16

Basics of Molecular Biology

Yinghui Li (yinghuidd@vip.sina.com), Dingsheng Zhao
State Key Laboratory of Space Medicine Fundamentals and Application, Astronaut Research and Training Center of China, Beijing 100094, China

16.1 Introduction

Molecular biology is the study of biology on molecular level. The field overlaps with areas of biology and chemistry, particularly genetics and biochemistry. Molecular biology chiefly concerns itself with understanding the interactions between the various systems of a cell, including the interactions between DNA (deoxyribonucleic acid), RNA (Ribonucleic acid) and protein biosynthesis as well as learning how these interactions are regulated[1].

Researchers in molecular biology use specific techniques native to molecular biology (see the techniques section), but they combine these with techniques and ideas from genetics and biochemistry. There is not a definite line between these disciplines. Today the terms *molecular biology* and *biochemistry* are nearly interchangeable.

Biochemistry is the study of chemical substances and vital processes occurring in living organisms. Biochemists focus heavily on the role, function, and structure of biomolecules. The study of chemistry behind biological processes and the synthesis of biologically active molecules are examples of biochemistry.

Genetics is the study of the effects of genetic differences on organisms. Often this can be inferred by the absence of a normal component (*e.g.* one gene). Mutants lack one or more functional components with respect to the so-called "wild type" or normal phenotype. Genetic interactions (epistasis) can often confound simple interpretations of such "knock-out" studies.

Molecular biology is the study of molecular underpinnings of the process of replication, transcription and translation of the genetic material. The central dogma of molecular biology where genetic material is transcribed into RNA and

J. Tian, *Molecular Imaging*

then translated into proteins, despite being an oversimplified picture of molecular biology, still provides a good starting point for understanding the field. This picture, however, is undergoing revision in light of emerging novel roles for RNA.

Much of the work in molecular biology is quantitative, and recently much work has been done at the interface of molecular biology and computer science in bioinformatics and computational biology. In the early 2000s, the study of gene structure and function, molecular genetics, has been amongst the most prominent sub-fields of molecular biology.

Increasingly, many other fields of biology focus on molecules, either by directly or indirectly studying their interactions such as in cell biology and developmental biology, where the techniques of molecular biology are used to infer historical attributes of populations or species, seen in fields in evolutionary biology such as population genetics and phylogenetics. There is also a long tradition of studying biomolecules "from the ground up" in biophysics.

16.1.1 History of Molecular Biology

The history of molecular biology begins in the 1930s with the convergence of various, previously distinct biological disciplines: biochemistry, genetics, microbiology, and virology. With the hope of understanding life at its most fundamental level, numerous physicists and chemists also took an interest in what would become molecular biology.

In its modern sense, molecular biology attempts to explain the phenomena of life starting from the macromolecular properties that generate them. Two categories of macromolecules in particular are the focus of the molecular biologist: (1) nucleic acids, among which the most famous is deoxyribonucleic acid (DNA), the constituent of genes, and (2) proteins, which are the active agents of living organisms. One definition of the scope of molecular biology therefore is to characterize the structure, function and relationships between these two types of macromolecules. This relatively limited definition will suffice to allow us to establish a date for the so-called "molecular revolution", or at least to establish a chronology of its most fundamental developments.

In its earliest manifestations, molecular biology—the name was coined by Warren Weaver of the Rockefeller Foundation in 1938—was an ideal of physical and chemical explanations of life, rather than a coherent discipline. Following the advent of the Mendelian-chromosome theory of heredity in the 1910s and the maturation of the atomic theory and quantum mechanics in the 1920s, such explanations seemed within reach. Weaver and others funded research at the intersection of biology, chemistry and physics, while prominent physicists such as Niels Bohr and Erwin Schrödinger turned their attention to biological speculation. However, in the 1930s and 1940s, it was by no means clear that—if any—cross-disciplinary research would bear fruit: work in colloid chemistry, biophysics and radiation biology, crystallography, and other emerging fields all seemed promising.

In 1940, George and Edward demonstrated the existence of a precise relationship between genes and proteins[2]. In the course of their experiments connecting genetics with biochemistry, they switched from the genetics mainstay *Drosophila* to a more appropriate model organism, the fungus *Neurospora*; the construction and exploitation of new model organisms would become a recurring theme in the development of molecular biology. In 1944, Oswald Avery demonstrated that genes are made up of DNA[3]. In 1952, Alfred Hershey and Martha Chase confirmed that the genetic material of the bacteriophage, the virus which infects bacteria, is made up of DNA[4]. In 1953, James Watson and Francis Crick discovered the double helical structure of the DNA molecule[5]. In 1961, Francois Jacob and Jacques Monod hypothesized the existence of an intermediary between DNA and its protein products, which they called messenger RNA (mRNA). Between 1961 and 1965, the relationship between the information contained in DNA and the structure of proteins was determined: there is a code, the genetic code, which creates a correspondence between the succession of nucleotides in the DNA sequence and a series of amino acids in proteins. At the beginning of the 1960s, Monod and Jacob also demonstrated how certain specific proteins, called regulative proteins, latch onto DNA at the edges of the genes and control the transcription of these genes into mRNA; they direct the "expression" of the genes[6].

The chief discoveries of molecular biology took place in a period of only about twenty-five years. Another fifteen years were required before new and more sophisticated technologies, united today under the name of genetic engineering, would permit the isolation and characterization of genes, in particular those of highly complex organisms.

16.1.2 Introduction of Genes and Gene Function

The gene is the basic unit of heredity in a living organism. All living things depend on genes. Genes hold the information to build and maintain their cells and pass genetic traits to offspring. In general terms, a gene is a segment of nucleic acid that contains some genetic information and specifies a trait. The colloquial usage of the term gene often refers to the scientific concept of an allele.

The existence of genes was first suggested by Gregor Mendel, who studied inheritance in pea plants in the 1860s and hypothesized a factor that conveys traits from the parents to the offspring. In the early 1900s, Mendel's work received renewed attention from scientists. In 1910, Thomas Hunt Morgan showed that genes reside on specific chromosomes. He later showed that genes occupy specific locations on the chromosome. With this knowledge, Morgan and his students began the first chromosomal map of the fruit fly *Drosophila*. In 1928, Frederick

Griffith showed that genes could be transferred. In what is now known as Griffith's experiment, injections into a mouse of a deadly strain of bacteria that had been heat-killed transferred genetic information to a safe strain of the same bacteria, killing the mouse. In 1941, George Wells Beadle and Edward Lawrie Tatum showed that mutations in genes caused errors in specific steps in metabolic pathways. This showed that specific genes code for specific proteins, leading to the "one gene, one enzyme" hypothesis[7]. In 1944, Oswald Avery, Colin Munro MacLeod, and Maclyn McCarty showed that DNA holds the gene's information[8]. In 1953, James D. Watson and Francis Crick demonstrated the molecular structure of DNA. Together, these discoveries established the central dogma of molecular biology, which states that proteins are translated from RNA which is transcribed from DNA. This dogma has been shown to have exceptions, such as reverse transcription in retroviruses. In 1972, Walter Fiers and his team the first determine the sequence of a gene: the gene for the Bacteriophage MS2 coat protein[9]. In 1977, Richard J. Roberts and Phillip Sharp discovered that genes can be split into segments. This leads to the idea that one gene can make several proteins. Recently, biological results let the notion of gene appear more dubious. In particular, genes do not seem to sit side by side on DNA like discrete beads. Instead, regions of the DNA producing distinct proteins may overlap, so that the idea emerges that "genes are one long continuum"[10].

In cells, a gene is a portion of DNA that contains both "coding" sequences that determine what the gene does, and "non-coding" sequences that determine when the gene is expressed. When a gene is active, the coding and non-coding sequences are copied in a process called transcription, and produce an RNA copy of the gene's information. This piece of RNA can then direct the synthesis of proteins via the genetic code. In other cases, the RNA is used directly as part of the ribosome. The molecules resulting from gene expression, whether RNA or protein, are known as gene products and are responsible for the development and functioning of all living things.

The vast majority of living organisms encode their genes in long strands of DNA which is made from four types of nucleotide subunits: adenine, cytosine, guanine, and thymine. Each nucleotide subunit consists of three components: a phosphate group, a deoxyribose sugar ring and a nucleobase. Thus, nucleotides in DNA or RNA are typically called "base" as a consequence, they are commonly referred to simply by their purine or pyrimidine original base components. Adenine and guanine are purines, and cytosine and thymine are pyrimidines. The most common form of DNA in a cell is a double helix structure, in which two individual DNA strands twist around each other in a right-handed spiral. In this structure, the base pairing rules specify that guanine pairs with cytosine and adenine pairs with thymine (each pair contains one purine and one pyrimidine). The base pairing between guanine and cytosine forms three hydrogen bonds, whereas the base pairing between adenine and thymine forms two hydrogen bonds. The two strands in a double helix must therefore be complementary, so that their bases must align such that the adenines of one strand are paired with the thymines of the other strand, and so on.

Due to the chemical composition of the pentose residues of the bases, DNA strands have directionality. One end of a DNA strand contains an exposed hydroxyl group on the deoxyribose; this is known as the 3' end of the molecule. The other end contains an exposed phosphate group; this is the 5' end. The directionality of DNA is vitally important to many cellular processes, since double helices are necessarily directional (a strand running 5'-3' pairs with a complementary strand running 3'-5'), and processes such as DNA replication occur in only one direction. All nucleic acid synthesis in a cell occurs in the 5'-3' direction, because new monomers are added via a dehydration reaction that uses the exposed 3' hydroxyl as a nucleophile.

The expression of genes begins by transcribing the gene into RNA, a second type of nucleic acid that is very similar to DNA, but whose monomers contain the sugar ribose rather than deoxyribose. RNA also contains the base uracil in place of thymine. RNA molecules are less stable than DNA and are typically single-stranded. Genes that encode proteins are composed of a series of three-nucleotide sequences called codons, which serve as the words in the genetic language. The genetic code specifies the correspondence during protein translation between codons and amino acids. The genetic code is nearly the same for all known organisms.

In some cases, RNA is an intermediate product in the process of manufacturing proteins from genes. However, for other gene sequences, the RNA molecules are the actual functional products. For example, RNAs known as ribozymes are capable of enzymatic function, and miRNAs have a regulatory role. The DNA sequences from which such RNAs are transcribed are known as RNA genes.

Some viruses store their entire genomes in the form of RNA, and contain no DNA at all. Because they use RNA to store genetic information, their cellular hosts may synthesize their proteins as soon as they are infected and without the delay in waiting for transcription. On the other hand, RNA retroviruses, such as HIV, require reverse transcription of their genome from RNA into DNA before their proteins can be synthesized. In 2006, French researchers came across a puzzling example of RNA-mediated inheritance in the mouse. Mice with a loss-of-function mutation in the gene Kit have white tails. Offspring of these mutants can have white tails despite having only normal Kit genes. The research team traced this effect back to mutated Kit RNA[11]. RNA is common as genetic storage material in viruses, and particularly in mammals, RNA inheritance has been rarely observed.

All genes have regulatory regions in addition to regions that explicitly code for a protein or RNA product. A regulatory region shared by almost all genes is known as the promoter, which provides a position that is recognized by the transcription machinery when a gene is about to be transcribed. A gene can have more than one promoter, resulting in RNAs that differ in how far they extend from the promoter[12]. Although promoter regions have a consensus sequence that is the most common sequence at this position, some genes have “strong” promoters that bind the transcription machinery well, and others have “weak” promoters that bind poorly. These weak promoters usually permit a lower rate of transcription than the strong promoters, because the transcription machinery binds to them and initiates transcription less frequently. Other possible regulatory regions include enhancers, which can compensate for a weak promoter. Most regulatory regions are "upstream"—that is, before or toward the 5' end of the transcription initiation site. Eukaryotic promoter regions are much more complex and difficult to identify than prokaryotic promoters.

Many prokaryotic genes are organized into operons, or groups of genes whose products have related functions and which are transcribed as a unit. By contrast, eukaryotic genes are transcribed only one at a time, but may include long stretches of DNA called "introns" which are transcribed but never translated into protein (they are spliced out before translation). Splicing can also occur in prokaryotic genes, but is less common than in eukaryotes.

The total of genes in an organism or cell is known as its genome, which may be stored on one or more chromosomes. The region of the chromosome at which a particular gene is located is called its locus. A chromosome consists of a single, very long DNA helix on which thousands of genes are encoded. Prokaryotes-bacteria and archaea-typically store their genomes on a single large, circular chromosome, which is sometimes supplemented by additional small circles of DNA called plasmids. These usually encode only a few genes and are easily transferable between individuals. For example, the genes for antibiotic resistance are usually encoded on bacterial plasmids and can be passed between individual cells, even those of different species via horizontal gene transfer. Although some simple eukaryotes also possess plasmids with a small number of genes, the majority of eukaryotic genes are stored on multiple linear chromosomes, which are packed within the nucleus in a complex with storage proteins called histones. The manner in which DNA is stored on the histone, as well as the chemical modifications of the histone itself, are regulatory mechanisms governing whether or not a particular region of DNA is accessible for transcription. The ends of eukaryotic chromosomes are capped by long stretches of repetitive sequences called telomeres, which do not code for any gene product but are present to prevent degradation of coding and regulatory regions during DNA replication. The length of the telomeres tends to decrease each time the genome is replicated in preparation for cell division; the loss of telomeres has been proposed as an explanation for cellular senescence, or the loss of the ability to divide[13].

Whereas the chromosomes of prokaryotes are relatively gene-dense, those of eukaryotes often contain so-called "junk DNA," or regions of DNA that serve no obvious function. Simple single-celled eukaryotes have relatively small amounts of such DNA, whereas the genomes of complex multi-cellular organisms contain an absolute majority of DNA without an identified function[14]. However, it now appears that although protein-coding DNA makes up barely 2% of the human genome, about 80% of the bases in the genome may be expressed, so the term "junk DNA" may be a misnomer[15].

16.2 Techniques of Molecular Biology

Since the late 1950s and early 1960s, molecular biologists have learned to characterize, isolate, and manipulate the molecular components of cells and organisms. These components include DNA, the repository of genetic information; RNA, a close relative of DNA whose functions range from serving as a temporary working copy of DNA to actual structural and enzymatic functions, as well as a functional and structural part of the translational apparatus; and proteins, the major structural and enzymatic type of molecule in cells.

16.2.1 Expression Qoning

One of the most basic techniques of molecular biology to study protein function is expression cloning which is a technique of DNA cloning to generate a library of clones, with each clone expressing one protein. This expression library is then screened for the property of interest gene and clones of interest are recovered for further analysis. An example would be using an expression library to isolate genes that could confer antibiotic resistance. Expression vectors are a specialized type of cloning vector which includes the transcriptional and translational signals needed for the regulation of the gene of interest. The transcriptional and translational signals may be synthetically created to make the expression of the gene of interest easier to regulate. Usually the ultimate aim of expression cloning is to produce large quantities of specific proteins. To this end, a bacterial expression clone may include a ribosome binding site (Shine-Dalgarno sequence) to enhance translation of messager RNA (mRNA) of the interested gene, a transcription termination sequence, or in eukaryotes, specific sequences to promote the post-translational modification of the protein product.

In this technique, DNA coding for a protein of interest is cloned (using PCR and/or restriction enzymes) into a plasmid (known as an expression vector). This plasmid may have special promoter elements to drive production of the protein of interest, and may also have antibiotic resistance markers to help to screen the plasmid.

This plasmid can be transferred into either bacterial or animal cells. Introducing DNA into bacterial cells can be done by transformation (via uptake of naked DNA), conjugation (via cell-cell contact) or by transduction (via viral vector). Introducing DNA into eukaryotic cells, such as animal cells, by physical or chemical means is called transfection. Several different transfection techniques are available, such as calcium phosphate transfection, electroporation, microinjection and liposome transfection. DNA can also be introduced into eukaryotic cells using viruses or bacteria as carriers; the latter is sometimes called bactofection and in particular uses *Agrobacterium tumefaciens*. The plasmid may be integrated into the genome, resulting in a stable transfection, or may remain independent of the genome, called transient transfection.

In either case, DNA coding for a protein of interest is now inside a cell, and the protein can be expressed. A variety of systems, such as inducible promoters and specific cell-signaling factors, are available to help express the protein of interest at high levels. Large quantities of a protein can then be extracted from the bacterial or eukaryotic cell. The protein can be tested for enzymatic activity under a variety of situations; the protein may be crystallized so its tertiary structure can be studied, or in the pharmaceutical industry, the activity of new drugs against the protein can be studied.[1].

16.2.2 Polymerase Chain Reaction

The polymerase chain reaction (PCR) is a widely used technique in molecular biology. It derives its name from one of its key components, a DNA polymerase used to amplify a piece of DNA by *in vitro* enzymatic replication. As PCR progresses, the DNA generated is used as a template for replication. This sets in motion a chain reaction in which the DNA template is exponentially amplified. With PCR, it is possible to amplify a single or few copies of a piece of DNA across several orders of magnitude, generating millions or more copies of the DNA piece. PCR can be extensively modified to perform a wide array of genetic manipulations.

Almost all PCR applications employ a heat-stable DNA polymerase, such as Taq polymerase, an enzyme originally isolated from the bacterium *Thermus aquaticus*[16]. This DNA polymerase enzymatically assembles a new DNA strand from DNA building blocks, the nucleotides, by using single-stranded DNA as a template and DNA oligonucleotides (also called DNA primers), which are required for initiation of DNA synthesis. The vast majority of PCR methods use thermal cycling, *i.e.*, alternately heating and cooling the PCR sample to a defined series of temperature steps. These thermal cycling steps are necessary to physically separate the strands (at high temperatures) in a DNA double helix (DNA melting) used as the template during DNA synthesis (at lower temperatures) by the DNA polymerase to selectively amplify the target DNA. The selectivity of PCR results from the use of primers that are complementary to the DNA region targeted for amplification under specific thermal cycling conditions.

Developed in 1984 by Kary Mullis[17], PCR is now a common and often indispensable technique used in medical and biological research labs for a variety of applications[18, 19]. These include DNA cloning for sequencing, DNA-based phylogeny, or functional analysis of genes; the diagnosis of hereditary diseases; the identification of genetic fingerprints (used in forensic sciences and paternity testing); and the detection and diagnosis of infectious diseases. In 1993, Mullis was awarded the Nobel Prize in Chemistry for his work on PCR.

PCR allows isolation of DNA fragments from genomic DNA by selective amplification of a specific region of DNA. This use of PCR augments many methods, such as generating hybridization probes for southern or northern hybridization and DNA cloning, which require larger amounts of DNA, representing a specific DNA region. PCR supplies these techniques with high amounts of pure DNA, enabling analysis of DNA samples even from very small amounts of starting material.

Other applications of PCR include DNA sequencing to determine unknown PCR-amplified sequences in which one of the amplification primers may be used in Sanger sequencing, which is isolation of a DNA sequence to expedite recombinant DNA technologies involving the insertion of a DNA sequence into a plasmid or the genetic material of another organism. Bacterial colonies (*E. coli*) can be rapidly screened by PCR for correct DNA vector constructs. PCR may also be used for genetic fingerprinting; a forensic technique used to identify a person or organism by comparing experimental DNAs through different PCR-based methods[20].

Some PCR 'fingerprints' methods have high discriminative power and can be used to identify genetic relationships between individuals, such as parent-child or between siblings, and are used in paternity testing. This technique may also be used to determine evolutionary relationships among organisms.

