

# DNA, Chromosomes, and the Central Dogma

Daniel F. Simola

with notes adapted from Miler and Praveen

22 May 2005

## 1 Simple DNA Biochemistry

**DNA:** Linear chain composed of a sugar (deoxyribose) phosphate backbone, with units of nitrogenous bases (A, T, C, G), which are planar and lie perpendicular to the length of the backbone, stacking on one another. DNA is double stranded due to hydrogen bonding between pairs of purine-pyrimidine bases (A-T, C-G), with an A-T pair sharing 2 Hydrogen bonds, and a G-C pair sharing 3 Hydrogen bonds. DNA is helical and every 10 base pairs makes one complete helical revolution. The B-form helix runs clockwise and has a narrow groove ( $\sim 12\text{\AA}$ ) and a wide groove ( $\sim 22\text{\AA}$ ). DNA changes form based on solution conditions, and can thus become more tightly wound (A-form) from high humidity or counter-clockwise, very tightly wound but elongated (Z-form) from high salt. DNA is rather negatively charged, moreso than proteins.

**Nucleic acid:** One unit of DNA composed of a phosphate group, a 5 Carbon sugar deoxyribose, and a nitrogenous base (purine or pyrimidine)

- Purines include A and G and have one six unit and one five unit C-N ring
- Pyrimidines include T and C and have one six unit C-N ring

**Point mutations:** Changes at a single DNA base

**Transitions:** occur when a Purine changes to a Purine or Pyrimidine to Pyrimidine (eg A-T  $\rightarrow$  G-C)

**Transversions:** occur when a Purine changes to a Pyrimidine and vice versa (eg A-T  $\rightarrow$  T-A or C-G)

## 2 DNA replication and chromosomes

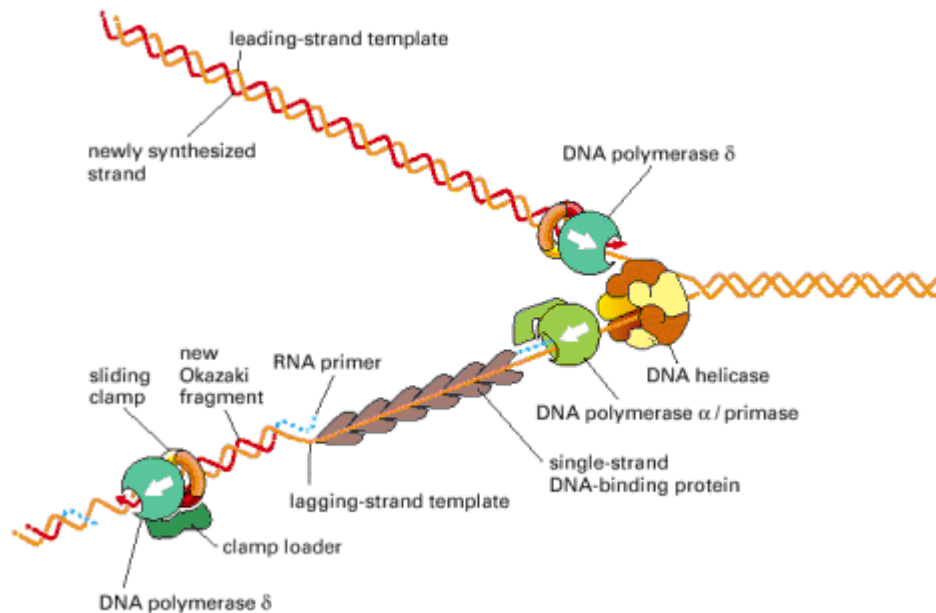
The goal of S-phase is to produce exact duplicates of the parent genome, realized as the duplication of each pair of chromosomes for diploid cells. During interphase, chromatin is extended and generally attached to the inner surface of the nucleus. Thus each individual chromatid must detach from the nuclear membrane, begin to condense, and align itself along a central axis in the nucleus. Eukaryotic DNA has three structural features which assist in replication: origins of replication, centromeres, telomeres.

Table 1: Miscellaneous figure, adapted from Molecular Biology of the Cell, Alberts et al.

Activated carrier	Group carried in high-energy linkage	Use
ATP	phosphate	energy for reactions, state switching
GTP	phosphate	state switching
NADH, NADPH, FADH <sub>2</sub>	electrons and hydrogens	biosynthesis, ATP generation
Acetyl CoA	acetyl group	biosynthesis

Each chromatid features several small DNA regions known as origins of replication, to which DNA transcriptional machinery first binds and begins the standard semi-conservative replication process. Eukaryotic cells must keep track of every replication origin so that the transcriptional machinery generates one and only one copy of each ORI-ORI segment. When replication is complete, a chromatid now has a pair, and this pair is known as a chromosome, with the pair bound together at the centromere via a protein complex.

Chromosomes initially in a form called euchromatin (roughly 1000x compaction, 30nm fiber on scaffold), condense into a very tightly packed structure called heterochromatin (10000x compaction). While this occurs throughout chromosomes, it is particularly important at chromosomal termini, the telomeres. Eukaryotic telomeres feature predominantly repeated and duplicated DNA sequences, which are thought to assist in efficient terminal replication and in forming tertiary structures to protect the termini from being digested by DNA nucleases. In addition telomeres are not replicated by DNA polymerase, rather by a telomerase, which binds to a series of G nucleotides, and extends the repeats in the 3' direction.



See Miler's notes for more detail on replication.

