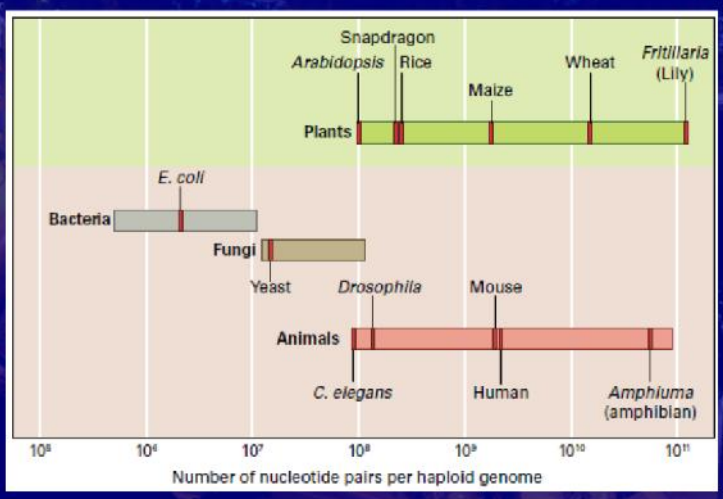


FIGURE 9.2 C values (haploid genome size in base pairs) from various organisms. Most eukaryotes have haploid genome sizes between 10^7 and 10^{11} bp of DNA.

- However their significance and contribution to organismal complexity are still not well understood.
- In plants, at least, some of the size variation can be attributed to one or more rounds of genome duplication, and often much more to the expansion of families of (retro) transposons, as well as to present-day polyploidy (see Section 9.3).
- DNA sequencing has provided more accurate estimates for the sizes of plant genomes.
- Estimation of DNA content in isolated nuclei stained with a fluorescent dye has become accepted as a reliable and quite precise measure of genome size.
- However, DNA sequence data of the many plant and animal genomes are not available yet.

II. GENOME ORGANIZATION

1. Basic Structure

- Before the advent of DNA sequencing, one method for studying a genome's sequence composition was Cot analysis, a technique used to measure the amount of repetitive DNA in a sample (see Box 9.2).
- A genome, based on Cot analysis, could be divided into three components: highly repetitive, intermediately repetitive, and single-copy DNA.
- These broad categories, even with the large amounts of DNA sequence now available, still provide a useful framework for thinking about plant genome structure, although the boundaries between each of these categories are fluid.

- By definition, single-copy DNA is present in only one copy per haploid genome; in practice, this is now considered low copy DNA.
- This class of DNA sequences contains almost all of the gene-coding sequences in a genome. However, in most plant genomes more than half of the genes belong to gene families of various sizes.
- Thus, depending upon the criterion of sequence similarity being considered (e.g., >80% nucleotide identity), closely related genes would hybridize in a reassociation experiment and align with high fidelity in an in silico analysis, and so would no longer be classified as single copy.

2. Genome Annotation

- Genome annotation is the process by which regions of the genome (DNA sequence) are recognized as encoding genes and the structure of the gene is defined (exons, introns, and other features) and its function inferred. Two general methods are used to define gene structures.
- One method is based on available sequences of mRNAs (sequenced as cDNA) and proteins (generally inferred from nucleotide sequences, but sometimes also from direct protein sequencing).
  - Various algorithms can be used to align pre-existing sequences to a new genome, and many of these algorithms can also recognize conserved splice sites (GT….AG; so-called spliced alignment) to define exon-intron boundaries.

- The most accurate gene structures are defined by full-length cDNAs that have captured the entire protein coding sequence of a particular gene.
  - Alignments of nucleotide and protein sequences from other species are also informative, but become progressively less so as the evolutionary distance between the new genome and available sequence increases.
- Overall alignments between two related genomes can also be highly informative, as sequences that are conserved between two genomes are thought to have functional significance and, thus, to encode either genes or regulatory sequences.
  - **A sequence alignment** is a way of arranging the **sequences** of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the **sequences**.

A sequence alignment, produced by ClustalO, of mammalian histone proteins. Sequences are the amino acids for residues 120-180 of the proteins.

Residues that are conserved across all sequences are highlighted in grey. Below the protein sequences is a key denoting conserved sequence (*), conservative mutations (:), semi-conservative mutations (.), and non-conservative mutations ( ). Source: *Clustal FAQ #Symbols,* 2014

- The second method for defining gene structures is strictly computational and relies on differences in base composition and adjacent nucleotide frequencies (di-, tri-, tetranucleotides, etc.) and their transition probabilities between exons, introns, and intergenic sequences.

- These sophisticated programs frequently involve hidden Markov models (HMMs; Box 9.5) and are generally trained on a small, well-defined set of structures that have been identified in the new genome (or that of a close relative) using existing sequences.