Because PCR amplifies the regions of DNA that it targets, PCR can be used to analyze extremely small amounts of the sample. This is often critical for forensic analysis, when only a trace amount of DNA is available as evidence. PCR may also be used in the analysis of ancient DNA that is tens of thousands of years old. These PCR-based techniques have been successfully used on animals, such as a forty-thousand-year-old mammoth, and also on human DNA, in applications ranging from the analysis of Egyptian mummies to the identification of a Russian Tsar.

Quantitative PCR methods allow the estimation of the amount of a given sequence present in a sample—a technique often applied to quantitatively determine levels of gene expression. Real-time PCR is an established tool for DNA quantification that measures the accumulation of DNA product after each round of PCR amplification.

PCR allows early diagnosis of malignant diseases such as leukemia and lymphomas, which is currently the highest developed in cancer research and is already being used routinely. PCR assays can be performed directly on genomic DNA samples to detect translocation-specific malignant cells at a sensitivity which is at least 10,000 fold higher than other methods[21].

PCR also permits identification of non-cultivatable or slow-growing microorganisms such as mycobacteria, anaerobic bacteria, or viruses from tissue culture assays and animal models. The basis for PCR diagnostic applications in microbiology is the detection of infectious agents and the discrimination of non-pathogenic from pathogenic strains by virtue of specific genes.

Viral DNA can likewise be detected by PCR. The primers used need to be specific to the targeted sequences in the DNA of a virus, and PCR can be used for diagnostic analyses or DNA sequencing of the viral genome. The high sensitivity of PCR permits virus detection soon after infection and even before the onset of disease. Such early detection may give physicians a significant lead in treatment. The amount of virus ("viral load") in a patient can also be quantified by PCR-based DNA quantitation techniques[21].

16.2.3 Other Techniques

Gel electrophoresis is one of the principal tools of molecular biology. The basic principle is that DNA, RNA, and proteins can all be separated by means of an electric field. In agarose gel electrophoresis, DNA and RNA can be separated on the basis of size by running the DNA through an agarose gel. Proteins can be separated on the basis of size by using an SDS-PAGE gel or on the basis of size and their electric charge by using what is known as a 2D gel electrophoresis.

The terms northern, western and eastern blotting are derived from what initially was a molecular biology joke that played on the term southern blotting, after the technique described by Edwin Southern for the hybridization of blotted DNA. Patricia Thomas, developer of the RNA blot which then became known as the northern blot actually didn't use the term[22]. Further combinations of these techniques produced such terms as southwesterns (protein-DNA hybridizations), northwesterns (to detect protein-RNA interactions) and farwesterns (protein-protein interactions), all of which are presently found in the literature.

Named after its inventor, biologist Edwin Southern, the southern blot is a method for probing for the presence of a specific DNA sequence within a DNA sample. DNA samples before or after restriction enzyme digestions are separated by gel electrophoresis and then transferred to a membrane by blotting via capillary action. The membrane is then exposed to a labeled DNA probe that has a complement base sequence to the sequence on the DNA of interest. Most original protocols used radioactive labels; however, non-radioactive alternatives are now available. Southern blotting is less commonly used in laboratory science due to the capacity of other techniques, such as PCR, to detect specific DNA sequences from DNA samples. These blots are still used for some applications; however, they measure the transgene copy number in transgenic mice or in the engineering of gene knockout embryonic stem cell lines.

The northern blot is used to study the expression patterns a specific type of RNA molecule has as a relative comparison among a set of different samples of RNA. It is essentially a combination of denaturing RNA gel electrophoresis and a blot. In this process, RNA is separated based on size and is then transferred to a membrane that is then probed with a labeled complement of a sequence of interest. The results may be visualized through a variety of ways depending on the label used; however, most result in the revelation of bands representing the sizes of the RNA detected in the sample. The intensity of these bands is related to the amount of the target RNA in the samples analyzed. The procedure is commonly used to study when and how much gene expression is occurring by measuring how much of that RNA is present in different samples. It is one of the most basic tools for determining at what time and under what conditions certain genes are expressed in living tissues.

Antibodies to most proteins can be created by injecting small amounts of the protein into an animal such as a mouse, rabbit, sheep, or donkey (polyclonal antibodies) or produced in cell culture (monoclonal antibodies). These antibodies can be used for a variety of analytical and preparative techniques. In western blotting, proteins are first separated by size, in a thin gel sandwiched between two glass plates in a technique known as SDS-PAGE (sodium dodecyl sulfate polyacrylamide gel electrophoresis)[23]. The proteins in the gel are then transferred to a PVDF, nitrocellulose, nylon or other support membrane. This membrane can then be probed with solutions of antibodies. Antibodies that specifically bind to the protein of interest can then be visualized by a variety of techniques, including colored products, chemiluminescence or autoradiography. Often, the antibodies are labeled with an enzyme. When a chemiluminescent substrate is exposed to the enzyme it allows detection. Using western blotting techniques allows not only detection but also quantitative analysis. Analogous methods to western blotting can be used to directly stain specific proteins in live cells or tissue sections. However, these immunostaining methods, such as FISH, are used more often in cell biology research.

The eastern blotting technique is used to detect post-translational modification of proteins. Proteins blotted onto the PVDF or nitrocellulose membrane are probed for modifications using specific substrates.

A DNA array is a collection of spots attached to a solid support such as a microscope slide where each spot contains one or more single-stranded DNA oligonucleotide fragments. Arrays make it possible to put down a large quantity of very small (100 mm diameter) spots on a single slide. Each spot has a DNA fragment molecule that is complementary to a single DNA sequence (similar to southern blotting). A variation of this technique allows the gene expression of an organism at a particular stage in development to be quantified (expression profiling). In this technique, the RNA in a tissue is isolated and converted to labeled cDNA. This cDNA is then hybridized to the fragments on the array and visualization of the hybridization can be done. Since multiple arrays can be made with the exact same position of fragments, they are particularly useful for comparing the gene expression of two different tissues, such as a healthy and cancerous tissue. Also, one can measure what genes are expressed and how that expression changes with time or with other factors. For instance, the common baker's yeast, *Saccharomyces cerevisiae*, contains about 7,000 genes; with a microarray, one can measure qualitatively how each gene is expressed, and how that expression changes, for example, with a change in temperature[24]. There are many different ways to fabricate microarrays; the most common are silicon chips, microscope slides with spots of –100 mm diameter, custom arrays, and arrays with larger spots on porous membranes (macroarrays). There can be anywhere from 100 spots to more than 10,000 on a given array[25]. Arrays can also be made with molecules other than DNA. For example, an antibody array can be used to determine what proteins or bacteria are present in a blood sample.

Allele specific oligonucleotide (ASO) is a technique that allows detection of single base mutations without the need for PCR or gel electrophoresis. Short (20 – 25 nucleotides in length) labeled probes are exposed to the non-fragmented target DNA. Hybridization occurs with high specificity due to the short length of the probes and even a single base change will hinder hybridization. The target DNA is then washed and the labeled probes that didn't hybridize are removed. The target DNA is then analyzed for the presence of the probe via radioactivity or fluorescence. In this experiment, as in most molecular biology techniques, a control must be used to ensure successful experimentation. The Illumina Methylation Assay is an example of a method that takes advantage of the ASO technique to measure one base pair differences in sequence.

In molecular biology, procedures and technologies are continually being developed and older technologies abandoned. For example, before the advent of DNA gel electrophoresis (agarose or polyacrylamide), the size of DNA molecules was typically determined by the sedimentation rate in sucrose gradients, a slow and labor-intensive technique requiring expensive instrumentation; prior to sucrose gradients, viscometry was used. Aside from their historical interest, it is often worth knowing about older technology, as it is occasionally useful to solve another new problem for which the newer technique is inappropriate.

16.3 Cells and Viruses

16.3.1 Cellular Organization

Cells are the smallest structural unit of living organisms, capable of maintaining life and reproducing. Viruses are not cells because they cannot maintain life and reproduce by themselves. There are certain kinds of cells, such as a nerve cell or a red blood cell, but their organizations are essentially the same. Even plant cells and animal cells share significant similarity in the overall organization. For animal cells, the cell surface consists of the plasma membrane only, but plant cells have an additional layer called the cell wall, which is made up of cellulose and other polymers[26].

All cells are divided into two types: prokaryotic cells and eukaryotic cells. The prokaryotic cell does not have a nucleus, shown as Fig. 16.1. The eukaryotic cell contains a nucleus, shown as Fig. 16.2. Eukaryotes are the organisms made up of eukaryotic cells. They include protista, fungi, animals and plants. Prokaryotes include archaebacteria and eubacteria. They are single-cell organisms. More recently, "archaebacteria" have been placed in a category outside "bacteria," because they are quite different from ordinary bacteria. According to the new classification, prokaryotes are divided into archaea and bacteria, where "archaea" is equivalent to "archaebacteria," and "bacteria" is the same as "eubacteria." Archaea live in extreme environments. They may be organized into three groups: Methanogens which live in anaerobic environment such as swamps. They produce methane and cannot tolerate exposure to oxygen. Extreme halophiles live in very high concentrations of salt (NaCl), *e.g.*, the Dead Sea and the Great Salt Lake. Extreme thermophiles[27] live in hot, sulfur-rich and low pH environments, such as hot springs, geysers and fumaroles as seen in the Yellowstone National Park.

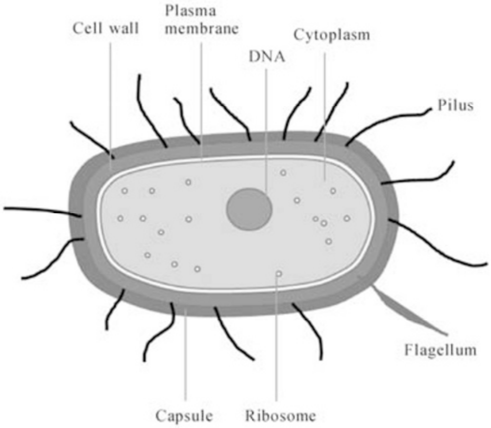


Fig. 16.1 Schematic drawing of a prokaryotic cell. The plasma membrane is surrounded by the cell wall, which is wrapped by the capsule

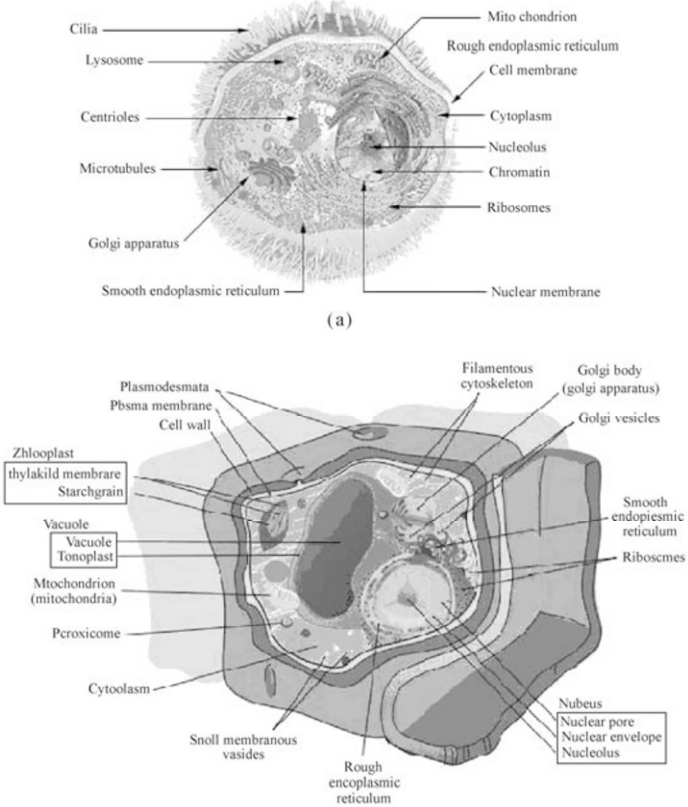


Fig. 16.2 Schematic drawing of eukaryotic cells. (a) An animal cell; (b) A plant cell[26] (with the pemission of Wikimedia)

All cells contain a cytoplasm, plasma membrane, and DNA. Cytoplasm is the viscous contents of a cell, including proteins, ribosomes, metabolites and ions. Ribosomes are the sites of protein synthesis. The plasma membrane is the cell membrane surrounding the cytoplasm, shown as Fig. 16.3. It consists of a phospholipid bilayer, associated proteins and carbohydrates. The phospholipid bilayer is also the basic constituent of other biomembranes. DNA is the genetic material. A eukaryotic cell contains several DNA molecules located in the nucleus and mitochondria which are membrane-bound organelles. A prokaryotic cell contains a single DNA molecule, which has no specific boundary with the cytoplasm.

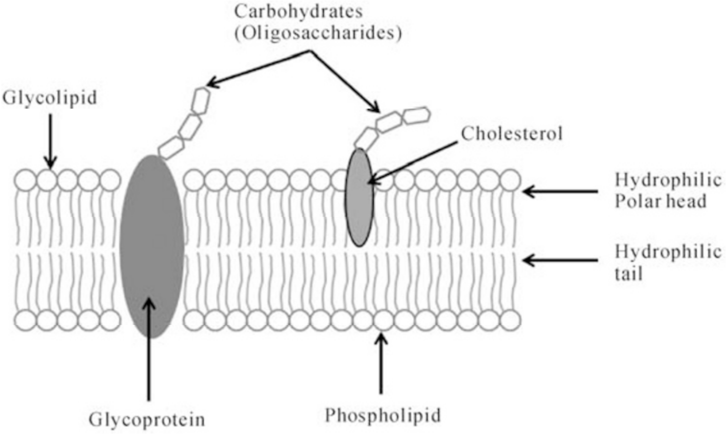


Fig. 16.3 Schematic drawing of a typical plasma membrane

The cell nucleus consists of a nuclear envelope, nucleolus and nucleoplasm. Most chromosomes are located in the nucleoplasm, but portions of several chromosomes containing clusters of rRNA genes may get together in the nucleolus, forming the nucleolar organizing region. The major role of the nucleolus is to produce rRNA. Chromosomes are the structures that hold DNA molecules. One chromosome contains a DNA molecule. Each chromosome has a *p* and *q* arm; *p* is the shorter arm and *q* is the longer arm. The arms are separated by a pinched region called the centromere. In order for chromosomes to be seen with a microscope, they need to be stained. Once stained, the chromosomes look like strings with light and dark "bands" and their picture can be taken. The picture, or chromosome map, is called a karyotype. The germ cell (sperm or egg) of a human being contains 23 chromosomes, labeled from 1 to 22 and is either X or Y. The somatic cell (cells other than germ cells) of a normal person has 46 chromosomes. For other species, the chromosome number varies from 1 to 1,260. A human somatic cell contains two chromosomes that determine the sex of a person. The two sex chromosomes are XY in males and XX in females. The gene (SRY) that is important for testes formation is located on the Y chromosome. It is possible that a person with testes still exhibits female characteristics[28-31].

16.3.2 Viruses

Viruses are the smallest organisms, with diameters ranging from 20 nm to 300 nm. Viruses are not cells. They consist of one or more molecules of DNA or RNA, which contain the virus's genes surrounded by a protein coat called the capsid. Some viruses also have an envelope surrounding the capsid. Viruses can be sphere-shaped or helical[26].

Viruses can be classified into different types. The Baltimore classification is based on genetic contents and replication strategies of viruses. The genetic material in all types of cells is double-stranded DNA, but some viruses use RNA or single-stranded DNA to carry genetic information. According to the Baltimore classification, viruses are divided into the following seven classes: dsDNA viruses, ssDNA viruses, dsRNA viruses, (+)-sense ssRNA viruses, (-)-sense ssRNA viruses, DNA reverse viruses, transcribing viruses, and DNA reverse transcribing viruses, where "ds" represents "double strands" and "ss" denotes "single strand".

The life cycle of viruses may be divided into the following stages: 1) Attachment is a specific binding between viral surface proteins and their receptors on the host cellular surface. This specificity determines the host range of a virus. For instance, the human immunodeficiency virus (HIV) attacks only human immune cells (mainly T cells), because its surface protein, gp120, can interact with CD4 and chemokine receptors on the T cell's surface. 2) Following attachment, viruses may enter the host cell through receptor mediated endocytosis or other mechanisms. 3) Uncoating is a process where the viral capsid is degraded by viral enzymes or host enzymes. 4) Replication involves assembly of viral proteins and genetic materials produced in the host cell. 5) Viruses may escape from the host cell by causing cell rupture (lysis). Enveloped viruses (*e.g.*, HIV) typically "bud" from the host cell. During the budding process, a virus acquires the phospholipid envelope containing the embedded viral glycoproteins[32].

Unlike most bacteria, most viruses cause disease because they invade living, normal cells, such as those in the human body. They then multiply and produce other viruses like themselves. Each virus is very particular about which cell it attacks. Various human viruses specifically attack particular cells in the body's organs, systems, or tissues, such as the liver, respiratory system, or blood cells.

Although viruses behave differently, most survive by taking over the machinery that makes a cell work. Briefly, when a single virus particle, a "virion," comes in contact with a cell it likes, it may attach to special landing sites on the surface of that cell. From there, the virus may inject molecules into the cell, or the cell may swallow up the virion. Once inside the cell, viral molecules such as DNA or RNA direct the cell to make new virus offspring. That's how a virus "infects" a cell.

Viruses can even "infect" bacteria. These viruses, called bacteriophages, may help researchers develop alternatives to antibiotics for wiping out bacterial infections.

Many viral infections do not result in disease. For example, by the time most people in the United States become adults, they have been infected by cytomegalovirus

(CMV). Most of these people, however, do not develop CMV disease symptoms. Other viral infections can result in deadly diseases, such as HIV which causes acquired immunodeficiency syndrome (AIDS) and coronaviruses which cause severe acute respiratory syndrome (SARS).

Bacteriophages are viruses infecting bacteria. Their genetic material can be DNA or RNA; single-stranded (ss) or double-stranded (ds). The life cycle of a bacteriophage may be lytic, lysogenic, or both. An important example is the λ phage which can have either a lytic or lysogenic cycle depending on the environment. The following are a few examples:

1) DsDNA phages with contractile tails, such as T4.
2) DsDNA phages with long flexible tails, such as l.
3) DsDNA phages with stubby tails, such as p22.
4) SsDNA phages, such as phi X 174.
5) SsRNA phages, such as MS2.

16.4 Transcription

Transcription is the process of creating an equivalent RNA copy of a sequence of DNA. Prokaryotic transcription occurs in the cytoplasm alongside translation. Unlike in eukaryotes, prokaryotic transcription and translation can occur simultaneously. This is impossible in eukaryotes, where transcription occurs in a membrane-bound nucleus while translation occurs outside the nucleus in the cytoplasm. In prokaryote, genetic material is not enclosed in a membrane-enclosed nucleus and has access to ribosomes in the cytoplasm[33].

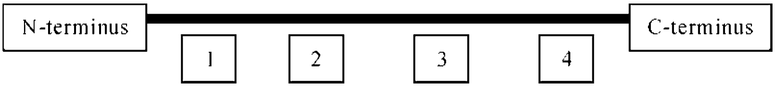
The following steps occur, in order, for transcription initiation: RNA polymerase (RNAP) binds to one of several specificity factors; σ, to form a holoenzyme. In this form, it can recognize and bind to specific promoter regions in the DNA. At this stage, the DNA is double-stranded ("closed"). This holoenzyme/wound-DNA structure is referred to as the closed complex.

