Biochemistry of Nucleic Acids

Lecture #21-22

Lecturer Alexander Koval

Content

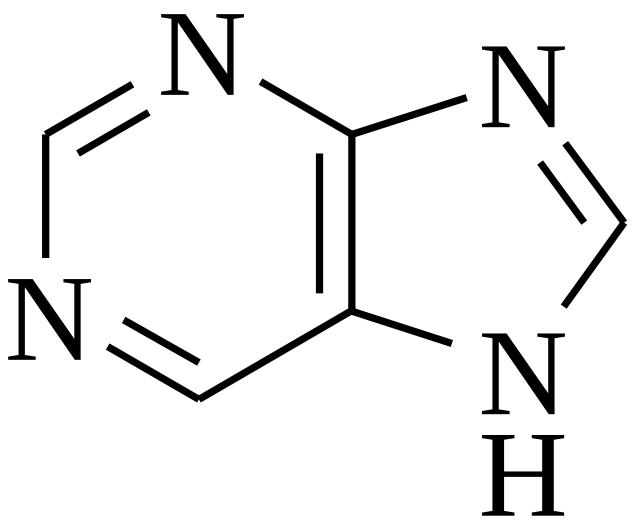
Introduction to Nucleic Acids

- As a class, the nucleotides may be considered one of the most important metabolites of the cell.
  - Nucleotides are found primarily as the **monomeric units** comprising the major **nucleic acids** of the cell, **RNA and DNA**.
  - They also are required for numerous other important functions within the cell.

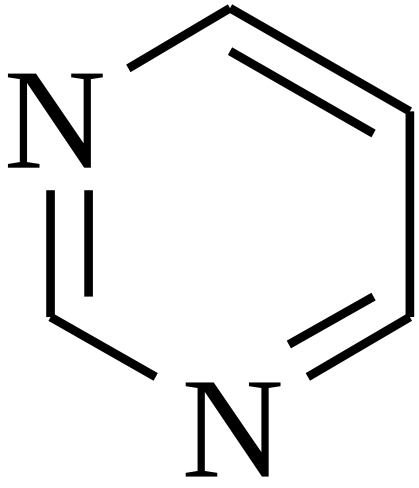
…These functions include:

1. serving as energy stores for future use in phosphate transfer reactions. These reactions are predominantly carried out by ATP.
2. forming a portion of several important coenzymes such as NAD^+, $NADP^+$, FAD and coenzyme A.
3. serving as mediators of numerous important cellular processes such as second messengers in signal transduction events.
    1. The predominant second messenger is cyclic-AMP (cAMP), a cyclic derivative of AMP formed from ATP.
4. controlling numerous enzymatic reactions through allosteric effects on enzyme activity.
5. serving as activated intermediates in numerous biosynthetic reactions.
    1. These activated intermediates include **S-adenosylmethionine (S-AdoMet)** involved in methyl transfer reactions as well as the many sugar coupled nucleotides involved in glycogen and glycoprotein synthesis.

Nucleoside and Nucleotide Structure and Nomenclature



purine



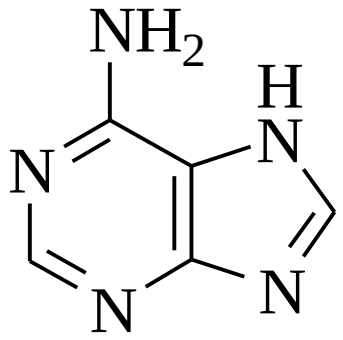
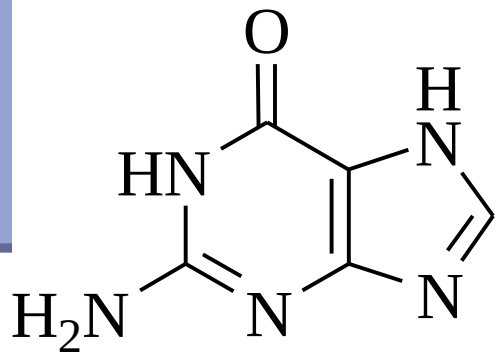
pyrimidine

- The nucleotides found in cells are derivatives of the heterocyclic highly basic, compounds, **purine and pyrimidine.**

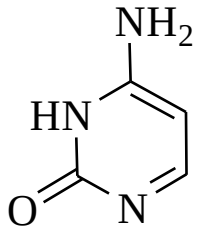
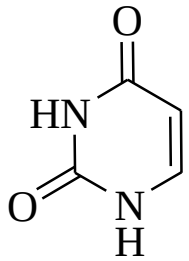
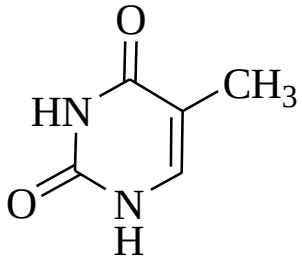
Five Major Bases

- It is the chemical basicity of the nucleotides that has given them the common term "bases" as they are associated with nucleotides present in DNA and RNA.
  - There are five major bases found in cells.
  - The derivatives of purine are called **adenine** and **guanine**, and the derivatives of pyrimidine are called **thymine**, **cytosine** and **uracil**.
  - The common abbreviations used for these five bases are, A, G, T, C and U

Purines

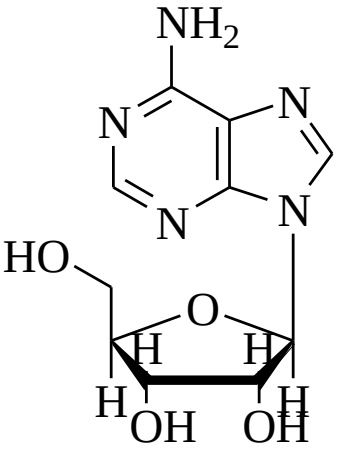
	Adenine, A	Adenosine, A	Adenosine monophosphate AMP
	Guanine, G	Guanosine, G	Guanosine monophosphate GMP

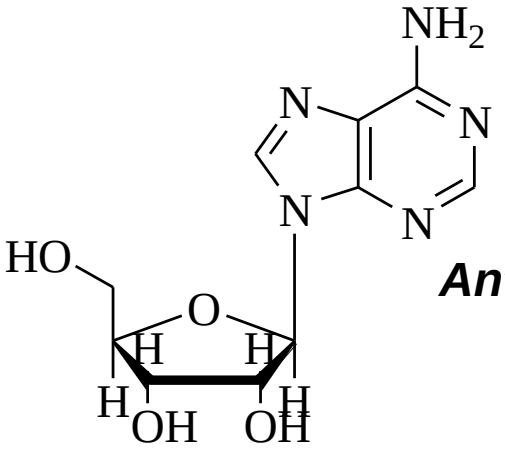
Pyrimidines

Base Formula	Base (X=H)	Nucleoside X=ribose or deoxyribose	Nucleotide X=ribose phosphate
	Cytosine, C	Cytidine, C	Cytidine monophosphate CMP
	Uracil, U	Uridine, U	Uridine monophosphate UMP
	Thymine, T	Thymidine, T	Thymidine monophosphate TMP

syn- and ***anti-*** conformers

Syn-Adenosine

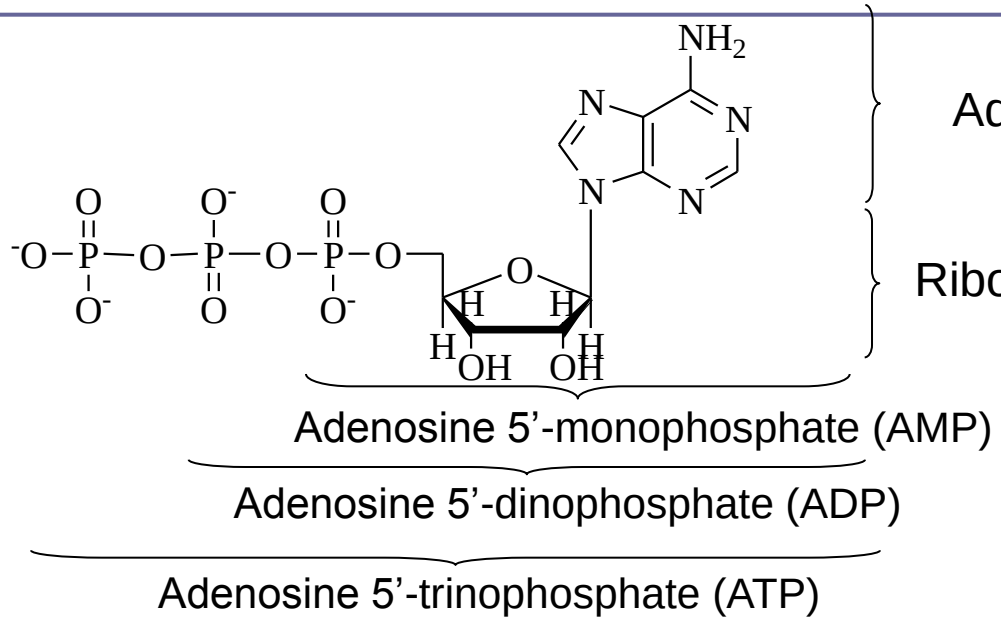




Anti-Adenosine

- The purine and pyrimidine bases in cells are linked to carbohydrate and in this form are termed, **nucleosides**.
  - The nucleosides are coupled to D-ribose or 2'-deoxy-D-ribose through a β-N-glycosidic bond between the anomeric carbon of the ribose and the N^9 of a purine or N^1 of a pyrimidine.
  - The base can exist in 2 distinct orientations about the N-glycosidic bond. These conformations are identified as, ***syn*** and ***anti***.
    - It is the **anti** conformation that predominates in naturally occurring nucleotides.

Nucleotides – AMP, ADP, ATP



- Nucleosides are found in the cell primarily in their phosphorylated form. These are termed **nucleotides**.
- The most common site of phosphorylation of nucleotides found in cells is the hydroxyl group attached to the 5'-carbon of the ribose The carbon atoms of the ribose present in nucleotides are designated with a prime (') mark to distinguish them from the backbone numbering in the bases.
- Nucleotides can exist in the mono-, di-, or tri-phosphorylated forms.

Phosphates of nucleotides

- Nucleotides are given distinct abbreviations to allow easy identification of their structure and state of phosphorylation.
  - The monophosphorylated form of adenosine (adenosine-5'-monophosphate) is written as, AMP.
  - The di- and tri-phosphorylated forms are written as, ADP and ATP, respectively.
  - The use of these abbreviations assumes that the nucleotide is in the 5'-phosphorylated form.
- The di- and tri-phosphates of nucleotides are linked by acid anhydride bonds.
  - Acid anhydride bonds have a high $\Delta G^{0'}$ for hydrolysis imparting upon them a high potential to transfer the phosphates to other molecules.
- It is this property of the nucleotides that results in their involvement in group transfer reactions in the cell.

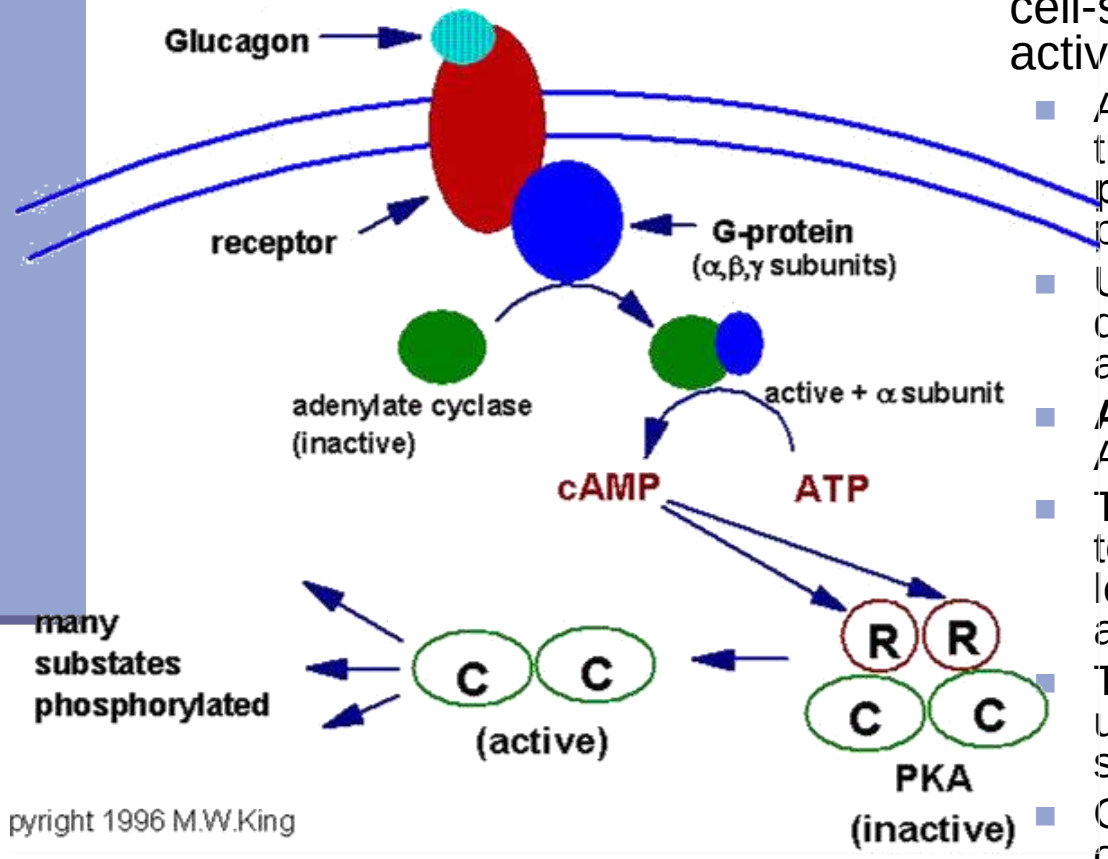
Deoxy- and minor nucleotides

- The nucleotides found in DNA are unique from those of RNA in that the ribose exists in the 2'-deoxy form and the abbreviations of the nucleotides contain a d designation. The monophosphorylated form of adenosine found in DNA (deoxyadenosine-5'-monophosphate) is written as dAMP.
  - The nucleotide **uridine** is never found in DNA and thymine is almost exclusively found in DNA.
  - Thymine is found in tRNAs but not rRNAs nor mRNAs. There are several less common bases found in DNA and RNA.
  - The primary modified base in DNA is **5-methylcytosine**.
  - A variety of modified bases appear in the tRNAs. Many modified nucleotides are encountered outside of the context of DNA and RNA that serve important biological functions.

Adenosine Derivatives

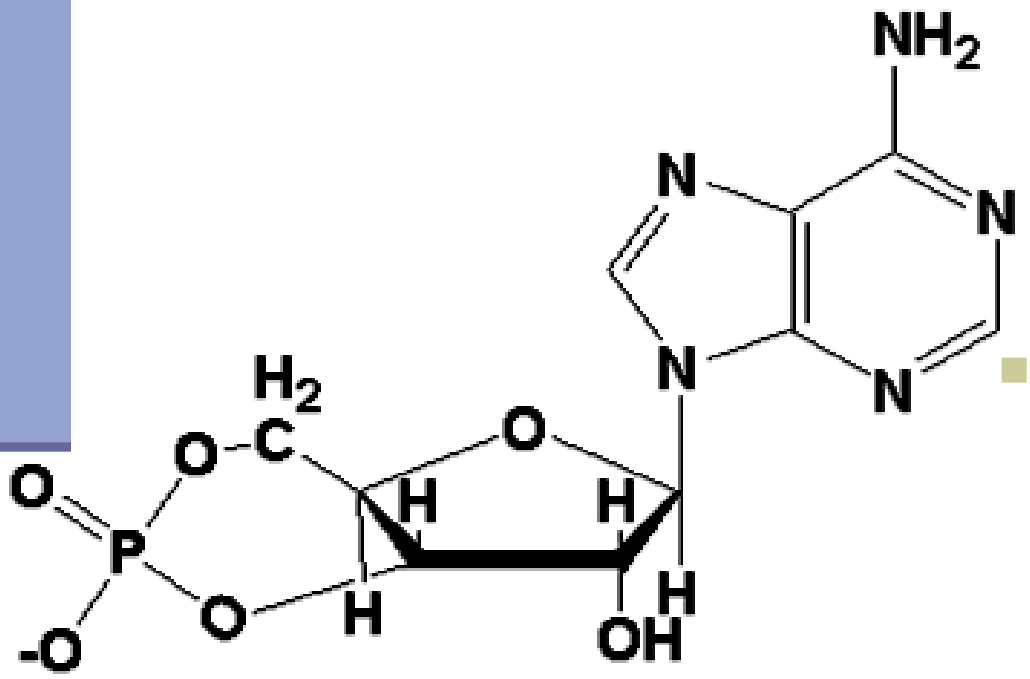
- The most common adenosine derivative is the cyclic form, **3'-5'-cyclic adenosine monophosphate, cAMP**.
  - This compound is a very powerful second messenger involved in passing signal transduction events from the cell surface to internal proteins, e.g. **cAMP-dependent protein kinase, PKA**.

