PART A

PLANT MOLECULAR BIOLOGY: AN OVERVIEW

1

Plant Nuclear Genome Organisation

The spherical or ovoidal plant cell DNA resides in a nucleus, occasionally when situation arises it develops lobes which increases its surface area. The majority of higher plant nuclei are 5–20 μm across, but exception is the nuclei of giant algae *Acetabularia* (up to 150 μm). The dynamic nature of the two-membrane nuclear envelop exhibits highly diversified pores loaded with transport proteins. Direct connection or sometimes slight association exists between nuclear envelop and organellar membrane of chloroplast and mitochondria.

A major nuclear organelle called the *nucleolus,* devoid of limiting membranes, contains DNA, granules of RNA and proteins. They are the sites of transcription of rRNA genes and processing. In addition, it involves assembly of 80S ribosomes destined for the plant cytosol.

Plant nuclei exhibit hectic activity due to the presence of unique macromolecules like structural proteins, tubulin, active polysomes, enzymes, acidic regulatory proteins, RNA and histone proteins. The ratio of nucleic acid to protein is very specific. The ratio of DNA to protein is 1:1, whereas for RNA and acidic protein is approximately 0.1:0.6. The association of five basic nature of histone proteins H1, H2A, H2B, H3 and H4 and amino acid sequence of H4 is identical in pea and cow.

1.1 DNA ORGANISATION IN NUCLEUS

The overall organisation of plant nuclear genome revealed that coding capacity is relatively constant among plants as seen in comparison of genome of *Arabidopsis* and maize. Comparing the genomic nature of these two plants also reveals genomic codes for same number of genes but differ in their genome size. Similarly, maize and sorghum plants contain 10 chromosomes but the maize genome is three times larger than the size of sorghum. Several striking similarities were observed on the arrangement of genes on chromosome of sorghum and maize. The extra DNA that account for differences in maize and sorghum genome size are mainly non-coding repetitive sequence between genes. This clearly indicates that in most of the organisms only 1% of the DNA is utilized for protein production and rest may have a significant role in structure and organisation of the genome.

1.2 GENERAL PROPERTIES OF DNA (OVERVIEW)

Naturally, the DNA molecules comprise of deoxyribose, a phosphate group and a base. The four nitrogenous bases in all types of DNA are purines—adenine and guanine, and pyrimidines—thymine and cytosine. Adenine is potential to form hydrogen bonds with guanine and cytosine. Sometimes modified bases are present in plant nuclear DNA, the cytosine residue is replaced by 5-methylcytosine. Base equivalence between adenine and thymine, and guanine and cytosine places methylcytosine which is unique in double stranded DNA and two strands are joined together by hydrogen bond to form a right handed B-form of DNA encloses ten base pairs in one complete helix, which extends up to 3.4 nm and width helix is 1.8 nm. Several other DNA configurations are possible like A and Z forms. The base composition of DNA is generally

expressed as the G + C in percentage. The base composition of a number of plant DNAs are given in Table 1.1. The most common method of fractioning DNA is to analyse differences in density between different types of concentrated salt using caesium chloride. Most nuclear DNAs from higher plants have buoyant density within the range of 1.69–1.71 gcm^{-3}.

Table 1.1 Base composition of plant DNA (moles per mole)

	A + T	**G + C**
Ascomycetes		
Neurospora sp.	(2.3 + 23.3) 25.6	(27.1 + 26.6) 53.7
Saccharomyces cervisiae	(31.7 + 32.6) 64.3	(18.3 + 17.4) 35.7
Algae		
Chlamydomonas globosa	(19 + 19.5) 38.5	(30.3 + 28.2) 58.5
Spirogyra sp.	(30.7 + 30.4) 61.1	(19.2 + 19.8) 39
Gymnosperms		
Ginko biloba	(31.6 + 33.5) 65.1	(17.2 + 17.7) 34.9
Angiosperms		
Daucus carota	(26.7 + 26.9) 53.6	(23.7 + 17.3) 41
Zea mays	(26.8 + 27.2) 54	(22.8 + 17 + 6.2) 46
Allium cepa	(31.8 + 31.3) 63.1	(18.4 + 12.8 + 5.4) 36.6
Arachis hypogea	(32.1 + 32.3) 64.4	(17.6 + 12.3 + 5.7) 35.6
Gossypium hirsutum	(32.8 + 32.9) 65.7	(16.9 + 12.7 + 4.6) 34.2