The steps of elongation are as follows: (1) The DNA is unwound and becomes single-stranded ("open") in the vicinity of the initiation site (defined as +1). This holoenzyme/unwound-DNA structure is called the open complex. (2) The RNA polymerase transcribes the DNA, but produces about 10 abortive (short, non-productive) transcripts which are unable to leave the RNA polymerase because the exit channel is blocked by the σ-factor. (3) The σ-factor eventually dissociates from the holoenzyme, and elongation proceeds. Promoters can differ in "strength;" that is, how actively they promote transcription of their adjacent DNA sequence. Promoter strength is in many (but not all) cases, a matter of how tightly RNA polymerase and its associated accessory proteins bind to their respective DNA sequences. The more similar the sequences are to a consensus sequence, the stronger the binding is. Additional transcription regulation comes from transcription factors that can affect the stability of the holoenzyme structure at initiation. Most transcripts originate using adenosine-5'-triphosphate (ATP) and to a lesser extent, guanosine-5'-triphosphate (GTP) (purine nucleoside triphosphates) at the +1 site. Uridine-5'-triphosphate (UTP) and cytidine-5'-triphosphate (CTP) (pyrimidine nucleoside triphosphates) are not favored at the initiation site.

Two termination mechanisms are well known: 1) Intrinsic termination (also called Rho-independent transcription termination) involves terminator sequences within the RNA that signal the RNA polymerase to stop. The terminator sequence is usually a palindromic sequence that forms a stem-loop hairpin structure that leads to the dissociation of the RNAP from the DNA template. 2) Rho-dependent termination uses a termination factor called ρ factor (rho factor) which is a protein to stop RNA synthesis at specific sites. This protein binds at a rho ultilization site on the nascent RNA strand and runs along the mRNA towards the RNAP. A stem loop structure upstream of the terminator region pauses the RNAP; when the ρ-factor reaches the RNAP, it causes RNAP to dissociate from the DNA, terminating transcription.

Other termination mechanisms include where RNAP comes across a region with repetitious thymidine residues in the DNA template, or where a GC-rich inverted repeat followed by 4 A residues. The inverted repeat forms a stable stem loop structure in the RNA, which causes the RNA to dissociate from the DNA template. The –35 region and the –10 ("Pribnow box") region comprise the basic prokaryotic promoter, and |T| stands for the terminator. The DNA on the template strand between the +1 site and the terminator is transcribed into RNA, which is then translated into protein[34].

A sigma factor (σ factor) is a prokaryotic transcription initiation factor that enables specific binding of RNA polymerase to gene promoters. Different sigma factors are activated in response to different environmental conditions. Every molecule of RNA polymerase contains exactly one sigma factor subunit, which in the model bacterium, *Escherichia coli,* is one of those listed below. *E. coli* has at least eight sigma factors; the number of sigma factors varies between bacterial species. Sigma factors are distinguished by their characteristic molecular weights. For example, σ^{70} refers to the sigma factor with a molecular weight of 70 kDa[35, 36]. Sigma factors have four main regions that are generally conserved:



The regions are further subdivided (*e.g.* 2 includes 2.1, 2.2, *etc.*).

(1) Region 1 is found only in "primary sigma factors" (RpoD and RpoS in *E. coli*). It is involved in ensuring the sigma factor will only bind the promoter when it is complexed with the RNA polymerase.

(2) Region 2.4 recognizes and binds to the Pribnow box.

(3) Region 4.2 recognizes and binds to the –35 promoter site.

The exception to this organization is in σ^{54}-type sigma factors. Proteins homologous to σ^{54} /RpoN are functional sigma factors, but they have significantly different primary amino acid sequences.

Different sigma factors are activated under different environmental conditions. These specialized sigma factors bind the promoters of genes appropriate to the environmental conditions, increasing the transcription of those genes. Sigma factors in *E.coli* are:

(1) σ^{70}(RpoD)—the "housekeeping" sigma factor, transcribes most genes in growing cells.

(2) σ^{54}(RpoN)—the nitrogen-limitation sigma factor.

(3) σ^{38}(RpoS)—the starvation/stationary phase sigma factor.

(4) σ^{32}(RpoH)—the heat shock sigma factor; it is turned on when exposed to heat.

(5) σ^{28}(RpoF)—the flagellar sigma factor.

(6) σ^{24}(RpoE)—the extracytoplasmic/extreme heat stress sigma factor.

(7) σ^{19}(FecI)—the ferric citrate sigma factor; regulates the fec gene for iron transport.

There are also anti-sigma factors that inhibit the function of sigma factors.

The core RNA polymerase (consists of 2 alpha (α), 1 beta (β), 1 beta-prime (β'), and 1 omega (ω) subunits) binds a sigma factor to form a complex called the RNA polymerase holoenzyme. It was previously believed that the RNA polymerase holoenzyme initiates transcription, while the core RNA polymerase alone synthesizes RNA. Thus, the accepted view was that the sigma factor must dissociate upon transition from transcription initiation to transcription elongation (this transition is called "promoter escape"). This view was based on analysis of purified complexes of RNA polymerase stalled at initiation and at elongation. Finally, structural models of RNA polymerase complexes predict that as the growing RNA product becomes longer than ~10 nucleotides, sigma must be "pushed out" of the holoenzyme since there is a steric clash between RNA and a sigma domain. However, a recent study[36] has shown that σ^{70} remains attached to the complex with the core RNA polymerase, at least during early elongation. Indeed, the phenomenon of promoter-proximal stalling suggests that sigma may play a role during early elongation. All studies are consistent with the assumption that promoter escape reduces the lifetime of the sigma-core interaction at initiation (too long to be measured in a typical biochemical experiment) to a shorter, measurable lifetime upon transition to elongation.

An Rho factor acts on an RNA substrate. Rho's key function is its helicase activity, for which energy is provided by RNA-dependent ATP hydrolysis. The initial binding site for Rho is an extended (~70 nucleotides, sometimes 80–100 nucleotides) single-stranded region, rich in cytosine and poor in guanine in the RNA being synthesized, upstream of the actual terminator sequence. Several Rho binding sequences have been discovered. No consensus is found among these, but the different sequences each seem specific, as small mutations in the sequence disrupt its function. Rho binds to RNA and then uses its ATPase activity to provide the energy to translocate along the RNA until it reaches the RNA-DNA helical region, where it unwinds the hybrid duplex structure. It is thought that the RNA polymerase pauses at the termination sequence, which allows the Rho factor to catch up. However, the kinetics are quite complex and have not been fully analyzed or verified[37].

In short, the Rho factor acts as an ATP-dependent unwinding enzyme, moving along the newly forming RNA molecule towards its 3' end and unwinding it from the DNA template as it proceeds.

A nonsense mutation in one gene of an operon prevents the translation of subsequent genes in the unit. This effect is called “polarity”. A common cause is the absence of the mRNA corresponding to the subsequent (distal) parts of the unit. Suppose that there are Rho-dependent terminators within the transcription unit before the terminator is used. Normally, these early terminators are not used, because the ribosome prevents Rho from reaching RNA polymerase. A nonsense mutation releases the ribosome, so that Rho is free to attach to and/or move along the RNA, enabling it to act on RNA polymerase at the terminator. As a result, the enzyme is released and the distal regions of the transcription unit are never transcribed.

The Pribnow box (also known as the Pribnow-Schaller box) is the sequence TATAAT of six nucleotides (thymine-adenine-thymine-*etc*.) that is an essential part of a promoter site on DNA for transcription to occur in bacteria[38, 39]. It is an idealized or consensus sequence—it shows the most frequently occurring base at each position in a large number of promoters analyzed; individual promoters often vary from the consensus at one or more positions. It is also commonly called the –10 sequence, because it is centered roughly 10 base pairs upstream from the site of initiation of transcription. The Pribnow box has a function similar to the TATA box which occurs in promoters in eukaryotes and archaea: it is recognized and bound by a subunit of RNA polymerase during initiation of transcription. This region of the DNA is also the first place where base pairs separate during prokaryotic transcription to allow access to the template strand. The AT-richness is important to allow this separation, since adenine and thymine pair together with only two hydrogen bonds (as opposed to three as with guanine and cytosine). They are easier to break apart.

16.5 Transcription and Translation in Eukaryotes

16.5.1 Translation in Eukaryotes

Eukaryotic translation is the process by which mRNA is translated into proteins in eukaryotes. It consists of initiation, elongation and termination[40].

Initiation of translation usually involves the interaction of certain key proteins with a special tag bound to the 5'-end of an mRNA molecule, the 5' cap. The protein factors bind the small ribosomal subunit (also referred to as the ^{40}S subunit), and these initiation factors hold the mRNA in place. The eukaryotic Initiation Factor 3 (eIF3) is associated with the small ribosomal subunit, and plays a role in keeping the large ribosomal subunit from prematurely binding. eIF3 also interacts with the eIF4F complex which consists of three other initiation factors: eIF4A, eIF4E and eIF4G. eIF4G is a scaffolding protein which directly associates with both eIF3 and the other two components. eIF4E is the cap-binding protein. It is the rate-limiting step of cap-dependent initiation, and is often cleaved from the complex by some viral proteases to limit the cell's ability to translate its own transcripts. This is a method of hijacking the host machinery in favor of the viral (cap-independent) messages. eIF4A is an ATP-dependent RNA helicase, which aids the ribosome in resolving certain secondary structures formed by the mRNA transcript. There is another protein associated with the eIF4F complex called the Poly(A)-binding protein (PABP), which binds the poly-A tail of most eukaryotic mRNA molecules. This protein has been implicated in playing a role in circularization of mRNA during translation.

This pre-initiation complex (^{43}S subunit, or the ^{40}S and mRNA) accompanied by the protein factors moving along the mRNA chain towards its 3'-end, scan for the 'start' codon (typically AUG) on the mRNA, which indicates where the mRNA will begin coding for the protein. In eukaryotes and archaea, the amino acid encoded by the start codon is methionine. The initiator tRNA charged with Met forms part of the ribosomal complex and thus all proteins start with this amino acid (unless it is cleaved away by a protease in subsequent modifications). The Met-charged initiator tRNA is brought to the P-site of the small ribosomal subunit by eukaryotic Initiation Factor 2 (eIF2). It hydrolyzes GTP and signals for the dissociation of several factors from the small ribosomal subunit which results in the association of the large subunit (or the ^{60}S subunit). The complete ribosome (^{80}S) then commences translation elongation during which the sequence between the 'start' and 'stop' codons is translated from mRNA into an amino acid sequence—thus a protein is synthesized.

The best studied example of the cap-independent mode of translation initiation in eukaryotes is the Internal Ribosome Entry Site (IRES) approach. Fig 16.4 gives the process of initiation of translation in eukaryotes. What differentiates cap-independent translation from cap-dependent translation is that cap-independent translation does not require the ribosome to start scanning from the 5' end of the mRNA cap until the start codon. The ribosome can be trafficked to the start site by ITAFs (IRES trans-acting factors) bypassing the need to scan from the 5' end of the untranslated region of the mRNA. This method of translation has been recently discovered, and has found to be important in conditions that require the translation of specific mRNAs, despite cellular stress or the inability to translate most mRNAs. Examples include factors responding to apoptosis and stress-induced responses[41].

Elongation is dependent on eukaryotic elongation factors[40]. At the end of the initiation step, the mRNA is positioned so that the next codon can be translated during the elongation stage of protein synthesis. The initiator tRNA occupies the P site in the ribosome, and the A site is ready to receive an aminoacyl-tRNA. During chain elongation, each additional amino acid is added to the nascent polypeptide chain in a three-step microcycle. The steps in this microcycle are (1) positioning the correct aminoacyl-tRNA in the A site of the ribosome, (2) forming the peptide bond and (3) shifting the mRNA by one codon relative to the ribosome. The translation machinery works relatively slowly compared to the enzyme systems that catalyze DNA replication. Proteins in procaryotes are synthesized at a rate of only 18 amino acid residues per second, whereas bacterial replisomes synthesize DNA at a rate of 1,000 nucleotides per second. This difference in rate reflects, in part, the difference between polymerizing four types of nucleotides to make nucleic acids and polymerizing 20 types of amino acids to make proteins. Testing and rejecting incorrect aminoacyl-tRNA molecules takes time and slows protein synthesis. The rate of transcription in prokaryotes is approximately 55 nucleotides per second, which corresponds to about 18 codons per second, or the same rate at which the mRNA is translated. In bacteria, translation initiation occurs as soon as the 5' end of an mRNA is synthesized, and translation and transcription are coupled. This tight coupling is not possible in eukaryotes because transcription and translation are carried out in separate compartments of the cell (the nucleus and cytoplasm). Eukaryotic mRNA precursors must be processeed in the nucleus (e.g. capping, polyadenylation, and splicing) before they are exported to the cytoplasm for translation.

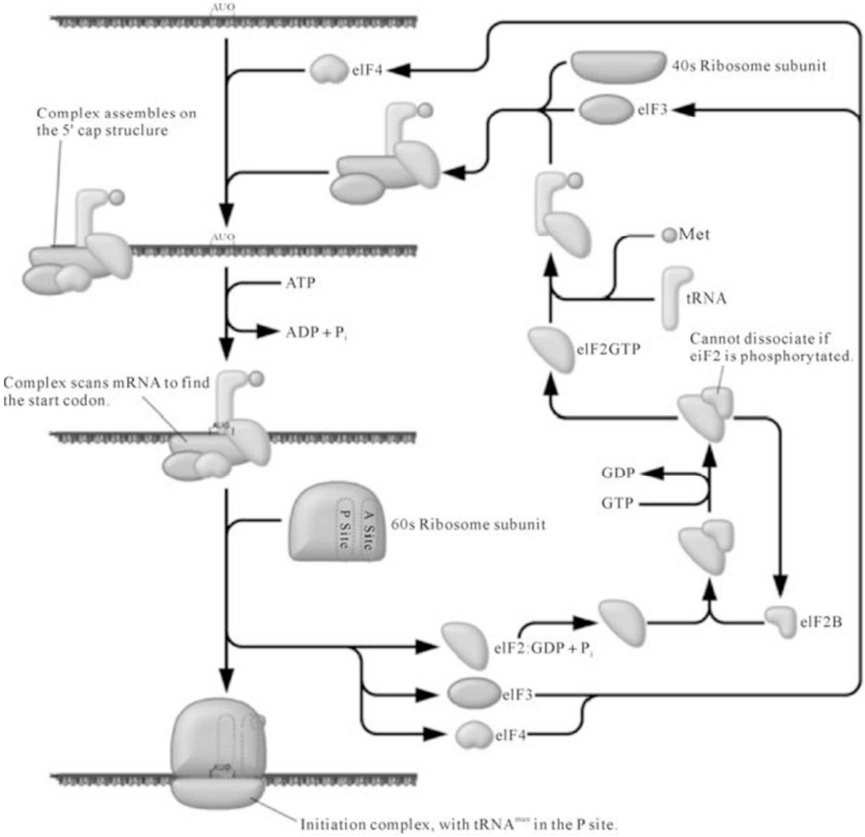

AUG
eIF4
40s Ribosome subunit
eIF3
Complex assembles on the 5' cap strucLure
AUG
Met
ATP
ADP + P$_i$
tRNA
eIF2GTP
Cannot dissociate if eiF2 is phosphorytated.
Complex scans mRNA to find the start codon.
GDP
GTP
A Site
P Site
60s Ribosome subunit
eIF2B
eIF2:GDP + P$_i$
eIF3
eIF4
Initiation complex, with tRNA^{met} in the P site.

Fig. 16.4 The process of initiation of translation in eukaryotes[42] (with the permission of Wikimedia)

Eukaryotic elongation factors are very similar to those in prokaryotes. Elongation in eukaryotes is carried out with two elongation factors: eEF-1 and eEF-2. The first is eEF-1, whose α and $\beta\gamma$ subunits act as counterparts to EF-Tu and EF-Ts respectively. The second is eEF-2, the counterpart to prokaryotic EF-G[43]

Termination of elongation is dependent on eukaryotic release factors. The process is similar to that of prokaryotic termination.

In eukaryotes, there is only one release factor, eRF, which recognizes all three stop codons in place of RF1, RF2, or RF3. However, the overall process of termination is similar to that of prokaryotes[44].

16.5.2 Eukaryotic Initiation Factor

Eukaryotic initiation factors are proteins used in eukaryotic translation. There exist many more eukaryotic initiation factors (eIF) than prokaryotic initiation factors due to greater biological complexity. Processes eIF is involved in: formation of initiation complexes with 5' mRNA and complexing with Met-tRNAi, binding mRNA-factor to Met-tRNAi, scanning mRNA for the initiator codon AUG, locating the binding site of initator tRNA to the AUG start site, and joining of the 60S subunit to create the 80S subunit[45].

The protein RLI is known to have an essential, probably catalytic role in the formation of initiation complexes as well.

The eIF4 initiation factors include eIF4A2, eIF4A3, eIF4B, eIF4E, and eIF4G. eIF4F is often referred to the complex of eIF4A, eIF4E, and eIF4G.

eIF4G is a scaffolding protein that interacts with eIF3 (see below), as well as the other members of the eIF4F complex. eIF4A-an RNA helicase-is important for resolving any secondary structures the mRNA transcript may form. eIF4E binds the 5' cap of the mRNA and the rate-limiting step for cap-dependent translation.

eIF4B contains two RNA binding domains-one non-specifically interacts with mRNA, while the second specifically binds the 18S portion of the small ribosomal subunit. It acts as an anchor, as well as a critical co-factor for eIF4A. It is a substrate of S6K and when phosphorylated, it promotes the formation of the pre-initiation complex.

eIF1, eIF1A, and eIF3 all bind to the ribosome subunit-mRNA complex. They have been implicated in preventing the large ribosomal subunit from binding the small subunit before it is ready to commence elongation.

In mammals, eIF3 is the largest scaffolding initiation factor made up of 13 subunits (a-m). It is roughly ~750 kDa and it controls the assembly of the 40S ribosomal subunit on mRNA that has a 5' cap or an IRES (Internal Ribosomal Entry Site). eIF3 uses the eIF4F complex or IRES from viruses to position the mRNA strand near the exit site of the 40S ribosome subunit, thus promoting the assembly of the pre-initiation complex.

In many cancers eIF3 is overexpressed. Under serum deprived conditions (inactive state), eIF3 is bound to S6K1. On stimulation either by mitogens, growth factors or drugs, the mTOR/Raptor complex gets activated and in turn binds and phosphorylates S6K1 on T389 (linker region) causing a conformational change that causes the kinase S6K1 to dissociate from eIF3. The T389 phosphorylated S6k1 is then further phosphorylated by PDK1 on T229. This second phosphorylation fully activates the S6K1 kinase which can then phosphorylate eIF4B, S6 and other protein targets.

eIF2 is a GTP-binding protein responsible for bringing the initiator tRNA to the P-site of the pre-initiation complex. It has specificity for the methionine-charged initiator tRNA, which is distinct from other methionine-charged tRNAs specific for elongation of the polypeptide chain. Once it has placed the initiator tRNA on the AUG start codon in the P-site, it hydrolyzes GTP into GDP, and dissociates. This hydrolysis, also signals for the dissociation of eIF3, eIF1, and eIF1A, and allows the large subunit to bind. This signals the beginning of elongation.

eIF2 has three subunits, eIF2–α, β, and γ. The former is of particular importance for cells which may need to turn off protein synthesis globally. When phosphorylated, it sequesters eIF2B (not to be confused with beta), a GEF. Without this GEF, GDP cannot be exchanged for GTP and translation is repressed.

eIF2α-induced translation repression occurs in reticulocytes when starved for iron. Additionally, protein kinase R (PKR) phosphorylates eIF2α when dsRNA is detected in many multi-cellular organisms, leading to cell death.

eIF5A is a GTPase activating protein, which helps the large ribosomal subunit associate with the small subunit. It is required for GTP-hydrolysis by eIF2 and contains the unusual amino acid hypusine[46].

eIF5B is a GTPase, and is involved in assembly of the full ribosome (which requires GTP hydrolysis).