## 2.1 Chromosome structure

**Nucleosome:** The core unit of chromosome structure is the nucleosome, composed of DNA, a complex of proteins called a histone, and associated non-histone proteins. A histone is an octamer of proteins: twice each of H2A, H2B, H3, H4. This forms a cylindrical unit around which approximately 146 bases of DNA are wound. Histones are the most conserved eukaryotic proteins. The nucleosome is the name given to this basic organizational structure, which is more or less permanent throughout the cell cycle. The resulting appearance of DNA is something like a necklace of beads: each bead is a histone, and histones are connected with 'unprotected' DNA called linker DNA. DNA interacts and binds to the histone octamers through 142 Hydrogen bonds, as well as salt bridges and hydrophobic interactions. Half of the hydrogen bonds form between the amino acid and phosphodiester backbones. This structure results in a 5x compression from 2nm (width of DNA) to 10nm.

**Modifications:** The characteristics of each histone are modulated via N-terminal tails extending from each histone subunit, at conserved residues of which covalent modifications are added, such as acetyl, methyl, and phosphate groups. Patterns of these modifications compose a "histone code", with the results of loosening the histone-DNA bonds, unravelling DNA, translocating a histone along the DNA, completely preventing other proteins from binding to DNA. Enzymes known as histone acetyl transferases (HATs) de-acetylases (HDACs) and histone methylases and de-methylases catalyze these modifications. Such enzymes are recruited

to specific regions of a chromosome by histone remodeling complexes, led by two proteins, Swi and Snf; there are a variety of such proteins. These proteins feature protein binding domains to which adaptor proteins bind, to which enzymes like HACs and methylases bind. In addition there exist variants of histone subunits, which are exchanged to alter the properties of a nucleosome, such as centromere structure, DNA repair, and naturally transcriptional regulation. The result of all of this is the orderly opening and closing of particular areas of DNA, commonly where promoters are found, so that particular genes may be transcribed when needed. The presence or absence of a particular modification is not necessarily indicative of expression or silencing, rather the particular combination and order of modifications which confers specificity. Thus at one histone the pattern Ac(1)-Ac(3)-Me(5) might open up a histone, but Ac(1)-Ac(3) might confer silence at another. However methylation events generally indicate transcriptional repression, and acetylation events activation. Of special importance is methylation of the Lysine residue at position 9 of histone H3.

There actually exist epigenetic mechanisms which control gene expression on a larger scale, by opening up a series of promoters at once. Such an example is the globin locus control region, which is used in hematopoietic differentiation.

**30nm fiber:** Following DNA replication, each chromosome continues to condense, with the help of the condensin protein, and the nucleosomal chromosome is structured into a 30nm diameter fiber, the structure of which is not completely known, but whose appearance is something similar to a lacing pattern one might use in sewing. This pattern is called the crossed-linker structure.

**Higher-order chromosome compression:** This 30nm fiber is then attached via a looping system to a protein scaffold which compresses the DNA another 10 fold, into a 300nm fiber. Little is known about this scaffold, but this is the level of compression typically found during interphase. Such fiber on a scaffold is called euchromatin, since it can be accessible to DNA and RNA transcriptional machinery as well as cis and trans acting transcription factors. Chromosomes compressed further are called heterochromatin, a state of chromosome where raw DNA is inaccessible. During interphase 10% of DNA is in the form of heterochromatin, and in general these areas are gene poor, thus including much of the telomeric and centromeric regions.

Transcriptional silencing via histone remodeling complexes regulates gene expression in three places:

1. Binding of transcription factors
2. Binding of the transcription pre-initiation complex at TATA and initiation at the transcription start site
3. Transcription elongation of RNA Pol II

The one additional layer of compression above heterochromatin is the mitotic chromosome itself, where the 300nm fiber-on-scaffold is coiled into what we can see during M-phase: the puffy arm of a chromosome, which is roughly 700nm in diameter. Since each chromosome is paired, this results in a final diameter of 1400nm, from the original 2nm of naked DNA. Thus there is a total length compression of 10000 fold, at the price of a total width increase of 700 fold.

An interesting side note is the effect this compression and transcriptional silencing has on phenotype. Since the cells discussed here are diploid, either maternal or paternal chromosome may be silenced or open. It would seem difficult for a metazoan to control which copy should be accessible on a large scale, and so often times some cells within a tissue or organ differ in which alleles are expressed, leading to some interesting visual effects, such as spotty colors in the eyes of flies. This is known as position effect variegation.

**Telomere silencing:** As mentioned above, repeated DNA sequences at chromosome termini foster protection. One of the reasons is because a protein called the silent information regulator (Sir2) protein, a histone deacetylase which binds the N-terminal tails of histones, creating a compact heterochromatic region in telomeres.

**Centromeres:** The centromere exhibits a similar transcriptional silencing as the telomeres, except repeated DNA takes the specific form of  $\alpha$ -satellite sequence. Thus the centromere becomes condensed by the beginning on M-phase, and its DNA is structured and bound by various protein scaffolds of the kinetochore, a specialized area to which spindle microtubules bind, in order to separate the sister chromatids from each other during mitosis.

### 3 Molecular basis of gene expression

The central dogma of molecular biology states that all proteins normally derive from mRNA messages transcribed from genes within a cell's genome: DNA  $\rightarrow$  RNA  $\rightarrow$  protein. The dogma is realized in the processes of RNA transcription and protein translation, which are both amplification steps. There are several intermediate steps which invite mechanisms of regulation and control of these two processes:

1. **Transcription:** cis, and trans controls to modulate concentration of pre-mRNA
2. **Epigenetic activity:** Meta-level control mechanisms that affect transcription on a larger scale: Chromatin structure and histone modifications
3. **RNA processing:** capping, poly-adenylation, and splicing
4. **RNA transport and localization:** nuclear export of mature mRNA and delivery to cellular destination
5. **Translation:** selection of mRNAs to be translated
6. **mRNA degradation:** How often to remove an RNA from activity
7. **Protein activity:** Post-translational modifications, activation, localization, and degradation

#### 3.1 Transcription

See Miler's notes on DNA transcription and repair.