Receptor-Mediated Activation of PKA



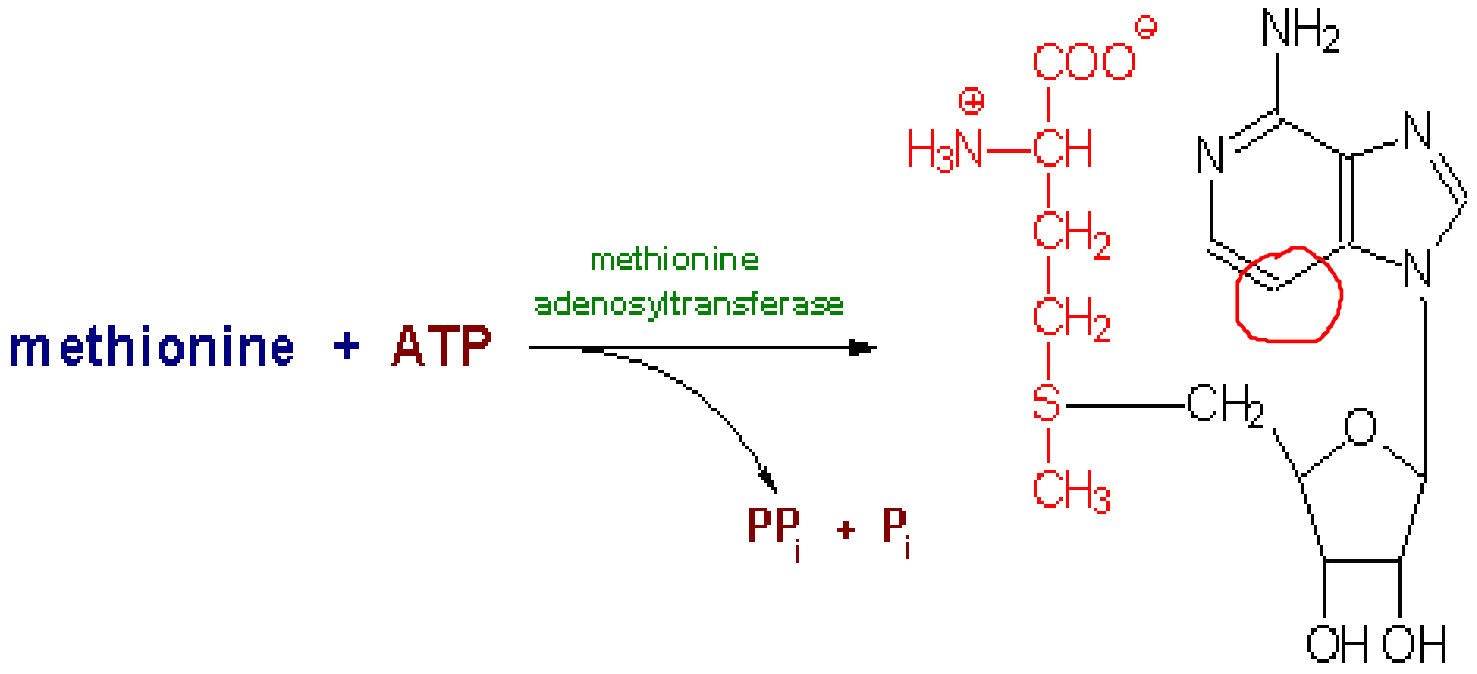
- In this example **glucagon** binds to its' cell-surface receptor, thereby activating the receptor.
  - Activation of the receptor is coupled to the activation of a receptor-coupled **G-protein** (GTP-binding and hydrolyzing protein) composed of 3 subunits.
  - Upon activation the **α-subunit** dissociates and binds to and activates adenylate cyclase.
  - **Adenylate cylcase** then converts ATP to cyclic-AMP (**cAMP**).
  - The cAMP thus produced then binds to the **regulatory subunits of PKA** leading to dissociation of the associated catalytic subunits.
  - The **catalytic subunits** are inactive until dissociated from the regulatory subunits.
  - Once released the catalytic subunits of PKA **phosphorylate numerous substrate** using ATP as the phosphate donor.

cAMP



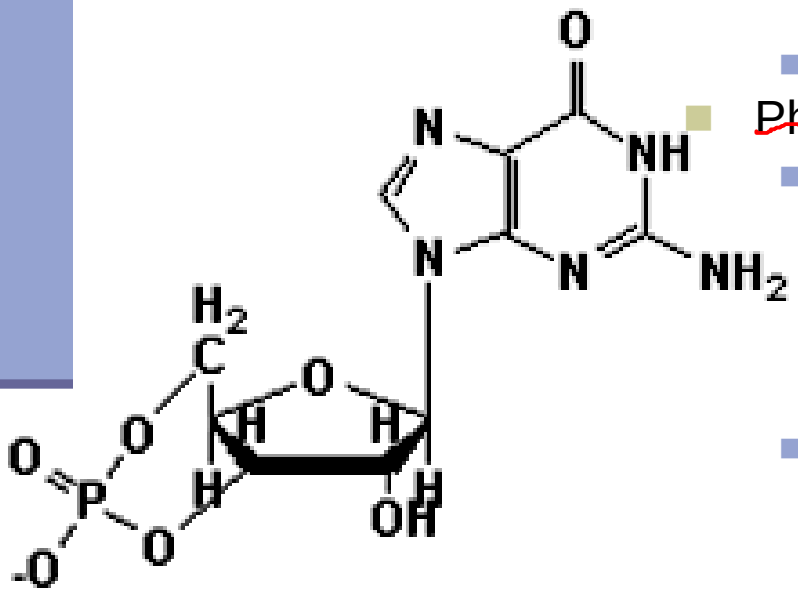
- PKA phosphorylates a number of proteins, thereby, affecting their activity either positively or negatively.
  - Cyclic-AMP is also involved in the regulation of ion channels by direct interaction with the channel proteins, e.g. in the activation of odorant receptors by odorant molecules.
- Formation of cAMP occurs in response to activation of receptor coupled **adenylate cyclase**.
  - These receptors can be of any type, e.g. hormone receptors or odorant receptors.

S-adenosylmethionine (Structure and Synthesis)



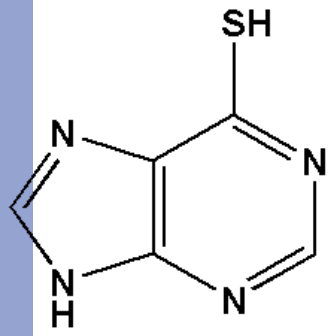
- **S-adenosylmethionine** is a form of activated methionine which serves as a methyl donor in methylation reactions and as a source of propylamine in the synthesis of polyamines.

Guanosine Derivatives

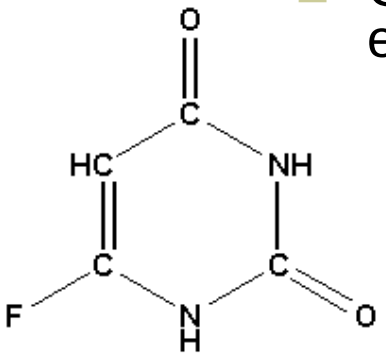


- A cyclic form of GMP (cGMP) also is found in cells involved as a second messenger molecule.
  - antagonize the effects of cAMP.
  - formation of cGMP occurs in response to receptor mediated signals.
  - guanylate cyclase is coupled to the receptor.
- Photoreception.
  - activation of rhodopsin (in the rods) by the absorption of a photon of light (through 11-cis-retinal covalently associated with rhodopsin and opsins) activates transducin which in turn activates a cGMP specific phosphodiesterase that hydrolyzes cGMP to GMP.
  - Concentration of cGMP bound to gated ion channels lowers resulting in their closure and a concomitant hyperpolarization of the cell.

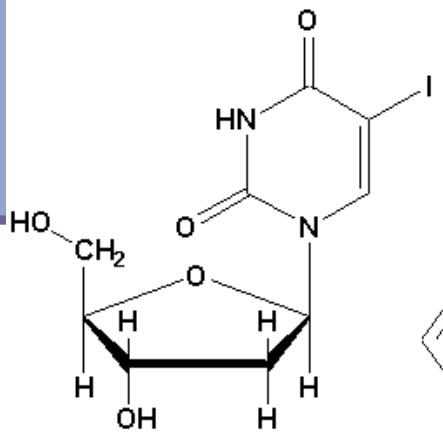
Nucleotide Analogs



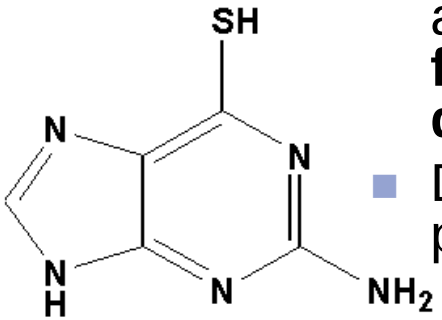
6-Mercaptopurine



Fluorouracil



5-Iododeoxyuridine

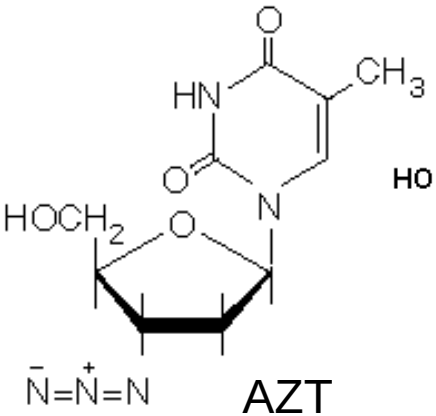


6-Thioguanine

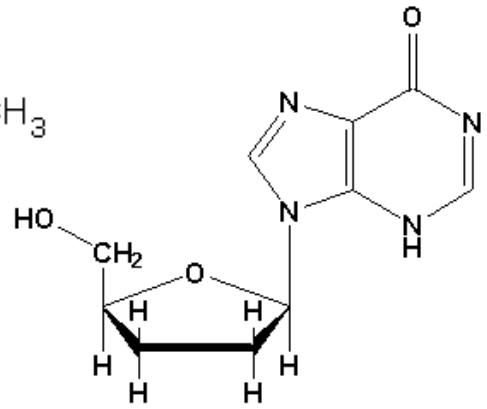
- Chemically synthesized.
- Can be utilized to inhibit specific enzymatic activities.
  - A large family of analogues are used as anti-tumor agents,
    - because they interfere with the synthesis of DNA and thereby preferentially kill rapidly dividing cells such as tumor cells.
  - Some of the nucleotide analogues commonly used in chemotherapy are **6-mercaptopurine**, **5-fluorouracil**, **5-iodo-2'-deoxyuridine** and **6-thioguanine**.
  - Disrupts the normal replication process.

Nucleotide Analogs as Antiviral Agents

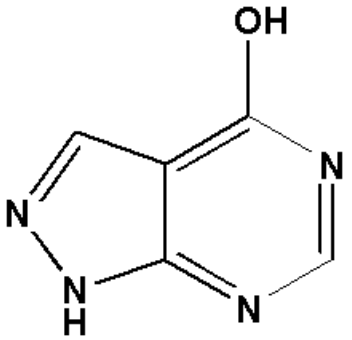
- Interfere with the replication of HIV
  - AZT (azidothymidine),
  - ddI (dideoxyinosine).
- Treatment of **gout:**
  - **Allopurinol** resembles hypoxanthine.
  - Inhibits the activity of *xanthine oxidase*.
- Several nucleotide analogues are used after organ transplantation in order to suppress the immune system and reduce the likelihood of transplant rejection by the host.



AZT



2',3'-Dideoxyinosine (ddI)



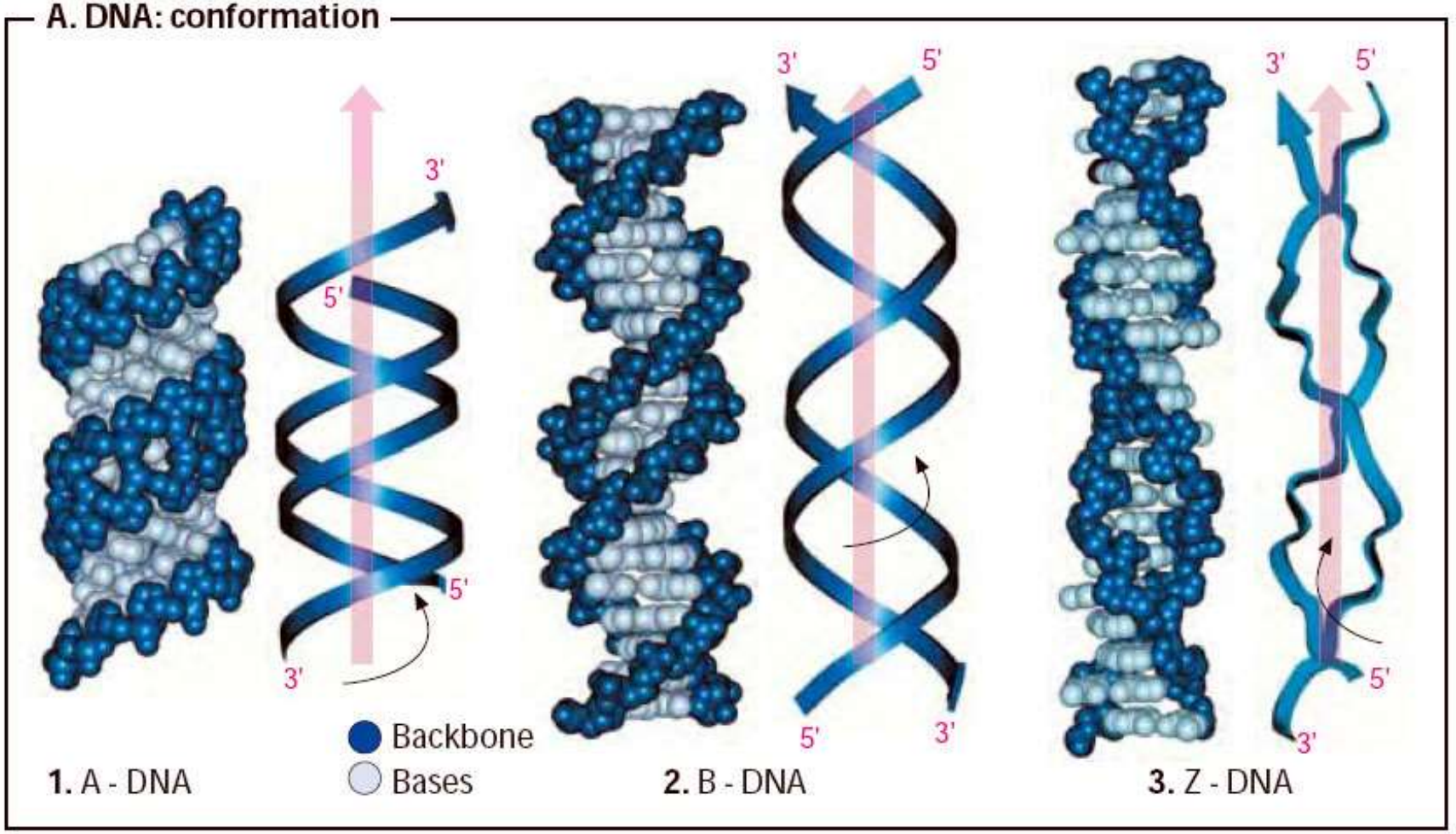
Allopurinol

Polynucleotides

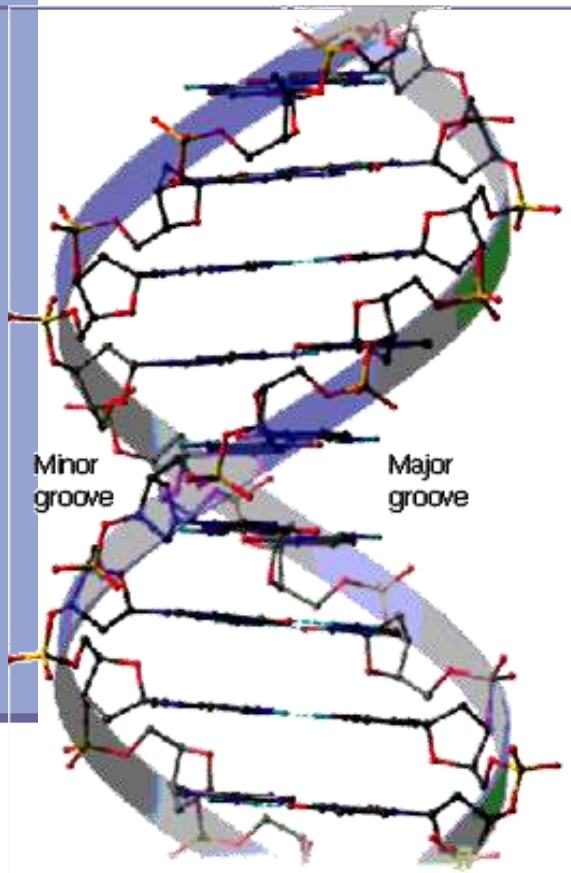
- **Polynucleotides** are formed by the condensation of two or more nucleotides.
  - Phosphodiester bond is formed.
  - The primary structure of DNA and RNA (the linear arrangement) proceeds in the 5' → 3' direction.
  - The common representation of the primary structure of DNA or RNA molecules :

5'-pGpApTpC-3'

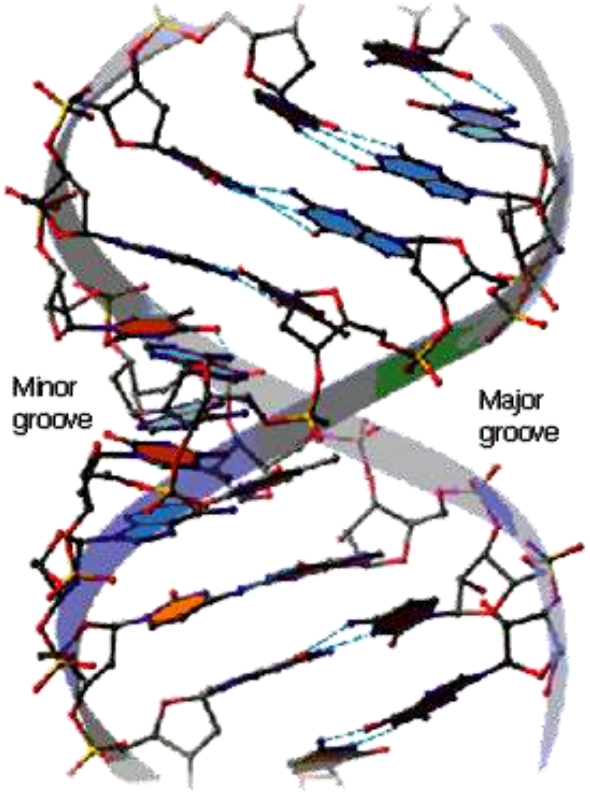
The Structure of DNA: B, A, Z forms



B-DNA and A-DNA



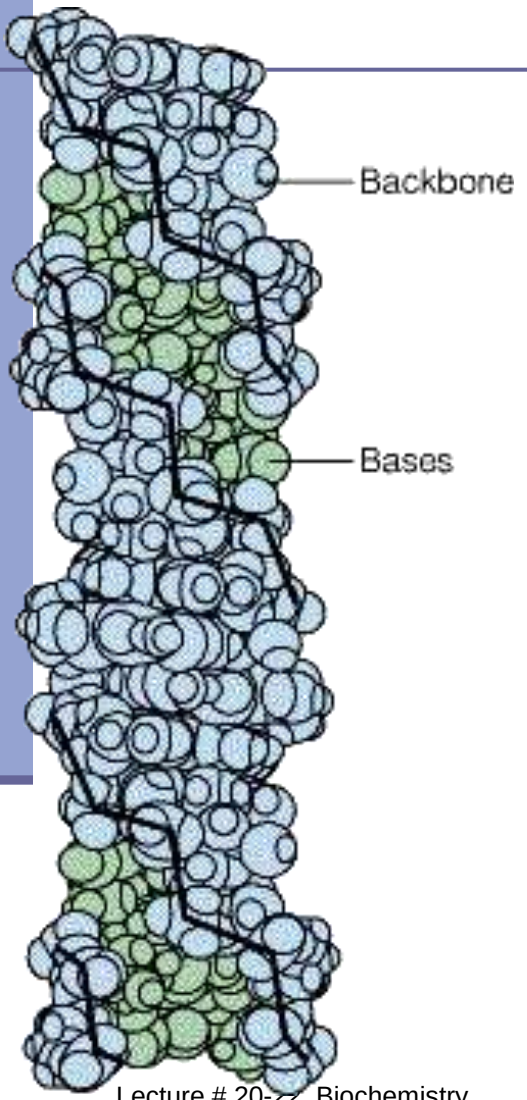
b) B-DNA, side view



(d) A-DNA, side view

- DNA is found in three forms:
  - B-DNA,
  - A-DNA,
  - Z-DNA.
- B-DNA form predominates.
- Both B- and A-DNA have major and minor grooves.

Z-DNA



- Both B- and A-DNA are right-handed helices
- Z-DNA is a left-handed helix.
  - Next table compares some of the physical parameters of B-, A-, and Z-DNA.