16.5.3 Eukaryotic RNA Polymerases

RNA polymerase (RNAP or RNApol) is an enzyme that produces RNA. In cells, RNAP is needed for constructing RNA chains from DNA genes as templates, a process called transcription. RNA polymerase enzymes are essential to life and are found in all organisms and many viruses. In chemical terms, RNAP is a nucleotidyl transferase that polymerizes ribonucleotides at the 3' end of an RNA transcript.

RNAP was discovered independently by Sam Weiss and Jerard Hurwitz in 1960[47]. By this time, the 1959 Nobel Prize in Medicine had been awarded to Severo Ochoa and Arthur Kornberg for the discovery of what was believed to be RNAP but instead it turned out to be polynucleotide phosphorylase. The 2006 Nobel Prize in Chemistry was awarded to Roger Kornberg for creating detailed molecular images of RNA polymerase during various stages of the transcription process.

Control of the process of gene transcription affects patterns of gene expression and thereby allows a cell to adapt to a changing environment, perform specialized roles within an organism, and maintain basic metabolic processes necessary for survival. Therefore, it is hardly surprising that the activity of RNAP is both complex and highly regulated. In *Escherichia coli* bacteria, more than 100 transcription factors have been identified which modify the activity of RNAP[48].

RNAP can initiate transcription at specific DNA sequences known as promoters. It then produces an RNA chain which is complementary to the template DNA strand. The process of adding nucleotides to the RNA strand is known as elongation; in eukaryotes, RNAP can build chains as long as 2.4 million nucleosides (the full length of the dystrophin gene). RNAP will preferentially release its RNA transcript at specific DNA sequences encoded at the end of genes known as terminators. Products of RNAP include[49]:

(1) mRNA—template for the synthesis of proteins by ribosomes.

(2) Non-coding RNA or "RNA genes"—a broad class of genes that encode RNA that is not translated into protein. The most prominent examples of RNA genes are transfer RNA (tRNA) and ribosomal RNA (rRNA), both of which are involved in the process of translation. However, since the late 1990s, many new RNA genes have been found, and thus RNA genes may play a much more significant role than previously thought. Transfer RNA (tRNA)—transfers specific amino acids to growing polypeptide chains at the ribosomal site of protein synthesis during translation. Ribosomal RNA (rRNA)—a component of ribosomes. Micro RNA—regulates gene activity. Catalytic RNA (Ribozymes)—are enzymatically active RNA molecules.

RNAP accomplishes de novo synthesis. It is able to do this because specific interactions with the initiating nucleotide hold RNAP rigidly in place, facilitating chemical attack on the incoming nucleotide. Such specific interactions explain why RNAP prefers to start transcripts with ATP (followed by GTP, UTP, and then CTP). In contrast to DNA polymerase, RNAP includes helicase activity; therefore, no separate enzyme is needed to unwind DNA.

Eukaryotes have several types of RNAP, characterized by the type of RNA they synthesize:

(1) RNA polymerase I synthesizes a pre-rRNA 45S, which matures into 28S, 18S and 5.8S rRNAs which will form the major RNA sections of the ribosome[50].

(2) RNA polymerase II synthesizes precursors of mRNAs and most snRNA and microRNAs[51]. This is the most studied type, and due to the high level of control required over transcription, a range of transcription factors are required for its binding to promoters.

(3) RNA polymerase III synthesizes tRNAs, rRNA 5S and other small RNAs found in the nucleus and cytosol[52].

(4) RNA polymerase IV synthesizes siRNA in plants[53].

There are other RNA polymerase types in mitochondria and chloroplasts. There are RNA-dependent RNA polymerases involved in RNA interference[54].

16.5.4 Chromatin and Its Effects on Transcription

Chromatin is the complex combination of DNA, RNA, and protein that makes up chromosomes. It is found inside the nuclei of eukaryotic cells and within the nucleoid in prokaryotic cells. It is divided between heterochromatin (condensed) and euchromatin (extended) forms[55]. The major components of chromatin are DNA and histone proteins, although many other chromosomal proteins have prominent roles too. The functions of chromatin are to package DNA into a smaller volume to fit in the cell, to strengthen the DNA to allow mitosis and meiosis, and to serve as a mechanism to control expression and DNA replication. Chromatin contains genetic material-instructions to direct cell functions. Changes in chromatin structure are affected by chemical modifications of histone proteins such as methylation (DNA and proteins) and acetylation (proteins), and by non-histone, DNA-binding proteins.

Simplistically, there are seven levels of chromatin organization (Fig. 16.5):

(1) DNA wrapping around nucleosomes—the "beads on a string" structure.

(2) A 30 nm condensed chromatin fiber consisting of nucleosome arrays in their most compact form.

(3) Higher level DNA packaging in the metaphase chromosome.

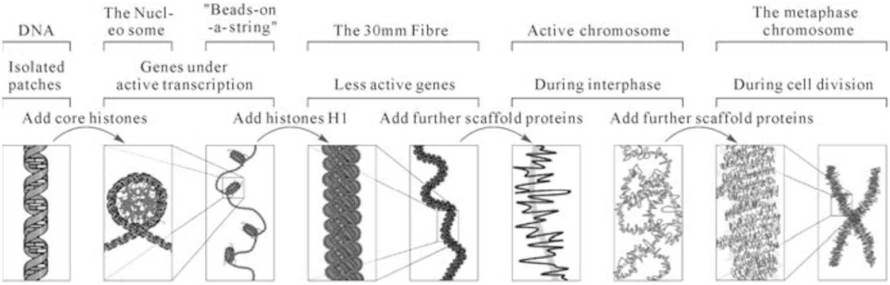


Fig. 16.5 The major structures in RNA compaction: DNA, the nucleosome, the 10nm "beads-on-a-string" fiber, the 30 nm fiber and the metaphase chromosome[56] (with the pemission of Wikimedia)

These structures do not occur in all prokaryotic cells. Examples of cells with more extreme packaging are spermatozoa and avian red blood cells.

During spermiogenesis, the spermatid's chromatin is remodelled into a more widely-spaced package, with almost a crystal-like structure. This process is associated with the cessation of transcription and involves nuclear protein exchange. The histones are mostly displaced and replaced by protamines (small, arginine-rich proteins).

It should also be noted that during mitosis, while most of the chromatin is tightly compacted, there are small regions that are not as tightly compacted. These regions often correspond to promoter regions of genes that were active in that cell type prior to entry into cromitosis. The lack of compaction of these regions is called bookmarking, which is an epigenetic mechanism believed to be important for transmitting to daughter cells the “memory” of which genes were active prior to entry into mitosis. This bookmarking mechanism is needed to help transmit this memory because transcription ceases during mitosis. It is found in a plant cell.

The structure of chromatin during interphase is optimized to allow easy access of transcription and DNA repair factors to the DNA while compacting the DNA into the nucleus. The structure varies depending on the access required to the DNA. Genes that require regular access by RNA polymerase require the looser structure provided by euchromatin.

Chromatin undergoes various forms of change in its structure. Histone proteins, the foundation blocks of chromatin, are modified by various post-translational modification to alter DNA packing. Acetylation results in the loosening of chromatin and lends itself to replication and transcription. When methylated, they hold DNA together strongly and restrict access to various enzymes. A recent study showed that there is a bivalent structure present in the chromatin: methylated lysine residues at location 4 and 27 on histone 3. It is believed that this may be involved in development; there is more methylation of lysine 27 in embryonic cells than in differentiated cells, whereas lysine 4 methylation positively regulates transcription by recruiting nucleosome remodeling enzymes and histone acetylases[57]. Polycomb-group proteins play a role in regulating genes through modulation of chromatin structure[58].

The vast majority of DNA within the cell is the normal DNA structure. However, in nature DNA can form three structures, A-, B- and Z-DNA. A and B chromosomes are very similar, forming right-handed helices, while Z-DNA is a more unusual left-handed helix with a zig-zag phosphate backbone. Z-DNA is thought to play a specific role in chromatin structure and transcription because of the properties of the junction between B- and Z-DNA. At the junction of B- and Z-DNA, one pair of bases is flipped out from normal bonding. These play a dual role of a site of recognition by many proteins and as a sink for torsional stress from RNA polymerase or nucleosome binding.

The basic repeat element of chromatin is the nucleosome, interconnected by sections of linker DNA, a far shorter arrangement than pure DNA in solution.

In addition to the core histones, there is the linker histone, H1, which contacts the exit/entry of the DNA strand on the nucleosome. The nucleosome, together with histone H1, is known as a chromatosome. Nucleosomes, connected by about 20 to 60 base pairs of linker DNA, form an approximately 10 nm “beads-on-a-string” fiber (Fig. 16.6).

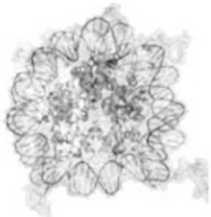


Fig. 16.6 A cartoon representation of the nucleosome structure[56] (with the permission of Wikimedia)

The nucleosomes bind DNA non-specifically, as required by their function in general DNA packaging. There is, however, some preference in the sequences the nucleosomes will bind. This is largely through the properties of DNA; adenosine and thymine are more favorably compressed into the inner minor grooves. This means nucleosomes bind preferentially at one position every 10 base pairs-where the DNA is rotated to maximize the number of A and T bases which will lie in the inner minor groove.

With addition to H1, the "beads-on-a-string" structure in turn coils into a 30 nm diameter helical structure known as the 30nm fiber or filament. The precise structure of the chromatin fiber in the cell is not known in detail, and there is still some debate over this.

This level of chromatin structure is thought to be in the form of euchromatin, which contains actively transcribed genes. EM studies have demonstrated that the 30 nm fiber is highly dynamic such that it unfolds into a 10 nm fiber ("beads-on-a-string") structure when transversed by an RNA polymerase engaged in transcription.

The existing models commonly accept that the nucleosomes lie perpendicular to the axis of the fiber, with linker histones arranged internally. A stable 30 nm fiber relies on the regular positioning of nucleosomes along DNA. Linker DNA is relatively resistant to bending and rotation. This makes the length of linker DNA critical to the stability of the fiber, requiring nucleosomes to be separated by lengths that permit rotation and folding into the required orientation without excessive stress to the DNA. In this view, the different length of the linker DNA should produce different folding topologies of the chromatin fiber. Recent theoretical work, based on electron-microscopy images of reconstituted fibers[59] support this view[60].

The layout of the genome within the nucleus is not random-specific regions of the genome are always found in certain areas. Specific regions of the chromatin are thought to be bound to the nuclear membrane, while other regions are bound together by protein complexes. The layout of this is not, however, well characterized apart from the compaction of one of the two X chromosomes into the Barr body in mammalian females. This serves the role of permanently deactivating these genes, which prevents females from getting a 'double dose' relative to the males.

The metaphase structure of chromatin differs vastly to that of interphase. It is optimized for physical strength and manageability, forming the classic chromosome structure seen in karyotypes. The structure of the condensed chromosome is thought to have loops of 30 nm fibers in a central scaffold of proteins. It is, however, not well characterized.

The physical strength of chromatin is vital for this stage of division to prevent shear damage to the DNA as the daughter chromosomes are separated. To maximize strength, the composition of the chromatin changes as it approaches the centromere, primarily through alternative histone H1 anologues.

The proteins that are found associated with isolated chromatin fall into several functional categories: chromatin-bound enzymes, high mobility group (HMG) proteins, transcription factors, scaffold proteins, transition proteins (testes specific proteins), and protamines (present in mature sperm).

Enzymes associated with chromatin are those involved in DNA transcription, replication and repair, and in post-translational modification of histones. They include various types of nucleases and proteases. Scaffold proteins encompass chromatin proteins such as insulators, domain boundary factors and cellular memory modules (CMMs)[61].

16.6 Post-Transcriptional Events

16.6.1 Post-Transcriptional Event: Splicing

In molecular biology, splicing is a modification of RNA after transcription, in which introns are removed and exons are joined (Fig. 16.7). This is needed for the typical eukaryotic mRNA before it can be used to produce a correct protein through translation. For many eukaryotic introns, splicing is done in a series of reactions which are catalyzed by the spliceosome, a complex of small nuclear ribonucleoproteins (snRNPs), but there are also self-splicing introns[62, 63].

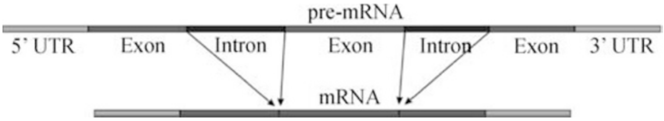


Fig. 16.7 Simple illustration of exons and introns in pre-mRNA and the formation of mature mRNA by splicing. The UTRs are non-coding parts of exons at the ends of the mRNA[64] (with the permission of Wikimedia)

Several methods of RNA splicing occur in nature: the type of splicing depends on the structure of the spliced intron and the catalysts required for splicing to occur.

Spliceosomal introns often reside in eukaryotic protein-coding genes. Within the intron, a 3' splice site, 5' splice site, and branch site are required for splicing. Splicing is catalyzed by the spliceosome which is a large RNA-protein complex composed of five small nuclear ribonucleoproteins (snRNPs, pronounced ‘snurps’). The RNA components of snRNPs interact with the intron and may be involved in catalysis. Two types of spliceosomes have been identified (the major and minor) which contain different snRNPs.

The major spliceosome splices introns containing GU at the 5' splice site and AG at the 3' splice site. It is composed of the U1, U2, U4, U5, and U6 snRNPs and is active in the nucleus.

E Complex-U1 binds to the GU sequence at the 5' splice site, along with accessory proteins/enzymes ASF/SF2, U2AF (binds at the Py-AG site), and SF1/BBP (BBP=Branch Binding Protein);

A Complex-U2 binds to the branch site, and ATP is hydrolyzed;

B1 Complex-U5/U4/U6 trimer binds, and the U5 binds exons at the 5' site, with U6 binding to U2;

B2 Complex-U1 is released, while U5 shifts from exon to intron and U6 binds at the 5' splice site;

C1 Complex-U4 is released, U6/U2 catalyzes trans-esterification, U5 binds exon at 3' splice site, and the 5' site is cleaved, resulting in the formation of the lariat;

C2 Complex-U2/U5/U6 remains bound to the lariat, the 3' site is cleaved and exons are ligated using ATP hydrolysis. The spliced RNA is released and the lariat debranches.

This type of splicing is termed canonical splicing or termed the lariat pathway, which accounts for more than 99% of splicing. In contrast, when the intronic flanking sequences do not follow the GU-AG rule, non-canonical splicing is said to occur[65].

The minor spliceosome is very similar to the major spliceosome; however, it splices out rare introns with different splice site sequences. While the minor and major spliceosomes contain the same U5 snRNP, the minor spliceosome has different, but functionally analogous snRNPs for U1, U2, U4, and U6, which are respectively called U11, U12, U4atac, and U6atac[66]. Like the major spliceosome, it is only found in the nucleus[67].

Trans-splicing is a form of splicing that joins two exons that are not within the same RNA transcript[68].

Self-splicing occurs for rare introns that form a ribozyme, performing the functions of the spliceosome by RNA alone. There are three kinds of self-splicing introns, Group I, Group II and Group III. Group I and II introns perform splicing similar to the spliceosome without requiring any protein. This similarity suggests that Group I and II introns may be evolutionarily related to the spliceosome. Self-splicing may also be very ancient, and may have existed in an RNA world that was present before protein. Although the two splicing mechanisms described below do not require any proteins to occur, 5 additional RNA molecules and over 50 proteins are used to hydrolyze many ATP molecules. The splicing mechanisms use ATP in order to accurately splice mRNAs. If the cell was to not use any ATPs, the process would be highly inaccurate and many mistakes would occur.

Two trans-esterifications characterize the mechanism in which group I introns are spliced:

(1) 3' OH of a free guanine nucleoside (or one located in the intron) or a nucleotide cofactor (GMP, GDP, or GTP) attacks phosphate at the 5' splice site.

(2) 3' OH of the 5' exon becomes a nucleophile and the second trans-esterification results in the joining of the two exons.

The mechanism in which group II introns are spliced (two trans-esterification reactions such as for group I introns) is as follows:

(1) The 2' OH of a specific adenosine in the intron attacks the 5' splice site, thereby forming the lariat.

(2) The 3' OH of the 5' exon triggers the second trans-esterification at the 3' splice site thereby joining the exons together.

tRNA (also tRNA-like) splicing is another rare form of splicing that usually occurs in tRNA. The splicing reaction involves a different biochemistry than the spliceosomal and self-splicing pathways. Ribonucleases cleave the RNA and ligases join the exons together.

Splicing occurs in all kingdoms or domains of life; however, the extent and types of splicing can be very different between the major divisions (Table 16.1). Eukaryotes splice many protein-coding mRNAs and some non-coding RNAs. Prokaryotes, on the other hand, splice rarely and mostly use non-coding RNAs. Another important difference between these two groups of organisms is that prokaryotes completely lack the correct spliceosomal pathway.

Because spliceosomal introns are not conserved in all species, there is debate concerning when spliceosomal splicing evolved. Two models have been proposed: the intron late and intron early models.

Table 16.1 Splicing diversity

	Eukaryotes	Prokaryotes
Spliceosomal	+	-
Self-splicing	+	+
tRNA	+	+

Spliceosomal splicing and self-splicing involves a two-step biochemical process. Both steps involve trans-esterification reactions that occur between RNA nucleotides. tRNA splicing, however, is an exception and does not occur by trans-esterification. Spliceosomal and self-splicing trans-esterification reactions occur via two sequential trans-esterification reactions. First, the 2' OH of a specific branch-point nucleotide within the intron is defined during spliceosome assembly and performs a nucleophilic attack on the first nucleotide of the intron at the 5' splice site forming the lariat intermediate. Second, the 3' OH of the released 5' exon then performs a nucleophilic attack on the last nucleotide of the intron at the 3' splice site thus joining the exons and releasing the intron lariat.

In many cases, the splicing process can create a range of unique proteins by varying the exon composition of the same mRNA. This phenomenon is then called alternative splicing. Alternative splicing can occur in many ways. Exons can be extended or skipped, or introns can be retained.

Splicing events can be experimentally altered[69] by binding steric-blocking antisense oligos, such as Morpholinos or Peptide nucleic acids to snRNP binding sites, or binding to the branchpoint nucleotide that closes the lariat[70] or to the splice-regulatory element binding sites[71].

Splicing errors may occur such as:

(1) Mutation of a splice site resulting in loss of function of that site. Results in exposure of a premature stop codon, loss of an exon, or inclusion of an intron.

(2) Mutation of a splice site reducing specificity. May result in variation in the splice location, causing insertion or deletion of amino acids, or most likely, a loss of the reading frame.

(3) Transposition of a splice site, leading to inclusion or exclusion of more RNA than expected which results in longer or shorter exons.

Many splicing errors are safeguarded by a cellular quality control mechanism termed nonsense-mediated mRNA decay [NMD][72].

Not only pre-mRNA but also proteins can undergo splicing. Although the biomolecular mechanisms are different, the principle is the same, so that parts of the protein, called inteins instead of introns, are removed. The remaining parts, called exteins instead of exons, are fused together. Protein splicing has been observed in all sorts of organisms, including bacteria, archaea, plants, yeast and humans[73].