##### 3.1.1 Transcription factor DNA-binding domains

1. Helix-turn-helix: binds at grooves in DNA (major/minor), eg homeodomain
2. Zinc fingers: eg  $\alpha$ -helix::Zn:: $\beta$ -sheet
3.  $\beta$ -sheets can also bind DNA
4. Leucine zipper: homodimer of  $\alpha$ -helices - neat because the DNA-binding portion of the motif is also the dimerizing portion
5. Helix-loop-helix: flexible chain because of the loop region - forms HLH homodimers as well as heterodimers
6. Subsequent dimers formed by combining pairs of motifs

The leucine zipper is an interesting example because it is an exception to typical TF modularity. Most TFs have more or less independent DNA-binding and activation domains. The former attaches the TF to the TF binding site motif, and the activation domain binds either a general transcription factor or the polymerase itself to stabilize the initiation of transcription.

##### 3.1.2 cis and trans acting transcription factors

Proteins utilizing the above DNA-binding domains acts as activators, repressors, enhancers, and silencers of gene expression. Proteins which assist these factors are co-factors, which include coactivators and corepressors. As an example as to how many TFs are used in a cell, yeast cells are thought to encode about 300 unique TFs, including Gal4, MCM1, SWI4, and SWI6. Combinations of several TFs can synergistically determine the concentration and rate of transcription.

### 3.1.3 Epigenetic factors

There are a few factors which influence transcription on a meta-level scale. Epigenetic change is defined as heritable change not in the form of DNA nucleotide alteration; thus heritable phenotypic change. Often times this includes the activation or silencing of a block of genes en masse. Factors that are able to influence transcription in this way include the structure of chromatin (histones), super-enhancers called locus control regions (LCR), and super-silencers called boundary elements.

Epigenetic control extends to the use of DNA methyltransferases to selectively *imprint* DNA with methyl groups, which inhibit TF binding and thus transcription. Such epigenetic factors are responsible for the mechanisms of X chromosome inactivation in females and heritable genomic imprinting.

Epigenetic effects also refer to heterochronic changes in the timing and magnitude of transcriptional regulatory events, and are supposedly responsible for drastic heritable changes in phenotypes among organisms within a species due to developmental changes.

See other notes for more details about chromatin remodeling, X-inactivation, and imprinting.

### 3.1.4 Experimental techniques for assessing control of gene expression

**Gel-Mobility shift assay:** Detect sequence-specific DNA-binding proteins. The idea is that proteins bound to DNA and run on a gel will slow down the migration rate of the DNA. Radioactively label a sample and run on a gel. If several proteins are bound to a stretch of DNA, the gel will show several discrete bands, each of which corresponds to a protein of a different weight. Cut these out, purify, and mass spec for identification.

**Affinity Chromatography:** Purify sequence-specific DNA-binding proteins. If you know the sequence of the DNA segment of interest, synthesize a short oligo complementary to the target and link to beads on chromatography column. Pour in sample, and proteins will be bound to the column via the oligos. Wash and purify.

**Chromatin immunoprecipitation:** An *in vivo* technique to identify a DNA region bound by a particular protein. Take a live cell at condition of interest and crosslink protein to DNA using formaldehyde. Sonicate cell and extract DNA, which will be fragments around 200bp due to histones bound during crosslinking. Use an antibody against the protein to extract those protein-DNA segments. Uncrosslink the complexes, and sequence the DNA segment.

**Microarray:** Assess on a genomic scale the relative or absolute levels of mRNA for genes of interest.

**ChIP-on-chip:** Microarrays meet ChIP. Perform ChIP as above, except probe with several antibodies of interest, purifying each sample separately. Linearly amplify purified DNA, and spot this on a normal 2-channel array where probes are not genes of interest but promoters of interest.

### 3.1.5 Transcription of other RNA classes

mRNAs are not the only type of RNA to be transcribed by RNA polymerase, there are actually several types which perform a variety of functions, mostly structural and catalytic. In fact mRNA constitutes only 3-5% of total cellular RNA mass. The following is a practical list of cellular RNA:

**mRNA:** messenger RNA encodes functional protein products, and is transcribed by RNA polymerase II.

**rRNA:** ribosomal RNA is the most abundant cellular RNA (80%), which forms the core of ribosomes. rRNA are transcribed by RNA polymerase I, which lacks a C-terminal tail domain, explaining the lack of capping or polyadenylation of rRNAs. Because rRNA is needed in such great quantity, all eukaryotic genomes include multiple copies in each genome: *E. coli* has seven, and humans have over 200. There are four kinds of eukaryotic rRNA, three of which (18S, 5.8S, 28S) result from the modification and cleaving of a single precursor, while the fourth (5S) requires no transcript modification and is transcribed by RNA polymerase III.

rRNA transcripts are really long, about 13000 bases, and require over 100 methylations and 100 uridine isomerizations, the positions of which are pointed out by hundreds of “guide” RNAs, the small nucleolar

RNAs (snoRNAs). Other snoRNAs guide cleavage of the premature transcripts. Most snoRNAs are encoded in the introns of ribosomal genes: fancy that.

**tRNA:** transfer RNAs are around 80bp, have a characteristic clover-leaf structure, and carry amino acids as a payload. There are three hairpin loops that form the 3 clover leaves: D, T, and the anticodon loop, which contains an anticodon of three nucleotides which recognizes a triple RNA codon specifically. The 3' end of the tRNA binds the amino acid appropriate for its anticodon. tRNAs are transcribed by RNA polymerase III, and are modified in some of 50 different ways before nuclear export.

**snoRNA:** small nucleolar RNAs guide cleavage of the premature transcripts. Most snoRNAs are encoded in the introns of ribosomal genes.