- These two methods are used to identify gene structures throughout a new genome, both in regions where there is supporting evidence from nucleotide and protein sequences and in those where there is not.
  - Gene structures supported only by computational methods are considered unreliable predictions in some plant genomic communities.
  - Prediction of gene structures (Box 9.6) results in the definition of a protein coding sequence (CDS) for each gene. Assignment of function to these genes then relies upon their matches to other genes in databases such as GenBank or SwissProt.

3. Genome Organization

- The organization of the Arabidopsis genome provides an excellent model and starting point to understand the overall structure of many plant genomes.
- The genome sequencing project produced approximately 105 Mb of DNA sequence, with the unsequenced portions around the centromeres estimated to be ≈15 Mb, for a total genome size of 120 Mb.
- Subsequent measures using flow cytometry have suggested that the amount of unsequenced DNA was larger, with a total genome size of ≈150 Mb, but this does not affect our description of the basic genome features, which are found in most of the other plant genomes.

- As with a number of plants and animals, the small genome size of Arabidopsis correlates with its short life cycle: Plants with larger genomes tend to reproduce more slowly.
- The main features of the chromosome can now be described at the DNA sequence level.
  1. **Most genes are found in the euchromatic chromosome arms**.
     - The chromosome arms stretching from the telomeres to the pericentromeric regions close to the centromeres comprise ≈100 Mb of sequence and are more or less uniformly gene rich.

     - The Arabidopsis genome encodes approximately 30,000 protein-coding genes, as well as many transfer RNA (tRNA) genes and various classes of noncoding RNA genes with a variety of regulatory functions.
     - These numbers continue to increase as more small genes (both protein-coding and regulatory) are discovered.

- With few exceptions, only one strand of any particular region of the double helix actually encodes a protein. Thus, protein-coding genes may lie on either strand and be transcribed in either direction (Fig. 9.3).

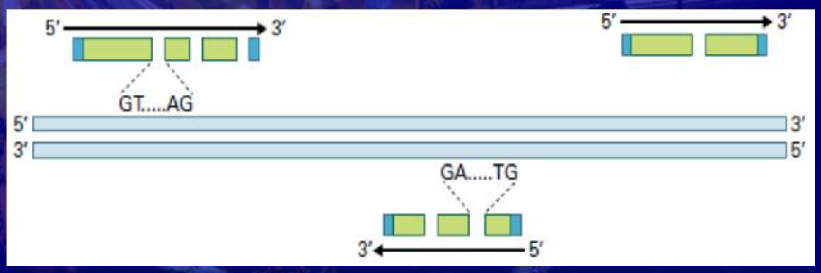


Fig. 9.3 Genes (protein-coding sequences) are found on either DNA strand, but rarely in the same region on both strands. Because transcription is always 5′ → 3′, genes on opposite strands are transcribed in opposite directions; transcription may be convergent or divergent. Note the orientation of the conserved splice site boundaries on the two strands. Green boxes represent the protein coding region of the transcripts (CDS), and blue depicts the 5′ and 3′ untranslated regions (UTRs). On rare occasions, two genes may occur in the same region of the DNA, but on opposite strands, with one gene lying within the intron of the other.

2. **Telomeres guard the ends of chromosomes**
   - Chromosome arms end in telomeres, which are specialized structures that protect chromosome ends, ensure their faithful replication, and prevent the shortening of chromosomes during successive rounds of DNA synthesis.

3. **Centromeres are responsible for the precise segregation of chromosomes into daughter cells during division**.
   - Cytologically, chromosome centromeres (Fig. 9.4) are heterochromatic (highly condensed) chromatin constrictions to which the spindle fibers attach to facilitate the separation of replicated chromatids in mitosis and meiosis.
   - Among the simplest of characterized centromeres are those of the budding yeast *Saccharomyces cerevisiae* and the fission yeast *Schizosaccaromyces pombe* (only tens to hundreds of nucleotides in length). Plant centromeres are almost invariably much larger and more complex.
   - The size of the centromeres varies between species and even across chromosomes within an individual species, ranging from 50 kb to many megabases. Despite a limited understanding of the functional elements of plant centromeres, both DNA sequence and the associated proteins, it is becoming possible to construct artificial chromosomes capable of sustained autonomous replication and of moving parts or all of important metabolic or developmental modules into other varieties or species for crop improvement (Fig. 9.5).