16.6.2 Post-Transcriptional Mbdification

Post-transcriptional modification is a process of molecular biology by which, in eukaryotic cells, primary transcript RNA is converted into mature RNA. A notable example is the conversion of precursor mRNA into mature mRNA, which includes splicing and occurs prior to protein synthesis. This process is vital for the correct translation of genomes of eukaryotes as the primary human RNA transcript that is produced as a result of transcription containing both exons, which are coding sections of the primary RNA transcript and introns, which are the noncoding sections of the primary RNA transcript[74-76].

The pre-mRNA molecule undergoes three main modifications. These modifications are 5' capping, 3' polyadenylation, and RNA splicing which occur in the cell nucleus before the RNA is translated.

Capping of the pre-mRNA involves the addition of 7-methylguanosine (m7G) to the 5' end. In order to achieve this, the terminal 5' phosphate requires removal, which is done with the aid of a phosphatase enzyme. The enzyme guanosyl transferase then catalyses the reaction which produces the diphosphate 5' end. The diphosphate 5' prime end then attacks the α phosphorus atom of a GTP molecule in order to add the guanine residue in a 5'5' triphosphate link. The enzyme S-adenosyl methionine then methylates the guanine ring at the N-7 position. This type of cap, with just the (m7G) in position is called a cap 0 structure. The ribose of the adjacent nucleotide may also be methylated to give a cap 1. Methylation of nucleotides downstream of the RNA molecule produce cap 2, cap 3 structures and so on. In these cases, the methyl groups are added to the 2' OH groups of the ribose sugar. The cap protects the 5' end of the primary RNA transcript from attack by ribonucleases that have a specificity to the 3'5' phosphodiester bonds[75].

The pre-mRNA processing at the 3' end of the RNA molecule involves cleavage of its 3' end and then the addition of about 200 adenine residues to form a poly(A) tail (Fig. 16.8). The cleavage and adenylation reactions occur if a polyadenylation signal sequence (5'- AAUAAA-3') is located near the 3' end of the pre-mRNA molecule, which is followed by another sequence, which is usually (5'-CA-3'). The second signal is the site of cleavage. A GU-rich sequence is also usually present further downstream on the pre-mRNA molecule. After the synthesis of the sequence elements, two multi-subunit proteins called cleavage and polyadenylation specificity factor (CPSF) and cleavage stimulation factor (CStF) are transferred from RNA Polymerase II to the RNA molecule. The two factors bind to the sequence elements. A protein complex form contains additional cleavage factors and the enzyme Polyadenylate Polymerase (PAP). This complex cleaves the RNA between the polyadenylation sequence and the GU-rich sequence at the cleavage site marked by the (5'-CA-3') sequences. Poly(A) polymerase then adds about 200 adenine units to the new 3' end of the RNA molecule using ATP as a precursor. As the poly(A) tails is synthesized, it binds multiple copies of the poly(A) binding protein, which protects the 3' end from ribonuclease digestion[75].

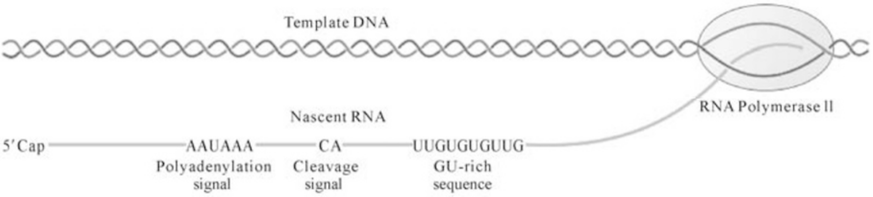


Fig. 16.8 Cleavage and polyadenylation[77] (with the permission of Wikimedia)

RNA splicing is the process by which introns, regions of RNA that do not code for protein, are removed from the pre-mRNA and the remaining exons connect to re-form a single continuous molecule. Although most RNA splicing occurs after the complete synthesis and end-capping of the pre-mRNA, transcripts with many exons can be spliced co-transcriptionally[76]. The splicing reaction is catalyzed by a large protein complex called the spliceosome, which is assembled from proteins and small nuclear RNA molecules that recognize splice sites in the pre-mRNA sequence. Many pre-mRNAs, including those encoding antibodies, can be spliced in multiple ways to produce different mature mRNAs that encode different protein sequences. This process is known as alternative splicing, and allows production of a large variety of proteins from a limited amount of DNA.

16.6.3 Post-Transcriptional Regulation

Post-transcriptional regulation is the control of protein synthesis by genes after synthesis of RNA has begun[78-80].

The first instance of regulation is at transcription (transcriptional regulation) where due to the chromatin arrangement and due to the activity of transcription factors, genes are differentially transcribed. After being produced, the stability and distribution of the different transcripts is regulated (post-transcriptional regulation) by means of the RNA binding protein (RBP) that controls the various steps and rates of the transcripts: events such as alternative splicing, nuclear degradation (exosome), processing, nuclear export (three alternative pathways), sequestration in DCP2-bodies for storage or degradation, and ultimately translation. These proteins achieve these events thanks to an RNA recognition motif (RRM) that binds a specific sequence or secondary structure of the transcripts, typically at the 5' and 3' UTR of the transcript.

This area of study has recently gained more importance due to the increasing evidence that post-transcriptional regulation plays a larger role than previously expected. Even though proteins with DNA binding domains are more abundant than proteins with RNA binding domains, a recent study by Cheadle *et al.*[81] showed that during T-cell activation, 55% of significant changes at the steady-state level had no corresponding changes at the transcriptional level, meaning they were a result of stability regulation alone.

RNA found in the nucleus is more complex than that found in the cytoplasm: more than 95% (bases) of the RNA synthesized by RNA polymerase II never reaches the cytoplasm. The main reason for this is due to the removal of introns which account for 80% of the total bases[82]. Some studies have shown that even after processing, the levels of mRNA between the cytoplasm and the nucleus differ greatly[83].

Developmental biology is a good source of models of regulation, but due to the technical difficulties, it was easier to determine the transcription factor cascades than regulation at the RNA level. In fact, several key genes such as nanos are known to bind RNA but often their targets are unknown[84]. Although RNA binding proteins may regulate post-transcriptionally large amounts of the transcriptome, the targeting of a single gene is of interest to the scientific community for medical reasons; this is from RNA interference and microRNAs which are both examples of post-transcriptional regulation, which regulate the destruction of RNA and change the chromatin structure. To study post-transcriptional regulation, several techniques are used such as RIP-Chip (RNA immunoprecipitation on chip)[85].

16.7 DNA Replication and Recombination

16.7.1 DNA Replication

DNA replication, the basis for biological inheritance, is a fundamental process occurring in all living organisms that want to copy their DNA. This process is "semiconservative" in that each strand of the original double-stranded DNA molecule serves as a template for the reproduction of the complementary strand. Hence, following DNA replication, two identical DNA molecules have been produced from a single double-stranded DNA molecule. Cellular proofreading and error-checking mechanisms ensure near perfect fidelity for DNA replication[86, 87].

In a cell, DNA replication begins at specific locations in the genome, called "origins"[88]. Unwinding of DNA at the origin and synthesis of new strands forms a replication fork. In addition to DNA polymerase, the enzyme that synthesizes the new DNA by adding nucleotides matched to the template strand, a number of other proteins are associated with the fork and assist in the initiation and continuation of DNA synthesis.

DNA replication can also be performed *in vitro* (outside a cell). DNA polymerases, isolated from cells and artificial DNA primers, are used to initiate DNA synthesis at known sequences in a template molecule. The polymerase chain reaction (PCR), a common laboratory technique, employs such artificial synthesis in a cyclic manner to amplify a specific target DNA fragment from a pool of DNA.

The pairing of bases in DNA through hydrogen bonding means that the information contained within each strand is redundant. The nucleotides on a single strand can be used to reconstruct nucleotides on a newly synthesized partner strand[89].

DNA polymerases are a family of enzymes that carry out all forms of DNA replication[90]. A DNA polymerase can only extend an existing DNA strand paired with a template strand; it cannot begin the synthesis of a new strand. To begin synthesis of a new strand, a short fragment of DNA or RNA, called a primer, must be created and paired with the template strand before DNA polymerase can synthesize new DNA, shown as Fig. 16.9.

Once a primer pairs with DNA to be replicated, DNA polymerase synthesizes a new strand of DNA by extending the 3' end of an existing nucleotide chain, adding new nucleotides matched to the template strand one at a time via the creation of phosphodiester bonds. The energy for this process of DNA polymerization comes from two of the three total phosphates attached to each unincorporated base (free bases with their attached phosphate groups are called nucleoside triphosphates). When a nucleotide is being added to a growing DNA strand, two of the phosphates are removed and the energy produced creates a phosphodiester (chemical) bond that attaches the remaining phosphate to the growing chain. The energetics of this process also help explain the directionality of synthesis—if DNA were synthesized in the 3' to 5' direction, the energy for the process would come from the 5' end of the growing strand rather than from free nucleotides.

DNA polymerases are generally extremely accurate, making less than one error for every 10^7 nucleotides added[92]. Even so, some DNA polymerases also have proofreading ability; they can remove nucleotides from the end of a strand in order to correct mismatched bases. If the 5' nucleotide needs to be removed during proofreading, the triphosphate end is lost. Hence, the energy source that usually provides energy to add a new nucleotide is also lost.

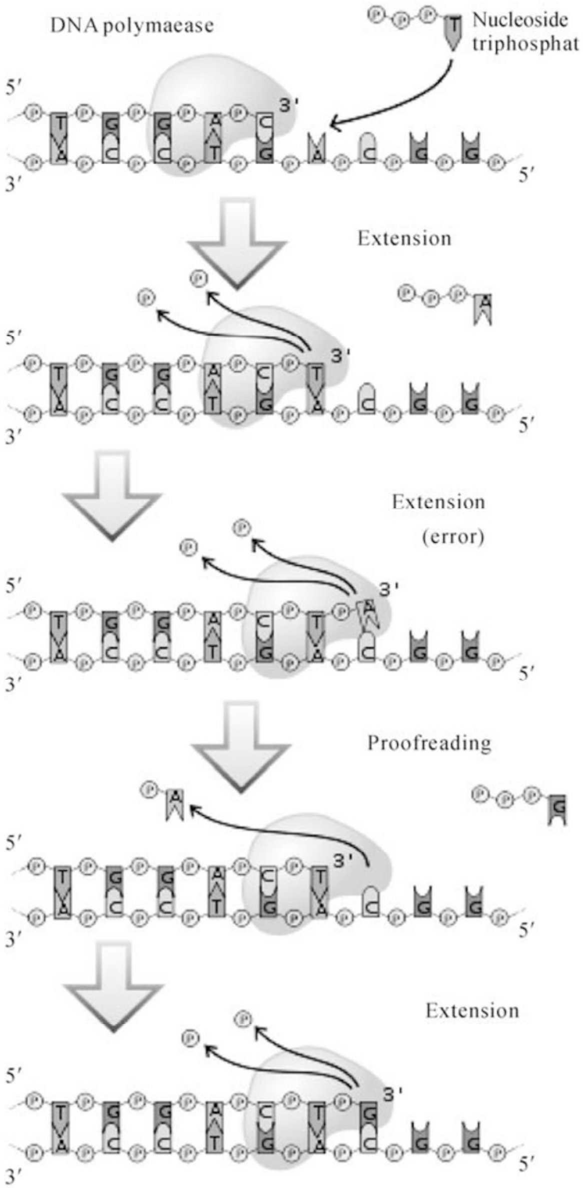


Fig. 16.9 Diagram of DNA polymerase extending a DNA strand and proof-reading[91] (with the permission of Wikimedia)

For a cell to divide, it must first replicate its DNA[93], shown as Fig. 16.10. This process is initiated at particular points within the DNA, known as “origins”, which are targeted by proteins that separate the two strands and initiate DNA synthesis[88]. Origins contain DNA sequences recognized by replication initiator proteins (*e.g*. DnaA in *E. coli* and the Origin Recognition Complex in yeast)[94]. These initiator proteins recruit other proteins to separate the two strands and initiate replication forks.

Initiator proteins recruit other proteins to separate the DNA strands at the origin forming a bubble. Origins tend to be "AT-rich" (rich in adenine and thymine bases) to assist this process because A-T base pairs have two hydrogen bonds (rather than the three formed in a C-G pair)—strands rich in these nucleotides are generally easier to separate[95]. Once strands are separated, RNA primers are created on the template strands and DNA polymerase extends these to create newly synthesized DNA.

As DNA synthesis continues, the original DNA strands continue to unwind on each side of the bubble, forming replication forks. In bacteria, which have a single origin of replication on their circular chromosome, this process eventually creates a "theta structure" (resembling the Greek letter theta: θ). In contrast, eukaryotes have longer linear chromosomes and initiate replication at multiple origins within these.

The replication fork is a structure which forms when DNA is being replicated. It is created through the action of helicase, which breaks the hydrogen bonds holding the two DNA strands together. The resulting structure has two branching "prongs," each one made up of a single strand of DNA.

In DNA replication, the leading strand is defined as the new DNA strand at the replication fork that is synthesized in the 5'→3' direction in a continuous manner. When the enzyme topoisomerase unwinds DNA, two single-stranded regions of DNA (the "replication fork") are formed by the enzyme helicase. On the leading strand, DNA polymerase III is able to synthesize DNA using the free 3'OH group donated by a single RNA primer, and continuous synthesis occurs in the direction in which the replication fork is moving.

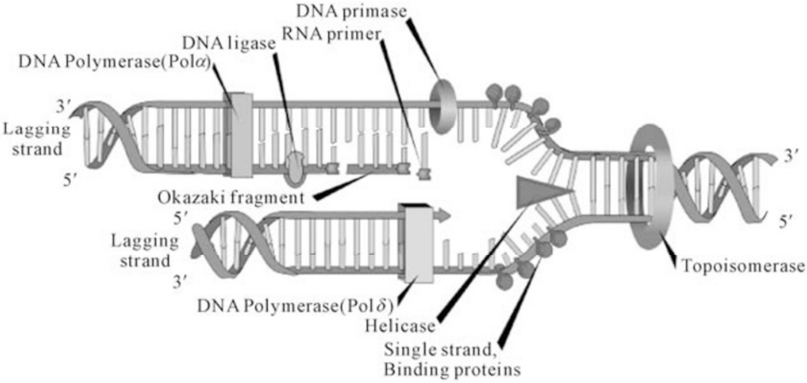


Fig. 16.10 DNA replication[96] (with the permission of Wikimedia)

The lagging strand is the DNA strand at the opposite side of the replication fork from the leading strand, running in the 3' to 5' direction. Because DNA polymerase III cannot synthesize in the 3'→5' direction, the lagging strand is synthesized in short segments known as Okazaki fragments. Along the lagging strand's template, primase builds RNA primers in short bursts. DNA polymerases are then able to use the free 3'OH groups on the RNA primers to synthesize DNA in the 5'→3' direction. The RNA fragments are then removed by DNA polymerase I for prokaryotes or DNA polymerase δ for eukaryotes (different mechanisms are used in eukaryotes and prokaryotes) and new deoxyribonucleotides are added to fill the gaps where the RNA was present. DNA ligase then joins the deoxyribonucleotides together, completing the synthesis of the lagging strand.

As helicase unwinds DNA at the replication fork, the DNA ahead is forced to rotate. This process results in a build-up of twists in the DNA ahead[97]. This build-up would form a resistance that would eventually halt the progress of the replication fork. DNA topoisomerases are enzymes that solve these physical problems in the coiling of the DNA. Topoisomerase I cuts a single backbone on the DNA, enabling the strands to swivel around each other to remove the build-up of twists. Topoisomerase II cuts both backbones, enabling one double-stranded DNA to pass through another, thereby removing knots and entanglements that can form within and between DNA molecules.

Bare single-stranded DNA has a tendency to fold back upon itself and form secondary structures; these structures can interfere with the movement of DNA polymerase. To prevent this, single-strand binding proteins bind to the DNA until a second strand is synthesized, preventing secondary structure formation[98].

Clamp proteins form a sliding clamp around DNA, helping the DNA polymerase maintain contact with its template and thereby assisting with processivity. The inner face of the clamp enables DNA to be threaded through it. Once the polymerase reaches the end of the template or detects double-stranded DNA, the sliding clamp undergoes a conformational change which releases the DNA polymerase. Clamp-loading proteins are used to initially load the clamp, recognizing the junction between template and RNA primers.

Within eukaryotes, DNA replication is controlled within the context of the cell cycle. As the cell grows and divides, it progresses through stages in the cell cycle; DNA replication occurs during the S phase (synthesis phase). The progress of the eukaryotic cell through the cycle is controlled by cell cycle checkpoints. Progression through checkpoints is controlled through complex interactions between various proteins, including cyclins and cyclin-dependent kinases[99].

The G1/S checkpoint (or restriction checkpoint) regulates whether eukaryotic cells enter the process of DNA replication and subsequent division. Cells which do not proceed through this checkpoint are quiescent in the "G0" stage and do not replicate their DNA.

Replication of chloroplast and mitochondrial genomes occur independent of the cell cycle, through the process of D-loop replication.

Most bacteria do not go through a well-defined cell cycle and instead continuously copy their DNA; during rapid growth this can result in multiple rounds of replication occurring concurrently[100]. Within *E. coli*, the most well-characterized bacteria, regulation of DNA replication can be achieved through several mechanisms including: the hemimethylation and sequestering of the origin sequence, the ratio of ATP to ADP, and the levels of protein DnaA. These all control the process of initiator proteins binding to the origin sequences.

Because *E. coli* methylates GATC DNA sequences, DNA synthesis results in hemimethylated sequences. This hemimethylated DNA is recognized by a protein (SeqA) which binds and sequesters the origin sequence; in addition, DnaA (required for initiation of replication) binds less well to hemimethylated DNA. As a result, newly replicated origins are prevented from immediately initiating another round of DNA replication[101].

ATP builds up when the cell is in a rich medium, triggering DNA replication once the cell has reached a specific size. ATP competes with ADP to bind to DnaA, and the DnaA-ATP complex is able to initiate replication. A certain number of DnaA proteins are also required for DNA replication—each time the origin is copied the number of binding sites for DnaA doubles, requiring the synthesis of more DnaA to enable another initiation of replication.

Because bacteria have circular chromosomes, termination of replication occurs when the two replication forks meet each other on the opposite end of the parental chromosome. *E. coli* regulate this process through the use of termination sequences which, when bound by the Tus protein, enable only one direction of the replication fork to pass through. As a result, the replication forks are constrained to always meet within the termination region of the chromosome[102].

Eukaryotes initiate DNA replication at multiple points in the chromosome, so replication forks meet and terminate at many points in the chromosome; these are not known to be regulated in any particular manner. Because eukaryotes have linear chromosomes, DNA replication often fails to synthesize to the very end of the chromosomes (telomeres), resulting in telomere shortening. This is a normal process in somatic cells—cells are only able to divide a certain number of times before the DNA loss prevents further division (this is known as the Hayflick limit). Within the germ cell line, which passes DNA to the next generation, the enzyme telomerase extends the repetitive sequences of the telomere region to prevent degradation. Telomerase can become mistakenly active in somatic cells, sometimes leading to cancer formation.

Another method of copying DNA, sometimes used *in vivo* by bacteria and viruses, is the process of rolling circle replication[103]. In this form of replication, a single replication fork progresses around a circular molecule to form multiple linear copies of the DNA sequence. In cells, this process can be used to rapidly synthesize multiple copies of plasmids or viral genomes.

