

Lecture 1: DNA, Genes, Gene Expression, transcription. 17/3/2005

Note: [figure numbers] refer to Alberts et al, 3rd edition or 4th or Genomes 2 (G)

1 DNA

1.1 Structure

Cell: *Eucaryotic* [caryon = nucleus] typical size (animals) 10 - 30 μ . Cytoplasm contains organelles (mitochondria, Golgi apparatus,...). [3P1-1] Nucleus bound by double membrane, contains DNA packed in chromosomes [NucToDNA] [3P3-2a]. Mitochondria also contain DNA (ring). 0.25 % of cell mass = DNA (70 % - water).

DNA = DeoxyriboNucleic Acid

A (chemically) boring molecule. Polymer (Oligomer): a long (short) chain of monomers bound to each other head to tail. Heteropolymers are made of many different units. Monopolymers - of the same repeated unit. DNA is made of two strands (backbones) of monopolymer sugar phosphates (hence boring) held together by pairs of bases, of 4 kinds. The sequence in which these 4 bases appear contains information, which is the subject of this lecture.

The basic unit is a nucleotide: made of a 5-carbon sugar (pentose) a phosphate and a base.

Sugar carbons: [3P3a] 1' ... 5'

1' - linked to base [3P3-2b]

5' - linked to phosphate [3P3-2c]

3' - linked to phosphate of adjacent nucleotide

the polynucleotide backbone is directional; it has 3' end and a 5' end. [2-10].

[4F4-4a]

the bases:	pyrimidines C, T	(single ring, 6)	(U)
	Purines G, A	(two rings, 6+5)	

Two strands are held together by hydrogen bonds between complementary base pairs G-C and A-T [4F4-4b]. The C-G bond is stronger - 3 hydrogen bonds, the A-T has two. Each strand/message/Watson is bound to its complementary (Crick).

Double helix, [4F4-3] in each mammalian cell $3 \cdot 10^9 = 3,000$ Mbp units, about 2 m long

Length: in e-coli: 4.7 M base pairs, 1 ring chromosome (DNA molecule)

yeast: 12 Mbp on 16 chromosomes

human: in each diploid cell $22 \cdot (2 \text{ copies}) + 2(\text{sex XX female, or XY male})$ chromosomes, 50 - 250 Mbp each, 1.7 - 8.5 cm long

Charge - : (One electronic charge per sugar or base, 0^- of the phosphate group.) chromatin, DNA wrapped on (+) charged core octamers, each made of 4×2 histones. [3F8-30] A DNA strand of 146 bp wrapped tightly on nucleosome; with 200 bp total/nucleosome. The nucleosomes form a string and these - a solenoid. Very compact entangled structure [4F4-21] of chromatin.

1.2 Information

Overview (rough): DNA contains genes - linear segments that code for proteins. 1st step - transcription to mRNA; second - translation of mRNA into protein dogma.

A strand has a particular sequence of bases; contains a message (4-letter alphabet)

Convention: message is read as English text, from 5' end towards 3' end.

message - *sense* (5')....ACGAATCGG...(3')
antisense (3')...TGCTTAGCC...(5')

The Content of the human (nuclear) genome

Focus on a 50kb segment of chromosome 7 [G1.14]; it contains the following elements:

Gene (TRY4) codes for protein trypsinogen, 5 exons 4 introns

Gene segments V28 and V29-1 parts of gene, need to be linked to other DNA

Pseudogene (TRY5) non-functional, unreadable copy of gene "evolutionary relic"

Genome-wide repeat sequences (52) Long Interspersed Nuclear Elements (**LINEs**), **SINEs** (e.g. Alu, 10^6 copies, 2×140 bp long), **LTRs** (Long Terminal Repeat) elements and **DNA transposons** (was copied with or without RNA) .

