LECTURE 20

Repeated DNA Sequences

Prokaryotes:

1) Most DNA is in the form of "unique" sequences. Exceptions are the genes encoding ribosomal RNA (rDNA, 10-20 copies) and various "recognition" sequences (e.g., promoters).

2)

Central Dogma is:	DNA	→	rRNA
		→	mRNA → protein
		→	tRNA
		→	Other RNA types (eukaryotes only)
			a) snRNA (Snerps)
			b) snoRNA (Snors)
			c) RNA_i (interference RNAs)
(non-coding)	DNA	→	Recognition sites
		→	Spacer bases (few)

3) DNA sequence organization is represented by a linear array of unique sequences (genes or recognition sites), with one or a few bases (spacer bases) separating the functional sequences.

4) Cesium-chloride (CsCl) density-gradient analysis of semi-sheared DNA yielded only a single "main band" of DNA, indicating that most "pieces" of DNA had similar buoyant densities (meaning base composition). In *E. coli*, for example, the buoyant-density profile indicates the base composition to be about 50% AT base pairs (and 50% GC base pairs).

Eukaryotes: Much more complex

A. CsCl analysis (initial discovery) of eukaryotic DNA revealed one large "main band" along with several, minor satellite bands (that presumably differed in AT/GC composition). So-called "heavy satellites had a higher percentage (proportion) of GC base pairs composition; whereas so-called "light" satellites had a higher percentage of AT base pairs.

1) The satellite bands were referred to as "satellite DNA" and turned out to be tandem (10^4 - 10^7) copies of short DNA sequences (from 10-300 bases pairs per monomer) that were rich in GC base pairs ("heavy" satellites), rich in AT base pairs ("light" satellites), or neither ("cryptic" satellites).

   a) Other names for satellite DNA include highly repeated (repetitive) DNA and simple sequence DNA.

2) A few, salient factoids:

a) Virtually all eukaryotes have one of more satellite DNAs. Some satellite DNAs are simple, some are complex; complex ones seem to be derived from more simple beginnings.

b) *In situ* hybridization experiments demonstrated that most satellite DNAs were localized (*clustered*) on chromosomes in constitutive heterochromatin. Chromosomal locations of constitutive heterochromatin are near centromeres, telomeres, and around NORs. In many organisms (e.g., plants), there are large blocks of constitutive heterochromatin (satellite DNAs) on chromosome arms. In several species, whole chromosome arms or even whole chromosomes can be entirely comprised of constitutive heterochromatin (satellite DNA).

c) Direct sequencing of satellite DNAs reveals an extreme pattern of sequence amplification, diminution, and diversification (mutation). Satellite DNAs thus appear to constitute a rapidly evolving (*dynamic*) set of DNA sequences in terms of base-pair composition and size.

d) Satellite DNAs are not transcribed, i.e., they are genetically inert in terms of serving as template sequences for synthesis of RNA molecules. This is not surprising, given the sequence pattern of short, tandemly repeated sequences that vary extensively in terms of specific sequence and size.

e) The function(s) of satellite DNA (heterochromatin) are unknown, but undoubtedly are <u>not</u> sequence specific). Past (present) suggestions are:

(i) structural, organizational, and/or protective role(s)
(ii) involvement in meiotic (homologue) pairing
(iii) coarse control of crossing over
(iv) modulation of genome size
(v) no function at all -- part of the "junk" DNA in eukaryotic genomes

B. Renaturation kinetics (DNA renaturation experiments)

1) Procedure: Shear DNA to small fragments (200 - 400 bp)
Heat to break into single-stranded pieces (ssDNA)
Remove heat to renature (reassociate) ssDNA into double-stranded (ds)DNA
Follow kinetics (rates of reassociation) spectrophotometrically

2) Eukaryotic genomes can be divided coarsely into three categories (largely for convenience):

Highly repeated sequences (>10^5 copies/genome) -- satellite DNA
Moderately repeated DNA sequences (100 - 10^5 copies/genome)
Unique or single-copy DNA sequences (1 - 10 copies/genome)

3) Comparisons among different species OVERHEAD

(a) Categories (classes) of repeated DNA represent a continuum where the categories are somewhat artificial. Use of three classes is a "generality" employed as a means to study potential function of sequences in different classes.

(b) The asterisk (*) after *E. coli* in the table refers to the 10-20 copies of rRNA genes (rDNA).

(c) Note that the total proportion of repeated DNAs in general is much higher in plants.

(d) Note also that the total quantity of single-copy DNA not related to organismal complexity and that much of the unique-sequence (single-copy) DNA is non-coding.