In the cell, rolling circle replication is initiated by an initiator protein encoded by the plasmid or virus DNA. This protein is able to nick one strand of double-stranded, circular DNA molecule at a site called the double-strand origin (DSO) and remains bound to the 5' phosphate end of the nicked strand. The free 3' hydroxyl end is released and can serve as a primer for DNA synthesis. Using the unnicked strand as a template, replication proceeds around the circular DNA molecule, displacing the nicked strand as single-stranded DNA. Continued DNA synthesis produces multiple single-stranded linear copies of the original DNA in a continuous head-to-tail series. *In vivo*, these linear copies are subsequently converted to double-stranded circular molecules.

Rolling circle replication can also be performed *in vitro* and has found wide uses in academic research and biotechnology, and is often used for amplification of DNA from very small amounts of starting material. Replication can be initiated by nicking a double-stranded circular DNA molecule or by hybridizing a primer to a single-stranded circle of DNA. The use of a reverse primer (or random primers) produces hyperbranched rolling circle amplification, resulting in exponential rather than linear growth of the DNA molecule.

Researchers commonly replicate DNA *in vitro* using the polymerase chain reaction (PCR). PCR uses a pair of primers to span a target region in template DNA, and then polymerizes partner strands in each direction from these primers using a thermostable DNA polymerase. Repeating this process through multiple cycles produces amplification of the targeted DNA region. At the start of each cycle, the mixture of template and primers is heated, separating the newly synthesized molecule and template. Then, as the mixture cools, both of these become templates for annealing new primers, and the polymerase extends from these. As a result, the number of copies of the target region doubles each round, increasing exponentially[104].

16.7.2 DNA Recombination

Recombinant DNA[105, 106] is a form of DNA that does not exist naturally, which is created by combining DNA sequences that would not normally occur together. In terms of genetic modification, recombinant DNA is produced through the addition of relevant DNA into an existing organismal genome, such as the plasmid of bacteria, to code for or alter different traits for a specific purpose, such as immunity. It differs from genetic recombination, in that it does not occur through processes within the cell or ribosome, but is exclusively engineered[107]. Recombinant protein is protein that is derived from recombinant DNA.

The recombinant DNA technique was engineered by Stanley Norman Cohen and others in 1973. They published their findings in a 1974 paper entitled "Construction of Biologically Functional Bacterial Plasmids *in vitro*"[108] which described a technique isolating and amplifing genes or DNA segments and inserting them into another cell with precision, creating a transgenic bacterium. Recombinant DNA technology was made possible by the discovery of restriction endonucleases by Werner Arber, Daniel Nathans, and Hamilton Smith for which they received the 1978 Nobel Prize in Physiology or Medicine.

The use of cloning is interrelated with recombinant DNA in classical biology, as the term "clone" refers to a cell or organism derived from a parental organism, with modern biology referring to the term as a collection of cells derived from the same cell that remain identical. In the classical sense, the use of recombinant DNA provides the initial cell from which the host organism is then expected to recapitulate when it undergoes further cell division, with bacteria remaining a prime example due to the use of viral vectors in medicine that contain recombinant DNA inserted into a structure known as a plasmid[107].

Plasmids are extrachromosomal self-replicating circular forms of DNA present in most bacteria, such as *E. coli*, containing genes related to catabolism and metabolic activity, and allowing the carrier bacterium to survive and reproduce in conditions present within other species and environments. These genes represent characteristics of resistance to bacteriophages and antibiotics and some heavy metals, but can also be easily removed or separated from the plasmid by restriction endonucleases, which regularly produce "sticky ends" and allow the attachment of a selected segment of DNA which codes for more "reparative" substances, such as peptide hormone medications including insulin, growth hormone, and oxytocin. When introducing useful genes into the plasmid, the bacteria are then used as a viral vector, which are encouraged to reproduce so as to recapitulate the altered DNA within other cells it infects, and increase the amount of cells with the recombinant DNA present within them[107].

The use of plasmids is also inportant in gene therapy, where their related viruses are used as cloning vectors or carriers, which are means of transporting and passing on genes in recombinant DNA through viral reproduction throughout an organism. Plasmids contain three common features—a replicator, selectable marker and a cloning site. The replicator or "ori" refers to the origin of replication with regard to location and bacteria where replication begins. The marker refers to a gene that usually contains resistance to an antibiotic, but may also refer to a gene that is attached alongside the desired one, such as that which confers luminescence to allow identification of successfully recombined DNA. The cloning site is a sequence of nucleotides representing one or more positions where cleavage by restriction endonucleases occurs[107]. Most eukaryotes do not maintain canonical plasmids; yeast is a notable exception[109]. In addition, the Ti plasmid of the bacterium *Agrobacterium tumefaciens* can be used to integrate foreign DNA into the genomes of many plants. Other methods of introducing or creating recombinant DNA in eukaryotes include homologous recombination and transfection with modified viruses.

When recombinant DNA is then further altered or changed to host additional strands of DNA, the molecule formed is referred to as a "chimeric" DNA molecule, with reference to the mythological chimera, which consisted as a composite of several animals. The presence of chimeric plasmid molecules is somewhat regular in occurrence throughout the lifetime of an organism, when the propagation by vectors ensures the presence of hundreds of thousands of organismal and bacterial cells that all contain copies of the original chimeric DNA[107].

In the production of chimeric plasmids, the processes involved can be somewhat uncertain, as the intended outcome of the addition of foreign DNA may not always be achieved and may result in the formation of unusable plasmids.

Initially, the plasmid structure is linearized to allow the addition by bonding of complementary foreign DNA strands to single-stranded "overhangs" or "sticky ends" present at the ends of the DNA molecule from staggered or "S-shaped" cleavages produced by restriction endonucleases[107].

A common vector used for the donation of plasmids originally was the bacterium *E.coli* and later, the EcoRI derivative, which was used for its versatility with addition of new DNA by "relaxed" replication when inhibited by chloramphenicol and spectinomycin, later replaced by the pBR322 plasmid. In the case of EcoRI, the plasmid can anneal with the presence of foreign DNA via the route of sticky-end ligation or with "blunt ends" via blunt-end ligation in the presence of the phage T4 ligase, which forms covalent links between 3-carbon OH and 5-carbon PO4 groups present on the blunt ends. Both sticky-end, or overhang ligation and blunt-end ligation can occur between foreign DNA segments, and cleaved ends of the original plasmid depending upon the restriction endonuclease used for cleavage[110].

16.8 DNA Damage and Repair

16.8.1 DNA Damage

DNA damage, due to environmental factors and normal metabolic processes inside the cell, occurs at a rate of 1,000 to 1,000,000 molecular lesions per cell per day[111]. While this constitutes only 0.000165% of the human genome's approximately 6 billion bases (3 billion base pairs), unrepaired lesions in critical genes (such as tumor suppressor genes) can impede a cell's ability to carry out its function and appreciably increase the likelihood of tumor formation[112].

The vast majority of DNA damage affects the primary structure of the double helix; the bases themselves are chemically modified. These modifications can in turn disrupt the molecules' regular helical structure by introducing non-native chemical bonds or bulky adducts that do not fit in the standard double helix. Unlike proteins and RNA, DNA usually lacks tertiary structure and therefore damage or disturbance does not occur at that level. DNA is, however, supercoiled and wound around "packaging" proteins called histones (in eukaryotes), and both superstructures are vulnerable to the effects of DNA damage.

DNA damage can be subdivided into two main types:

(1) Endogenous damage such as an attack by reactive oxygen species produced from normal metabolic byproducts (spontaneous mutation), especially the process of oxidative deamination. It also includes replication errors.

(2) Exogenous damage caused by external agents such as ultraviolet [UV 200 – 300 nm] radiation from the sun, other radiation frequencies, including X-rays and γ-rays, hydrolysis or thermal disruption, certain plant toxins, human-made mutagenic

chemicals, especially aromatic compounds that act as DNA intercalating agents, cancer chemotherapy and radiotherapy,and viruses[113, 114].

The replication of damaged DNA before cell division can lead to the incorporation of wrong bases opposite damaged ones. Daughter cells that inherit these wrong bases carry mutations from which the original DNA sequence is unrecoverable (except in the rare case of a back mutation, for example, through gene conversion).

In human cells and eukaryotic cells in general, DNA is found in two cellular locations-inside the nucleus and inside the mitochondria. Nuclear DNA (nDNA) exists as chromatin during non-replicative stages of the cell cycle and is condensed into aggregate structures known as chromosomes during cell division. In either state, DNA is highly compacted and wound up around bead-like proteins called histones. Whenever a cell needs to express the genetic information encoded in its nDNA the required chromosomal region is unravelled, genes located therein are expressed, and then the region is condensed back to its resting conformation. Mitochondrial DNA (mtDNA) is located inside mitochondria organelles, exists in multiple copies, and is also tightly associated with a number of proteins to form a complex known as the nucleoid. Inside mitochondria, reactive oxygen species (ROS) or free radicals, byproducts of the constant production of adenosine triphosphate (ATP) via oxidative phosphorylation, create a highly oxidative environment that is known to damage mtDNA. A critical enzyme in counteracting the toxicity of these species is superoxide dismutase, which is present in both the mitochondria and cytoplasm of eukaryotic cells.

Senescence, an irreversible state in which the cell no longer divides, is a protective response to the shortening of the chromosome ends. The telomeres are long regions of repetitive noncoding DNA that cap chromosomes and undergo partial degradation each time a cell undergoes division (see Hayflick limit)[115]. In contrast, quiescence is a reversible state of cellular dormancy that is unrelated to genome damage. Senescence in cells may serve as a functional alternative to apoptosis in cases where the physical presence of a cell for spatial reasons is required by the organism[116], which serves as a “last resort” mechanism to prevent a cell with damaged DNA from replicating inappropriately in the absence of pro-growth cellular signaling. Unregulated cell division can lead to the formation of a tumor, which is potentially lethal to an organism. Therefore, the induction of senescence and apoptosis is considered to be part of a strategy of protection against cancer.

It is important to distinguish between DNA damage and mutation, the two major types of error in DNA. DNA damages and mutation are fundamentally different. Damages are physical abnormalities in the DNA, such as single- and double-strand breaks, 8-hydroxydeoxyguanosine residues and polycyclic aromatic hydrocarbon adducts. DNA damages can be recognized by enzymes, and thus they can be correctly repaired if redundant information, such as the undamaged sequence in the complementary DNA strand or in a homologous chromosome, is available for copying. If a cell retains DNA damage, transcription of a gene can be prevented and thus translation into a protein will also be blocked. Replication may also be blocked and/or the cell may die.

In contrast to DNA damage, a mutation is a change in the base sequence of the DNA. A mutation cannot be recognized by enzymes once the base change is present in both DNA strands, and thus a mutation cannot be repaired. At the cellular level, mutations can cause alterations in protein function and regulation. Mutations are replicated when the cell replicates. In a population of cells, mutant cells will increase or decrease in frequency according to the effects of the mutation on the ability of the cell to survive and reproduce. Although distinctly different from each other, DNA damages and mutations are related because DNA damages often cause errors of DNA synthesis during replication or repair and these errors are a major source of mutation.

Given these properties of DNA damage and mutation, it can be seen that DNA damages are a special problem in non-dividing or slowly dividing cells, where unrepaired damages will tend to accumulate over time. Besides, in rapidly dividing cells, unrepaired DNA damages that do not kill the cell by blocking replication will tend to cause replication errors and thus mutation. The great majority of mutations that are not neutral in their effect are deleterious to a cell's survival. Thus, in a population of cells comprising a tissue with replicating cells, mutant cells will tend to be lost. However, infrequent mutations that provide a survival advantage will tend to clonally expand at the expense of neighboring cells in the tissue. This advantage to the cell is disadvantageous to the whole organism, because such mutant cells can give rise to cancer. DNA is damaged in frequently dividing cells because they give rise to mutations and are a prominent cause of cancer. In contrast, DNA damages in infrequently dividing cells are a prominent cause of aging.

16.8.2 DNA Repair Mechanisms

Cells cannot function if DNA damage corrupts the integrity and accessibility of essential information in the genome (but cells remain superficially functional when so-called "non-essential" genes are missing or damaged). Depending on the type of damage inflicted on the DNA's double helical structure, a variety of repair strategies have evolved to restore lost information. If possible, cells use the unmodified complementary strand of the DNA or the sister chromatid as a template to recover the original information. Without access to a template, cells use an error-prone recovery mechanism known as translesion synthesis as a last resort.

Damage to DNA alters the spatial configuration of the helix and such alterations can be detected by the cell. Once damage is localized, specific DNA repair molecules bind at or near the site of damage, inducing other molecules to bind and form a complex that enables the actual repair to take place. The types of molecules involved and the mechanism of repair that is mobilized depend on the type of damage that has occurred and the phase of the cell cycle that the cell is in.

Cells are known to eliminate three types of damage to their DNA by chemically reversing it. These mechanisms do not require a template, since the types of damage they counteract can only occur in one of the four bases. Such direct reversal mechanisms are specific to the type of damage incurred and do not involve breakage of the phosphodiester backbone. The formation of thymine dimers (a common type of cyclobutyl dimer) upon irradiation with UV light results in an abnormal covalent bond between adjacent thymidine bases. The photoreactivation process directly reverses this damage by the action of the enzyme photolyase, whose activation is obligately dependent on energy absorbed from blue/UV light (300 – 500 nm wavelength) to promote catalysis[117]. Another type of damage, methylation of guanine bases, is directly reversed by the protein methyl guanine methyl transferase (MGMT), the bacterial equivalent of which is called ogt. This is an expensive process because each MGMT molecule can only be used once; the reaction is stoichiometric rather than catalytic[118]. A generalized response to methylating agents in bacteria is known as the adaptive response and confers a level of resistance to alkylating agents upon sustained exposure by upregulation alkylation repair enzymes[119]. The third type of DNA damage reversed by cells is certain methylation of the bases cytosine and adenine.

When only one of the two strands of a double helix has a defect, the other strand can be used as a template to guide the correction of the damaged strand. In order to repair damage to one of the two paired molecules of DNA, there exist a number of excision repair mechanisms that remove the damaged nucleotide and replace it with an undamaged nucleotide complementary to that found in the undamaged DNA strand[118].

Base excision repair (BER) repairs damage to a single base caused by oxidation, alkylation, hydrolysis, or deamination. The damaged base is removed by a DNA glycosylase, resynthesized by a DNA polymerase, and a DNA ligase performs the final nick-sealing step.

Nucleotide excision repair (NER) recognizes bulky, helix-distorting lesions such as pyrimidine dimers and 6,4 photoproducts. A specialized form of NER known as transcription-coupled repair deploys NER enzymes to genes that are being actively transcribed.

Mismatch repair (MMR) corrects errors of DNA replication and recombination that result in mispaired (but undamaged) nucleotides.

Double-strand breaks (DSBs), in which both strands in the double helix are severed, are particularly hazardous to the cell because they can lead to genome rearrangements. Two mechanisms exist to repair DSBs: non-homologous end joining (NHEJ) and recombinational repair (also known as template-assisted repair or homologous recombination repair)[118].

In NHEJ, DNA Ligase IV, a specialized DNA Ligase that forms a complex with the cofactor XRCC4, directly joins the two ends[120]. To guide accurate repair, NHEJ relies on short homologous sequences called microhomologies present on the single-stranded tails of the DNA ends to be joined. If these overhangs are compatible, repair is usually accurate[120-123]. NHEJ can also introduce mutations during repair. Loss of damaged nucleotides at the break site can lead to deletions, and joining of non-matching termini forms translocations. NHEJ is especially important before the cell has replicated its DNA, since there is no template available for repair by homologous recombination. There are "backup" NHEJ pathways in higher eukaryotes[124]. Besides its role as a genome caretaker, NHEJ is required for joining hairpin-capped double-strand breaks induced during V(D)J recombination, the process that generates diversity in B-cell and T-cell receptors in the vertebrate immune system[125, 126].

Recombinational repair requires the presence of an identical or nearly identical sequence to be used as a template for repair of the break. The enzymatic machinery responsible for this repair process is nearly identical to the machinery responsible for chromosomal crossover during meiosis. This pathway allows a damaged chromosome to be repaired using a sister chromatid (available in G2 after DNA replication) or a homologous chromosome as a template. DSBs caused by the replication machinery which attempt to synthesize across a single-strand break or unrepaired lesion cause collapse of the replication fork and are typically repaired by recombination.

Topoisomerases introduce both single- and double-strand breaks in the course of changing DNA's state of supercoiling, which is especially common in regions near an open replication fork. Such breaks are not considered as DNA damage because they are a natural intermediate in the topoisomerase biochemical mechanism and are immediately repaired by the enzymes that created them.

A team of French researchers bombarded *Deinococcus radiodurans* to study the mechanism of double-strand break DNA repair in that organism. At least two copies of the genome, with random DNA breaks, can form DNA fragments through annealing. Partially overlapping fragments are then used for synthesis of homologous regions through a moving D-loop that can continue extension until they find complementary partner strands. In the final step there is crossover by means of RecA-dependent homologous recombination[127].

Translesion synthesis is a DNA damage tolerance process that allows the DNA replication machinery to replicate past DNA lesions such as thymine dimers or AP sites. It involves switching out regular DNA polymerases for specialized translesion polymerases (*e.g.* DNA polymerase V), often with larger active sites that can facilitate the insertion of bases opposite damaged nucleotides. The polymerase switching is thought to be mediated by, among other factors, the post-translational modification of the replication processivity factor PCNA. Translesion synthesis polymerases often have low fidelity (high propensity to insert wrong bases) relative to regular polymerases. However, many are extremely efficient at inserting correct bases opposite specific types of damage. For example, Pol η mediates error-free bypass of lesions induced by UV irradiation, whereas Pol ζ introduces mutations at these sites. From a cellular perspective, risking the introduction of point mutations during translesion synthesis may be preferable to resorting to more drastic mechanisms of DNA repair, which may cause gross chromosomal aberrations or cell death.

16.8.3 Global Response to DNA Damage

Cells exposed to ionizing radiation, ultraviolet light or chemicals are prone to acquire multiple sites of bulky DNA lesions and double-strand breaks. Moreover, DNA damaging agents can damage other biomolecules such as proteins, carbohydrates, lipids and RNA. The accumulation of damage, specifically double-strand breaks or adducts stalling the replication forks, are among known stimulation signals for a global response to DNA damage[128]. The global response to damage is an act directed toward the cell's own preservation and triggers multiple pathways of macromolecular repair, lesion bypass, tolerance or apoptosis. The common features of global response are induction of multiple genes, cell cycle arrest, and inhibition of cell division.

After DNA damage, cell cycle checkpoints are activated. Checkpoint activation pauses the cell cycle and gives the cell time to repair the damage before continuing to divide. DNA damage checkpoints occur at the G1/S and G2/M boundaries. An intra-S checkpoint also exists. Checkpoint activation is controlled by two master kinases, ATM and ATR. ATM responds to DNA double-strand breaks and disruptions in chromatin structure[129], whereas ATR primarily responds to stalled replication forks. These kinases phosphorylate downstream targets in a signal transduction cascade, eventually leading to cell cycle arrest. A class of checkpoint mediator proteins including BRCA1, MDC1, and 53BP1 has also been identified[130]. These proteins seem to be required for transmitting the checkpoint activation signal to downstream proteins.

p53 is an important downstream target of ATM and ATR, as it is required for inducing apoptosis following DNA damage[131]. At the G1/S checkpoint, p53 functions by deactivating the CDK2/cyclin E complex. Similarly, p21 mediates the G2/M checkpoint by deactivating the CDK1/cyclin B complex.