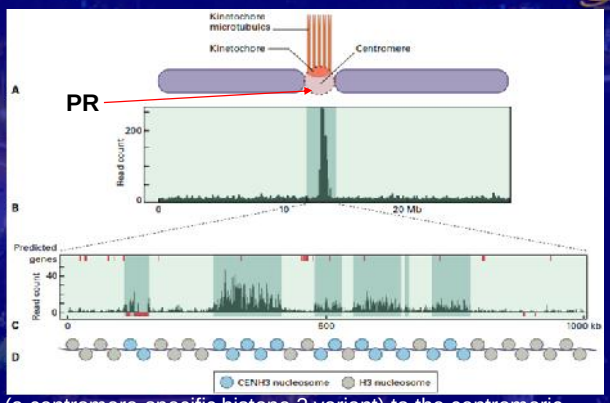
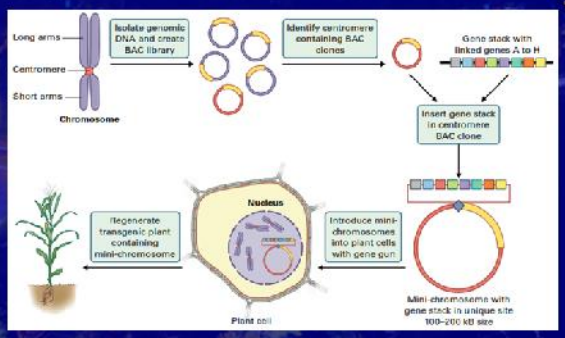


Fig. 9.4 Structure of the centromere of chromosome 8 of rice (Oryza sativa). (A) General structure of a chromosome displaying chromosome arms, centromere, kinetochore, and microtubules. Only one sister chromatid is shown for simplicity. (B) Localization of CENH3 (a centromere-specific histone 3 variant) to the centromeric region deduced from ChIP-Seq (see Box 9.10) against CENH3 across the entire chromosome. The gray box indicates the crossing over-suppressed region of the chromosome. (C) Expanded view of the CENH3-binding domain. Predicted genes (red boxes) are located predominantly outside CENH3 binding domains (gray boxes), and the 155-bp centromeric repeat (CentO) arrays are shown as red boxes beneath the mapped reads. (D) Representation of nucleosome pattern across Cen8 in rice. Gray circles represent nucleosomes with canonical histone H3, and cyan circles represent CENH3-containing nucleosomes. Not drawn to scale.

Fig. 9.5 Artificial mini-chromosomes carrying sets of genes can be constructed for plant genetic engineering.



4. **Genes encoding the RNA components of ribosomes are localized to a small number of genome regions**
   - Nucleolar organizer regions (NORs) are cytologically identifiable regions of the chromosome around which ribosomal RNA transcription, biogenesis, and assembly occur. Ribosomal RNA (rRNA) is needed in such large quantities for ribosome structure and protein synthesis in active cells.
   - At the DNA sequence level, these regions consist of many tandem repeats of an approximately 10-kb sequence that encodes the 28S, 18S, and 5.8S RNA components of the ribosome (Fig. 9.6).

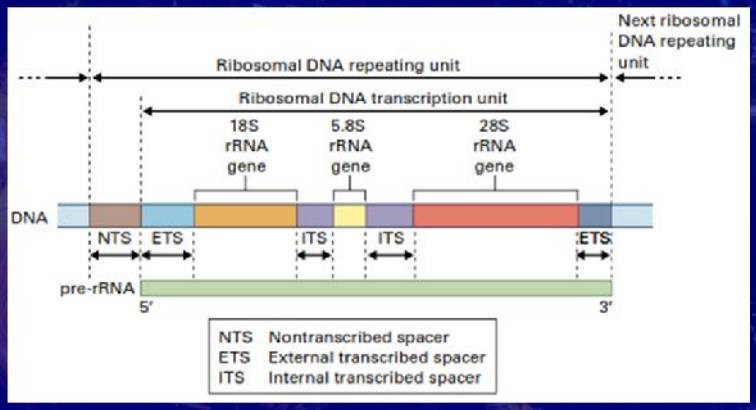
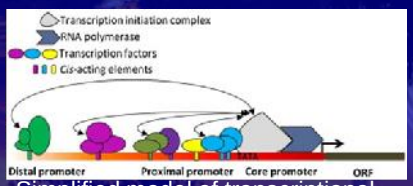


Fig. 9.6 The eukaryotic ribosomal DNA repeating unit. Most plant ribosomal genes lie within a repeating unit ranging from 7,800 to 185,000 bp long. The repeating unit is composed of highly conserved rRNA genes separated by short stretches of spacer sequences that are not transcribed (NTS, ETS, and ITS). Ribosomes contain four distinct RNA molecules of sizes 18S, 5.8S, 28S, and 5S that contribute to their structure. Genes for the first three are contained on a rDNA repeating unit, whereas the 5S gene is found elsewhere in the genome.

2. TRANSCRIPTION

1. Definition

- The expression of genetic information is the central process that controls the growth and development of plants and their ability to respond to changes in the environment.
- This process requires the transcription of genes into RNA and in many cases the translation of these RNAs into the ultimate gene products—proteins.
- Transcription is the biochemical process of transferring the information in a DNA sequence to an RNA molecule.