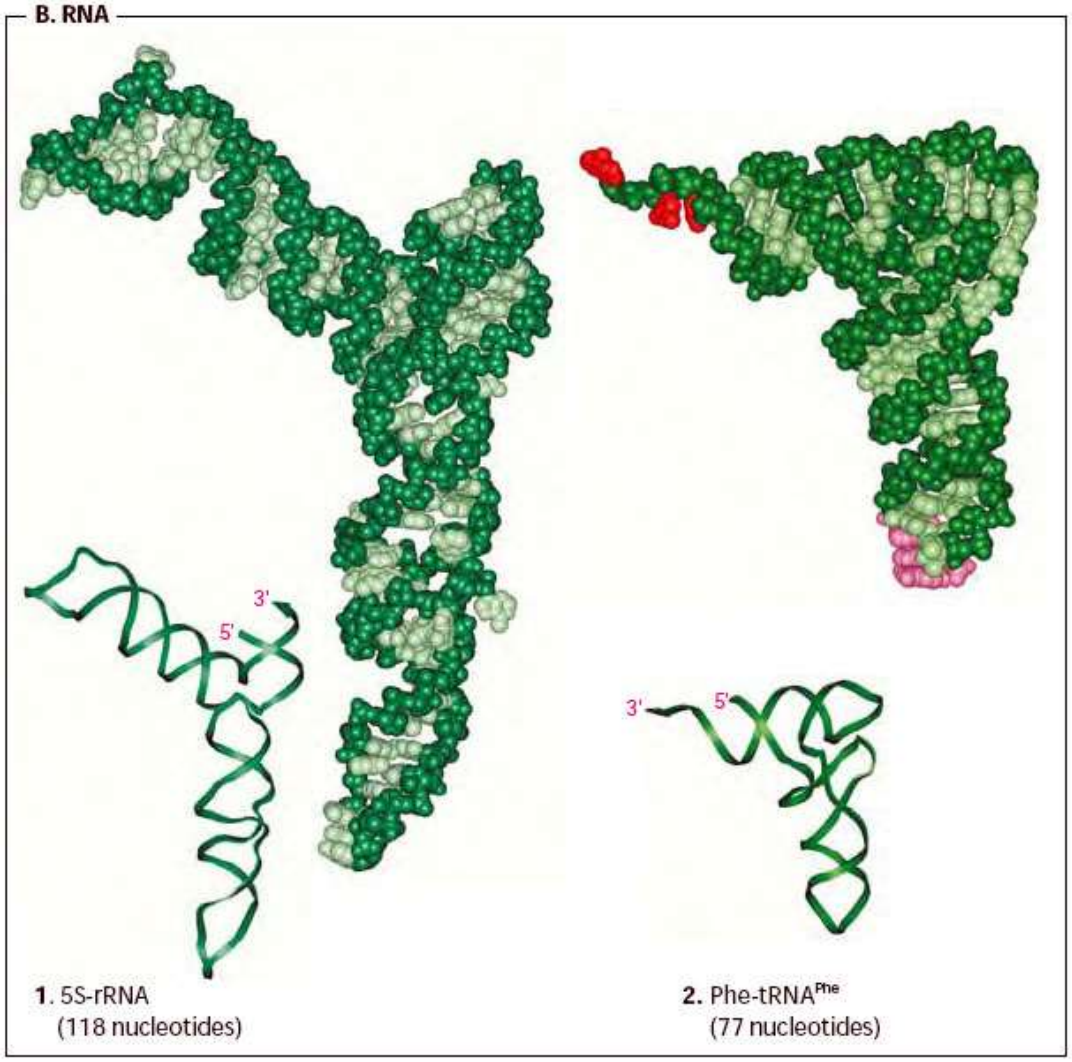
Parameters of Major DNA Helices

Parameters	A Form	B Form	Z-Form
Direction of helical rotation	Right	Right	Left
Residues per turn of helix	11	10	12 base pairs
Rotation of helix per residue (in degrees)	33	36	-30
Base tilt relative to helix axis (in degrees)	20	6	7
Major groove	narrow and deep	wide and deep	Flat
Minor groove	wide and shallow	narrow and deep	narrow and deep
Orientation of N-glycosidic Bond	Anti	Anti	Anti for Py, Syn for Pu
Comments		most prevalent within cells	occurs in stretches of alternating purine-pyrimidine base pairs

Thermal Properties of the Double Helix

- Separation of two strands of DNA: **denaturation**.
  - can be carried out in vitro. At high temperature: **thermal denaturation**.
  - Regions that have predominantly A-T base-pairs are less thermally stable than those rich in G-C base-pairs.
- When thermally melted DNA is cooled, the complementary strands will again re-form the correct base pairs, in a process is termed **annealing** or **hybridization**.
  - The rate of annealing is dependent upon the nucleotide sequence of the two strands of DNA.

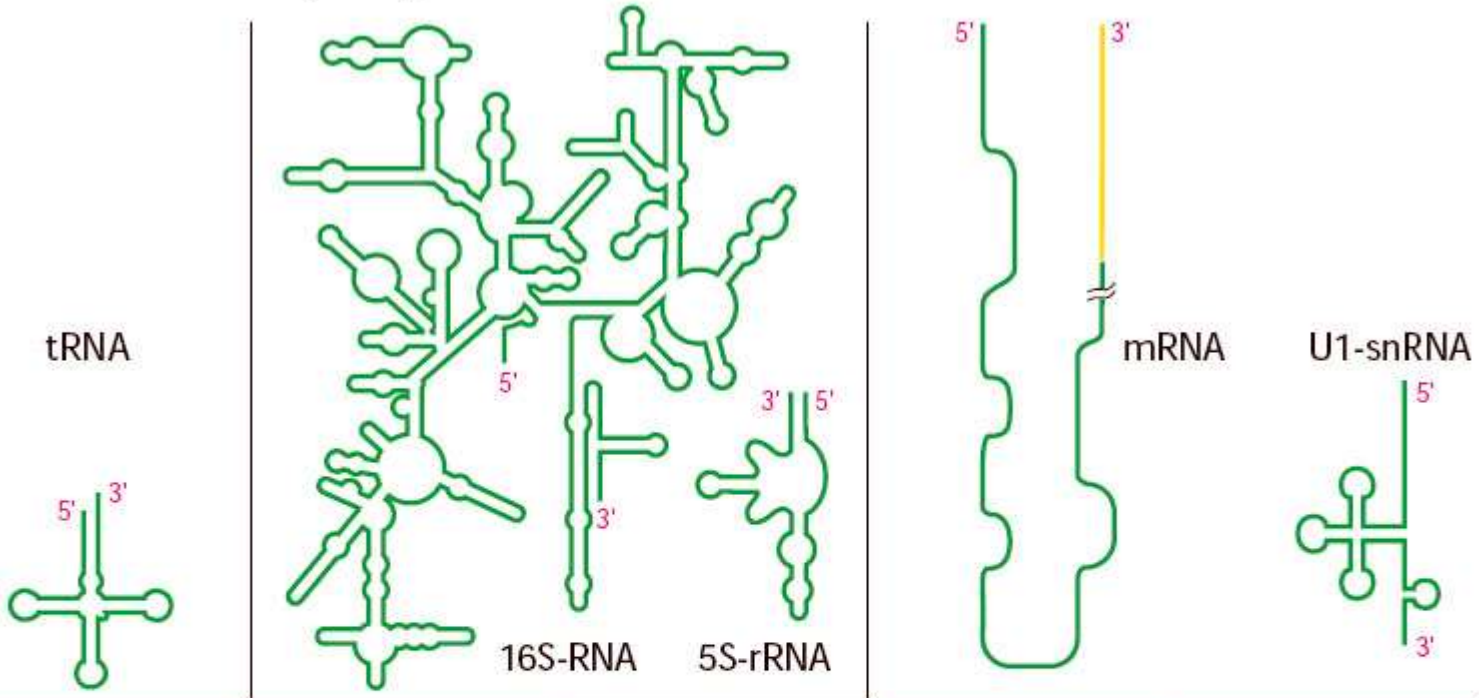
RNA Molecules



- RNA molecules are unable to form extended double helices, and are therefore less highly ordered than DNA molecules.
  - they have defined secondary and tertiary structures, and most of the nucleotides are base paired. Examples:
    - 5S-rRNA, in ribosomes,
    - tRNA molecule from yeast that is specific for phenylalanine.
  - molecules are folded, 3' and the 5'end are close together.
  - three bases of the anticodon (pink) interact with mRNA
  - lie on the surface of the molecule.
- The bases of the CCA triplet at the 3' end (red) also jut outward.
- During amino acid activation, they are recognized and bound by the ligases.

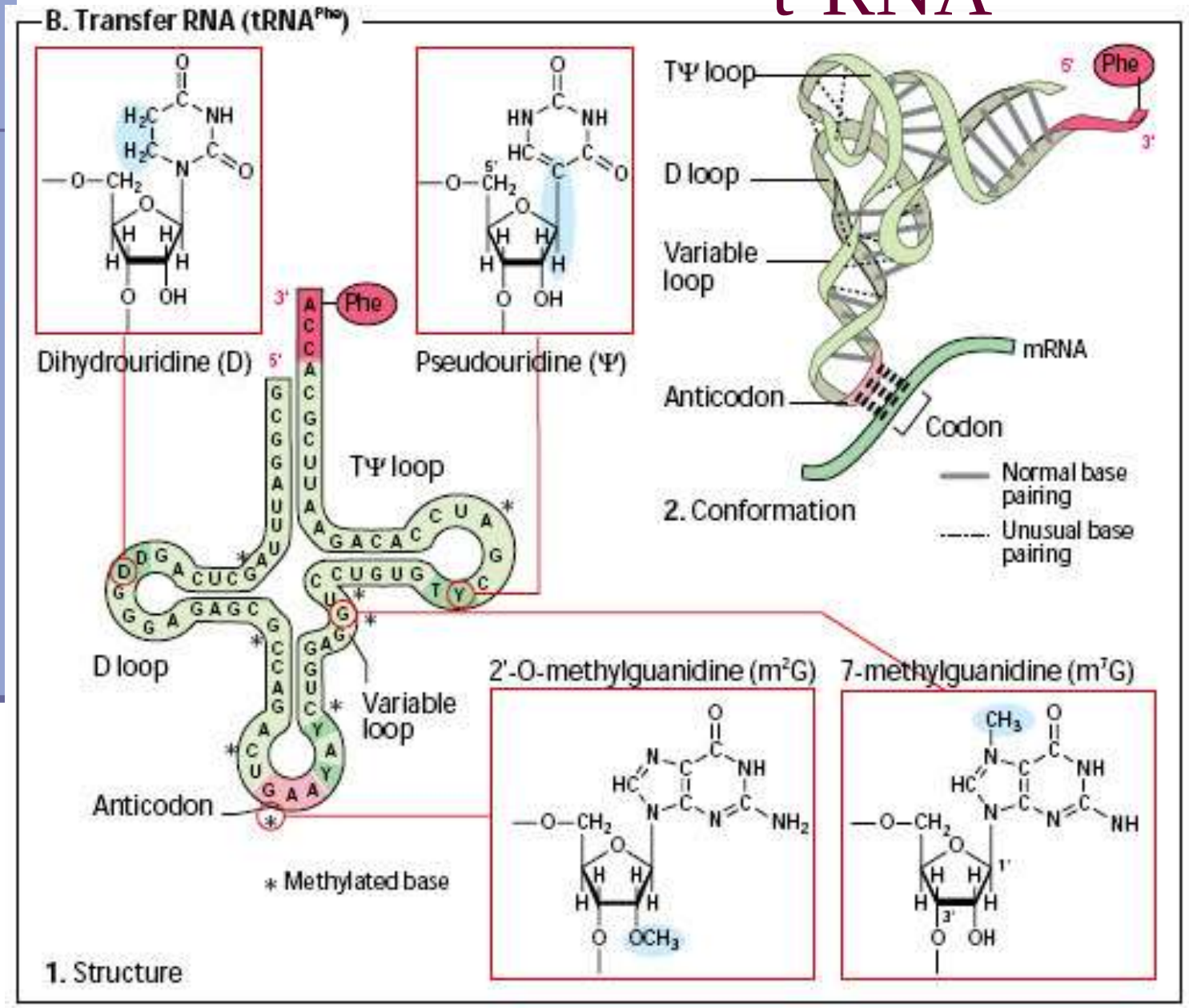
RNA Types

A. Ribonucleic acids (RNAs)



tRNA	rRNA	Type	mRNA	snRNA
>50	4	Species per cell	> 1000	~ 10
74 - 95	120 - 5000	Length (b)	400 - 6000	100 - 300
10-20%	80%	Proportion	5%	< 1%
Long	Long	Lifespan	Short	Long
Translation	Translation	Function	Translation	Splicing

t-RNA


B. Transfer RNA (tRNA^Phe)
Dihydrouridine (D)
Pseudouridine (Ψ)
3'
Phe
5'
TΨ loop
D loop
Variable loop
Anticodon
* Methylated base
1. Structure
TΨ loop
D loop
Variable loop
mRNA
Anticodon
Codon
2. Conformation
Normal base pairing
Unusual base pairing
2'-O-methylguanidine (m²G)
7-methylguanidine (m⁷G)

RNA Synthesis. Introduction

- Transcription is the mechanism by which a template strand of DNA is utilized to generate one of the three different classes of RNA:
  1. **Messenger RNAs** (mRNAs):
     1. the genetic **coding** templates to determine the order of amino acids incorporated into a polypeptide in the process of **translation**.
  2. **Transfer RNAs** (tRNAs):
     1. small RNAs, form covalent attachments to individual amino acids and recognize the encoded sequences of the mRNAs for correct insertion of AAs into the polypeptide chain.
  3. **Ribosomal RNAs** (rRNAs):
     1. Assembled together with ribosomal proteins to form the **ribosomes**.
     2. Ribosomes engage the mRNAs and form a catalytic domain into which the tRNAs enter with their attached amino acids.
     3. The **proteins of the ribosomes** catalyze all of the functions of polypeptide synthesis.

RNA vs DNA

- All RNA polymerases are dependent upon a DNA template in order to synthesize RNA.
- The resultant RNA is, therefore, complimentary to the template strand of the DNA duplex and identical to the non-template strand.
- The non-template strand is called the coding strand because its' sequences are identical to those of the mRNA. However, in RNA, U is substituted for T.

Classes of RNA Polymerases

- In prokaryotic cells, all 3 RNA classes are synthesized by a single polymerase.
- In eukaryotic cells there are 3 distinct classes of RNA polymerase, **RNA polymerase (pol) I, II and III**. Each polymerase is responsible for the synthesis of a different class of RNA.
- The capacity of the various polymerases to synthesize different RNAs was shown with the toxin α-amanitin.
  - At low concentrations of α-amanitin synthesis of mRNAs are affected but not rRNAs nor tRNAs.
  - At high concentrations, both mRNAs and tRNAs are affected.
- These observations have allowed the identification of which polymerase synthesizes which class of RNAs.
- RNA pol I is responsible for rRNA synthesis (excluding the 5S rRNA).
  - There are 4 major rRNAs in eukaryotic cells designated by there sedimentation size.
  - The 28S, 5S 5.8S RNAs are associated with the large ribosomal subunit and the 18S rRNA is associated with the small ribosomal subunit.
- RNA pol II synthesizes the mRNAs and some of the small nuclear RNAs (snRNAs) involved in RNA splicing.
- RNA pol III synthesizes the tRNAs, the 5S rRNA and some snRNAs.

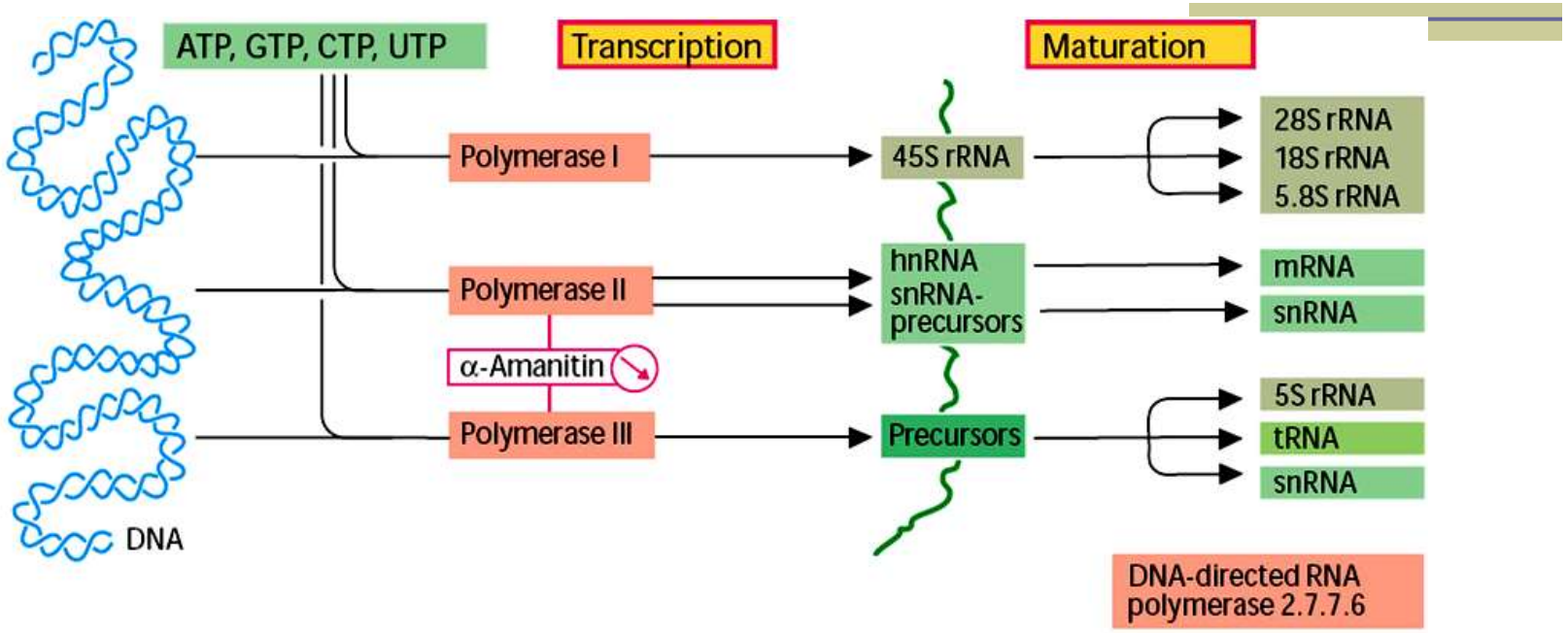
Mechanism of RNA Polymerases

- Synthesis of RNA exhibits several features that are synonymous with DNA replication.
  - RNA synthesis requires accurate and efficient initiation,
  - elongation proceeds in the 5' ---> 3' direction (i.e. the polymerase moves along the template strand of DNA in the 3' ----> 5' direction),
  - and RNA synthesis requires distinct and accurate termination.

Transcription exhibits several features

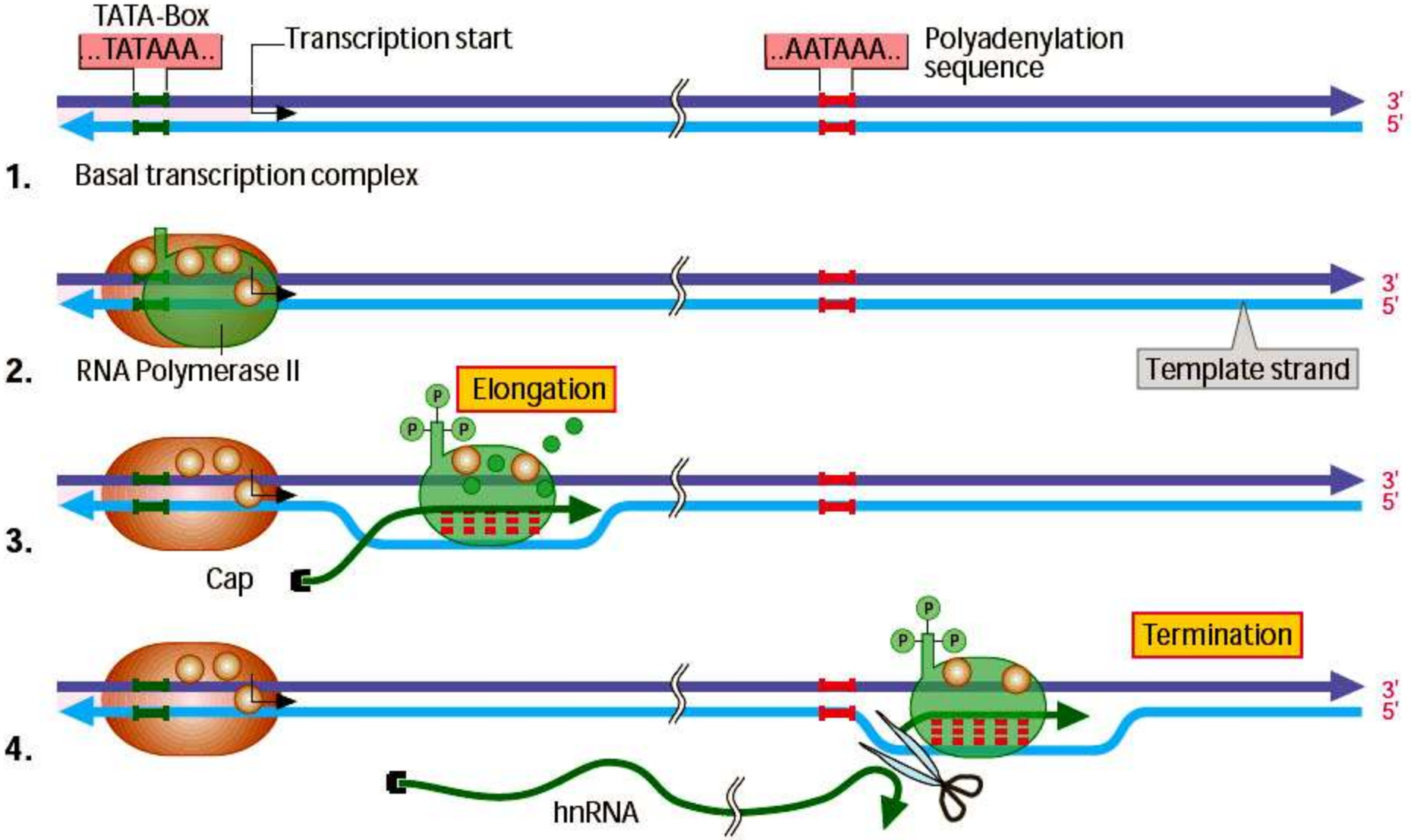
- Transcription exhibits several features that are distinct from replication.
  1. Transcription initiates, both in prokaryotes and eukaryotes, from many more sites than replication.
  2. There are many more molecules of RNA polymerase per cell than DNA polymerase.
  3. RNA polymerase proceeds at a rate much slower than DNA polymerase (approximately 50-100 bases/sec for RNA versus near 1000 bases/sec for DNA).
  4. Finally the fidelity of RNA polymerization is much lower than DNA. This is allowable since the aberrant RNA molecules can simply be turned over and new correct molecules made.