The SOS response is the term used to describe changes in gene expression in *Escherichia coli* and other bacteria in response to extensive DNA damage. The prokaryotic SOS system is regulated by two key proteins: LexA and RecA. The LexA homodimer is a transcriptional repressor that binds to operator sequences commonly referred to as SOS boxes. It is known that LexA regulates transcription of approximately 48 genes including the lexA and recA genes[132]. The most common cellular signals activating the SOS response are regions of single-stranded DNA (ssDNA), arising from stalled replication forks or double-strand breaks, which are processed by DNA helicase to separate the two DNA strands[129]. In the initiation step, RecA protein binds to ssDNA in an ATP hydrolysis driven reaction creating RecA-ssDNA filaments. RecA-ssDNA filaments activate LexA autoprotease activity which ultimately leads to cleavage of the LexA dimer and subsequent LexA degradation. The loss of LexA repressor induces transcription of the SOS genes and allows for further signal induction, inhibition of cell division and an increase in levels of proteins responsible for damage processing[133].

SOS boxes are 20-nucleotide long sequences near promoters with a palindromic structure and a high degree of sequence conservation. This distinction in promoter sequences causes differential binding of LexA to different promoters and allows for timing of the SOS response. Logically, the lesion repair genes are induced at the beginning of the SOS response. The error prone translesion polymerases, s, for example: UmuCD’2 (also called DNA polymerase V), are induced later as a last resort[134]. Once the DNA damage is repaired or bypassed using polymerases or through recombination, the amount of single-stranded DNA in cells is decreased, lowering the amounts of RecA filaments decreases cleavage activity of LexA homodimer which subsequently binds to the SOS boxes near promoters and restores normal gene expression.

Eukaryotic cells exposed to DNA damaging agents also activate important defensive pathways by inducing multiple proteins involved in DNA repair, cell cycle checkpoint control, protein trafficking and degradation. Such genome wide transcriptional response is very complex and tightly regulated, thus allowing coordinated global response to damage. Exposure of the yeast *Saccharomyces cerevisiae* to DNA damaging agents results in overlapping but distinct transcriptional profiles. Similarities to environmental shock responses indicate that a general global stress response pathway exists at the level of transcriptional activation. In contrast, different human cell types respond to damage differently indicating an absence of a common global response. The probable explanation for this difference between yeast and human cells may be in the heterogeneity of mammalian cells. In an animal, different types of cells are distributed amongst different organs which have evolved different sensitivities to DNA damage[134].

In general, the global response to DNA damage involves the expressions of multiple genes responsible for post-replication repair, homologous recombination, nucleotide excision repair, DNA damage checkpoint, global transcriptional activation, genes controlling mRNA decay and many others. The vast amount of damage to a cell leaves it with an important decision; undergo apoptosis and die, or survive at the cost of living with a modified genome. An increase in tolerance to damage can lead to an increased rate of survival which will allow a greater accumulation of mutations. Yeast Rev1 and human polymerase η are members of the Y family translesion DNA polymerases present during a global response to DNA damage and are responsible for enhanced mutagenesis during a global response to DNA damage in eukaryotes[129].

16.8.4 DNA Repair and Aging

Experimental animals with genetic deficiencies in DNA repair often show a decreased lifespan and increased cancer incidence. For example, mice deficient in the dominant NHEJ pathway and in telomere maintenance mechanisms get lymphoma and infections more often and consequently have shorter lifespans than wild-type mice[135]. Similarly, mice deficient in a key repair and transcription protein that unwinds DNA helices have a premature onset of aging-related diseases and consequent shortening of their lifespan[136]. However, not every DNA repair deficiency creates the exact predicted effects; mice deficient in the NER pathway exhibited a shortened lifespan without correspondingly higher rates of mutation[137].

If the rate of DNA damage exceeds the capacity of the cell to repair it, the accumulation of errors can overwhelm the cell and result in early senescence, apoptosis or cancer. Inherited diseases associated with a faulty DNA repair function result in premature aging, increased sensitivity to carcinogens, and correspondingly increased cancer risk (see below). On the other hand, organisms with enhanced DNA repair systems, such as *Deinococcus radiodurans* which is the most radiation-resistant known organism, exhibit remarkable resistance to the double-strand break-inducing effects of radioactivity, likely due to enhanced efficiency of DNA repair especially NHEJ[138].

A number of individual genes have been identified as influencing variations in lifespan within a population of organisms. The effects of these genes are strongly dependent on the environment, particularly on the organism's diet. Caloric restriction reproducibly results in extended lifespan in a variety of organisms, likely via nutrient sensing pathways and decreased metabolic rate. The molecular mechanisms by which such restriction results in a lengthened lifespan are not yet unclear (see[139] for discussion); however, the behavior of many genes known to be involved in DNA repair is altered under conditions of caloric restriction.

For example, increasing the gene dosage of the gene SIR-2, which regulates DNA packaging in the nematode worm *Caenorhabditis elegans*, can significantly extend lifespan[140]. The mammalian homolog of SIR-2 is known to induce downstream DNA repair factors involved in NHEJ, an activity that is especially promoted under conditions of caloric restriction[141]. Caloric restriction has been closely linked to the rate of base excision repair in the nuclear DNA of rodents[142], although similar effects have not been observed in mitochondrial DNA[143].

Interestingly, the *C. elegans* gene AGE-1, an upstream effector of DNA repair pathways, confers a dramatically extended lifespan under free-feeding conditions but leads to a decrease in reproductive fitness under conditions of caloric restriction[144]. This observation supports the pleiotropy theory of the biological origins of aging, which suggests that genes conferring a large survival advantage early in life will be selected for even if they carry a corresponding disadvantage later in life.

16.8.5 Medicine and DNA Repair Modulation

Defects in the NER mechanism are responsible for several genetic disorders including:

(1) Xeroderma pigmentosum: hypersensitivity to sunlight/UV resulting in

increased skin cancer incidence and premature aging.

(2) Cockayne syndrome: hypersensitivity to UV and chemical agents.

(3) Trichothiodystrophy: sensitive skin and brittle hair and nails.

(4) Mental retardation often accompanies the latter two disorders, suggesting increased vulnerability of developmental neurons.

Other DNA repair disorders include:

(1) Werner's syndrome: premature aging and retarded growth.

(2) Bloom's syndrome: sunlight hypersensitivity and a high incidence of malignancies (especially leukemias).

(3) Ataxia telangiectasia: sensitivity to ionizing radiation and some chemical agents.

All of the above diseases are often called "segmental progerias" ("accelerated aging diseases") because their victims appear elderly and suffer from aging-related diseases at an abnormally young age.

Other diseases associated with reduced DNA repair function include Fanconi's anemia, hereditary breast cancer and hereditary colon cancer.

Inherited mutations that affect DNA repair genes are strongly associated with high cancer risks in humans. Hereditary nonpolyposis colorectal cancer (HNPCC) is strongly associated with specific mutations in the DNA mismatch repair pathway. BRCA1 and BRCA2, two famous mutations conferring a significantly increased risk of breast cancer in carriers, are both associated with a large number of DNA repair pathways, especially NHEJ and homologous recombination.

Cancer therapy procedures such as chemotherapy and radiotherapy work by overwhelming the capacity of the cell to repair DNA damage, resulting in cell death. Cells that are most rapidly dividing—most typically cancer cells—are preferentially affected. The side effect is that other non-cancerous but rapidly dividing cells such as stem cells in the bone marrow are also affected. Modern cancer treatments attempt to localize the DNA damage to cells and tissues only associated with cancer, either by physical means (concentrating the therapeutic agent in the region of the tumor) or by biochemical means (exploiting a feature unique to cancer cells in the body).

16.8.6 DNA Repair and Evolution

The basic processes of DNA repair are highly conserved among both prokaryotes and eukaryotes and even among bacteriophage (viruses that infect bacteria); however, more complex organisms with more complex genomes have correspondingly more complex repair mechanisms[145]. The ability of a large number of protein structural motifs to catalyze relevant chemical reactions has played a significant role in the elaboration of repair mechanisms during evolution. For an extremely detailed review of hypotheses relating to the evolution of DNA repair see[146].

The fossil record indicates that single-celled life began to proliferate on the planet at some point during the Precambrian period, although exactly when recognizable modern life first emerged is unclear. Nucleic acids became the sole and universal means of encoding genetic information, requiring DNA repair mechanisms that in their basic form have been inherited by all extant life forms from their common ancestor. The emergence of Earth's oxygen-rich atmosphere (known as the "oxygen catastrophe") due to photosynthetic organisms, as well as the presence of potentially damaging free radicals in the cell due to oxidative phosphorylation, necessitated the evolution of DNA repair mechanisms that acted specifically to counter the types of damage induced by oxidative stress.

On some occasions, DNA damage is not repaired, or is repaired by an error-prone mechanism, which results in a change from the original sequence. When this occurs, mutations may propagate into the genomes of the cell's progeny. Should such an event occur in a germ line cell that will eventually produce a gamete, the mutation has the potential to be passed on to the organism's offspring. The rate of evolution in a particular species (or more narrowly, in a particular gene) is a function of the rate of mutation. Consequently, the rate and accuracy of DNA repair mechanisms have an influence over the process of evolutionary change[147].

16.9 Translation

Translation is the first stage of protein biosynthesis (part of the overall process of gene expression). Translation is the production of proteins by decoding mRNA produced in transcription. Translation occurs in the cytoplasm where the ribosomes are located. Ribosomes are made of small and large subunits which surround the mRNA. In translation (Fig. 16.11), mRNA is decoded to produce a specific polypeptide according to the rules specified by the genetic code. This uses an mRNA sequence as a template to guide the synthesis of a chain of amino acids that form a protein. Many types of transcribed RNA, such as transfer RNA, ribosomal RNA, and small nuclear RNA are not necessarily translated into an amino acid sequence. Translation proceeds in four phases: activation, initiation, elongation and termination (all describing the growth of the amino acid chain or polypeptide that is the product of translation). Amino acids are brought to ribosomes and assembled into proteins[148, 149].

During activation, the correct amino acid is covalently bonded to the correct transfer RNA (tRNA). While this is not technically a step in translation, it is required for translation to proceed. The amino acid is joined by its carboxyl group to the 3'OH of the tRNA by an ester bond. When the tRNA has an amino acid linked to it, it is termed "charged." Initiation involves the small subunit of the ribosome binding to 5' end of mRNA with the help of initiation factors (IF). Termination of the polypeptide happens when the A site of the ribosome faces a stop codon (UAA, UAG, or UGA). When this happens, no tRNA can recognize it, but a releasing factor can recognize nonsense codons and causes the release of the polypeptide chain. The 5' end of the mRNA gives rise to the protein's N-terminal and the direction of translation can therefore be stated as N->C.

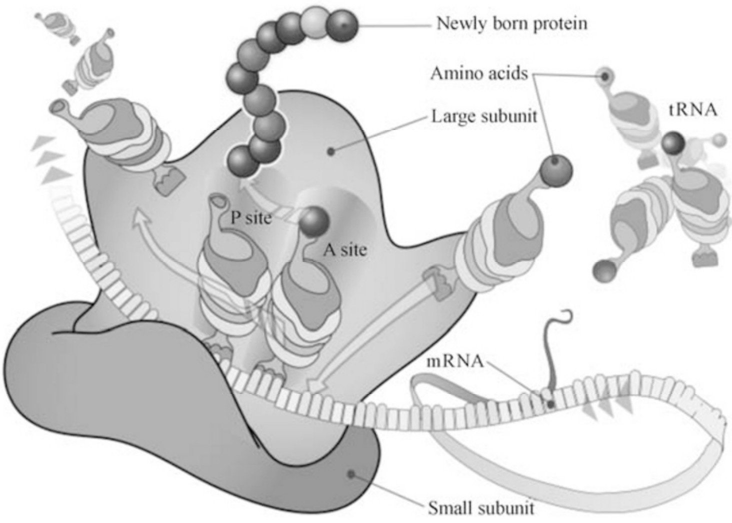


Fig. 16.11 Diagram showing how the translation of the mRNA and the synthesis of proteins is made by ribosomes

The capacity of disabling or inhibiting translation in protein biosynthesis is used by antibiotics such as: anisomycin, cycloheximide, chloramphenicol, tetracycline, streptomycin, erythromycin, puromycin and others. Prokaryotic ribosomes have a different structure than eukaryotic ribosomes, and thus antibiotics can specifically target bacterial infections without any detriment to the host's cells.

The mRNA carries genetic information encoded as a ribonucleotide sequence from the chromosomes to the ribosomes. The ribonucleotides are "read" by translational machinery in a sequence of nucleotide triplets called codons. Each of those triplets codes for a specific amino acid.

The ribosome and tRNA molecules translate this code to a specific sequence of amino acids. The ribosome is a multi-subunit structure containing rRNA and proteins. It is the "factory" where amino acids are assembled into proteins. tRNAs are small noncoding RNA chains (74-93 nucleotides) that transport amino acids to the ribosome. tRNAs have a site for amino acid attachment, and a site called an anticodon. The anticodon is an RNA triplet complementary to the mRNA triplet that codes for their cargo amino acid.

Aminoacyl tRNA synthetase (an enzyme) catalyzes the bonding between specific tRNAs and the amino acids that their anticodons sequences call for. The product of this reaction is an aminoacyl-tRNA molecule. This aminoacyl-tRNA travels inside the ribosome, where mRNA codons are matched through complementary base pairing to specific tRNA anticodons. The amino acids that the tRNAs carry are then used to assemble a protein. The energy required for translation of proteins is significant. For a protein containing n amino acids, the number of high-energy phosphate bonds required to translate it is 4n-1. The rate of translation varies; it is significantly higher in prokaryotic cells (up to 17-21 amino acid residues per second) than in eukaryotic cells (up to 6-7 amino acid residues per second).

References

[1] Webpage link: http://en.wikipedia.org/wiki/Molecular_biology.

[2] Beadle, G. W. & E. L. Tatum (1941). "Genetic control of biochemical reactions in neurospora", Proceedings of the National Academy of Sciences 27: 499-506.

[3] Avery, O. T., C. M. Macleod & M. McCarty (1944). "Studies on the chemical nature of the substance inducing transformation of pneumococcal types: Induction of transformation by A desoxyribonucleic acid fraction isolated from pneumococcus type III", Journal of Experimental Medicine 79(2): 137-158.

[4] Hershey, A. D. & M. Chase (1952). "Independent functions of viral protein and nucleic acid in growth of bacteriophage", Journal of General Physiology 36(1): 39-56.

[5] Watson, J. D. & F. H. C. Crick (1953). "Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid", Nature 171: 737-738.

[6] Jacob, F. & J. Monod (1961). "Genetic regulatory mechanisms in the synthesis of proteins", Journal of Molecular Biology 3: 318-356.

[7] Gerstein, M. B., C. Bruce, J. S. Rozowsky, D. Zheng, J. Du, J. O. Korbel, O. Emanuelsson, Z. D. Zhang, S. Weissman & M. Snyder (2007). "What is a gene, post-ENCODE history and updated definition", Genome Research 17(6): 669-681.

[8] Steinman, R. M. & C. L. Moberg (1994). "A triple tribute to the experiment that transformed biology", Journal of Experimental Medicine 179(2): 379-384.

[9] Min Jou, W., G. Haegeman, M. Ysebaert & W. Fiers (1972). "Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein", Nature 237(5350): 82-88.

[10] Pearson, H (2006). "Genetics: What is a gene?" Nature 441(7092): 398-401.

[11] Rassoulzadegan, M., V. Grandjean, P. Gounon, S. Vincent, I. Gillot & F. Cuzin (2006). "RNA-mediated non-mendelian inheritance of an epigenetic change in the mouse", Nature 441(7092): 469-474.

[12] Mortazavi, A., B. A. Williams, K. McCue, L. Schaeffer & B. Wold (2008).

"Mapping and quantifying mammalian transcriptomes by RNA-Seq", Nature Methods 5: 621.

[13] Braig, M. & C. Schmitt (2006). "Oncogene-induced senescence: Putting the brakes on tumor development", Cancer Research 66(6): 2881-2884.

[14] International Human Genome Sequencing Consortium (2004). "Finishing the euchromatic sequence of the human genome", Nature 431(7011): 931-945.

[15] Elizabeth, P (2007). "DNA study forces rethink of what it means to be a gene", Science 316(5831): 1556-1557.

[16] Chien, A., D. B. Edgar & J. M. Trela (1976). "Deoxyribonucleic acid polymerase from the extreme thermophile thermus aquaticus", Journal of Bacteriology 127(3): 1550-1557.

[17] Bartlett, J. M. & D. Stirling (2003). "A short history of the polymerase chain reaction", Methods in Molecular Biology 226: 3-6.

[18] Saiki, R. K., S. Scharf, F. Faloona, K. B. Mullis, G. T. Horn, H. A. Erlich & N. Arnheim (1985). "Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia", Science 230(4732): 1350-1354.

[19] Saiki, R. K., D. H. Gelfand, S. Stoffel, S. J. Scharf, R. Higuchi, G. T. Horn, K. B. Mullis & H. A. Erlich (1988). "Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase", Science 239: 487-491.

[20] Pavlov, A. R., N. V. Pavlova, S. A. Kozyavkin & A. I. Slesarev (2006). "Thermostable DNA polymerases for a wide spectrum of applications: Comparison of a robust hybrid topoTaq to other enzymes. Kieleczawa J. DNA sequencing II: optimizing preparation and cleanup", Jones and Bartlett: 241-257.

[21] Webpage link: http://en.wikipedia.org/wiki/Polymerase_chain_reaction.

[22] Thomas, P. S. (1980). "Hybridization of denatured RNA and small DNA fragments transferred to nitrocellulose", Proceedings of the National Academy of Sciences 77(9): 5201-5205.

[23] Joseph S., *et al.* (2001). A Laboratory Manual. Cold Spring Harbor Laboratory Press.

[24] Schade, B., G. Jansen, M. Whiteway, K. D. Entian & D. Y. Thomas (2004). "Cold adaptation in budding yeast", Molecular Biology of the Cell 15(12): 5492-5502.

[25] Pérez-Ortín, J. E., J. García-Martínez & T. M. Alberola (2002). "DNA chips for yeast biotechnology. The case of wine yeasts", Journal of Biotechnology 98(2-3): 227-241.

[26] Webbook: "Moecular biology web book", Chapter 1: Cells and Viruses—Overview. Webpage link: at http://www.web-books.com/MoBio/Free/Ch1A1.htm.

[27] Madigan, M. T. & J. M. Martino (2006). Brock Biology of Microorganisms. Pearson.

[28] Webpage link: http://en.wikipedia.org/wiki/File:Phylogenetic_Tree_of_Life.png.

[29] Pike, L. J. (2004). "Lipid rafts: heterogeneity on the high seas", Biochemical Journal 378(Pt2): 281-292.

[30] Goldman, R. D., Y. Gruenbaum, R. D. Moir, D. K. Shumaker & T. P. Spann

(2002). “Nuclear lamins: Building blocks of nuclear architecture”, Genes and Development 16(5): 533-547.

[31] Rout, M. P. & J. D. Aitchison (2001). “The nuclear pore complex as a transport machine”, Journal of Biological Chemistry 276(20): 16593-16596.

[32] Chazal, N. & D. Gerlier (2003). “Virus entry, assembly, budding, and membrane rafts”, Microbiology and Molecular Biology Reviews 67(2): 226-237.

[33] Webpage link: http://www.tutornext.com/ws/prokaryotic-gene-structure.

[34] Webpage link: http://en.wikipedia.org/wiki/Prokaryotic_transcription.

[35] Gruber, T. M. & C. A. (2003). “Gross multiple sigma subunits and the partitioning of bacterial transcription space”, Annual Review of Microbiology 57: 441-466.