Simplified model of transcriptional regulation of protein-encoding genes. Source: Hernandez-Garcia & Finer (2014)

- The RNA molecule can be the final product, or in the case of messenger RNA (mRNA), it can be used in the process of translation to produce proteins.

2. RNA Polymerase

- RNA polymerases (RNAPs) are the enzymes that read DNA templates and synthesize the different types of RNA found in plant cells.
- In plants, up to five different RNAPs mediate the transcription of genes in the nuclear genome, and additional RNAPs are responsible for transcription of organellar genomes.
- The nuclear RNAPs (designated RNAP I–V) have distinct functions and synthesize different types of RNAs (Table 9.2).

Table 9.2 Plant nuclear RNA polymerases and their products.

RNAP	RNA synthesized
I	5.8S, 18S and 28S rRNAs
II	mRNAs, miRNAs, snRNAs, ta-siRNAs
III	5S rRNA, tRNAs, snRNAs
IV	siRNAs
V	noncoding RNAs

- **RNAPI**, a highly specific in gene expression, only transcribes the 45S ribosomal RNA (rRNA) precursor, which is processed into 5.8S, 18S, and 28S rRNAs. Despite this limitation, RNAPI-mediated transcription accounts for up to 50% of RNA synthesis in growing cells.
- RNAPII is responsible for the transcription of protein-coding genes and produces messenger RNAs (mRNAs). In addition, it transcribes microRNA (**miRNA**) genes and synthesizes small nuclear RNAs (**snRNAs**), as well as transacting small interfering RNAs (**ta-siRNA**).

- Due to its central role in the expression of proteins, RNAPII has been studied extensively and its structure has been resolved in atomic detail (Fig. 9.16).

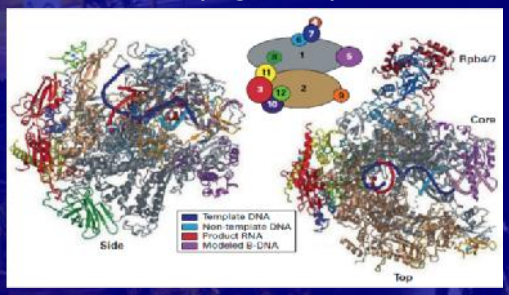


Fig. 9.16 Crystal structure of the **RNAPI**I elongation complex from yeast (*Saccharomyces cerevisiae*). The polymerase subunits are color-coded according to the key between the top and side views. Template DNA, nontemplate DNA, and product RNA are shown in blue, cyan, and red, respectively.

- **RNAPIII** transcribes genes coding for transfer RNAs (tRNA), 5S rRNA, and a variety of other small RNAs.
- **RNAPIV** and **RNAPV** have been discovered only relatively recently, and in contrast to RNAP I to III, which are present in all eukaryotes, appear to be plant-specific (RNAPV is only found in angiosperms, or flowering plants).

- Furthermore, these RNAPs are not essential for the viability of a plant, unlike RNAPs I, II, and III.
- The function of RNAPIV is to synthesize small interfering RNAs (siRNAs).
- RNAPV transcribes noncoding RNAs involved in siRNA-mediated gene silencing.

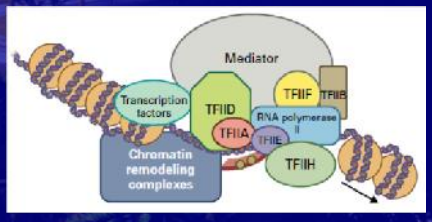
- Because of structural differences, RNAPs differ in their sensitivity to inhibitors of transcription.
  - For example, RNAPII is highly sensitive to α-amanitin, a cyclic octapeptide found in the Amanita genus of mushrooms (one of the main sources of mushroom poisoning). RNAPI is not sensi-tive to α-amanitin, and RNAPIII has intermediate sensitivity.

3. Transcriptional Machinery

- The process of transcription can be divided into three consecutive phases: (i) initiation of transcription, (ii) an elongation step that leads to the formation of an RNA product through RNAP activity, and (iii) termination.

- The initiation of transcription is an important control step that depends on the formation of a transcription preinitiation complex (Fig. 9.17).

Fig. 9.17 Schematic representa-tion of the RNAPII preinitiation complex and its interactions with transcription factors and chromatin remodeling complexes on DNA. Light-brown beads symbolize nucleosomes.



- This multisubunit protein complex is composed of an RNAP and associated proteins (Table 9.3) that modulate RNAP activity and recruit the enzyme to the promoters of genes.
  - For example, RNAPII cannot bind to DNA alone; rather, it requires the presence of several general transcription factors (TFIIA, B, D, E, F, and H).