1.3 VARIATION OF GENOMIC SIZE AMONG PLANTS (C-VALUE PARADOX)

Nuclear genome content of plant cells has been estimated by applying various parameters like micro-densitometre. The DNA content of haploid eukaryotic cell ranges from 10^7 to 10^{11} base pairs. These values are expressed as C-value. Generally, nuclear DNA content of higher plants range from 0.5 to 200 picograms and sometimes above this level. Interestingly, genome size of the different higher plants are highly diversified. *Arabidopsis*, smallest of known plants, comprises of 7×10^7 bp; whereas largest member, *Fritillaria*, contains genomic size of 1×10^{11} bp. The most closely-related plants—rice, wheat, having genome size of 5×10^9 and 6.5×10^9 bp respectively. However, there is a lacuna in the direct relationship between genomic size and organelles called the *C-value paradox*. Comparison of genomic size between *E. coli* and pea plant exhibits considerable difference. It is estimated that pea plant contains 1000 times as much DNA as *E. coli*, which is presumed to contain above 3000 genes. Presence of large number of genes in higher plants are susceptible to mutation and may consequently be responsible for genetic diversity and speciation. According to several striking evidences, a reasonable estimate of number of plant genes code for proteins would be in the range between 40 and 100,000. Since genomics of *Arabidopsis* have been completed in the year 2000 and rice genome is expected to

be completed in 2004, several unravelling functioning of each gene and its C-value paradox can be assessed.

The fact that number of genes in *Arabidopsis* is also present in other plants in terms of basic functions. The typical genes cover up approximately 4000 bp. Based on this, it can be presumed that *Arabidopsis* genome can accommodate about 15,000 genes.

1.4 PLANT REPETITIVE/SATELLITE DNA

Satellite DNA has a unique place in the structural organisation of chromosome. Satellite DNA is non-coding, tandem repetitive DNA sequence associated with centromere or other heterochromatic regions of chromosomes. The term satellite can be referred to any DNA sequence which can be physically separated by isopycnic centrifuge and does not correlate to any specific functions. Plant chromosomes are of the size of many satellites. Several satellites have been described in higher plants (Ingle et al., 1973). They contain long stretches of repeated sequence and can represent small proportion of the total DNA. The plant satellite DNA comprises of repeat length ranging from six to several hundreds of base pairs. They are often dominant components in centromere. Satellite repeats or tandem repeats associated with centromere have been reported in several plant species, like alfa satellite in plants are organised into huge length of several mega bases long (Jiong et al., 2003). (Repetitive DNA contributed to the functions of specificities and play a role in genome organisation.) For example, each of the five *Arabidopsis* centromere can be 2–4 mega bases of the 180 bp centromere satellite repeat and similarly, centromeric regions of the chromosome in maize contains up to 9 mb of the B-specific repeat. Satellite DNA of plants and animals differ in their base ratio. Plant DNA is GC rich whereas animal and yeast DNA is AT rich. The organisation of arrangement of repetitive DNA in centromere region is highly conserved among organisms (Table 1.2).

Table 1.2 Repetitive DNA elements reported in plants (after Jiang et al., 2003)

Plant species	Repeat
Arabidopsis thaliana	pAL et al.: 180 bp tandem repeat
Brassica campestris	pBT, 175 bp tandem repeat
Brassica oleracia	pBOKB 171 bp tandem repeat
Hordeum vulgare (*barley*)	(AGGGAG) satellite DNA
Oryza sativa specific retrotransposons	Cent 0: 155 bp tandem repeat, CRR: centromere
Petunia hybrida	pBS-SB1-2: 666 bp tandem repeat
Sacharum officinarum	PsG1-2: 140 bp tandem repeat
Triticum aestivum	Tail: 570 bp tandem repeat
Zea mays	Cent C: 156 bp tandem repeat
	Cent 4: 74 bp tandem repeat

1.5 TELOMERE REPEAT SEQUENCE

The end of centromeric structure is termed as telomere, also contains tandem repeats. The repeated telomere sequence of *Arabidopsis* is TTTAGGC, which is different from human telomere repeats by only one base (TTAGGC). Telomere plays a significant role in replication of genome. At the end of chromosome lies a portion of single stranded DNA composed of only two

to three copies of sequence. An enzyme called *telomerase* maintains the single-stranded DNA and prevents it from shortening after each round of replication.