Transcription and Maturation of RNA: Overview



- Eukaryotic cells contain at least three different types of RNA polymerase. *RNA polymerase I* synthesizes an RNA with a sedimentation coefficient of 45 S, which serves as precursor for three ribosomal RNAs. The products of *RNA polymerase II* are hnRNAs, from which mRNAs later develop, as well as precursors for snRNAs. Finally, *RNA polymerase III* transcribes genes that code for tRNAs, 5S rRNA, and certain snRNAs. These precursors give rise to functional RNA molecules by a process called **RNA maturation**. Polymerases II and III are inhibited by *amanitin,* a toxin in the *Amanita phalloides* mushroom.

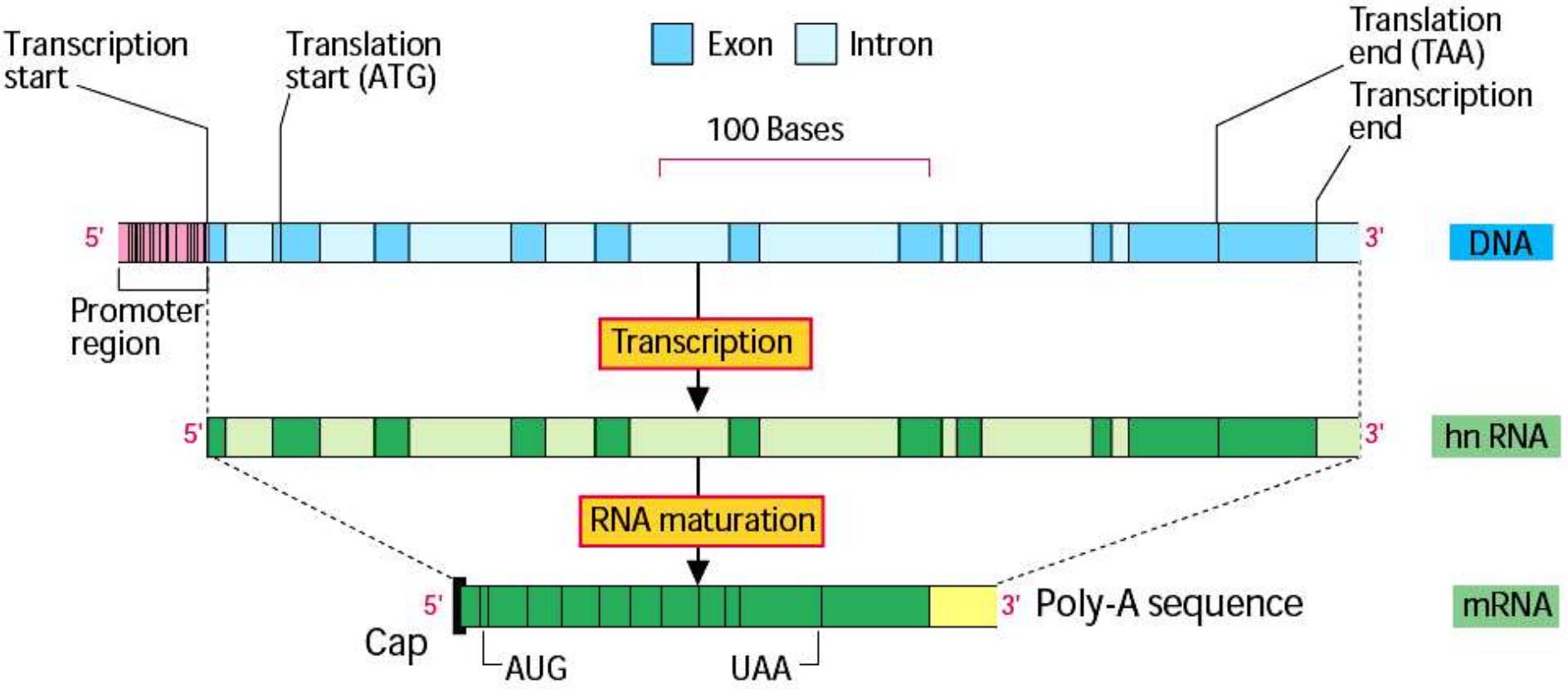
Process of transcription



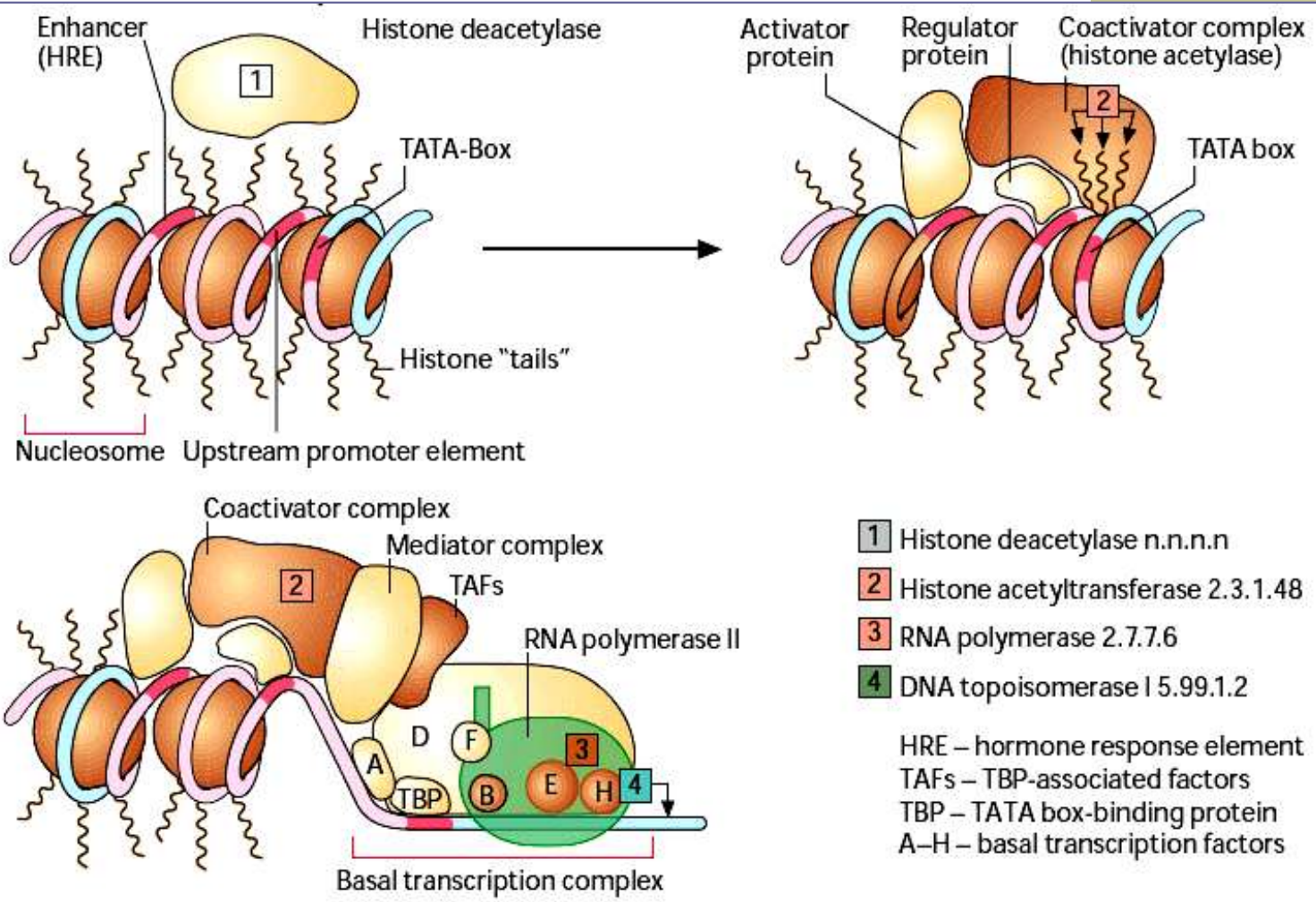
Processes of Transcription: Initiation

- The genome sequence elements (**promoters)** promote the ability of RNA polymerases to recognize the nucleotide at which initiation begins.
- Additional sequence elements are present within genes that enhance polymerase activity. These sequence elements are termed **enhancers**.
- Transcriptional promoter and enhancer elements are important sequences used in the control of gene expression.

Organization of the PEP-CK gene



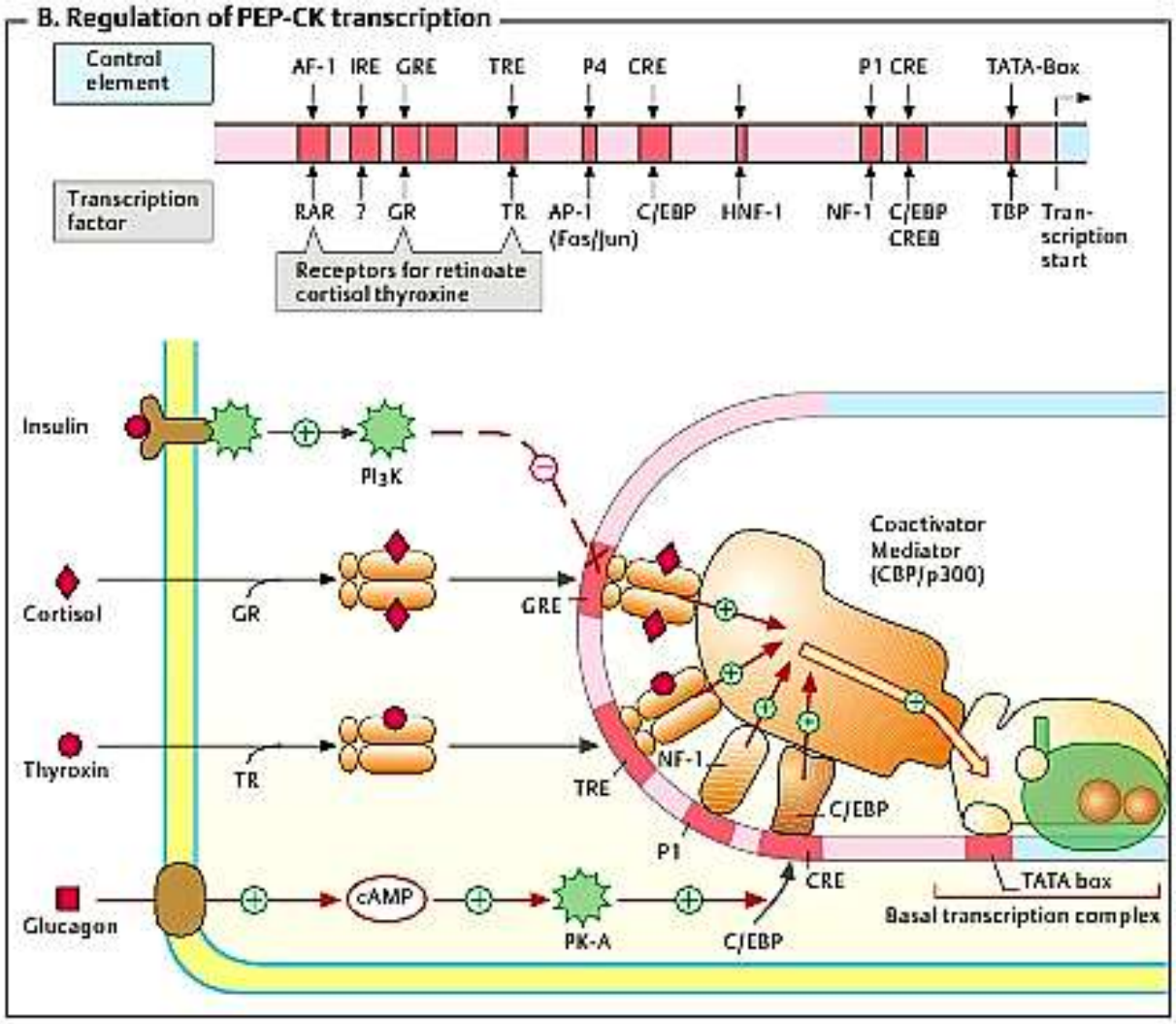
Initiation of transcription



Basal Transcription Complex

- This consists of DNA-dependent RNA polymerase II and basal transcription factors (TFIIX, X = A – H). First, the basal factor TFIID binds to the promoter. TFIID, a large complex of numerous proteins, contains TATA boxbinding protein (TBP) and so-called TAFs (TBP-associated factors). The polymerase is attached to this core with the help of TFIIB. Before transcription starts, additional TFs have to bind, including TFIIH, which has helicase activity and separates the two strands of DNA during elongation. In all, some 35 different proteins are involved in the basal complex.
- This alone, however, is still not suficient for transcription to start. In addition, positive signals have to be emitted by more distant trans-active factors, integrated by the coactivator/mediator complex, and passed on to the basal complex (see B). The actual signal for starting elongation consists of the multiple phosphorylation of a domain in the C-terminal region of the polymerase. In phosphorylated form, it releases itself from the basal complex along with a few TFs and starts to synthesize hnRNA.

Regulation of PEP-CK transcription



Phosphoenolpyruvate carboxykinase (PEP-CK), a key enzyme in gluconeogenesis, is regulated by several hormones, all of which affect the transcription of the PEP-CK gene. Cortisol, glucagon, and thyroxin induce PEP-CK, while insulin inhibits its induction.

Regulation of PEP-CK transcription (cont'd)

- More than ten **control elements** (dark red), distributed over approximately 1 kbp, have been identified in the promoter of the PEP-CK gene.
- These include: response elements
  - for the glucocorticoid receptor (GRE),
  - for the thyroxin receptor (TRE),
  - and for the steroid-like retinoic acid (AF-1).
- Additional control elements (CRE, cAMP-responsive element) bind the transcription factor C/EBP, which is activated by cAMP-dependent protein kinase A through phosphorylation.
  - This is the way in which **glucagon**, which raises the cAMP level, works.
- Control element P1 binds the hormone-independent factor NF-1 (nuclear factor-1).

Regulation of PEP-CK transcription (cont'd)

- All the proteins are in contact with a **coactivator/mediator complex** (CBP/p300):
  - integrates their input and transmits the result in the form of stronger or weaker signals to the basal transcription complex.
- Inhibition of PEP-CK transcription by insulin is mediated by an insulin-responsive element (IRE) in the vicinity of the GRE.
- Binding of an as yet unknown factor takes place here, inhibiting the binding of the glucocorticoid receptor to the GRE.

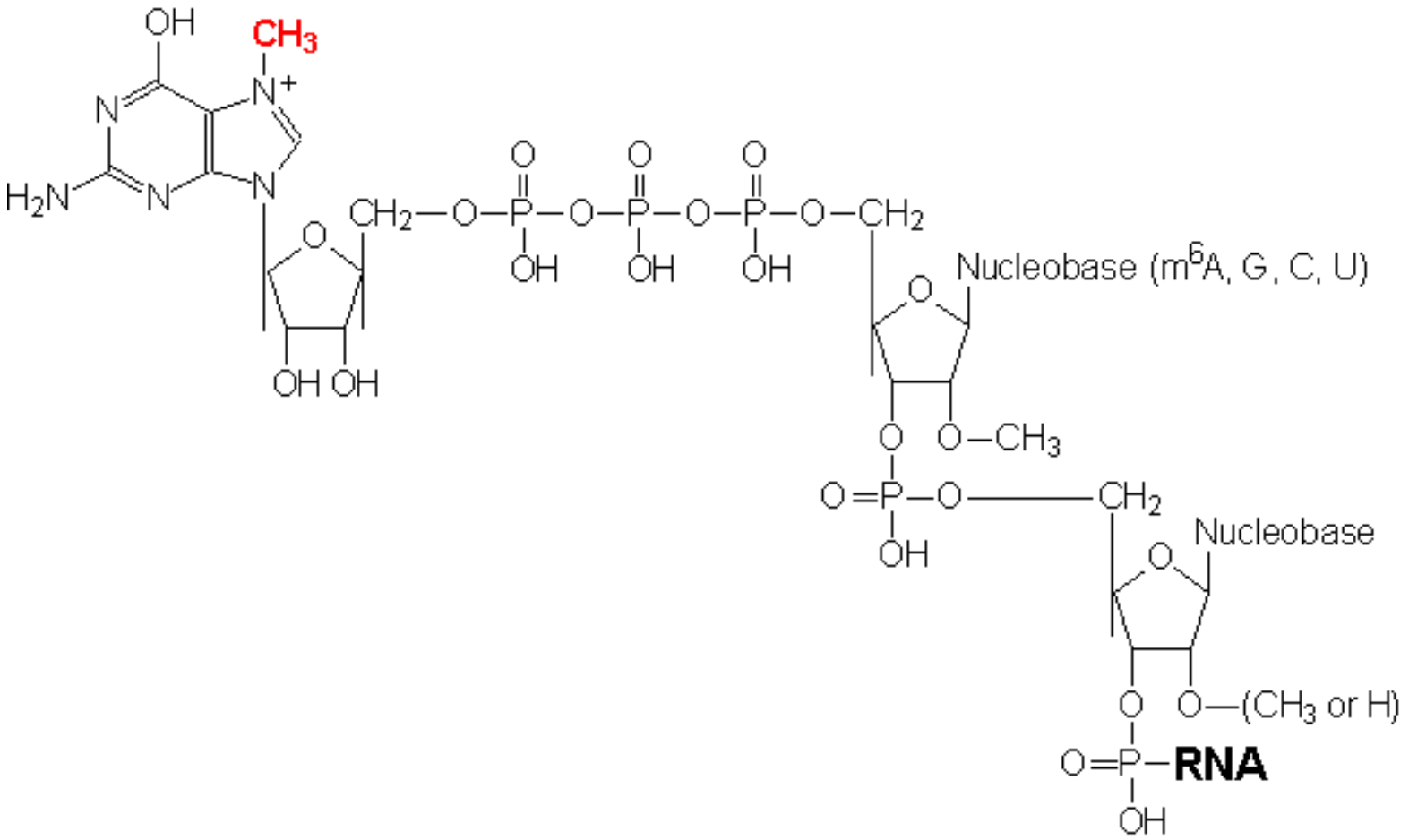
Posttranscriptional Processing of RNAs

- When transcription of bacterial rRNAs and tRNAs is completed they are immediately ready for use in translation. No additional processing takes place.
- Translation of bacterial mRNAs can begin even before transcription is completed due to the lack of the nuclear-cytoplasmic separation that exists in eukaryotes.
- The ability to initiate translation of prokaryotic RNAs while transcription is still in progress affords a unique opportunity for regulating the transcription of certain genes.
- An additional feature of bacterial mRNAs is that most are polycistronic.
  - This means that multiple polypeptides can be synthesized from a single primary transcript.
  - This does not occur in eukaryotic mRNAs.