Table 9.3 Components of the RNAPII preinitiation complex in Arabidopsis.

Factor	Number of subunits	Function
TFIIA	3	Stabilizes TBP-TATA interaction
TFIIB	1	Selection of the transcription initiation site
TFIID	TBP and 12-15 TAFs	Recognition of promoter elements
TFIIE	2	Formation of the initiation complex
TFIIF	2	Recruits TFIIE and TFIIH
TFIIH	10	Allows RNAPII promoter escape and elongation
RNAPII	12	Initiation, elongation and termination of transcription
Mediator	≈34	Relays information of DNA-bound transcription factors to RNAPII

TAF: TBP-Associated Factor. TBP: TATA binding protein

- TFIID, itself a multiprotein complex, constitutes the key set of proteins in the transcription preinitiation complex.
- One of the TFIID components, TATA binding protein (TBP) recognizes the TATA box (Fig. 9.17).

Transcriptional Machinery continued 2

- The TBP is a DNA sequence motif located in the promoters of many protein-coding genes around position –30 (i.e., 30 nucleotides upstream from the transcription initiation site, which corresponds to the first nucleotide of the RNA).

- TBP-mediated binding of the TFIID complex to a promoter leads to the sequential recruitment of additional general transcription factors and RNAPII into the preinitiation complex and subsequently to the initiation of transcription.

4. Promoters

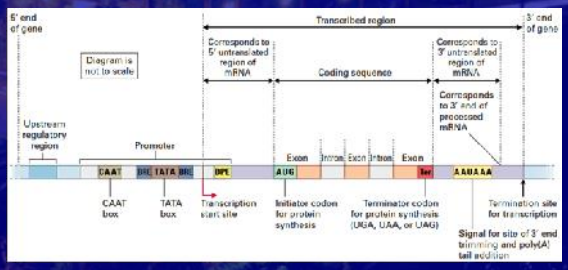
- A promoter contains various sequence elements that collectively recruit the protein factors that regulate the transcription of a gene.
  - Each class of RNA polymerase binds to its cognate promoters to initiate transcription.

- Promoters consist of two types of regulatory elements, basal elements and *cis*-elements.
  - Basal elements are necessary for RNA polymerase binding and positioning and generally lie within 50 bp up- or downstream of the transcription initiation site. They are often relatively conserved between different genes and together form the **core promoter**.
- TATA box is an example for a basal element, and other examples for elements in core promoters are the TFIIB recognition elements (BREs) found immediately upstream and downstream of TATA boxes (Fig. 9.18).
- The BRESs are recognized by the general transcrip-tion factor TFIIB, and the downstream promoter element (DPE), around 30 bp downstream of the transcriptional start site and bound by subunits of the TFIID complex.
  - Unlike basal elements, which show some degree of conservation between different genes, *cis*-elements vary greatly from gene to gene.

Promoters 2

Fig. 9.18 **Structure and organization of a eukaryotic gene**.

A gene is divided into several functional units. The transcribed region acts as a template for synthesis of mRNA, which is then edited and translated into the protein product of the gene.



The transcribed region is interspersed with noncoding sequences that partition the region into coding sections (exons) and noncoding sections (introns). The transcribed region is flanked on either side by noncoding sequences that play a role in regulation of the gene. Most of the regulatory sequence elements are in the 5′ flanking region. The first 1,000 bp or so of the 5′ flanking region is referred to as the gene promoter, as it contains sequence motifs important for the “promotion” of transcription. One of the most highly conserved regulatory elements is the TATA box, which is usually found within the first 50 bp of the transcription start site. The TATA box coordinates the recruitment of RNAPII to the gene.

Promoters 3

- "Cis" (from Latin, meaning "on the same side") refers to the fact that these regulatory elements are located on the same DNA molecule as the genes they control.
  - *Cis*-elements are responsible for the activation, repression or modulation of gene expression bound by transacting factors.
- The transacting factors are proteins encoded elsewhere ("in trans") in the genome that can interact directly with the basal transcriptional machinery (e.g. with components of the TFIID general transcription factor complex) or indirectly through mediator, a large protein complex that acts as a transcriptional coregulator (Fig. 9.17).
- Most *cis*-elements in plants lie within 1–2 kb of sequence upstream of the transcription start site (Fig. 9.18).
- However, they may also be found further upstream, in a gene's introns or exons, or downstream of the site where termination of transcription occurs.
- *Cis*-elements are usually short (<10 bp) in sequence and vary greatly in their base composition (Table 9.4)

Promoters 5

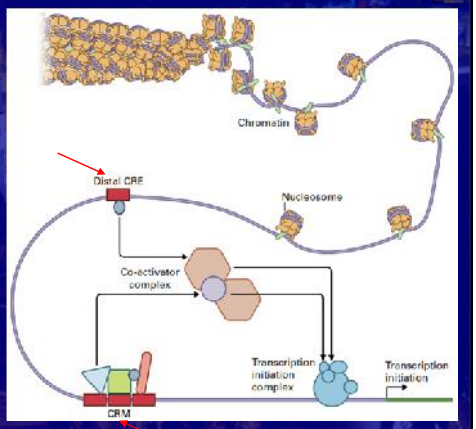
TABLE 9.4 Examples for DNA sequence motifs bound by Arabidopsis transcription factors, and their functions in plant growth and development.