Plant genome contains another significant repeat portion called *dispersed repeat sequences*. Its tandem repeat differ from other tandem repeats by dispersing through the genome. Plant transposons in several plant species make up the dispersed repeat sequence. For example, dozens of transposons derived repeat sequence occupies the region surrounding the Adh gene of maize.

1.6 SINGLE COPY SEQUENCE

It refers to DNA sequence present in only one copy per haploids genome. In addition to repetitive DNA, nuclear genome contains unique single copy sequence. In tobacco, biochemical analysis indicates that nearly 40% of the single copy sequence are not transcribed.

1.7 FINE STRUCTURE OF PLANT GENE

Organisation of plant gene is like that of any other eukaryotes. Fine structure of plant genes are described under two headings. One is protein coding and other is RNA coding gene sequence. In plant genome, transcription can start 40–80 nucleotides upstream of the ATG initiation codon. The non-coding sequence it carries, is removed successfully at the post-transcriptional level to form mature mRNA. The genes have controlling sequence at the 5′—3′ end of DNA. The 5′ region has many regulatory elements and are unique for specific genes. These unique sequences are called *core sequences* of promoter. The specialised core sequences are commonly found in all plant genes called TATA and CAAT box regions. The TATA sequence in plants are recognised by RNA polymerase II. Maize *zein gene* contains two TATA boxes. One is 900 bp upstream from the first regulatory sequence. Transcription of the gene results in the production of two different mRNA of 900 and 800 bases long. The second probable transcription regulatory is AGGA or CAAT box located upstream from the TATA box.

All plant protein coding genes contain one or two poly A signals in the 3′ region. The mRNAs are terminated and polyadenylated some nucleotides after first or the second poly A signal. The introns of plants contain AT rich sequence and its border sequence always ends in well conserved 5′GT and 3′AGA. In leghaemoglobin gene, the introns are inserted between DNA sequence, encoding specialised domains. However, no correlation has been attributed to the distribution of introns and functional domains of active gene. Similarly, the *zein gene* encoding storage protein in *Zea mays* also contains different domains but totally devoid of any introns in the gene.

The TATA box is situated 25–40 nucleotide upstream of the transcription initiation site. The consensus sequence of the TATA box is TCACTATATATAG. The CAAT box is located further upstream from the start of transcription. In certain plant genes, CAAT box sequence is replaced by AGGA box. Generally, the coding region of the mRNA begins with AUG codon as a part of the conserved sequence AACAAUGGCT. The 40S ribosome sub-unit can scan mRNA from 5′ directions and select this conserved sequence for initiation of translation. All plant mRNA genes have one or two poly A signals 3′ untranslated regions. Plant intron contains AT rich DNA and its border always ends in 5′ GT and 3′ AG, and this sequence is well conserved on eukaryotes. Other salient features of gene are as follows :

- Genes reside on the chromosome.
- Genes are made up of DNA but never from protein. The message present in the gene can only direct the synthesis of protein.

- Gene is the unit of recombination. Genes can be recombined into novel combinations through chromosomal exchange referred as *crossing over*. The crossing over takes place between and also within a single gene. One of the sub-units of gene is referred as *recon*. More precisely, recon is a smallest unit capable of recombining genetically. Recon may consist of one or two pairs of nucleotide. There is a strong evidence that crossing over takes place within a single gene against earlier concept of crossing over only between genes.
- Gene is a unit of function, otherwise known as *cistron*. The one gene–one enzyme hypothesis proposed originally by Beadle and Tatum was well recognised and expressed in diversified way. Once genes are expressed and synthesized mRNA then it becomes one gene-one messenger RNA. When several genes express together and produce mRNA as single units they are referred as *poly-cistronic mRNA*. The one gene-one enzyme or one gene-one protein hypothesis cannot be justified. This hypothesis is becoming presently modified as one gene-one polypeptide. The genes present on the DNA not only code for protein synthesis but also code for other RNA like ribosomal RNA and tRNA.
- Gene is a unit of mutation otherwise known as *muton*. Higher plants are very much prone to mutation due to large number of genes and genomic size.
- Gene undergoes duplication and amplification.
- Genes are mapped by restriction cleavage. A restriction map represents a linear sequence of the site at which restriction enzyme cut at the specific target sites.
- Plant genes or in general eukaryotic cells are interruptive due to the presence of non-coding intervening sequence referred as *introns*. These interrupted genes are also known as *split genes*.