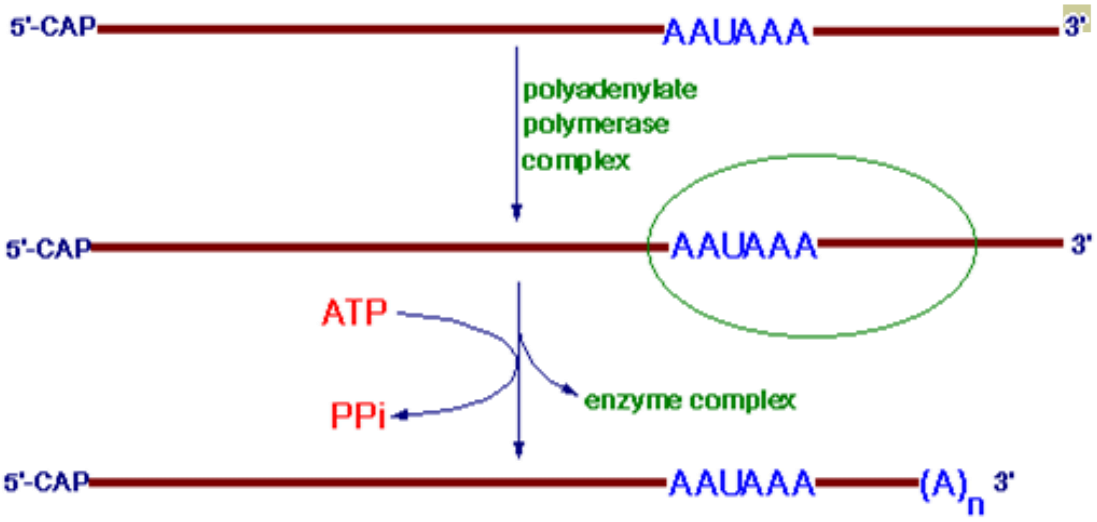
Eukaryotic RNAs

- In contrast to bacterial transcripts, eukaryotic RNAs (all 3 classes) undergo significant post-transcriptional processing.
  - All 3 classes of RNA are transcribed from genes that contain introns.
  - The sequences encoded by the intronic DNA must be removed from the primary transcript prior to the RNAs being biologically active.
  - The process of intron removal is called RNA **splicing**.
  - Additional processing occurs to mRNAs. The 5' end of all eukaryotic mRNAs are capped with a unique 5' -----> 5' linkage to a 7-methylguanosine residue.
  - The capped end of the mRNA is thus, protected from exonucleases and more importantly is recognized by specific proteins of the translational machinery.

Structure of the 5'-Cap of Eukaryotic mRNAs



Processes of Polyadenylation of mRNA



- Messenger RNAs also are polyadenylated at the 3' end.
- A specific sequence, AAUAAA, is recognized by the endonuclease activity of by polyadenylate polymerase which cleaves the primary transcript approximately 11 - 30 bases 3' of the sequence element.
- A stretch of 20 - 250 A residues is then added to the 3' end by the polyadenylate polymerase activity.

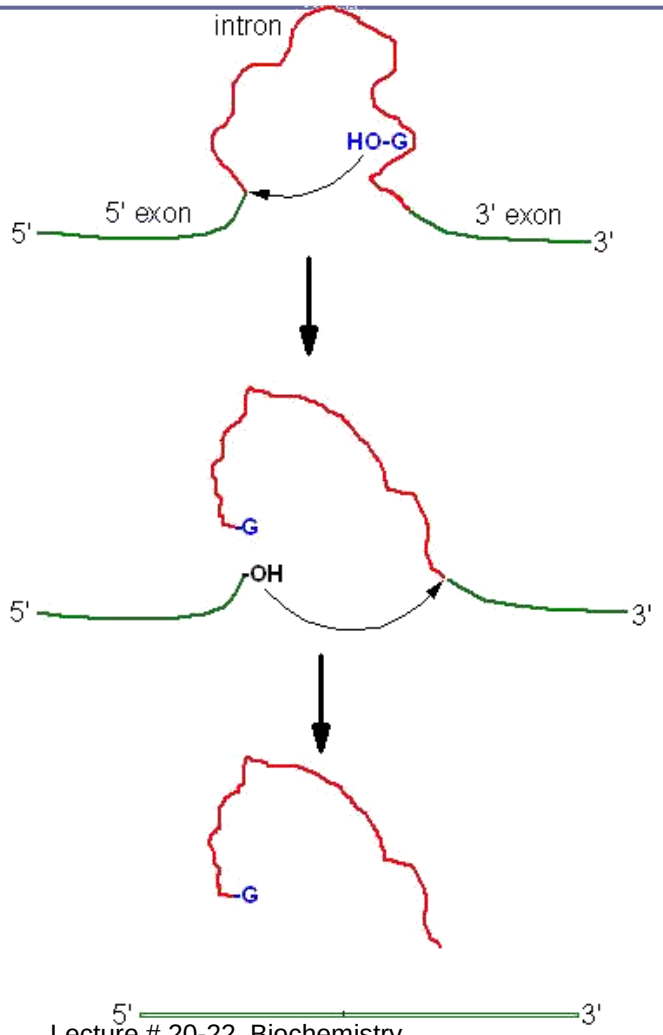
Modifications in tRNAs and rRNAs

- In addition to intron removal in tRNAs, extra nucleotides at both the 5' and 3' ends are cleaved, the sequence 5'-CCA-3' is added to the 3' end of all tRNAs and several nucleotides undergo modification.
  - There have been more than 60 different modified bases identified in tRNAs.
- Both prokaryotic and eukaryotic rRNAs are synthesized as long precursors termed preribosomal RNAs.
  - In eukaryotes a 45S preribosomal RNA serves as the precursor for the 18S, 28S and 5.8S rRNAs.

Splicing of RNAs

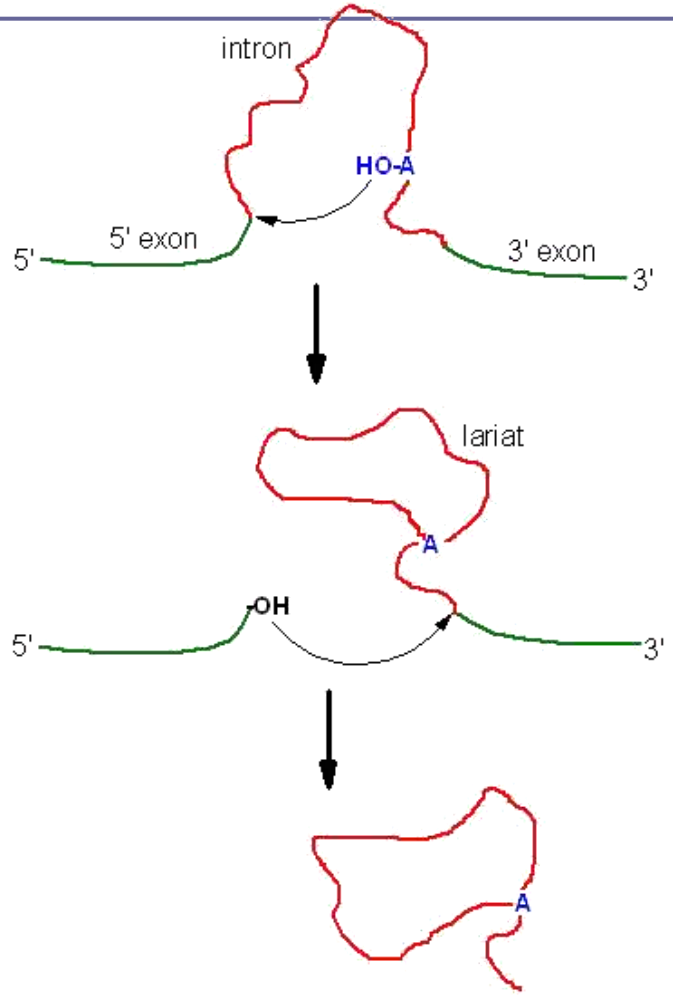
- There are several different classes of reactions involved in intron removal.
- The 2 most common are the group I and group II introns.
  - Group I introns are found in nuclear, mitochondrial and chloroplast rRNA genes,
  - group II in mitochondrial and chloroplast mRNA genes.
  - Many of the group I and group II introns are self-splicing, i.e. no additional protein factors are necessary for the intron to be accurately and efficiently spliced out.

Group I introns



- Group I introns require an external guanosine nucleotide as a cofactor.
- The 3'-OH of the guanosine nucleotide acts as a nucleophile to attack the 5'-phosphate of the 5' nucleotide of the intron.
- The resultant 3'-OH at the 3' end of the 5' exon then attacks the 5' nucleotide of the 3' exon releasing the intron and covalently attaching the two exons together.
- The 3' end of the 5' exon is termed the **splice donor site** and the 5' end of the 3' exon is termed the **splice acceptor site**.

Group II introns

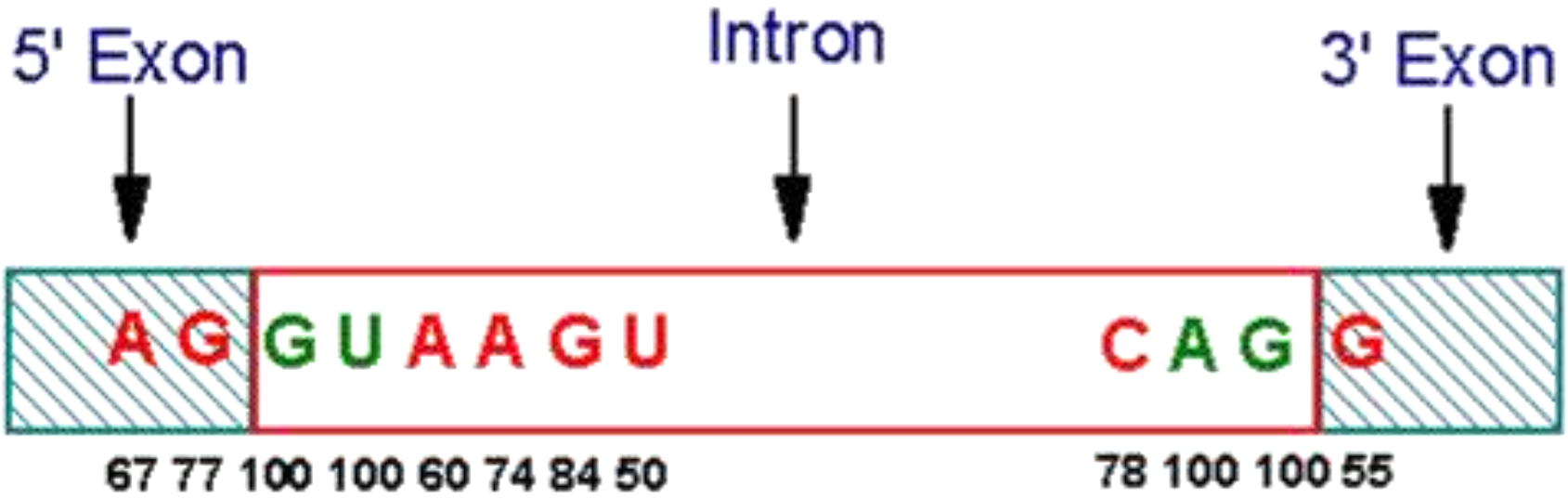


- Group II introns are spliced similarly except that instead of an external nucleophile the 2'-OH of an adenine residue within the intron is the nucleophile.
- This residue attacks the 3' nucleotide of the 5' exon forming an internal loop called a **lariat structure**.
- The 3' end of the 5' exon then attacks the 5' end of the 3' exon as in group I splicing releasing the intron and covalently attaching the two exons together.

Small Nuclear Ribonucleoprotein Particles

- The third class of introns is also the largest class found in nuclear mRNAs. This class of introns undergoes a splicing reaction similar to group II introns in that an internal lariat structure is formed.
- However, the splicing is catalyzed by specialized RNA-protein complexes called small nuclear ribonucleoprotein particles (snRNPs, pronounced *snurps*).
  - The RNAs found in snRNPs are identified as U1, U2, U4, U5 and U6.
  - The genes encoding these snRNAs are highly conserved in vertebrate and insects and are also found in yeasts and slime molds indicating their importance.

Splicing consensus sequences

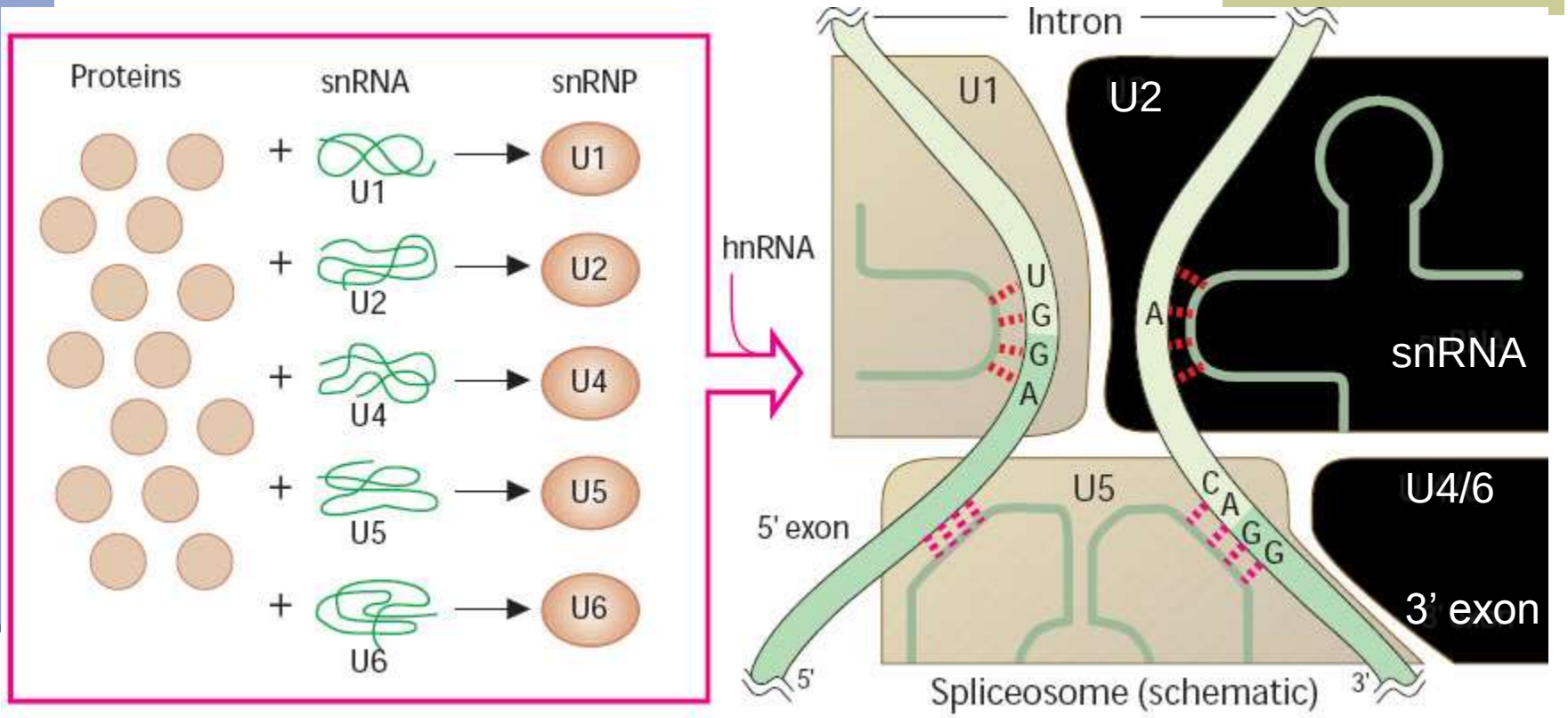


- Analysis of a large number of mRNA genes has led to the identification of highly conserved consensus sequences at the 5' and 3' ends of essentially all mRNA introns.

Splicesome

- The U1 RNA has sequences that are complimentary to sequences near the 5' end of the intron.
  - The binding of U1 RNA distinguishes the GU at the 5' end of the intron from other randomly placed GU sequences in mRNAs.
  - The U2 RNA also recognizes sequences in the intron, in this case near the 3' end.
  - The addition of U4, U5 and U6 RNAs forms a complex identified as the **spliceosome** that then removes the intron and joins the two exons together.

Splicesome


Proteins
snRNA
snRNP
U1
U2
U4
U5
U6
hnRNA
Intron
U1
U2
U G G A
A
snRNA
U5
C A G G
U4/6
5' exon
3' exon
5'
3'
Spliceosome (schematic)

Additional Mechanism of Intron Removal

- An additional mechanism of intron removal is the process of tRNA splicing. These introns are spliced by a specific splicing endonuclease that involves a cut-and-paste mechanism.
- In order for tRNA intron removal to occur the tRNA must first be properly folded into its characteristic cloverleaf shape.
  - Misfolded precursor tRNAs are not processed which allows the splicing reaction to serve as a control step in the gneration of mature tRNAs.

Clinical Significances of Alternative and Aberrant Splicing

- The presence of introns in eukaryotic genes would appear to be an extreme waste of cellular energy when considering the number of nucleotides incorporated into the primary transcript only to be removed later as well as the energy utilized in the synthesis of the splicing machinery.
- However, the presence of introns can protect the genetic makeup of an organism from genetic damage by outside influences such as chemical or radiation.
- An additionally important function of introns is to allow **alternative splicing** to occur, thereby, increasing the genetic diversity of the genome without increasing the overall number of genes.
- By altering the pattern of exons, from a single primary transcript, that are spliced together different proteins can arise from the processed mRNA from a single gene.
- Alternative splicing can occur either at specific developmental stages or in different cell types.

Alternative Splicing

- This process of **alternative splicing** has been identified to occur in the primary transcripts from at least 40 different genes.
  - Depending upon the site of transcription, the **calcitonin gene** yields an RNA that synthesizes **calcitonin** (thyroid) or **calcitonin-gene related peptide** (CGRP, brain).
  - Even more complex is the alternative splicing that occurs in the **α-tropomyosin** transcript.
    - At least 8 different alternatively spliced α-tropomyosin mRNAs have been identified.

β-Talassemias

- **Abnormalities in the splicing process** can lead to various disease states.
  - Many defects in the β-globin genes are known to exist leading to **β-thalassemias**.
  - Some of these defects are caused by mutations in the sequences of the gene required for intron recognition and, therefore, result in abnormal processing of the β-globin primary transcript.

Systemic Lupus Erythematosis

- Patients suffering from a number of different connective tissue diseases exhibit **humoral auto-antibodies** that recognize cellular RNA-protein complexes.
  - Patients suffering from **systemic lupus erythematosis** have auto-antibodies that recognize the U1 RNA of the spliceosome.