Motif name	Sequence	Bound by	Function
ABA-responsive element (ABRE)	CACGTGGC	ABRE-binding factors (bZIP proteins)	ABA response
Auxin response element (ARE)	TGTCTC	Auxin response factors	Auxin response
CArG box	CC(A/T)6GG	MADS domain proteins	Flower development, etc.
Evening element	AAAATATCT	Certain MYB proteins	Gene regulation by the circadian clock
G box	CACGTG	bZIP and bHLH proteins	Light response, etc.
GATA promoter motif	(A/T)GATA(G/A)	GATA transcription factors	Various

- A promoter can contain several different *cis*-elements, and more than one copy of the same element can be present. These elements often cluster together to form *cis*-regulatory modules.
  - The function of these modules likely allows the cooperative action of several (different or identical) transcription factors in the regulation of gene expression by bringing them in close proximity to one another (Fig. 9.19).

Promoters 6

Fig. 9.19 Example for the function of *cis*-regulatory modules and enhancers. A *cis*-regulatory module in the promoter of a gene is bound by several transcription factors. This transcription factor complex, together with an input from a distal *cis*-regulatory element (CRE), which acts as an enhancer, leads to the recruitment of a co-activator complex and ultimately, to the promotion of transcription.



cis-regulatory modules

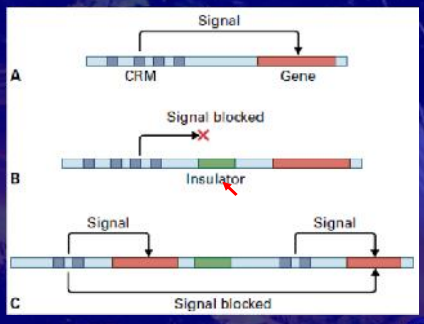
- *Cis*-regulatory modules are often classified as enhancers or silencers, depending on whether they promote or inhibit the expression of a gene.

Promoters 7

- The terms enhancer and silencer refer to regulatory elements that are located a great distance away from the transcribed region of a gene to distinguish them from proximal promoter elements (regulatory elements found in the vicinity of the transcription start site).

- *Cis*-regulatory modules can also function as insulators to block signals from enhancers and silencers and prevent them from influencing gene expression activities of certain genes (Fig. 9.20).

Fig. 9.20 Function of genetic insulators. The effect of a *cis*-regulatory module (CRM) on a gene (A) can be blocked by the presence of an insulator (B). Insulators can also limit the effects of a CRM on a specific gene, preventing neighboring genes from responding to CRM activity (C).



Promoters 8

- *Cis*-elements can mediate expression of specific genes during development or in response to internal or external signals.
  - One example is the auxin response element (**AuxRE**; sequence: 5′-TGTCTC-3′), which is found in the promoters of genes controlled by the phytohormone auxin.
- Transcription factors are typically composed of at least two domains with different functions. The DNA-binding domain mediates the recognition of and binding to a *cis*-element, while the transactivating domain interacts with the transcriptional machinery and determines the activity of the protein.
- Transcription factors also typically contain a nuclear localization sequence (NLS), which allows the proteins to enter the nucleus after they have been synthesized in the cytoplasm.

Promoters 9

- Several different types of DNA-binding domains have been identified, and their structures have been resolved in atomic detail, yielding detailed insights into how they interact with DNA (Fig. 9.21).

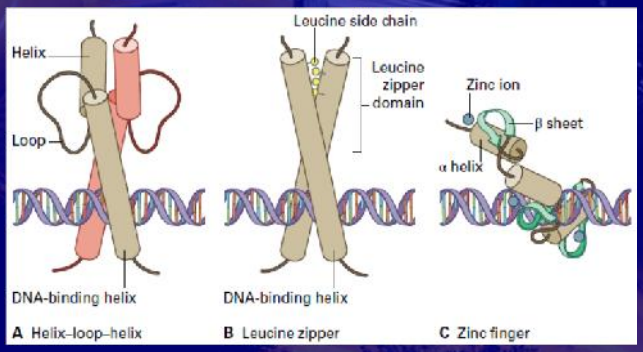


Fig. 9.21 Three major types of DNA binding domains. (A) Helix–loop–helix motif, (B) Basic leucine zipper & (C) Zinc finger domain. Each model illustrates the DNA/protein complex.