**snRNA:** small nuclear RNAs are short RNA transcripts of 100-300 bp that associate with proteins to form small nuclear ribonucleoprotein particles (snRNPs), which participate in RNA processing.

**hnRNA:** heteronuclear RNAs bind along the length of the mRNA during transcription and remove helical 2° structures for better signal recognition. After histones, hnRNPs are the most abundant proteins in the nucleus.

**miRNAs and siRNAs:** These guys bind to mRNA and recruit nucleases and helicases which triggers a negative amplification reaction. See the section on mRNA degradation below.

**Other:** There are several other classes of RNA, including ribonuclease pRNA, X-inactivation specific RNA (XIST), and small modulatory RNA (smRNA).

## 3.2 RNA processing

Transcription produces a premature mRNA transcript, or pre-mRNA, containing 5' and 3' untranslated regions (UTR), introns, and a poly-adenylation signal. Before a pre-mRNA can become a mature genetic instruction and exported from the nucleus, it must receive a special 5' cap composed of a modified methyl guanosine, introns must be spliced and exons fused together, and a polyA tail must be ligated onto the 3' end, after the 3' polyA signal found in the 3' UTR. It is important to realize that these three processes are not independent or serially performed, but are highly coupled and dependent on each other temporally and spatially; this is especially true for splicing and poly-adenylation. The regulatory use of both 5' and 3' UTR tends to be negative. eg several RNA binding proteins bind to mRNAs at the UTRs and either inhibit or localize the mRNA.

### 3.2.1 5' capping

Three enzymes operate in succession to add the guanosine cap:

1. Phosphatase removes one phosphate from the 5' nucleotide
2. Guanyl transferase adds a guanosine monophosphate (GMP) to the 5' end
3. Methyl transferase adds a methyl group to the guanosine (and sometimes a second methyl). This nucleotide is called 7-methyl guanosine.

These enzymes all bind the phosphorylated tail of RNA polymerase II, so as soon as transcription begins, the nascent transcript is capped.

This cap serves as a binding site for a complex called the cap-binding complex (CBC), which greatly aids in the recognition of a mature mRNA and its ability to be exported into the cytosol. In addition the distance from the cap to the translation start site (length of 5' UTR) influences the efficiency of translation, since the 40S ribosomal subunit binds towards the 5' cap.

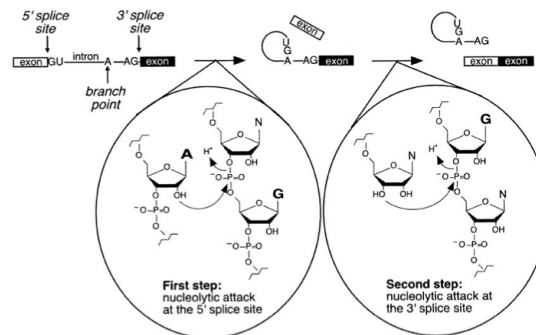
### 3.2.2 Intron splicing

A eukaryotic transcript may contain several introns which separate short ( $\sim 100$ bp) coding segments called exons. The combination of individual exons results in the generation of novel proteins with specific functions. This combinatorial effect explains why humans and flies have similar raw count of genes; in humans about 60% of all genes are alternatively spliced, meaning there is more than one protein product is generated from a single pre-mRNA.

Thus splicing is somewhat modular in nature, and a protein complex called the spliceosome focuses on removing introns one at a time, although multiple spliceosomes may operate in parallel. The general procedure is to identify an intron, remove it from the transcript by forming an intron loop called a lariat and coincidentally religate the exon ends using a special phosphorylation reaction called transesterification. Many **small nuclear RNAs** (snRNA) of about 200bp (U1, U2, U4, U5, U6) couple with several proteins to form a ribonucleoprotein(snRNP), and these snRNPs compose the spliceosome. These protein bind to the C-terminal domain tail of Pol II for efficient intron boundary identification. Splicing is an ATP-dependent process.

Intron sequence removal requires identification and binding to **three conserved sequences**: the two splice sites found at the beginning and end of an intron, and the branchpoint consensus sequence, which contains a critical Adenine residue and forms the joining site of the intron lariat.

The U1RNP and U2RNP bind to the 5' site and branchpoint site respectively, and bring the two ends of the intron together by moving towards each other. Many complicated reactions and exchanges take place where the 5' splice site is cleaved and ligated to the branchpoint sequence, the 5' exon end is carried to the 3' splice site, which is subsequently cleaved, allowing the 5' exon end to be ligated onto the 3' exon end. The intron lariat is then carried away by the remaining snRNPs and is degraded in the nucleus.



**Experimental techniques:** In vitro splicing assays (which introns), site-directed mutagenesis of conserved splice site sequences, RNase protection assays (what RNAs, proteins bind), and gel shift assays (sequential complex formation) were used to determine the actions and components of splicing. Mutations in the splice sites result in genetic mutations.