DNA Synthesis. Introduction

- All cells undergo a division cycle during their life span.
- Some cells are continually dividing (e.g. stem cells), others divide a specific number of times until cell death (**apoptosis**) occurs, and still others divide a few times before entering a terminally differentiated or quiescent state.
- Most cells of the body fall into the latter category of cells.
- During the process of cell division everything within the cell must be duplicated in order to ensure the survival of the two resulting daughter cells.
- Of particular importance for cell survival is the accurate, efficient and rapid duplication of the cellular genome.
- This process is termed **DNA replication**.

Eukaryotic Genomes

- The size of eukaryotic genomes is vastly larger than those of prokaryotes. This is partly due to the complexity of eukaryotic organisms compared to prokaryotes.
- However, the size of a particular eukaryotic genome is not directly correlated to the organism' complexity.
- This is the result of the presence of a large amount of non-coding DNA.
- The functions of these non-coding nucleic acid sequences are only partly understood.
  - Some sequences are involved in the control of gene expression while others may simply be present in the genome to act as an evolutionary buffer able to withstand nucleotide mutation without disrupting the integrity of the organism.

Repetitive DNA

- One abundant class of DNA is termed **repetitive DNA**.
- There are 2 different sub-classes of repetitive DNA, **highly repetitive** and **moderately repetitive**.
  - Highly repetitive DNA consists of short sequences 6-10 bp long reiterated from 100,000- 1,000,000 times.
  - The DNA of the genome consisting of the **genes** (coding sequences) is identified as **non-repetitive** DNA since most genes occur but once in an organism' haploid genome.
- However, it should be pointed out that several genes exist as tandem clusters of multiple copies of the same gene ranging from 50 to 10,000 copies such as is the case for the rRNA genes and the histone genes.

Introns and Exons

- Another characteristic feature that distinguishes eukaryotic from prokaryotic genes is the presence of introns.
  - Introns are stretches of nucleic acid sequences that separate the coding exons of a gene.
    - The existence of introns in prokaryotes is extremely rare.
  - Essentially all humans genes contain introns.
  - A notable exception are the histone genes which are intronless.
  - In many genes the presence of introns separates exons into coding regions exhibiting distinct functional domains.

Chromatin Structure

- Chromatin is a term designating the structure in which DNA exists within cells.
  - The structure of chromatin is determined and stabilized through the interaction of the DNA with DNA-binding proteins.
  - There are 2 classes of DNA-binding proteins.
    - The histones are the major class of DNA-binding proteins involved in maintaining the compacted structure of chromatin.
    - There are 5 different histone proteins identified as H1, H2A, H2B, H3 and H4.
- The other class of DNA-binding proteins is a diverse group of proteins called simply, non-histone proteins. This class of proteins includes the various transcription factors, polymerases, hormone receptors and other nuclear enzymes.
  - In any given cell there are greater than 1000 different types of non-histone proteins bound to the DNA.

DNA Structure

- The binding of DNA by the histones generates a structure called the **nucleosome**.
  - The nucleosome core contains an octamer protein structure consisting of 2 subunits each of H2A, H2B, H3 and H4.
  - Histone H1 occupies the internucleosomal DNA and is identified as the linker histone.
  - The nucleosome core contains approximately 150 bp of DNA. The linker DNA between each nucleosome can vary from 20 to more than 200 bp.
  - These nucleosomal core structures would appear as beads on a string if the DNA were pulled into a linear structure.
- The nucleosome cores themselves coil into a solenoid shape which itself coils to further compact the DNA.
- These final coils are compacted further into the characteristic chromatin seen in a karyotyping spread.
- The protein-DNA structure of chromatin is stabilized by attachment to a non-histone protein scaffold called the nuclear matrix.

Eukaryotic Cell Cycles

- The cell cycle is defined as the sequence of events that occurs during the lifetime of a cell.
  - The eukaryotic cell cycle is divided into 4 major periods. During each period a specific sequence of events occurs.
  - The ultimate conclusion of one cell cycle is cytokinesis resulting in two identical daughter cells.
- The 4 phases of a typical cell cycle and the events occurring during each phase are outlined:
  - M phase is the period when cells prepare for and then undergo cytokinesis.
  - M phase stands for mitotic phase or mitosis.
  - During mitosis the chromosomes are paired and then divided prior to cell division.
  - The events in this stage of the cell cycle leading to cell division are prophase, metaphase, anaphase and telophase.

G_1 Phase

- G_1 phase corresponds to the gap in the cell cycle that occurs following cytokinesis.
  - During this phase cells make a decision to either exit the cell cycle and become quiescent or terminally differentiated or to continue dividing.
  - Terminal differentiation is identified as a non-dividing state for a cell.
  - Quiescent and terminally differentiated cells are identified as being in G_0 phase.
  - Cells in G_0 can remain in this state for extended periods of time. Specific stimuli may induce the G_0 cell to re-enter the cell cycle at the G_1 phase or alternatively may induce permanent terminal differentiation.
- During G_1 cells begin synthesizing all the cellular components needed in order to generate two identically complimented daughter cells. As a result the size of cells begins to increase during G1.

S and G_2 phases

- S phase is the phase of the cell cycle during which the DNA is replicated.
  - This is the DNA synthesis phase.
  - Additionally, some specialized proteins are synthesized during S phase, particularly the histones.
- G_2 phase is reached following completion of DNA replication.
  - During G_2 the chromosomes begin condensing, the nucleoli disappear and two microtubule organizing centers begin polymerizing tubulins for eventual production of the spindle poles.
- Typical eukaryotic cell cycles occupy approximately 16 - 24 hrs when grown in culture. However, in the context of the multicellular organization of organisms the cell cycles can be as short as 6 - 8 hrs to greater than 100 days.
- The high variability of cell cycle times is due to the variability of the G_1 phase of the cycle.

DNA Replication

- Replication of DNA occurs during the process of normal cell division cycles. Because the genetic complement of the resultant daughter cells must be the same as the parental cell, DNA replication must possess a very high degree of fidelity. The entire process of DNA replication is complex and involves multiple enzymatic activities.
- The mechanics of DNA replication was originally characterized in the bacterium, E. coli which contains 3 distinct enzymes capable of catalyzing the replication of DNA. These have been identified as DNA polymerase (pol) I, II, and III. Pol I is the most abundant replicating activity in E. coli but has as its primary role to ensure the fidelity of replication through the repair of damaged and mismatched DNA. Replication of the E. coli genome is the job of pol III. This enzyme is much less abundant than pol I, however, its activity is nearly 100 times that of pol I.

DNA polymerases

- There have been 5 distinct eukaryotic DNA polymerases identified, α, β γ, δ and ε.
- The identity of the individual enzymes relates to its subcellular localization as well as its primary replicative activity.
- The polymerase of eukaryotic cells that is the equivalent of E. coli pol III is pol-a. The pol I equivalent in eukaryotes is pol-ß.
- Polymerase-g is responsible for replication of mitochondrial DNA.
- The ability of DNA polymerases to replicate DNA requires a number of additional accessory proteins. The combination of polymerases with several of the accessory proteins yields an activity identified as DNA polymerase holoenzyme. These accessory proteins include (not ordered with respect to importance):
    1. Primase
    2. Processivity accessory proteins
    3. Single strand binding proteins
    4. Helicase
    5. DNA ligase
    6. Topoisomerases
    7. Uracil-DNA N-glycosylase

Process of DNA Replication

- The process of DNA replication begins at specific sites in the chromosomes termed origins of replication, requires a primer bearing a free 3'-OH, proceeds specifically in the 5' -----> 3' direction on both strands of DNA concurrently and results in the copying of the template strands in a semiconservative manner.
- The ***semiconservative*** nature of DNA replication means that the newly synthesized daughter strands remain associated with their respective parental template strands.

Replication Fork

- The large size of eukaryotic chromosomes and the limits of nucleotide incorporation during DNA synthesis, make it necessary for multiple origins of replication to exist in order to complete replication in a reasonable period of time.
- The precise nature of origins of replication in higher eukaryotic organisms is unclear. However, it is clear that at a replication origin the strands of DNA must dissociate and unwind in order to allow access to DNA polymerase.
- Unwinding of the duplex at the origin as well as along the strands as the replication process proceeds is carried out by helicases.
- The resultant regions of single-stranded DNA are stabilized by the binding of single-strand binding proteins. The stabilized single-stranded regions are then accessible to the enzymatic activities required for replication to proceed.
- The site of the unwound template strands is termed the **replication fork**.

DNA: reparation and replication

- In order for DNA polymerases to synthesize DNA they must encounter a free 3'-OH which is the substrate for attachment of the 5'-phosphate of the incoming nucleotide.
- During repair of damaged DNA the 3'-OH can arise from the hydrolysis of the backbone of one of the two strands.
- During replication the 3'-OH is supplied through the use of an **RNA primer**, synthesized by the **primase activity**.
- The primase utilizes the DNA strands as templates and synthesizes a short stretch of RNA generating a primer for DNA polymerase.

Synthesis of DNA

- Synthesis of DNA proceeds in the 5' -----> 3' direction through the attachment of the 5'-phosphate of an incoming dNTP to the existing 3'-OH in the elongating DNA strands with the concomitant release of pyrophosphate.
  - Initiation of synthesis, at origins of replication, occurs simultaneously on both strands of DNA.
  - Synthesis then proceeds bidirectionally, with one strand in each direction being copied continuously and one strand in each direction being copied discontinuously.
- During the process of DNA polymerases incorporating dNTPs into DNA in the 5' -----> 3' direction they are moving in the 3' -----> 5' direction with respect to the template strand.
- In order for DNA synthesis to occur simultaneously on both template strands as well as bidirectionally one strand appears to be synthesized in the 3' ------> 5' direction. In actuality one strand of newly synthesized DNA is produced discontinuously.

Leading and Lagging Strands. Okazaki Fragments

- The strand of DNA synthesized continuously is termed the **leading strand** and the discontinuous strand is termed the **lagging strand**.
  - The lagging strand of DNA is composed of short stretches of RNA primer plus newly synthesized DNA approximately 100-200 bases long (the approximate distance between adjacent nucleosomes).
  - The lagging strands of DNA are also called **Okazaki fragments**. The concept of continuous strand synthesis is somewhat of a misnomer since DNA polymerases do not remain associated with a template strand indefinitely. The ability of a particular polymerase to remain associated with the template strand is termed its' **processivity**. The longer it associates the higher the processivity of the enzyme.
  - DNA polymerase processivity is enhanced by additional protein activities of the replisome identified as ***processivity accessory proteins***.

How is it that DNA polymerase can copy both strands of DNA

- How is it that DNA polymerase can copy both strands of DNA in the 5' ----> 3' direction simultaneously?
  - A model has been proposed where DNA polymerases exist as dimers associated with the other necessary proteins at the replication fork and identified as the **replisome**.
  - The template for the lagging strand is temporarily looped through the replisome such that the DNA polymerases are moving along both strands in the 3' ----> 5' direction simultaneously for short distances, the distance of an Okazaki fragment.
  - As the replication forks progress along the template strands the newly synthesized daughter strands and parental template strands reform a DNA double helix.
  - The means that only a small stretch of the template duplex is single-stranded at any given time.

DNA Topoisomerases

- The progression of the replication fork requires that the DNA ahead of the fork be continuously unwound.
  - Due to the fact that eukaryotic chromosomal DNA is attached to a protein scaffold the progressive movement of the replication fork introduces severe torsional stress into the duplex ahead of the fork.
  - This torsional stress is relieved by **DNA topoisomerases**.
    - Topoisomerases relieve torsional stresses in duplexes of DNA by introducing either double- (topoisomerases II) or single-stranded (topoisomerases I) breaks into the backbone of the DNA.
    - These breaks allow unwinding of the duplex and removal of the replication-induced torsional strain.
    - The nicks are then resealed by the topoisomerases.

DNA Ligases

- The RNA primers of the leading strands and Okazaki fragments are removed by the repair DNA polymerases simultaneously replacing the ribonucleotides with deoxyribonucleotides.
- The gaps that exist between the 3'-OH of one leading strand and the 5'-phosphate of another as well as between one Okazaki fragment and another are repaired by **DNA ligases** thereby, completing the process of replication.

Additional DNA Polymerase Activities

- The main enzymatic activity of DNA polymerases is the 5' ----> 3' synthetic activity. However, DNA polymerases possess two additional activities of importance for both replication and repair.
- These additional activities include a 5' ----> 3' exonuclease function and a 3' ------> 5' exonuclease function.
  - The 5' -----> 3' exonuclease activity allows the removal of ribonucleotides of the RNA primer, utilized to initiate DNA synthesis, along with their simultaneous replacement with deoxyribonucleotides by the 5' ------> 3' polymerase activity.
  - The 5' -----> 3' exonuclease activity is also utilized during the repair of damaged DNA.
  - The 3' -----> 5' exonuclease function is utilized during replication to allow DNA polymerase to remove mismatched bases.
  - It is possible (but rare) for DNA polymerases to incorporate an incorrect base during replication.
  - These mismatched bases are recognized by the polymerase immediately due to the lack of Watson-Crick base-pairing.
  - The mismatched base is then removed by the 3' ------> 5' exonuclease activity and the correct base inserted prior to progression of replication.

Post-Replicative Modification of DNA, Methylation

- One of the major post-replicative reactions that modifies the DNA is **methylation**.
  - The sites of natural methylation (i.e. not chemically induced) of eukaryotic DNA is always on cytosine residues that are present in CpG dinucleotides.
  - However, it should be noted that not all CpG dinucleotides are methylated at the C residue.
  - The cytidine is methylated at the 5 position of the pyrimidine ring generating 5-methylcytidine.
- Methylation of DNA in prokaryotic cells also occurs.
  - The function of this methylation is to prevent degradation of host DNA in the presence of enzymatic activities synthesized by bacteria called **restriction endonucleases**.
    - These enzymes recognize specific nucleotide sequences of DNA.
    - The role of this system in prokaryotic cells (called the restriction-modification system) is to degrade invading viral DNAs. Since the viral DNAs are not modified by methylation they are degraded by the host restriction enzymes.
    - The methylated host genome is resistant to the action of these enzymes.

Methylated DNA Would Be Less Transcriptionally Active

- The precise role of methylation in eukaryotic DNA is unclear.
- It was originally thought that methylated DNA would be less transcriptionally active.
  - Indeed, experiments have been carried out to demonstrate that this is true for certain genes. For example, under-methylation of the MyoD gene (a master control gene regulating the differentiation of muscle cells through the control of the expression of muscle-specific genes) results in the conversion of fibroblasts to myoblasts.
  - The experiments were carried out by allowing replicating fibroblasts to incorporate 5-azacytidine into their newly synthesized DNA. This analog of cytidine prevents methylation. The net result is that the maternal pattern of methylation is lost and numerous genes become under methylated.
  - However, lack of methylation nor the presence of methylation is a clear indicator of whether a gene will be transcriptionally active or silent.

Maintenance Methylase System

- The pattern of methylation is copied post-replicatively by the **maintenance methylase system**.
- This activity recognizes the pattern of methylated C residues in the maternal DNA strand following replication and methylates the C residue present in the corresponding CpG dinucleotide of the daughter strand.

Examples of Methylation

- The phenomenon of **genomic imprinting** refers to the fact that the expression of some genes depends on whether or not they are inherited maternally or paternally.
- **Insulin-like growth factor-2 (Igf2)** is a gene whose expression is required for normal fetal development and growth. Expression of Igf2 occurs exclusively from the paternal copy of the gene. Imprinted genes are "marked" by their state of methylation.
  - In the case of Igf2 an element in the paternal locus, called an **insulator element**, is methylated blocking its function.
  - The function of the un-methylated insulator is to bind a protein that when bound blocks activation of Igf2 expression.
  - When methylated the protein cannot bind the insulator thus allowing a distant enhancer element to drive expression of the Igf2 gene.
  - In the maternal genome, the insulator is not methylated, thus protein binds to it blocking the action of the distant enhancer element.

DNA Recombination

- DNA recombination refers to the phenomenon whereby two parental strands of DNA are spliced together resulting in an exchange of portions of their respective strands. This process leads to new molecules of DNA that contain a mix of genetic information from each parental strand.
- There are 3 main forms of genetic recombination. These are
  - **homologous recombination**,
  - **site-specific recombination** and
  - **transposition**.

Homologous recombination

- Homologous recombination is the process of genetic exchange that occurs between any two molecules of DNA that share a region (or regions) of homologous DNA sequences.
- This form of recombination occurs frequently while sister chromatids are paired during meiosis. Indeed, it is the process of homologous recombination between the maternal and paternal chromosomes that imparts genetic diversity to an organism.
- Homologous recombination generally involves exchange of large regions of the chromosomes.

Site-specific Recombination

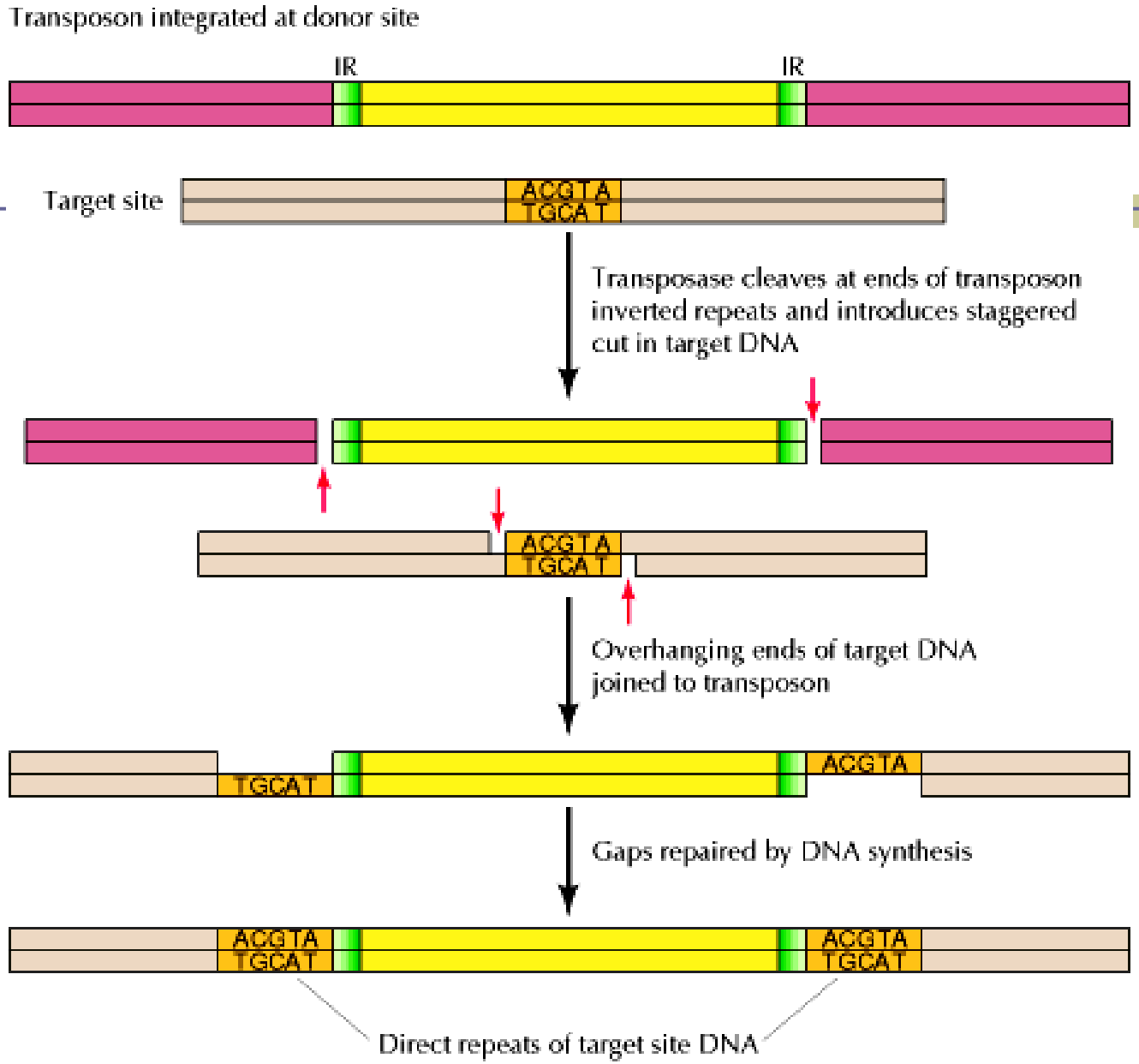
- Site-specific recombination involves exchange between much smaller regions of DNA sequence (approximately 20 - 200 base pairs) and requires the recognition of specific sequences by the proteins involved in the recombination process.
- Site-specific recombination events occur primarily as a mechanism to alter the program of genes expressed at specific stages of development.