(i) 1-2% of the single-copy DNA in garden peas forms double-stranded molecules {DNA/RNA heteroduplexes} with bulk-isolated mRNA – the percentage in humans may be a bit higher, perhaps 3-5% of all DNA)

(ii) the proportions (as shown in the overhead) are misleading, as the species differ in total genome size (DNA content per cell), i.e., the total amount of a given class of repeated DNA also is a function of total genome size

(iii) this is related in a way to the "C-value paradox" where the total quantity of DNA is not related to organismal or biological complexity; consider, for example, several salamander species who have genome sizes 7 - 10 times the genome size of humans)

C. What is known about the cellular function(s) of sequences in these repeat classes?

1. Satellite DNA – already discussed

2. Unique DNA sequences (1-10 copies)

a) A significant fraction of unique sequence DNA is "coding" for proteins and various types of RNA sequences.

(i) included are transcriptionally active units such as enzymes, structural proteins, and proteins involved in normal cellular function (structural genes, e.g., enzymes in metabolic pathways), gene regulation (regulatory genes, e.g., transcription factors), and (recently discovered) retrogenes

(ii) also included are potentially functional retroviruses and intact transposons and retroposons (transposable elements)

b) Not all unique sequences are transcribed or even functional. What's in the remainder of unique sequence DNA (perhaps well more than 50% of all DNA) is not known for all species. Where there are considerable data (e.g., *Drosophila*, humans, mouse), these non-coding unique sequences are comprised variously of…

(i) intervening sequences (introns)

(ii) fossil repeats (degenerating or decaying repeated DNA sequences)
(iii) retroviral-like, transposon-like, and retroposon-like DNAs (decaying retroviruses, transposons, and retroposons)
(iv) pseudogenes
(v) retroprocessed pseudogenes (reverse transcripts of processed mRNAs)

c) Except for the introns (*within* genes), the remainder of these unique sequences are located *between* genes and can be called "spacer bases." Of interest is to compare the distribution of genes/spacer bases between prokaryotes and eukaryotes. S. Ohno called the eukaryotic genome a "vast desert with oases of genes"

3. Moderately repeated DNAs ($100 – 10^5$ copies)

a) An extremely heterogeneous class in terms of sequence complexity, with relatively little known about possible functional roles. Can be subdivided into:

(i) coding DNAs that contain structural gene sequences, SINEs, and LINEs

(ii) non-coding DNAs that contain common regulatory sequences, fossil repeats, and introns of moderately-repeated structural gene sequences

b) Structural gene sequences are genes whose products needed in large quantity in cells. Examples include histones, rRNAs, tRNAs, immune system proteins, seed storage proteins, some transcription factors

(i) the proportion of the genome containing these sequences is hard to estimate, but could represent up to 10% of all moderately repeated DNAs

c) SINEs [**S**hort **I**nterspersed **N**uclear **E**lement(s)]

(i) 150-300 base pair (bp) repeated elements that are found in the "short interspersion pattern" – typically possess an 8-20 bp inverted repeat (characteristic of "insertion" sequences) called 'target-site duplications'

(ii) exhibit a highly variable pattern among organisms

(a) in sea urchins, for example, there are more than 100 families, with from $10 - 10^4$ members per family, and with 10 – 20 distinct subfamilies per (large) family

(b) in primates and rodents, alternatively, there are only a few prominent families, with on family usually outnumbering all other families; in humans, for example, the *Alu*I family consists of up to 500,000 copies of the *Alu*I repeat (300 bp) and comprises up to 5% of the human genome

(iii) SINE sequences are transcribed but are not translated -- in humans, *Alu*I sequences are found in 20% of hn (pre-)mRNA but are removed during mRNA processing

(iv) the function of SINE sequences are unknown; *ad hoc* suggestions include transcription regulation and regulation of mRNA processing

(v) another suggestion is that SINEs represent remnants of transposition; this suggestion is based on the small, inverted repeat that flanks SINE sequences and on finding SINEs sequences inserted in various locations (including functional genes) throughout the genome

(vi) now thought to be possibly 'mobilized' by retroposons (LINES)

c) LINEs [**L**ong **I**nterspersed **N**uclear **E**lements]

1. An interesting and heterogeneous class of sequences comprised in part of transposons and retrotransposons

   a) Transposons: mobile genetic elements, jumping genes, nomadic sequences, etc.

   b) Retrotransposons: mobile elements depending on reverse transcription

   c) Will devote a lecture on these types of sequences later in the semester

2. Elements that are 3,000 - 5,000 bp in length that are dispersed (interspersed) throughout genomes (hence LINE)

3. Clearly mobile (able to "move" from location to location within a genome) and inducible. The latter accounts for the phenomenon of "hybrid dysgenesis."

4. Definite involvement of transposable elements in <u>mutation</u> and <u>chromosomal rearrangement</u>.