5. Control of Transcription Factors

- *Cis*-elements are bound by transcription factors that in turn promote or inhibit gene expression by interacting with components of the transcriptional machinery.
- These transcription factors usually bind DNA in a sequence-specific manner so that a given transcription factor targets a specific *cis*-element; for example, the ARE mentioned above is bound by members of a family of transcription factors called auxin response factors (ARFs).
- The pivotal role of transcription factors for the regulation of plant growth and development is demonstrated by the large number of transcription factor-coding genes in plant genomes. For example, the Arabidopsis genome contains approximately 2,000 genes that encode transcription factors.

- This number corresponds to roughly 6% of all Arabidopsis genes, considerably higher than that observed in many other model organisms.
- Many plant transcription factor families are also found in other eukaryotes (e.g., bZIP and MYB proteins) (Table 9.5); however, the number of family members is often considerably enlarged or reduced in plants when compared to those of animals or fungi.

Table 9.5. Examples for transcription factor families in Arabidopsis and rice. The number of family members, and whether the families are plant specific, is indicated.

Family	Arabidopsis	Rice (Oryza sativa)	Plant specific
AP2/ERF	147	161	no*
bZIP	74	94	No
GRAS	32	57	Yes
MADS	108	77	No
MYB	198	182	No
NAC	100	149	Yes
WRKY	71	109	No*

5. Transcription Factors

- Diverse transcription factor families control gene expression. As discussed above, *cis*-elements are by transcription factors that in turn promote or inhibit gene expression by interacting with components of the transcriptional machinery.
- These transcription factors usually bind DNA in a sequence-specific manner so that a given transcription factor targets a specific *cis*-element such as ARE bound by members of a family of transcription factors called auxin response factors (ARFs).
- The pivotal role of transcription factors for the regulation of plant growth and development is demonstrated by the large number of transcription factor-coding genes in plant genomes.

Control of Gene Expression 2

- For example, the Arabidopsis genome contains approximately 2,000 genes that encode transcription factors. This number corresponds to roughly 6% of all Arabidopsis genes, considerably higher than that observed in many other model organisms.
- The different transcription factors found in higher plants can be grouped into 50 or so families, depending mainly on sequence similarities.
  - These families sometimes are named after common structural features [e.g., basic helix–loop–helix proteins (bHLH) contain a DNA-binding domain composed of two α-helical segments separated by an intervening loop] (Fig. 9.21), but also often after their physiological function in the plant (e.g., the ARFs mentioned above) or after the founding member(s) of a family (e.g., the AP2-like transcription factors, which constitute one of the largest transcription factor families in plants, were named after the APETALA2 (AP2) protein, a key regulator of flower development).

Control of Gene Expression 3

- The large number of transcription factors in plants implies a high degree of complexity in the regulation of gene expression activities.
- This complexity is further increased by the fact that many transcription factors, such as the bZIP proteins, act as dimers.
- In addition to homodimers (pairs of identical proteins), many transcription factors also form heterodimers (pairs of different transcriptional regulators).
- Because a single transcription factor can form heterodimers with many different partner proteins, the number of possible transcription factor complexes is staggering.
- A key problem in regulatory biology is identifying the genes that a transcription factor controls and the *cis*-elements to which it binds.

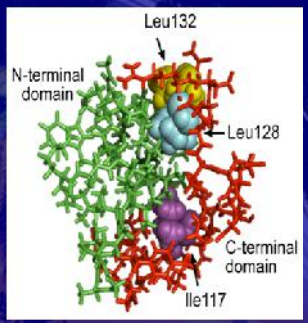
II. PROTEIN

1. Definition

What are proteins?

- **Proteins** are large biomolecules (macromolecules) made up of one or more long chains of amino acid residues (polypeptide molecule) which are amino acids linked covalently by peptide bonds.
- Proteins are the most complex and abundant of the macro molecules, and many proteins within cells function as enzymes in the catalysis of metabolic reactions.
- Within any one cell there may be thousands of different proteins having a variety of sizes, structures, and functions.