Microsatellites - short motif is repeated in tandem e.g. CGCGCGCG.... [GBox1.4], [4F4-17]

1.2.1 Genes:

Segments of DNA that are transcribed (copied) into RNA. Mainly to mRNA which is processed and translated into protein, but also to non-coding RNA (rRNA,tRNA,miRNA,...). The gene contains 5'-UTR (UnTranslated Region) followed by coding regions (exons) separated by introns, closing with 3'UTR (few 10s to few 100s bp long). [myslid2a]

In *human genome* on average 9 exons/gene (titin - 178, largest, 81Kbp); mean length 145bp, longest exon - 17Kbp. Current estimate - 25,000 genes; the exons constitute 1.5 % of the genome, introns ~ 25 % !! - [Myslid2b]

Apparently **Complexity** resides in the intergenic text:

organism	genome length (Mbp)	No. genes	chromosomes
E.coli	4.7	4639	1
yeast	12	6000	16
C. elegans	97	19,000	6
Drosophila	137	14,000	4
mouse	1,250	25,000	20
man	3,200	25,000	23

1.3 Replication

The genetic information is transferred to the next (cell) generation during the process of cell division and DNA replication. Each strand serves as a template for the assembly of its complementary strand [4F5-5]. DNA polymerase - the enzyme that synthesizes DNA (with the template read from 3' to 5' and the complementary synthesized strand assembled from 5'). [4F5-3] Exact copies are made - errors in replication give rise to mutations. This process of cell division (mitosis) CellCyc is the main "reproduction mechanism" used by single cell organisms as well as the cells of a multicellular organisms. To produce an offspring, higher organisms use sexual reproduction, in which two haploid cells, a sperm and an egg, merge and the fertilized egg contains one chromosome from each parent.

All cells of a multicellular organism contain the same information (DNA). Every cell contains all the information: the instructions needed for (a) synthesis of proteins (chemical formulae) and non-coding RNA, and (b) the regulation of their production.

2 FROM DNA TO PROTEIN

2.1 Proteins

These are the working molecules of cells: Catalize chemical reactions (enzymes)

Provide structural rigidity (cytoskeleton; actin filaments, microtubules)

Control membrane permeability (channels, pumps)

Regulate metabolite concentration

Recognize and bind molecules (receptors)

Cause motion and transport of material (molecular motors)

Signaling between and within cells, pathways

Control gene function (transcription factors)

All proteins have similar chemical structure; linear chain of amino acids [chain]. Nature uses 20 amino acids: a central carbon atom (C_α) is bound to an amino group (NH_2) a carboxyl group (COOH) a hydrogen atom (H) and a residue or side chain ($R_i, i = 1, 2, \dots, 20$). They form a chain with each amino head connected to the carboxyl tail ahead of it by a covalent bond [bond]. To synthesize a protein one needs the sequence of amino acids from head to tail. Short proteins = peptides. The polypeptide chain folds to a complex 3-dimensional structure. [crambin] Secondary structure elements: alpha helix and beta sheet. The chemical composition drives this folding and the folded structure, together with the chemical identity of the constituents determines which residues are on the surface of the protein and in what shape, thus governing the biological function.

The chemical formula of a protein of length N amino acids is an N-letter word, with each letter taken from a 20-letter alphabet

(crambin, N=46) TTCCPSIVARSNFNVCRLLPGTPEALCATYTGCIIPGATCPGDYAN

Such a formula is encoded in gene on the DNA.

2.2 Transcription

The process of copying the instructions from the DNA onto RNA. [3F3-19]

Protein synthesis takes place outside the nucleus, at the ribosomes on the basis of the information coded in the DNA. To get the information from the DNA to the ribosomes, cells use a strategy of transcription, which is safe and also works as a tremendous amplifier. [Hard disk to diskette] The gene on the DNA is TRANSCRIBED into a *RNA*. Transcription resembles DNA replication [3F3-19a]: the double helix opens, one strand serves as template, read from 3' to 5' and a matching string of the transcribed gene is assembled (growing in the 3' direction). The process is catalyzed by RNA polymerase [4F6-8] enzyme [4F6-9]. RNA resembles a single-stranded DNA molecule; the differences are

(a) DNA has DeOxy sugar ring

(b) the base T (DNA) is replaced by U (A of DNA template is matched to U in RNA transcription).

(c) RNA is single stranded, folds up on itself [4F6-6].

Transcription initiation: in bacteria the RNA-polymerase recognizes special sequences on the DNA, just upstream from the transcription start site [G9.17]. The binding sequences are not fixed. In eucaryotes an initiation complex assembles near the start site (core promoter) and there are also upstream promoter elements, extending to several Kb from the TSS [4F6-19].

A variety of Transcription Factors control the levels at which various genes are expressed.