- identify introns (Hybridization Techniques)
  - hybridize mRNA to DNA and using electron microscopy, can see that only some regions of DNA are hybridized, where as others remain single stranded. These single stranded portions of the DNA are the introns that were spliced out during the production of mRNA.
- use in vitro splicing assays to implicate factors and steps in spliceosome assembly and catalysis
- importance of various intronic regions in gene function
  - Site-directed mutagenesis at intronic splice site consensus (5' and 3') and branchpoint consensus
  - Splice site mutations result in genetic diseases
- what regions are hybridized to other RNAs or bound by proteins (RNase H degradation/protection method)

- cuts at double stranded RNA adjacent to bound region
- Used to support RNA secondary structure prediction, find snRNP binding sites (how this affects splicing)
- order of spliceosome assembly on to pre-mRNA (Gel Shift assay)
  - sequential formation of complex over time
  - two step intron splicing mechanism
- determine spliceosome protein composition
  - fuse MS2 with another heterologous protein of interest to direct protein of interest (POI) to specific regions of RNA
  - run affinity column with bead hooked to POI to pull down protein complexes
  - purify protein and mass spec
- how are splice sites chosen (in vivo splicing assays)
  - many genes have exonic/intronic splicing enhancers/silencers
  - design a “minigene”, which contains all chosen regulatory elements
  - transfect minigene into live cells (of variety of cell types)
  - RT-PCR to determine type of mRNA constructed (alt splicing due to cell type)
  - deletion/mutagenesis of minigene regions indicates importance

**Alternative splicing:** To determine the selection of particular introns, in vivo splicing assays are performed using minigenes, which contain selected splicing enhancers/silencers needed for the selection process. Minigenes are transfected into live cells, and RT-PCR is used to determine the type of mRNA constructed. Deletion and mutagenesis of minigene regions indicates their importance.

Alternative splicing is affected by a protein binding to a downstream element DCS, which influences the positioning of RNP complexes. Other things which affect splicing are the cap structure and CBP affinity, exon and intron lengths, the length of the polyA tail, and splice site consensus sequences.

### 3.2.3 Poly-adenylation

There are four key elements for poly-adenylation: an upstream element (USE), the polyA signal motif, which is typically AAUAAA, followed by a 10-30 base spacing region, a CA dimer, and a GU-rich downstream element (DSE) of at most 30 bases. Cleavage and ligation occur at the CA dimer, and the GU segment is degraded in the nucleus.

Cleavage occurs via the cleavage and polyA specificity factor (CPSF), which binds the AAUAAA hexamer. polyA addition efficiency is enhanced by the U1A snRNA which can bind CPSF as it binds an intron, but too much U1A actually inhibits polyA addition. The CPSF and another protein, the cleavage stimulatory factor F (CStF) ride along the CTD tail of polymerase, just like the snSRNPs and capping enzymes. The general transcription factor TFIID mitigates loading of these two factors onto the CTD tail. The CStF binds the DSE and is a trans factor affecting the splice product. In some cases low concentrations of CStF result in membrane bound protein products, whereas high concentrations result in a secreted protein, eg in the production of antibodies.

Once the CA dimer is cleaved, the poly-A polymerase (PAP) binds and adds Adenines to the end of the transcript by catalyzing ATP into AMP; it requires no template to operate. A sequence of about 200 Adenines is polymerized onto the primary transcript, and a group of polyA binding proteins modulates exactly how many are added. In addition these protein remain associated with the mature mRNA and help it get exported into the cytosol.

Note that RNA polymerase II continues to operate as it passes these 3' elements, perhaps for a few hundred bases after the DSE. It is unclear why the polymerase dissociates from the DNA.

### Experimental techniques:

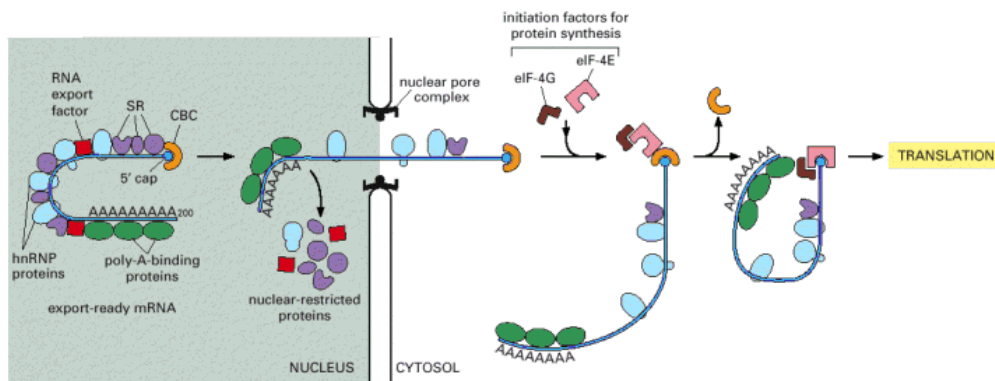
- what binds to these sites? (RNase protection assay)
  - protection due to hybridization
  - delete DSE's to determine their requirement in properly processing mRNA
- efficiency of cleavage (ie polyadenylation) of USE's (deletion analysis)
- alter cleavage site (in vitro cleavage reactions)
  - add substrate with AAUAA cut site
  - add analog chain terminator = cleavage product
  - add ATP = poly-adenylated product
  - => cleavage smaller than substrate smaller than polyA-product
- remove USE (linker substitution mutagenesis (LSM))
  - more than point mutations (you make 2 cuts, insert your LSM and re-ligate)
  - => GU rich USE and DSE regions allow for many RNA 2° structures - important for polyadenylation
- RNA sequencing and structure analysis
  - OH ladder - OH group causes hydrolysis at RNA binding point
  - RNase followed by sequence analysis (ss- and ds-RNases)
    - \* G cleavage is done at high temperature
    - \* structure probing at normal temperature
  - PbOAc cleaves ss-RNA
  - alter/block 2° structures via oligo hybridization to ssRNA (create directed oligos)
- U1A-RNA binding studies
  - determine that increasing U1A increases CPSF binding, increasing cleavage and polyadenylation
  - too much U1A inhibits CPSF
  - U1A conc. affects splicing and polyA efficiency
- alter coupling of splicing and polyadenylation (Mg<sup>++</sup> titration assay)
  - Mg<sup>++</sup> stabilizes ATP

### 3.3 RNA Nucleocytoplasmic transport and export pathways

Now that the mRNA is mature, it needs to be exported to the endoplasmic reticulum whence it will be translated by ribosomes. The mRNA is assisted in export by its entourage of the CBP and PABPs, and the spliceosome components are notably absent. The predominant factor that binds is actually heteronuclear ribonuclear proteins (hnRNP), which bind along the length of the mRNA during transcription and remove helical 2° structures for better signal recognition. After histones, hnRNPs are the most abundant proteins in the nucleus. Other proteins that bind are the RNA export factor which binds towards the 5' and 3' ends, and SR proteins, which bind the 5' end after the CBP. hnRNP and SR proteins actually enforce a specific packed 2° structure for the mRNA, which is necessary for export.