Site-specific Recombination (cont'd)

- The most significant site-specific recombinational events in humans are the somatic cell gene rearrangements that take place in the **immunoglobulin genes** during B-cell differentiation in response to antigen presentation.
  - These gene rearrangements in the immunoglobulin genes result in an extremely diverse potential for antibody production.
  - A typical antibody molecule is composed of both heavy and light chains. The genes for both these peptide chains undergo somatic cell rearrangement yielding the potential for approximately 3000 different light chain combinations and approximately 5000 heavy chain combinations.
  - Then because any given heavy chain can combine with any given light chain the potential diversity exceeds 10,000,000 possible different antibody molecules.

DNA Transposition

- Transposition is a unique form of recombination where **mobile genetic elements** can virtually move from one region to another within one chromosome or to another chromosome entirely.
- There is no requirement for sequence homology for a transpositional event to occur.
- Because the potential exists for the disruption of a vitally important gene by a transposition event this process must be tightly regulated.
- The exact nature of how transpositional events are controlled is unclear.


Transposon integrated at donor site
IR
IR
Target site
ACGTA
TGCAT
Transposase cleaves at ends of transposon inverted repeats and introduces staggered cut in target DNA
ACGTA
TGCAT
Overhanging ends of target DNA joined to transposon
TGCAT
ACGTA
Gaps repaired by DNA synthesis
ACGTA
TGCAT
ACGTA
TGCAT
Direct repeats of target site DNA

Processed Genes

- Transposition occurs with a higher frequency in bacteria and yeasts than it does in humans.
  - The identification of the occurrence of transposition in the human genome resulted when it was found that certain processed genes were present in the genome.
  - These processed genes are nearly identical to the mRNA encoded by the normal gene.
  - The processed genes contain the poly(A) tail that would have been present in the RNA and they lack the introns of the normal gene.
  - These particular forms of genes must have arisen through a reverse transcription event, similar to the life cycle of retroviral genomes, and then been incorporated into the genome by a transpositional event.
  - Since most of the processed genes that have been identified are non-functional they have been termed **pseudogenes**.

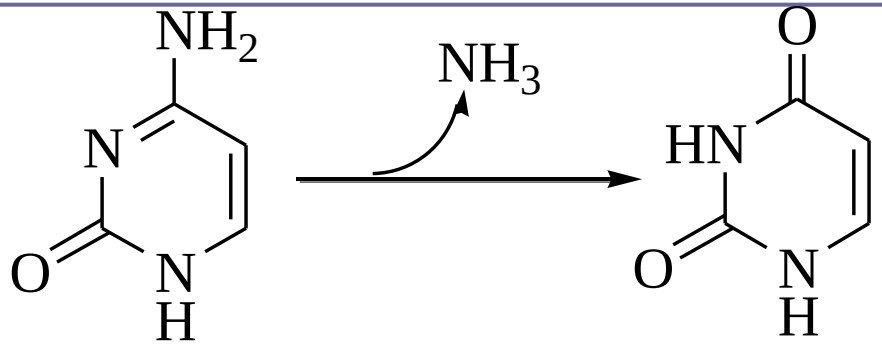
DNA Damaging

- **Cancer**, in most non-viral induced cases, is the severe medically relevant consequence of the inability to repair damaged DNA.
  - It is clear that multiple somatic cell mutations in DNA can lead to the genesis of the transformed phenotype. Therefore, it should be obvious that complete understanding of DNA repair mechanisms would be invaluable in the design of potential therapeutic agents in the treatment of cancer.

Repair of Damaged DNA

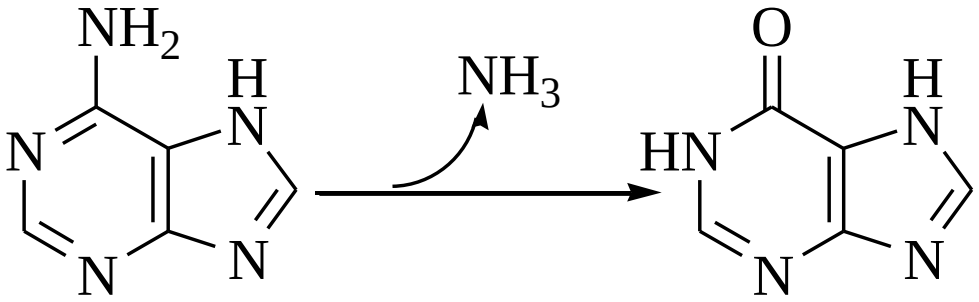
- DNA damage can occur as the result of exposure to environmental stimuli such as **alkylating chemicals** or **ultraviolet or radioactive irradiation** and **free radicals** generated spontaneously in the oxidizing environment of the cell.
  - These phenomena can, and do, lead to the introduction of **mutations** in the coding capacity of the DNA.
  - Mutations in DNA can also, but rarely, arise from the **spontaneous tautomerization** of the bases.

Deamination of adenine, cytosine, and guanine



Cytosine

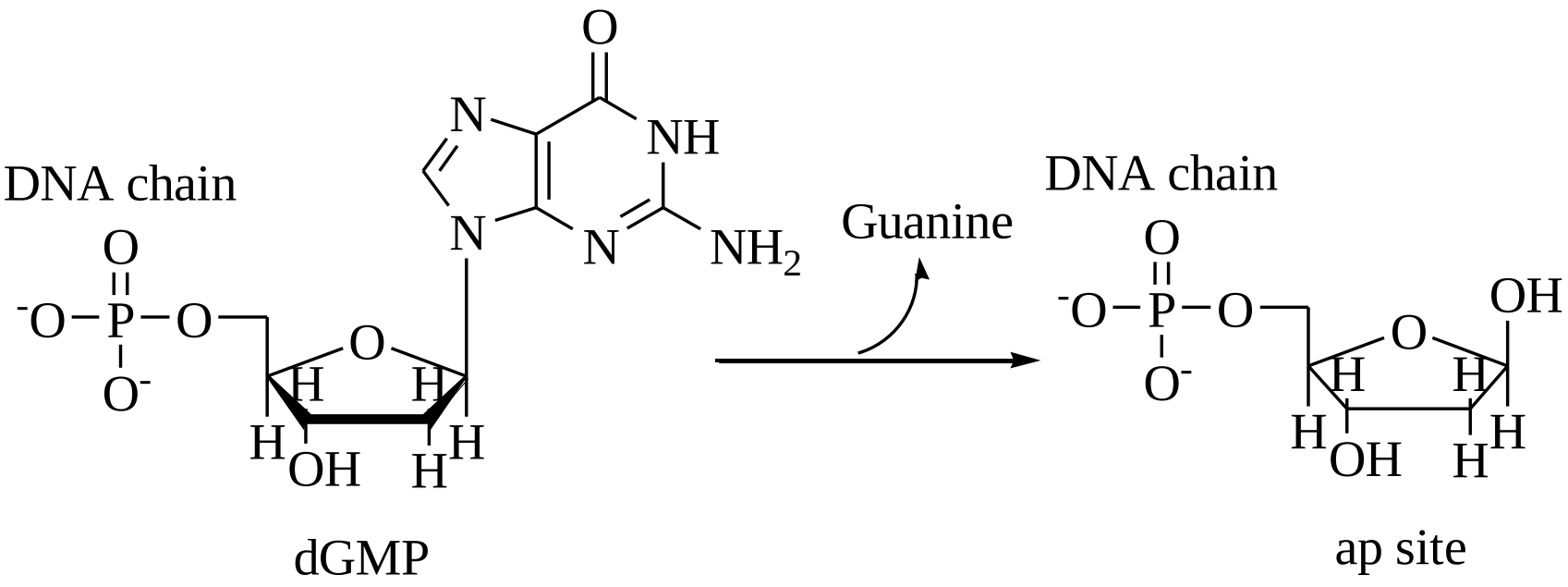
Uracil



Adenine

Hypoxanthine

Depurination



- **Depurination** (loss of purine bases) resulting from cleavage of the bond between the purine bases and deoxyribose, leaving an apurinic (AP) site in DNA.
- dGMP = deoxyguanosine monophosphate

Modification of the DNA Bases

- Modification of the DNA bases by **alkylation** (predominately the incorporation of -CH_3 groups) predominately occurs on **purine residues**.
  - **Methylation of G residues** allows them to base pair with T instead of C. A unique activity called **O^6-alkylguanine transferase** removes the alkyl group from G residues.
    - The protein itself becomes alkylated and is no longer active, thus, a single protein molecule can remove only one alkyl group.

Mutations in DNA

- Mutations in DNA are of two types.
  - **Transition mutations** result from the exchange of one purine, or pyrimidine, for another purine, or pyrimidine.
  - **Transversion mutations** result from the exchange of a purine for a pyrimidine or visa versa.

Repair of Thymine Dimers

- The prominent by-product from ***uv*-irradiation** of DNA is the formation of **thymine dimers**.
  - These form from two adjacent T residues in the DNA.
  - Repair of thymine dimers is most understood from consideration of the mechanisms used in *E. coli.*
  - However, several mechanism are common to both prokaryotes and eukaryotes.

Thymine Dimers Removal

- Thymine dimers are removed by several mechanisms.
  - Specific **glycohydrolases** recognize the dimer as abnormal and cleave the N-glycosidic bond of the bases in the dimer.
    - This results in the base leaving and generates an apyrimidinic site in the DNA.
    - This is repaired by DNA polymerase and ligase.
  - Glycohydrolases are also responsible for the removal of other abnormal bases, not just thymine dimers.

Thymine Dimers Removal (cont'd)

- Another, widely distributed activity, is **DNA photolyase** or **photoreactivating enzyme**.
  - This protein binds to thymine dimers in the dark.
  - In response to visible light stimulation the enzyme cleaves the pyrimidine rings.
  - The chromophore associated with this enzyme that allows visible light activation is $FADH_2$.

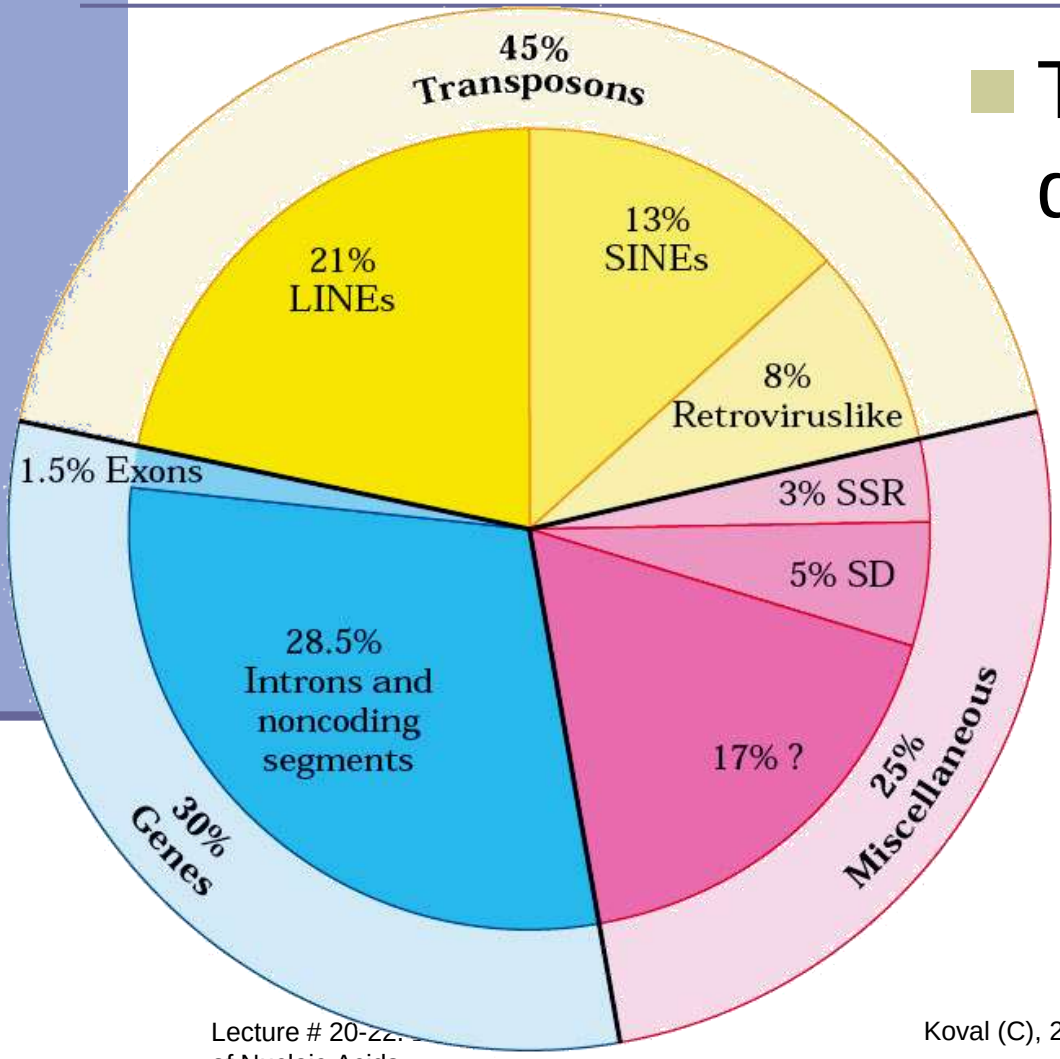
DNA Repair Defects: Xeroderma Pigmentosum

- Humans defective in DNA repair, (in particular the repair of *uv*-induced thymine dimers), due to autosomal recessive genetic defects suffer from the disease **Xeroderma pigmentosum (XP)**.
  - There are at least nine distinct genetic defects associated with this disease.
    - One of these is due to a defect in the gene coding for the glycohydrolase that cleaves the N-glycosidic bond of the thymine dimers.
    - There are two major clinical forms of XP, one which leads to progressive degenerative changes in the eyes and skin and the other which also includes progressive neurological degeneration.

DNA Repair Defects: Cockayne Syndrome and Ataxia Telangiectasia

- Another inherited disorder affecting DNA repair in which patients suffer from sun sensitivity, short stature and progressive neurological degeneration without an increased incidence of skin cancer is **Cockayne Syndrome**.
- **Ataxia telangiectasia (AT)** is an autosomal recessive disorder resulting in neurological disability and suppressed immune function.
  - Patients develop a disabling cerebellar ataxia early in life and have recurrent infections.
  - Patients suffering from AT have an increased sensitivity to *X*-irradiation suggesting a role for the AT gene in DNA repair.

Human Genome Organization



- The genome is divided into:
  - **Transposons** (transposable elements) – **45%,**
  - **Genes – 30%,**
  - **Miscellaneous sequences – 25%**.

Human Genome Organization: Transposons

- 4 main classes of transposons.
  - **Long interspersed elements (LINEs)**, 6 to 8 kbp long.
    - include genes encoding proteins that catalyze transposition.
    - about 850,000 LINEs.
  - **Short interspersed elements (SINEs):** about 100 to 300 bp long.
    - Of the 1.5 million in the human genome more than 1 million are **Alu** elements.
  - **Retroviruslike transposons**: 450,000 copies of, 1.5 to 11 kbp long.
    - “trapped” in the genome, cannot move from one cell to another.
  - **transposon remnants**, differ in length. (< 1%, not shown).

Human Genome Organization: Genes + Miscellaneous Sequences

- About 30% of the genome consists of sequences included in **genes for proteins**, but only a small fraction of this DNA is in **exons** (coding sequences).
- **Miscellaneous sequences:**
  - **simple-sequence repeats (SSR),**
  - **large segmental duplications (SD):** segments that appear more than once in different locations.
- unlisted genes (denoted by "?"):
  - encoding RNAs: harder to identify than genes for proteins)
  - remnants of transposons: evolutionarily altered, now hard to identify.

Major Findings Reported in the Rough Drafts of the Human Genome

- More than 90% of the genome has been sequenced; gaps, large and small, remain to be filled in.
- Estimated number of protein-coding genes ranges from 30,000 to 40,000.
- Only 1.1–1.5% of the genome codes for proteins.
- There are wide variations in features of individual chromosomes (eg, in gene number per Mb, SNP density, GC content, numbers of transposable elements and CpG islands, recombination rate).

Major Findings Reported in the Rough Drafts of the Human Genome (cont'd)

- Human genes do more work than those of the roundworm or fruit fly (eg, alternative splicing is used more frequently).
- The human proteome is more complex than that found in invertebrates.
- Repeat sequences probably constitute more than 50% of the genome.
- Approximately 100 coding regions have been copied and moved by RNA-based transposons.
- Approximately 200 genes may be derived from bacteria by lateral transfer.
- More than 3 million SNPs have been identified.

(from Marry et. al., 2003, p. 636)

Human Genome: Implications for Medicine

- Practically every area of medicine will be affected by the new information accruing from knowledge of the human genome.
  - **Tracking of disease genes**: SNP maps assist determination of genes involved in complex diseases.
  - **Improved diagnostic testing** for disease susceptibility genes.
  - **Pharmacogenomics:** drugs will be tailored to accommodate the variations in enzymes and other proteins.
  - Genes involved in **behavior:** possible treatment of psychiatric disorders.
  - But there are **many ethical issues** – eg, privacy concerns and the use of genomic information.
  - Medical and economic benefits to **Third World countries**.

(from Marry et. al., 2003, p. 638)

Human Genome: Summary

- Determination of the complete sequence of the **human genome** is one of the most significant scientific achievements of all time.
- Many important findings: the number of human genes may be only two to three times that estimated for the roundworm and the fruit fly.
- Information from the Human Genome Project is having major influences in proteomics, bioinformatics, biotechnology, and pharmacogenomics as well as all areas of biology and medicine.
- The knowledge will be used wisely and fairly and that the benefits that will ensue regarding health, disease, and other matters will be made available to all people everywhere.

(from Marry et. al., 2003, p. 638)

Immunoglobulin Diversity

Immunoglobulin Diversity

- **Antibodies** are proteins manufactured by vertebrate immune systems that aid in defense against infectious agents and other substances foreign to the animal.
  - It is estimated that a human is capable of synthesizing more than 10 million distinct antibodies. Most of this great diversity is generated through the action of precisely controlled **gene rearrangements** which occur during differentiation of many individual clones of cells, each clone specialized for the synthesis of one and only one antibody.
  - The immune response involves proliferation of clones of cells that produce antibodies reacting with a specific antigen, or **immunogen**.
  - This clonal expansion allows large-scale production of the specific antibodies needed to combat infection or other challenges to the immune system

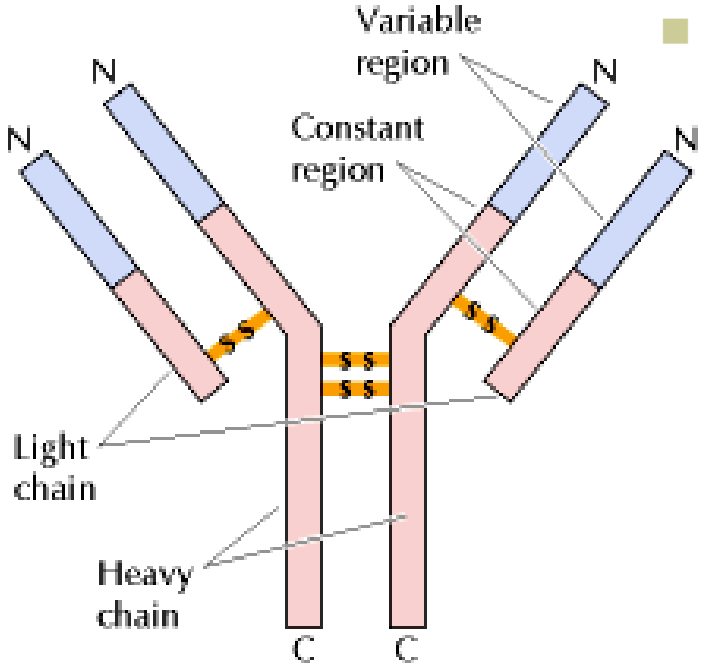
Site-specific recombination

- Site-specific recombination is important in programmed gene rearrangements within cell genomes.
- In vertebrates, site-specific recombination is critical to the development of the immune system, which recognizes foreign substances (antigens) and provides protection against infectious agents.
- There are two major classes of immune responses, which are mediated by B and T lymphocytes.
  - B lymphocytes secrete antibodies (immunoglobulins) that react with soluble antigens;
  - T lymphocytes express cell surface proteins (called T cell receptors) that react with antigens expressed on the surfaces of other cells.