Transcriptional Networks - proteins control the expression levels of genes (which produce [Alon], [Cervix] proteins).

RNA processing steps:

RNA is produced and processed in the nucleus: **capped** for protection at the 5' end, marked for export from the nucleus by addition of (100 - 200 long) **poly-A tail** to the 3'end [8-49], **spliced** (transcribed introns are spliced out). When all this is completed the RNA has matured to mRNA, [4F6-21] is transported out of the nucleus to the cytoplasm, is found by and read by ribosomes.

Locating and reading genes:

ORF Open Reading Frame - a stretch of DNA from the beginning of the first exon to the end of the last. Every amino acid is coded for by three consecutive bases. It is a many-to-one code (64 triplets, 20 amino acids + stop). [3F3-16] There are three reading frames, that generate three different protein sequences. Usually only one of these yields a functional protein.

Special sequences code for exon/intron and intron/exon boundaries, used by spliceosomes (RNA molecules, NOT proteins!)

upstream - exon to intron (5')AG |**GUAAGU**...(3')
upstream - intron to exon (5')...**YYYYYNCAG** |.....(3')

Y=U or C, N = any; only the bold are invariant [4F6-28].

The mRNA molecule contains the same information as it's matching DNA; a 3-letter code from base triplets (codons) to amino acids. Note: START code is AUG (codes also for methionin).

The reading frame is set by the start and stop *codons* [betaglobin].

2.3 Translation:

The code is read and a protein is produced. This process, of translating a codon to an amino acid is carried out by a *transfer RNA - tRNA* [3RtRNA]. There are 20 different tRNA molecules, corresponding to the 20 amino acids. [3F6-8] [3F6-9]

The ribosome slides down the mRNA ribbon. Say the polypeptide chain of covalently bound amino acids has already been assembled, up to its present location, and is held at the ribosome. The codon of the next amino acid is in place for reading. The "foot" of the correct tRNA molecule fits the codon, sticks to it; this tRNA has the corresponding amino

acid connected to its arm. This amino acid is now attached to the polypeptide chain by a covalent bond, the tRNA falls off and the ribosome moves a step along the mRNA. [4F6-65] The tRNA molecule bears a 3-letter code on its foot, that matches the codon of a particular amino acid. This codon is recognized/read and the corresponding aa is attached to the arm of the tRNA, forming an aminoacyl-tRNA. This process of reading and attachment is the most basic process of transcription/translation; it is carried out by "the molecule that knows" [Cusack Curr. Op. Struct. Biol. 1997, 7:881-889], aminoacyl-tRNA synthetase. On one end of the process we have the digital message on the DNA - at the other end a corresponding particular aa is added to the polypeptide chain. One has to differentiate between copying a text and understanding it. The transcription process from DNA to mRNA is one of copying. During translation, the recognition of the codon on the mRNA (by the foot of the aminoacyl-tRNA) is also one of copying. Where is "understanding"? e.g. the step that attaches the particular digital message (codon) to its "meaning". i.e. the correct amino acid? This step is carried out by the aminoacyl-tRNA synthetase enzyme. [4F6-58,6-60]

```
DNA = cookbook;  codons = text(meat);
copy letters onto notepad = transcription to mRNA;
supermarket = ribosome; real meat = amino acid;
package of meat with letters "meat" on it = aminoacyl-tRNA;
comparing letters on notepad with letters on package = recognition of codon
                                                         on mRNA by tRNA
```

Where is "understanding"? In the packaging stage, when someone stuck the letters "meat" onto a package with meat. This is done by the aminoacyl-tRNA synthetase.

Two steps:

1. Recognition of the specific amino acid
2. Recognition of the corresponding tRNA

For step 2 two different recognition mechanisms are used:
for 17/20 - read the anticodon
for 3/20 by shape of tRNA [6-13,14]

Synthesis of a single protein takes 20 - 60 sec.

Central Dogma: unidirectional flow of information, from DNA to protein.

Amplification: Polyribosomes - several ribosomes attach to the same mRNA (at intervals of 60 nucleotides) [3F6-28] [3F6-29] In each silk gland cell a single gene of fibroin makes 10^4 copies of mRNA; each mRNA generates 10^5 fibroin molecules; 10^9 molecules in 4 days.