The concept of a mature vs. nascent transcript probably exists because of the tight regulatory control a cell would want to have on the translation of RNAs. Many attempts at transcription probably result in broken or incomplete transcripts, and the cell would not want these being released into the cytosol. Thus the concept of 'licensing' a mRNA for export through **nuclear pore complexes**. Different RNA classes are recognized by different nuclear export receptors via nuclear export signals on the mRNA. One pore factor is the exon junction complex (EJC) which recognizes whether an mRNA has been properly spliced by binding to a 3' region of the RNA which forms a characteristic hairpin loop.

**Exosome:** A quality control mechanism assesses whether a mRNA is normal and mature before export. If for any reason the mRNA fails, it is cleaved by 3'-5' exonucleases as part of the exosome in the nucleus. Introns are also degraded by this complex.



**Nuclear import:** Just as an addendum, many proteins must be imported into the nucleus. These require a nuclear localization signal and are guided by nuclear chaperone proteins. There are shuttling proteins which can travel in and out of the nucleus. These typically utilize the Ran GTP cycle:

- GDP in cytoplasm -> import
- GTP in nucleus -> export
- GAP in cytoplasm, GEF in nucleus
- uRNA and tRNA use this

**Experimental techniques:**

- find chaperone proteins for nuclear import (in vitro import assay)
  - label protein substrate with NLS to observe localization due to import via purified chaperone
  - measure nuclear fluorescence, adding purified protein - fluorescence increases if there is chaperoning
- which nuclear proteins depend on Ran GTPase cycle for export/import
  - Nuclear/Cytoplasmic injections of RAN-GAP in Xenopus oocyte
  - inject protein and species of labeled RNA, fractionate eggs -> N or C, and probe for localization
  - U6 snRNA always nuclear - good control as target
  - fuse NLS/NES to protein to look at localization effects
- RNA classes have different export receptors
  - Saturation/RNA competition assay in Xenopus oocyte
  - saturate nucleus with exportable labeled RNA to show presence of chaperone/receptor protein (import/export)
  - cross-competition between different RNA types can be determined
- find shuttling protein/protein complexes (Heterokaryon assay)
  - transfect labeled protein into one cell type, fuse with another cell type
  - if Ab picks up labeled protein in nucleus of second cell => shuttling

- determine whether in situ splicing is required as “licensing” for export
  - Nuclear injections of pre-transcribed mRNA (with polyA et al)
- determine protection of mRNA at exon junction complex (Footprinting assay (RNase H))
- determine whether introns or EJC that is enhancing expression (Tethering assay)
  - making MS2:EJC fusion to direct EJC to other places along RNA and measuring transcriptional ability with reporter

### 3.4 Translation

**Genetic code:** Translation is the step which maps nucleotide sequences, which in some sense are arbitrary strings, to amino acid sequences, whose identities and order are biochemically critical, using a set of rules called the genetic code. This code has evolved so that triplets of nucleotides called codons represent a single amino acid, of which 20 generally pervade the tree of life. Note that using triplets of 4 characters, one can encode up to 64 unique elements. In terms of the genetic code, there is actually some redundancy, since more than one input triple maps to the same amino acid. One effect of this is to allow genetic robustness, so that mutational events do not affect a gene as easily. Several amino acids are encoded by two or three codons, which differ at the third position in the triplet; this is known as genetic degeneracy, and is a biochemical effect of the tRNAs which actually recognize each codon. The third position degeneracy of a tRNA is called wobble. Thus 20 amino acids are mapped to the 61 codons by a minimum of 31 tRNAs, although the number of tRNA genes varies from species to species, humans having about 500 tRNA genes, representing 48 anticodons.

**tRNAs** are adaptors which map codons to amino acids, but themselves need an adaptor to load an amino acid onto them. This role is satisfied by typically 20 enzymes called aminoacyl-tRNA synthetases, which hydrolyzes ATP to catalyze a covalent bond between the C-terminus of an amino acid and the 3' OH end of a tRNA. These synthetases are as important for genetic decoding as the tRNAs, and so they feature an automatic error-correcting mechanism using two active sites. Loaded tRNAs are called aminoacyl tRNAs.

**start and stop codons:** There are two specially designated codons, START and STOP, which delineate the initiation and termination points along an mRNA sequence. Thus the presence of a START codon, AUG, which codes for a methionine amino acid, specifies that the first position, A, is the first position in an *open reading frame* or *coding region* of nucleotides which code for a functional protein product. There are three stop codons, UAG, UAA, UGA, which specify that the end of this gene is the last position of this codon. Again note that the choice of these triplets is arbitrary, but the sequence and identity of amino acids which are encoded in the mRNA are physically important.

Any area outside of the open reading frame (ORF) of a mRNA transcript is known as an *untranslated region*; thus there may be an UTR 5' of the start codon, and 3' of the stop codon, which includes the polyadenylation sequence.

**Frameshift:** Just to mention it here, mutations which result in a nucleotide insertion or deletion within an ORF are called frameshift mutations, because they change the starting index of a codon, which completely alters the sequence of encoded amino acids following the mutation. Thus if an insertion occurs at the 5' end of a gene, that gene will probably not be functional, and in the worst case it will be a dysfunctional (neomorphic) protein.