1.8 GENES FOR RIBOSOMAL RNA

Many plant rRNA gene sequences have been purified, characterised and restriction maps determined. There is a considerable variation of rRNA gene number within species (Table 1.3). They also contain repeat units ranging from 8 to 11 kb in various plants. Studies on rRNA synthesis show that rRNA is transcribed as a polycistronic RNA, which contain sequences for 18S and 25S rRNA. There are differences in repeat length and sequences due to variation in spacer DNA. Similar gene structures are found in other higher plants. Three different repeat units are present in wheat with a length of 9, 9.15 and 9.45 kb. Requirement of multiple rRNA genes is indispensable for plants in order to facilitate rate of ribosome production, as plant cell requires several hundred thousands of ribosomes per hour. To meet the above demand, plant cell probably contains multiple sites of rRNA synthesis and partly by transcribing each gene several times simultaneously as evidenced in *Acetabularia*. In wheat, methylation of some rRNA genes specifically residing condensed chromatin are transcriptionally inactive.

Plant also contains multiple 5S rRNA genes in tandem arrays, which are not linked to 18S rRNA and 25S genes. The two variable lengths differ mainly in the spacer DNA of 410 and 500 bp has been detected in one of the variety of wheat. The 70 bp region upstream from the coding sequence of these two genes is strongly conserved and play a crucial role in regulating transcription. Similarly, AT rich region 3′ end of coding sequence is presumed to act as termination signal for RNA polymerase.

Table 1.3 Number of ribosomal RNA genes in higher plants

Plant	Number of genes/haploid
Maize	3,100
Swiss chard (*Beta vulgaris*)	1,150
Wheat	2,100
Cucumber	4,400
Onion	6,650
Pea	3,900
Artichoke	260

1.9 PALINDROME SEQUENCES

Palindrome sequences are reverse repeats and are known to be distributed through a large part of the genome. This comes from the zero time binding of DNA to hydroxyl apatite.

--A-T-A-G-G-C-G-C-C-T-A-T

--T-A-T-C-C-G-C-G-G-A-T-A

There are several possible functions for palindrome sequences. Any RNA molecule transcribed from DNA intersequence of this type would be expected to have double stranded hairpin regions.

CONCLUDING REMARKS

The typical spherical or oval shaped plant cell DNA resides in a nucleus. It measures up to 2–20 μM. The nuclear membrane is loaded with transport proteins. Plant nuclei exhibit hectic activity due to the presence of macromolecules. The overall organisation of nuclear genome to plant revealed that coding capacity is relatively constant among plants as seen in comparison of genome of *Arabidopsis* and maize.

One of the unique features of plant genome is the presence of satellite DNA. They contain long stretches of repeated sequences and can represent small proportion of the total DNA. Satellite repeats associated with centromere have been reported in several plant species.

Fine structure of plant genes are described under two headings. One is protein coding gene sequence and other is RNA coding genes. Plant genes also contain regulatory sequences called core sequences of promoter. The specialised sequences are called TATA and CAAT box regions. All plant protein coding genes contain one or two poly A signals in 3′ regions. Some of the most salient features of plant genes are as follows:

Genes reside on the chromosome. Genes are made up of DNA but never from protein. Genes can be recombined into novel combination through chromosomal exchange known as crossing over. Gene is a unit of function. Gene is a unit of mutation otherwise known as muton. Plants are very much prone to mutation due to large number of genes and genomic DNA.

Several plant rRNA gene sequences have been characterised and restriction maps have been determined. These genes also contain repeat units. Studies show that rRNA is transcribed as a polycistronic RNA.

Plant requires multiple RNA genes due to considerable demand by the cell. Plant also contains multiple 5S rRNA genes in tandem arrays. The palindrome sequences are present in plant DNA. These are reverse repeats known to be distributed through a large part of the genome. There are several possible functions for palindrome sequences. Any RNA molecule transcribed from DNA intersequence of the type would be expected to have double stranded hairpin regions.