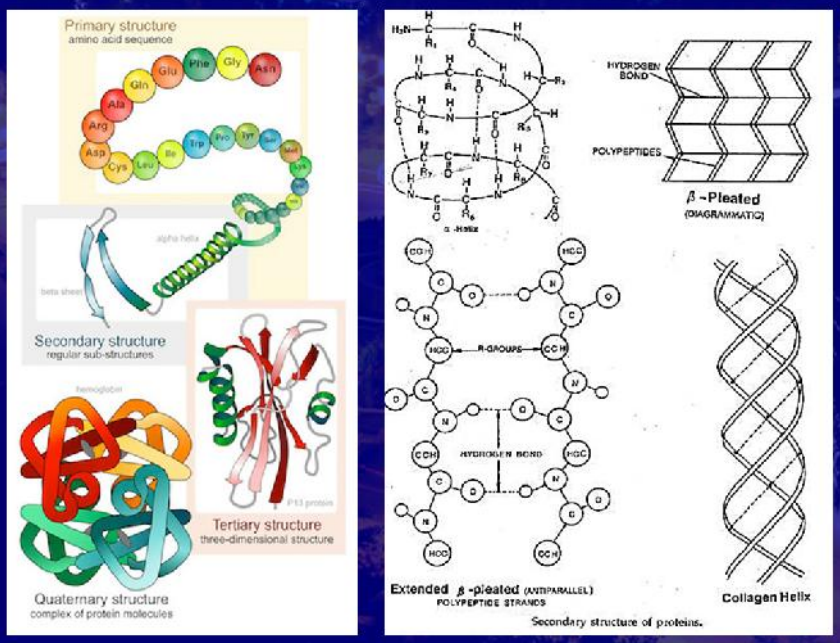
- Others serve as transport molecules, storage proteins, electron carriers, structural components of the cell, constituent of hormones, receptor proteins, protection against diseases, molecules used for growth and repair, visual pigments and molecules in urea formation.
- For isntance, Mms6 is a protein isolated from *Magnetospirillum magneticum* AMB-1 that promotes the formation of superparamag-netic magnetite particles in vitro, and is found associated with the magnetites of magnetosomes when isolated from these bacteria.



Predicted Mms6 Monomer Structure

2. Levels of Structure

- A protein can have up to four levels of organisation primary, secondary, tertiary and quaternary.
  - **Primary structure.** The primary structure of a protein refers to the linear sequence of amino acids in the polypeptide chains and location of disulphide bridges, if there are any
  - **Secondary structure**. The folding of a linear polypeptides chain into a specific coiled structure is referred to as the secondary structure of the protein. Such folding is obtained by hydrogen bond. There are three types of secondary structures α-helix, β-pleated, and collagen helix.
  - **Tertiary structure**. The arrangement and inter-relationship of the twisted chains of protein into specific loops and bends is called the tertiary structure of the protein.
  - **Quaternary structure**. Two or more polypeptides may be associated to give rise to a complex macromolecule, a functional unit, which is referred as the quaternary structure.

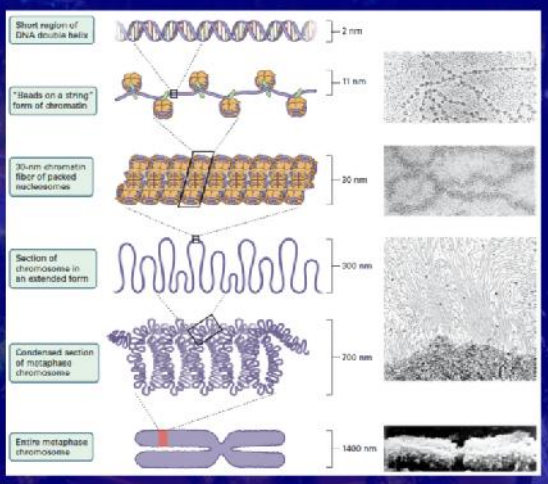


2. Genetic Maps

- **Genetic maps** are essentially ordered lists of the distinctive features (markers) along each chromosome and the relative distance between them (determined by the frequency of recombination between them during meiosis).
- These markers may be based upon either DNA sequence or phenotypic traits (e.g., plant height, leaf shape, disease resistance, stress tolerance). The DNA sequence alone is not particularly informative.
- Where genetic maps have been aligned with sequenced genomes, excellent correspondence exists between the order of sequence-based markers in the genetic map and the occurrence of these marker sequences in the genome itself.

- Chromosomes were first visualized at the turn of the 20th century, and they were recognized as the bearers of a cell's hereditary material.
- As more became known about DNA, chromosomes of higher eukaryotes were found to contain many centimeters, even meters, of double-stranded DNA molecules in compact chromosomal structures that, when visualized in cells at the metaphase of mitosis, are only microns in length.
- Increasing microscopic resolution allowed visualization of different features of chromosomes at a level that can now, to some extent, be described in terms of DNA sequence (Fig. 9.1).

Fig. 9.1 Organization of DNA and proteins into chromatin at different levels of condensation. The cartoons on the left show DNA,



DNA wrapped around nucleosomes, packed nucleosomes, extended loops of packed nucleosomes, DNA as it might be condensed into metaphase chromosomes, and a complete metaphase chromosome. Images at right are scanning electron micrographs at different levels of resolution that can be inferred from the scales on the cartoons.