#### 3.4.1 Protein synthesis:

Translation actually carried out predominantly by specialized RNAs, including tRNAs and ribozymes. (A ribozyme is an RNA with catalytic activity, and some believe these guys created a pre-DNA life called RNA-world. Ribozymes can dually serve as a basic heritable information template and can interact stably with other RNAs and proteins to catalyze reactions. Ribozymes can also autocatalyze, specifically having the

ability to replicate themselves.) The process which translates an RNA sequence into an amino acid sequence occurs in three steps:

1. **Initiation:** preparation of mRNA by loading proteins
2. **Elongation:** ribosomal processing of mRNA transcript
3. **Termination:** conclusion at stop codon and dismantling of translational machinery

### 3.4.2 Ribosomes

The ribosome is a complex of several rRNAs and 50 proteins which form a locus in which tRNAs can match their anticodons to the codons of an mRNA strand, and which catalyzes an amino-carboxyl peptide bond through its peptidyl transferase activity. The eukaryotic ribosome is an 80S unit composed of 60S and 40S subunits. These major subunits are assembled in the nucleolus within the nucleus. The ribosome looks something like a gumball atop a spindle, between which an mRNA is threaded; the gumball part has three active sites, A (amino-acyl), P (peptidyl), and E (exit), where loaded, chain-bound, and used tRNAs are positioned, respectively. All of the important functions of a ribosome are handled by its RNA constituents. In fact it is the 23S RNA which actually contains the transferase ability, and this is why the ribosome is a ribozyme.

Several molecular inhibitors of protein synthesis exist, which act on the ribosome. These include chloramphenicol, cycloheximide, tetracycline, and  $\alpha$ -Amanitin.

For more information see any biology text.

### 3.4.3 Initiation

The prerequisites for initiation are similar to those of mRNA nuclear export. The cap-binding complex (CBC) must be bound to the 5' cap, composed of eukaryotic initiation factors (eIF). These factors unwind the 5' hairpin structure at the cap. This unwinding exposes the transcript for interaction with the scanning complex. Proteins can bind to individual proteins in the CBC, inhibiting the CBC from binding to the 5' cap and unwinding secondary structure. Also, certain post-translational modifications (i.e. phosphorylation, acetylation, etc) of CBC proteins will cause dissociation of the CBC and inhibit binding of the complex at the 5' cap.

**Binding of the scanning complex (pre-initiation complex):** A scanning complex, composed of eIF2, met-tRNA, and a 40S ribosomal subunit, binds to the mRNA-CBC complex. eIF2 is composed of three subunits ( $\alpha$ ,  $\beta$ ,  $\gamma$ ). The  $\alpha$  subunit is targeted by several kinases, and when it is phosphorylated, it inhibits the formation of eIF2 and prevents translation.

**Ribosomal Scanning:** Once bound to the mRNA-CBC complex, the scanning complex then proceeds to scan the transcript until encountering an AUG (start) codon. At this time, the 60S ribosomal subunit binds to the transcript and initiates translation. The 60S ribosomal subunits initiate translation at the first AUG codon approximately 90-95% of the time. 5-10% of the time, the first AUG codon is bypassed in favor of another one downstream. This indicates that 60S is not blindly searching for the first AUG. Statistical analyses reveal that the nucleotide sequence surrounding the AUG codon is a determining factor. This surrounding sequence is referred to as the Kozak sequence [ACCAUGG] and it influences the efficiency of translation.

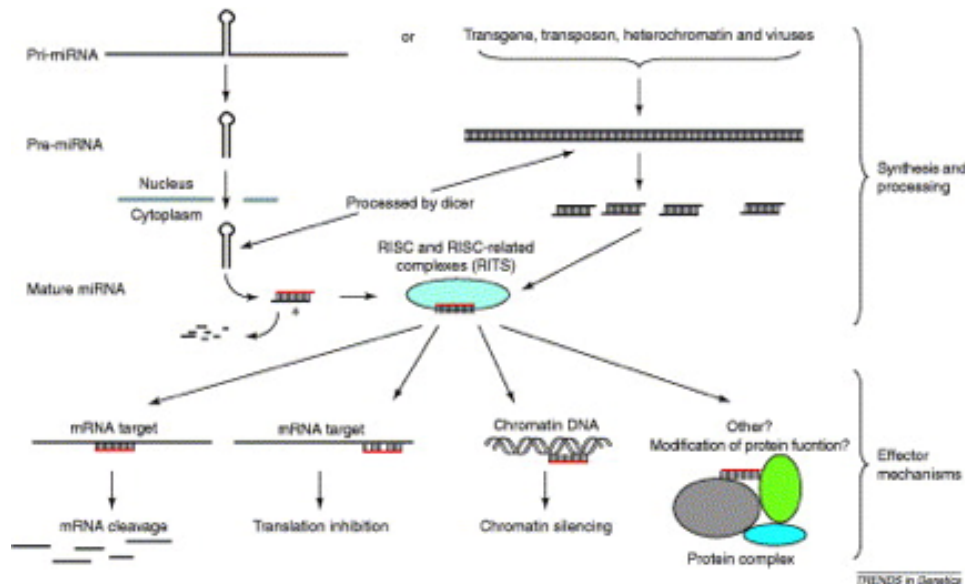
Most sites of translation in eukaryotic cells are found along the endoplasmic reticulum near the nucleus, in places called rough endoplasmic reticulum (RER), which is studded with ribosomes, yielding a grainy microscopic appearance. Many mRNAs are actually simultaneously translated by a slew of ribosomes in a polyribosome structure. Note that all through translation, the polyA binding proteins remain bound to the polyadenylated 3' tail of the mRNA.