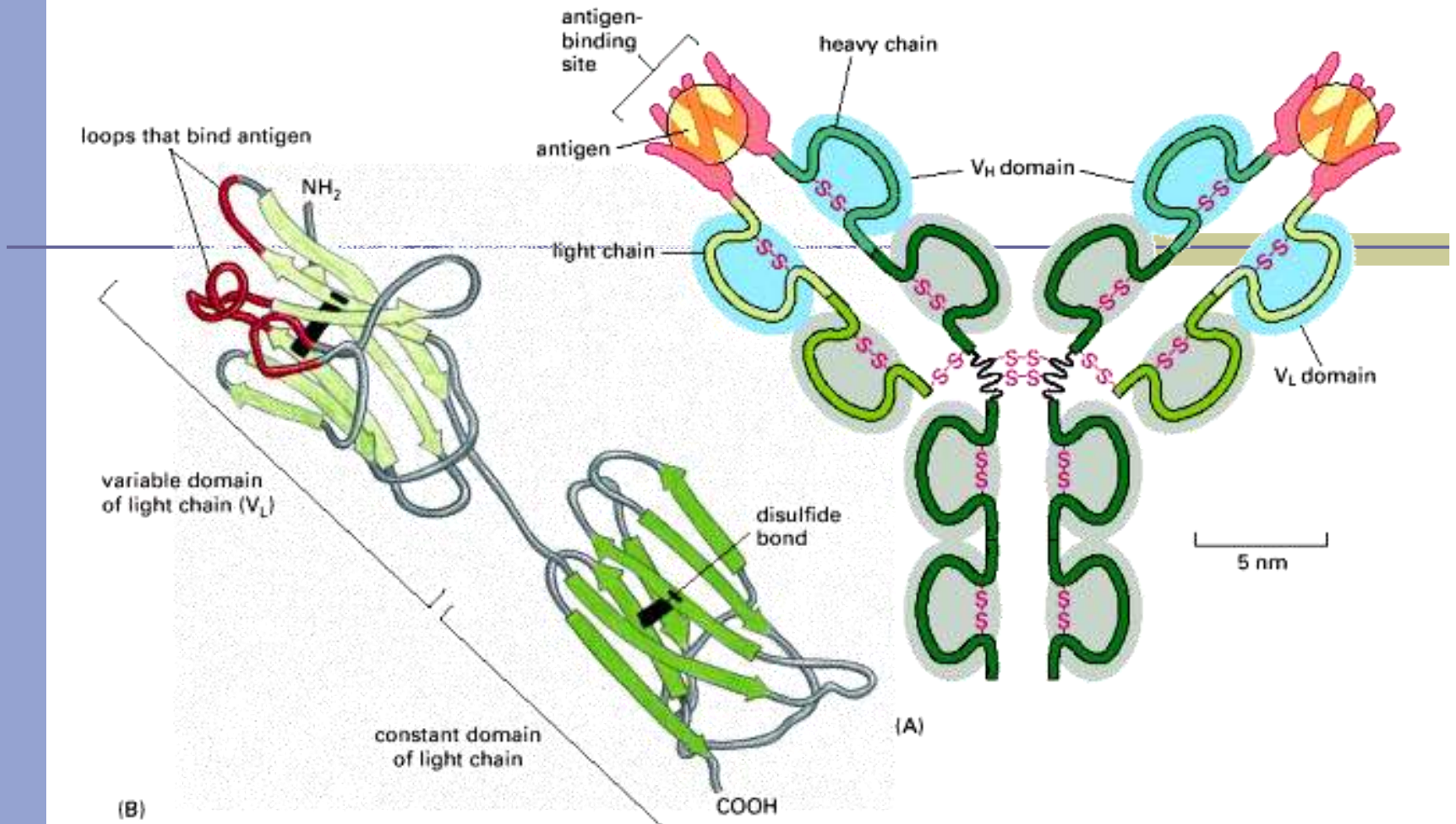
Site-specific recombination (cont'd)

- The key feature of both immunoglobulins and T cell receptors is their enormous diversity, which enables different antibody or T cell receptor molecules to recognize a vast array of foreign antigens.
  - For example, each individual is capable of producing more than 10^{11} different antibody molecules, which is far in excess of the total number of genes in the human genome (approximately 10^5).
  - Rather than being encoded in germ-line DNA, these diverse antibodies (and T cell receptors) are encoded by unique lymphocyte genes that are formed during development of the immune system as a result of sitespecific recombination between distinct segments of immunoglobulin and T cell receptor genes.

Common Structure of an Immunoglobulin



- The role of site-specific recombination in the formation of immunoglobulin genes was first demonstrated by **Susumu Tonegawa** in 1976.
  - Immunoglobulins consist of pairs of identical heavy and light polypeptide chains.
  - Both the heavy and light chains are composed of C-terminal **constant** regions and N-terminal **variable** regions.
  - The **variable** regions, which are responsible for antigen binding, and it is the diversity of variable region amino acid sequences that allows different individual antibodies to recognize unique antigens.



- (A) The antigen-binding site is formed where a heavy-chain variable domain (V_H) and a light-chain variable domain (V_L) come close together. These are the domains that differ most in their sequence and structure in different antibodies.
- (B) This ribbon model of a light chain shows the parts of the V_L domain most closely involved in binding to the antigen in red. They contribute half of the fingerlike loops that fold around each of the antigen molecules in (A).

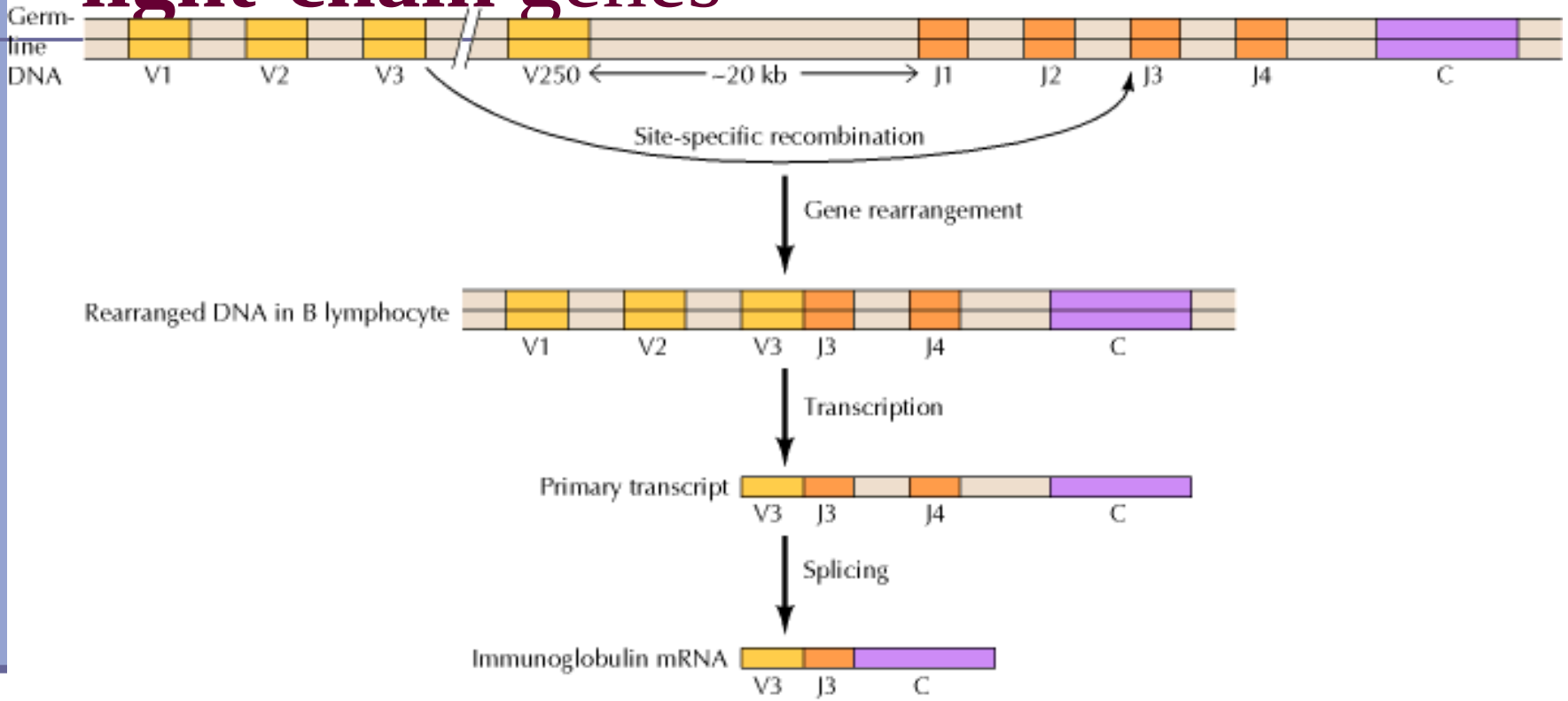
Gene Rearrangements

- Although every individual is capable of producing a vast spectrum of different antibodies, each B lymphocyte produces only a single type of antibody.
- Tonegawa's key discovery was that each antibody is encoded by unique genes formed by site-specific recombination during B lymphocyte development.
- These gene rearrangements create different immunoglobulin genes in different individual B lymphocytes, so the population of approximately 10^{12} B lymphocytes in the human body includes cells capable of producing antibodies against a diverse array of foreign antigens.

Immunoglobulin Light Chains

- The genes that encode immunoglobulin **light chains** consist of three regions:
  - a V region that encodes the 95 to 96 N-terminal amino acids of the polypeptide variable region;
  - a joining (J) region that encodes the 12 to 14 C-terminal amino acids of the polypeptide variable region; and
  - a C region that encodes the polypeptide constant region.
- The major class of light-chain genes in the mouse are formed from combinations of approximately 250 V regions and four J regions with a single C region.
- Site-specific recombination during lymphocyte development leads to a gene rearrangement in which a single V region recombines with a single J region to generate a functional light-chain gene.
- Different V and J regions are rearranged in different B lymphocytes, so the possible combinations of 250 V regions with 4 J regions can generate approximately 1000 (4 × 250) unique light chains.

Rearrangement of immunoglobulin **light-chain** genes



- Each light-chain gene (mouse κ light chains are illustrated) consists of a constant region (C), a joining region (J), and a variable region (V).
- There are approximately 250 different V regions, which are separated from J and C by about 20 kb in germ-line DNA.

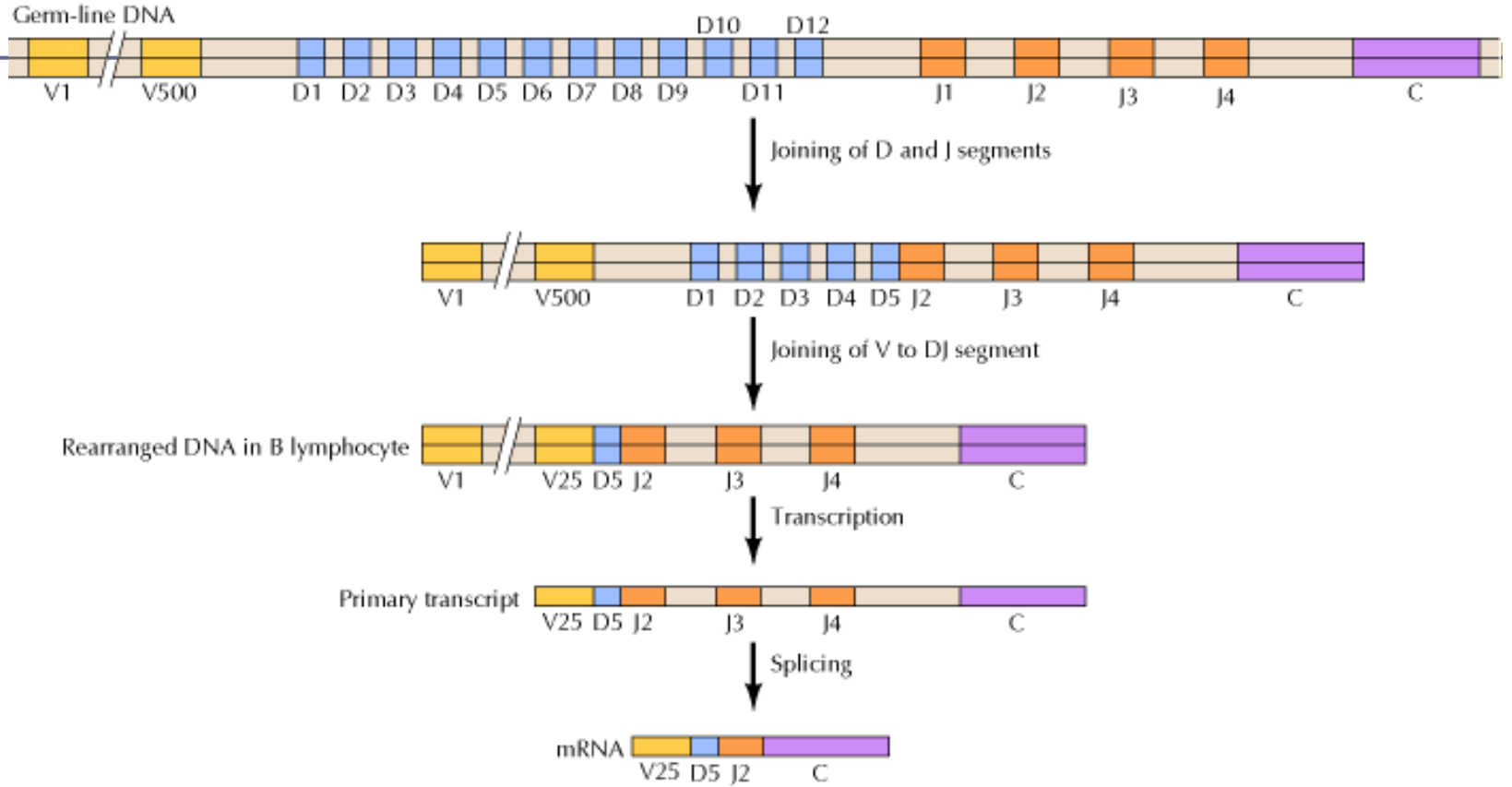
Rearrangement of immunoglobulin **light-chain** genes (cont'd)

- During the development of B lymphocytes, site-specific recombination joins one of the V regions to one of the four J regions.
  - This rearrangement activates transcription, resulting in the formation of a primary transcript containing the rearranged VJ region together with the remaining J regions and C.
  - The remaining unused J regions and the introns between J and C are then removed by splicing, yielding a functional mRNA.

Immunoglobulin Heavy Chains

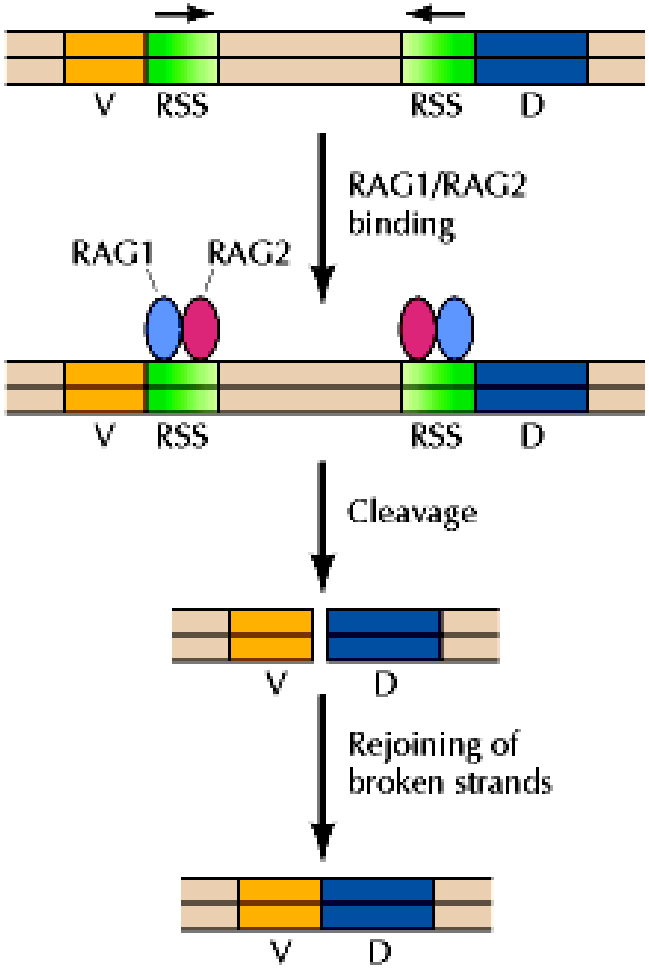
- The heavy-chain genes include a fourth region (known as the diversity, or D, region), which encodes amino acids lying between V and J.
- Assembly of a functional heavy-chain gene requires two recombination events:
  - A D region first recombines with a J region, and
  - a V region then recombines with the rearranged DJ segment.
- In the mouse, there are about 500 heavy-chain V regions, 12 D regions, and 4 J regions, so the total number of heavy chains that can be generated by the recombination events is: 24,000 (500 × 12 × 4).

Rearrangement of immunoglobulin **heavy-chain** genes



- The heavy-chain genes contain D regions in addition to V, J, and C regions. First the D and J segments join.
- Then a V segment is joined to the rearranged DJ region. The introns between J and C are removed by splicing to yield heavy-chain mRNA.

VDJ recombination



- The coding segments of immunoglobulin and T cell receptor genes (e.g., a V and D segment) are flanked by short **recombination signal sequences (RSS)**, which are in opposite orientations at the 5' and 3' ends of the coding sequences.
- The RSS are recognized by a complex of the **lymphocytespecific recombination proteins RAG1** and **RAG2**, which cleave the DNA between the coding sequence and the RSS.
- The broken coding strands are then rejoined to yield a rearranged gene segment.

Antibody Diversity

- Combinations between the 1000 different light chains and 24,000 different heavy chains formed by site-specific recombination can generate approximately **2×10^7** different immunoglobulin molecules.
- This diversity is further increased because the joining of immunoglobulin gene segments is often imprecise, with one to several nucleotides frequently being lost or gained at the sites of joining.
- The mutations resulting from these deletions and insertions increase the diversity of immunoglobulin variable regions approximately a hundredfold, corresponding to the formation of about **10^5** different light chains and **2×10^6** heavy chains, which can then combine to form more than **10^{11}** distinct antibodies.
- Still further antibody diversity is generated after the formation of rearranged immunoglobulin genes by a process known as **somatic hypermutation**, which results in the introduction of frequent mutations into the variable regions of both heavy-chain and light-chain genes.


Thank you
for your attention