### 3.5 mRNA degradation

A cell wants to control how much protein is created from a given mRNA transcript, and so a mechanism for the degradation of mRNA has evolved. Different mRNAs display a wide range of lifetimes in a given cell. For example, beta-globin mRNA has a half-life of seventeen hours, while fos mRNA has a half-life of approximately thirty minutes. Half-lives can be controlled by stabilizing and destabilizing elements within the transcript.

**There are four mechanisms of RNA degradation:**

1. Deadenylation-dependent: The shorter the tail, the less stable the protein. The majority of eukaryotic mRNAs are degraded using this slow-acting process, where a deadenylase enzyme slowly eats away at the tail, resulting in a loss of polyA-binding proteins which are bound to the 3' end. Deadenylase also binds 5' at the cap, and results in the dissociation of the CBC. mRNA is degraded simultaneously from both ends.
2. Deadenylation-independent: While not certain it is thought decapping is followed by endonucleolytic cleavage, followed by degradation from both ends.
3. Nonsense mediated decay: The basic idea is that directly preceding the export of a mature mRNA from the nucleus, a test round of translation is attempted, at or near the nuclear pore. If an in-frame stop codon is reached before the end of the exon boundary, surveillance proteins trigger a decay mechanism.
4. RNAi: RNA interference probably evolved as a defense against retrotransposon replication. The presence of free double stranded RNA attracts a protein complex of a RNA nuclease (Dicer) and helicase, which open and cut the RNA into 23bp fragments of siRNA, which then remain associated with the enzymes and actually help recruit them to other free dsRNA strands because they hybridize with them. This complex also recruits the RNA-induced silencing complex (RISC), in which the siRNA hybridize specifically to mature mRNA transcripts to prevent their translation. Thus RNAi performs a potent but specific temporary negative amplification as a safeguard against future encounters. This mechanism is similar to that of miRNAs.



**Stabilizing elements of transcripts:**

- 5' cap
- PolyA tail

- Few examples of protein binding regions in the 3' UTR that stabilize the transcript, i.e. binding site for  $\alpha$ CP in the 3' UTR of many mRNA transcripts

#### Destabilizing elements:

- Sequences that are particularly susceptible to temperature changes and will misfold upon heat shock
- AU-rich elements in the 3' UTR that are known to confer instability (the mechanism is still being flushed out and may include miRNA activity)
- Sequences in both the ORF and the 3' UTR that recruit proteins, miRNAs, or siRNAs to induce degradation of transcript.

### 3.6 Additional mechanisms of translational regulation

**mRNA masking:** Oocytes contain maternally contributed mRNA. However, some of these transcripts are not translated efficiently until fertilization of the oocyte. So prior to fertilization, these transcripts are bound by RNPs and are **masked** from translation.

**mRNA secondary structure:** Double stranded hairpin structures in the 5' cap of the transcript impede translation initiation. Destabilizing factors can bind to the 5' cap and unwind these hairpins (such as the CBC). Stabilizing factors can also bind to the 5' cap and prevent CBC binding, thus preventing the initiation of translation. In some instances a translational repressor protein, such as aconitase, binds to a hairpin loop created in the 5' UTR to prevent translation during periods of high iron concentration. Another example of a 5' hairpin use is with internal ribosomal entry sites (IRES), where binding of eIF-4G to a 5' hairpin results in the downstream translation of a transcript somewhere within the coding region.

**RNA localization within cells:** Why: Cellular compartment-specific expression and function. This is essential for polarity (i.e. such as in mitosis or in migration) How: Elements typically found in the 3' UTR (300-600 bases) are known to recruit proteins that shuttle the transcripts to the appropriate subcellular locations. Some of these proteins have been identified by UV crosslinking and RNA gel shift assays.

### 3.7 Post-translational modifications

Post-Translational Regulation of Expression occurs primarily through post-translation modifications (PTMs).

#### Functions of PTMs:

1. Activate/Inactivate
2. Alter functionality
3. Affect lifespan
4. Specify subcellular localization
5. Provide a mechanism for multiple functions, such as protein-protein binding interactions

#### Types of PTMs:

1. Phosphorylation
  - The addition or removal of a phosphate group changes the charge in a localized area and has a quick and dramatic effect on a protein's activity level
  - There are over 1000 known kinases and 500 known phosphatases

- One can map sites of phosphorylation by exposing a protein to proteases that cleave at sites of phosphorylation. The peptide fragments can then be sequenced by mass spectrometry.

## 2. Acetylation

- The addition or removal of an acetyl group can affect DNA binding, protein stability, subcellular localization, and protein-protein interaction.
- Histones are well-known examples of proteins that are acetylated by acetyltransferases and affect DNA:histone interaction.

## 3. Ubiquitination/Ubiquitylation

- Ubiquitin is a small protein that can be conjugated to exposed lysines in larger proteins.
- Single-ubiquitin-like modifications (SUMO) can be used to alter the localization of a protein within the cell.
- Poly-ubiquitination is used to target proteins to the proteasome for degradation (a very specific manner of regulating protein activity).  
Pathway to poly-ubiquitination: E1-ubiquitin activating enzyme. E2-ubiquitin conjugating enzyme. E3-ubiquitin ligase

## 4. Methylation

- In prokaryotes, there are many regulatory roles for protein methylation.
- In eukaryotes, there are many proteins that are methylated, but few examples of biological significance. Examples: Homo or Hetero-dimer formation sometimes requires methylation of subunits.
- The methyl acceptor is the carboxyl group of glutamate.

## 5. Others

- Protein cleavage: For example, consider the protease cascade necessary for proper blood clotting
- Lipid Modifications: Useful in localizing proteins to cellular membranes
- Glycosylation: Most proteins that pass through the secretory pathway undergo glycosylation.