[36] .Kapanidis, A. N., E. Margeat, T. A. Laurence, S. Doose, S. O. Ho, J. Mukhopadhyay, E. Kortkhonjia, V. Mekler, R. H. Ebright & S. Weiss (2005). “Retention of transcription initiation factor sigma70 in transcription elongation: single-molecule analysis”, Molecular Cell 20(3): 347-356.

[37] Webpage link: http://en.wikipedia.org/wiki/Rho_factor.

[38] David, P. (1975). “Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter”, Proceedings of the National Academy of Sciences of the United States of America 72: 784-788.

[39] Heinz, S., G. Christopher & H. Karin (1975). “Nucleotide sequence of an RNA polymerase binding site from the DNA of bacteriophage fd”, Proceedings of the National Academy of Sciences of the United States of America 72: 737-741.

[40] Webpage link: http://en.wikipedia.org/wiki/Eukaryotic_translation.

[41] López-Lastra, M., A. Rivas & M. I. Barría (2005). “Protein synthesis in eukaryotes: The growing biological relevance of cap-independent translation initiation”, Biological Research 38(2-3): 121-146.

[42] Webpage link: http://en.wikipedia.org/wiki/File:Eukaryotic_Translation_Initiation.Png.

[43] Webpage link: http://en.wikipedia.org/wiki/Eukaryotic_elongation_factors.

[44] Kisselev, L., M. Ehrenberg & L. Frolova (2003). “Termination of translation: interplay of mRNA, rRNAs and release factors?” EMBO Journal 22: 175-182.

[45] Webpage link: http://en.wikipedia.org/wiki/Eukaryotic_initiation_factor.

[46] Park, M. H. (2006). “The post-translational synthesis of a polyamine-derived amino acid, hypusine, in the eukaryotic translation initiation factor 5A (eIF5A)”, Journal of Biochemistry 139(2): 161-169.

[47] Jerard, H. (2005). “The discovery of RNA polymerase”, Journal of Biological Chemistry 280(52): 42477-42485.

[48] Ishihama, A. (2000). “Functional modulation of *Escherichia coli* RNA polymerase”, Annual Review of Microbiology 54: 499-518.

[49] Webpage link: http://en.wikipedia.org/wiki/RNA_polymerase.

[50] Grummt, I. (1999). “Regulation of mammalian ribosomal gene transcription by RNA polymerase I”, Progress in Nucleic Acid Research & Molecular

Biology 62: 109-154.

[51] Lee, Y., M. Kim, J. Han, K. H. Yeom, S. Lee, S. H. Baek & V. N. Kim (2004). “MicroRNA genes are transcribed by RNA polymerase II”, EMBO Journal 23(20): 4051-4060.

[52] Willis, I. M. (1993). “RNA polymerase III. Genes, factors and transcriptional specificity”, European Journal of Biochemistry 212(1): 1-11.

[53] Herr, A. J., M. B. Jensen, T. Dalmay & D. C. Baulcombe (2005). “RNA polymerase IV directs silencing of endogenous DNA”, Science 308(5718): 118-120.

[54] Makeyev, E. V. & D. H. Bamford (2002). “Cellular RNA-dependent RNA polymerase involved in posttranscriptional gene silencing has two distinct activity modes”, Molecular Cell 10(6): 1417-1427.

[55] Dame, R. T. (2005). “The role of nucleoid-associated proteins in the organization and compaction of bacterial chromatin”, Molecular Microbiology 56(4): 858-870.

[56] Webpage link: http://en.wikipedia.org/wiki/Chromatin.

[57] Bernstein, B. E., T. S. Mikkelsen, X. Xie, M. Kamal, D. J. Huebert, J. Cuff, B. Fry, A. Meissner, *et al.* (2006). “A bivalent chromatin structure marks key developmental genes in embryonic stem cells”, Cell 125(2): 315-326.

[58] Portoso, M. & G. Cavalli (2008). “The Role of RNAi and Noncoding RNAs in polycomb mediated control of gene expression and genomic programming”, in RNA and the Regulation of Gene Expression: A Hidden Layer of Complexity. Caister Academic Press.

[59] Robinson, P. J., L. Fairall, V. A. Huynh & D. Rhodes (2006). “EM measurements define the dimensions of the ‘30-nm’ chromatin fiber: evidence for a compact, interdigitated structure”, Proceedings of the National Academy of Sciences 103(17): 6506-6511.

[60] Wong, H., J. M. Victor & J. Mozziconacci (2007). “An all-atom model of the chromatin fiber containing linker histones reveals a versatile structure tuned by the nucleosomal repeat length”, PLoS ONE 2(9): e877.

[61] Webpage link: http://en.wikipedia.org/wiki/Chromatin_structure.

[62] Lodish, H., A. Berk, S. L. Zipursky, P. Matsudaira, D. Baltimore & E. J. Darnell (1999). Molecular Cell Biology. W. H. Freeman & Co.

[63] Daniel, L. H. & W. J. Elizabeth (2005). Genetics: Analysis of Genes and Genomes. Jones & Bartlett Publishers.

[64] Webpage link: http://en.wikipedia.org/wiki/RNA_splicing.

[65] Ng, B., F. Yang, D. P. Huston, Y. Yan, Y. Yang, Z. Xiong, L. E. Peterson, H. Wang & X. F. Yang (2004). “Increased noncanonical splicing of autoantigen transcripts provides the structural basis for expression of untolerized epitopes”, Journal of Allergy and Clinical Immunology 114(6): 1463-1470.

[66] Patel, A. A. & J. A. Steitz (2003). “Splicing double: insights from the second spliceosome”, Nature Reviews Molecular Cell Biology. 4(12): 960-970.

[67] Friend, K., N. G. Kolev, M. D. Shu & J. A. Steitz (2008). “Minor-class splicing occurs in the nucleus of the Xenopus oocyte”, Ribonucleic Acid 14(8): 1459-1462.

[68] Di, S. G., S. Gastaldi & G. P. Tocchini-Valentini (2008). “Cis- and trans-splicing of mRNAs mediated by tRNA sequences in eukaryotic cells”, Proceedings of the National Academy of Sciences of the United States of America 105(19): 6864-6869.

[69] Draper, B. W., P. A. Morcos & C. B. Kimmel (2001). “Inhibition of zebrafish fgf8 pre-mRNA splicing with morpholino oligos: a quantifiable method for gene knockdown”, Genesis 30(3): 154-156.

[70] Sazani, P., S. H. Kang, M. A. Maier, C. Wei, J. Dillman, J. Summerton, M. Manoharan & R. Kole (2001). “Nuclear antisense effects of neutral, anionic and cationic oligonucleotide analogs”, Nucleic Acids Research 29(19): 3965-3974.

[71] Morcos, P. A. (2007). “Achieving targeted and quantifiable alteration of mRNA splicing with Morpholino oligos”, Biochemical and Biophysical Research Communications 358(2): 521-527.

[72] Bruno, I. G., W. Jin & G. J. Cote (2004). “Correction of aberrant FGFR1 alternative RNA splicing through targeting of intronic regulatory elements”, Human Molecular Genetic 13(20): 2409-2420.

[73] Danckwardt, S., G. Neu-Yilik, R. Thermann, U. Frede, M. W. Hentze & A. E. Kulozik (2002). “Abnormally spliced beta-globin mRNAs: a single point mutation generates transcripts sensitive and insensitive to nonsense-mediated mRNA decay”, Blood 99(5): 1811-1816.

[74] Hanada, K. & J. C. Yang (2005). “Increased Novel biochemistry: post-translational protein splicing and other lessons from the school of antigen processing”, Journal of Molecular Medicine 83(6): 420-428.

[75] Berg, J. M., L. J. Tymoczko & L. Stryer (2007). Biochemistry (6th edition), W. H. Freeman & Co.

[76] Hames, D. & Nigel H. (2006). Instant Notes Biochemistry (3rd edition). Taylor and Francis.

[77] Webpage link: http://en.wikipedia.org/wiki/Post-transcriptional_modification #Cleavage_and_Polyadenylation.

[78] Lodish, H. F., A. Berk, C. Kaiser, M. Krieger, M. P. Scott, A. Bretscher, H. Ploegh & P. T. Matsudaira (2007). “Post-transcriptional gene control”, in Molecular Cell Biology. W. H. Freeman.

[79] Bruce, A., J. Alexander, L. Julian, R. Martin, R. Keith & W. Peter (2007). Molecular Biology of the Cell (5th edition). Garland Science.

[80] Weaver, R. J. (2007). “Part V: Post-transcriptional events”, in Molecular Biology. McGraw Hill Higher Education.

[81] Cheadle, C., J. Fan, Y. S. Cho-Chung, T. Werner, J. Ray, L. Do, M. Gorospe & K. G. Becker (2005). “Control of gene expression during T cell activation: alternate regulation of mRNA transcription and mRNA stability”, BMC Genomics 6(1): 75.

[82] Jackson, D. A., A. Pombo & F. Iborra (2000). “The balance sheet for transcription: an analysis of nuclear RNA metabolism in mammalian cells”, FASEB Journal 14(2): 242-254.

[83] Schwanekamp, J. A., M. A. Sartor, S. Karyala, D. Halbleib, M. Medvedovic

& C. R. Tomlinson (2006). "Genome-wide analyses show that nuclear and cytoplasmic RNA levels are differentially affected by dioxin", Biochimica et Biophysica Acta 1759(8-9): 388-402.

[84] Scott, F. G.(2003). Developmental Biology. Sinauer.

[85] Keene, J. D., J. M. Komisarow & M. B. Friedersdorf (2006). "RIP-Chip: the isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts", Nature Protocols 1(1): 302-307.

[86] Berg, J. M., J. L. Tymoczko, L. Stryer & N. D. Clarke (2002). "DNA replication, recombination, and repair", in Biochemistry. W. H. Freeman and Company.

[87] Alberts, B., A. Johnson, J. Lewis, M. Raff, K. Roberts & P. Walter (2002). "DNA replication, repair, and recombination", in Molecular Biology of the Cell. Garland Science.

[88] Berg, J. M., J. L. Tymoczko, L. Stryer & N. D. Clarke (2002). "DNA replication of both strands proceeds rapidly from specific start sites", in Biochemistry. W. H. Freeman and Company.

[89] Alberts, B., A. Johnson, J. Lewis, M. Raff, K. Roberts & P. Wlater (2002). Molecular Biology of the Cell (4th edition). Garland Science.

[90] Berg, J. M., J. L. Tymoczko, L. Stryer & N. D. Clarke (2002). "DNA polymerases require a template and a primer", in Biochemistry. W. H. Freeman and Company.

[91] Webpage link: http://en.wikipedia.org/wiki/File:DNA_polymerase.svg.

[92] McCulloch, S. D. & T. A. Kunkel (2008). "The fidelity of DNA synthesis by eukaryotic replicative and translesion synthesis polymerases", Cell Research 18: 148-161.

[93] Alberts, B., A. Johnson, J. Lewis, M. Raff, K. Roberts & P. Walter (2002). "DNA replication mechanisms", in Molecular Biology of the Cell. Garland Science.

[94] Weigel, C., A. Schmidt, B. Rückert, R. Lurz & W. Messer (1997). "DnaA protein binding to individual DnaA boxes in the Escherichia coli replication origin, oriC", EMBO Journal 16(21): 6574-6583.

[95] Lodish, H., A. Berk, L. S. Zipursky, P. Matsudaira, D. Baltimore & J. Darnell (2000). "General features of chromosomal replication: Three common features of replication origins", in Molecular Cell Biology. W. H. Freeman and Company.

[96] Webpage link: http://en.wikipedia.org/wiki/DNA.

[97] Alberts, B., A. Johnson, J. Lewis, M. Raff, K. Roberts & P. Walter (2002). "DNA replication mechanisms: DNA topoisomerases prevent DNA tangling during replication", in Molecular Biology of the Cell. Garland Science.

[98] Alberts, B., A. Johnson, J. Lewis, M. Raff, K. Roberts & P. Walter (2002). "DNA replication mechanisms: Special proteins help to open up the DNA double helix in front of the replication fork", in Molecular Biology of the Cell. Garland Science.

[99] Alberts, B., A. Johnson, J. Lewis, M. Raff, K. Roberts & P. Walter (2002). "Intracellular control of cell-cycle events: S-phase cyclin-Cdk complexes (S-Cdks) initiate DNA replication once per cycle", in Molecular Biology of the Cell. Garland Science.

[100] Tobiason, D. M. & H. S. Seifert (2006). "The obligate human pathogen, neisseria gonorrhoeae, is polyploid", PLoS Biology 4(6): e185.

[101] Slater, S., S. Wold, M. Lu, E. Boye, K. Skarstad & N. Kleckner (1995). "*E. coli* SeqA protein binds oriC in two different methyl-modulated reactions appropriate to its roles in DNA replication initiation and origin sequestration", Cell 82(6): 927-936.

[102] Brown, T. A. (2002). "Termination of replication", in Genomes. BIOS Scientific Publishers Ltd.

[103] Griffiths, A. J. F., J. H. Miller, D. T. Suzuki, R. C. Lewontin & W. M. Gelbart (2000). "Replication of DNA: Rolling-circle replication", in An Introduction to Genetic Analysis. W. H. Freeman.

[104] Saiki, R. K., D. H. Gelfand, S. Stoffel, S. J. Scharf, R. Higuchi, G. T. Horn, K. B. Mullis & H. A. Erlich (1988). "Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase", Science 239: 487-91.

[105] Garret, R. H. & C. M. Grisham (2000). Biochemistry. Saunders College Publishers.

[106] Colowick, S. P. & O. N. Kapian (1980). "Recombinant DNA", in Methods in Enzymology 68. Academic Press.

[107] Jeremy, M. B., L. T. John & L. Stryer (2002). Biochemistry. W. H. Freeman.

[108] Cohen, S. N., A. C. Chang, H. W. Boyer & R. B. Helling (1973). "Construction of biologically functional bacterial plasmids *in vitro*", Proceedings of the National Academy of Sciences 70(11): 3240-3244.

[109] San Diego State University. 2007. "Plasmids in eukaryotic microbes: An example", Webpage link: http://www.sci.sdsu.edu/~smaloy/MicrobialGenetics/topics/plasmids/yeast-plasmid.html.

[110] Nathan, P. K., P. C. Nathan & W. Ray (1980). "Recombinant DNA", Volume 68: Recombinant Dna Part F (Methods in Enzymology). Academic Press.

[111] Lodish, H., A. Berk, P. Matsudaira, C. A. Kaiser, M. Krieger, M. P. Scott, S. L. Zipursky & J. Darnell (2004). Molecular Biology of the Cell. W. H. Freeman.

[112] Browner, W. S., A. J. Kahn, E. Ziv, A. P. Reiner, J. Oshima, R. M. Cawthon, W. C. Hsueh & S. R. Cummings (2004). "The genetics of human longevity", American Journal of Medicine 117(11): 851-860.

[113] Roulston, A., R. C. Marcellus & P. E. Branton (1999). "Viruses and apoptosis", Annual Review of Microbiology 53: 577-628.

[114] Ohta, T., S. Tokishita, K. Mochizuki, J. Kawase, M. Sakahira & H. Yamagata (2006). "UV Sensitivity and Mutagenesis of the Extremely Thermophilic Eubacterium Thermus thermophilus HB27", Genes and Environment 28(2): 56-61.

[115] Braig, M. & C. A. Schmitt (2006). "Oncogene-induced senescence: putting the brakes on tumor development", Cancer Research 66: 2881-2884.

[116] Lynch, M. D. (2006). "How does cellular senescence prevent cancer?" DNA and Cell Biology 25(2): 69-78.

[117] Sancar, A. (2003). "Structure and function of DNA photolyase and cryptochrome blue-light photoreceptors", Chemical Reviews 103(6): 2203-2237.

[118] Watson, J. D., T. A. Baker, S. P. Bell, A. Gann, M. Levine & R. Losick (2004). Molecular Biology of the Gene. CSHL Press.

[119] Volkert, M. R. (1988). "Adaptive response of Escherichia coli to alkylation damage", Environmental and Molecular Mutagenesis 11(2): 241-255.

[120] Wilson, T. E., U. Grawunder & M. R. Lieber (1997). "Yeast DNA ligase IV mediates non-homologous DNA end joining", Nature 388: 495-498.

[121] Moore, J. K. & J. E. Haber. (1996) "Cell cycle and genetic requirements of two pathways of nonhomologous end-joining repair of double-strand breaks in Saccharomyces cerevisiae", Molecular Cell Biology 16(5): 2164-2173.

[122] Boulton, S. J. & S. P. Jackson (1996). "Saccharomyces cerevisiae Ku70 potentiates illegitimate DNA double-strand break repair and serves as a barrier to error-prone DNA repair pathways", EMBO Journal 15(18): 5093-5103.

[123] Wilson, T. E. & M. R. Lieber (1999). "Efficient processing of DNA ends during yeast nonhomologous end joining. Evidence for a DNA polymerase beta (Pol4)-dependent pathway", Journal of Biological Chemistry 274: 23599-23609.

[124] Budman, J. & G. Chu (2005). "Processing of DNA for nonhomologous end-joining by cell-free extract", EMBO Journal 24(4): 849-860.

[125] Wang, H., A. R. Perrault, Y. Takeda, W. Qin, H. Wang & G. Iliakis (2003). "Biochemical evidence for Ku-independent backup pathways of NHEJ", Nucleic Acids Research 31(18): 5377-5388.

[126] Jung, D. & F. W. Alt (2004). "Unraveling V(D)J recombination: Insights into gene regulation", Cell 116(2): 299-311.

[127] Zahradka, K., D. Slade, A. Bailone, S. Sommer, D. Averbeck, M. Petranovic, A. B. Lindner & M. Radman (2006). "Reassembly of shattered chromosomes in Deinococcus radiodurans", Nature 443(7111): 569-573.

[128] Friedberg, E. C., G. C. Walker, W. Siede, R. D. Wood, R. A. Schultz & T. Ellenberger (2006). DNA Repair and Mutagenesis. ASM Press.

[129] Bakkenist, C. J. & M. B. Kastan (2003). "DNA damage activates ATM through intermolecular autophosphorylation and dimer dissociation", Nature 421(6922): 499-506.

[130] Wei, Q. Y., L. Li & D. Chen (2007). DNA Repair, Genetic Instability, and Cancer. World Scientific.

[131] Schonthal, A. H. (2004). Checkpoint Controls and Cancer. Humana Press.

[132] Janion, C. (2001). "Some aspects of the SOS response system-a critical survey", Acta Biochimica Polonica 48(3): 599-610.

[133] Schlacher, K., P. Pham, M. M. Cox & M. F. Goodman (2006). "Roles of

DNA polymerase V and RecA protein in SOS damage-induced mutation", Chemical Reviews 106(2): 406-419.

[134] Espejel, S., M. Martin, P. Klatt, J. Martin-Caballero, J. M. Flores & M. A. Blasco (2004). "Shorter telomeres, accelerated ageing and increased lymphoma in DNA-PKcs-deficient mice", EMBO Reports 5(5): 503-509.

[135] De Boer, J., J. O. Andressoo, J. de Wit, J. Huijmans, R. B. Beems, H. van Steeg, G. Weeda, G. T. van der Horst, W. van Leeuwen, A. P. Themmen, M. Meradji & J. H. Hoeijmakers (2002). "Premature aging in mice deficient in DNA repair and transcription", Science 296(5571): 1276-1279.

[136] Dolle, M. E., R. A. Busuttil, A. M. Garcia, S. Wijnhoven, E. van Drunen, L. J. Niedernhofer, G. van der Horst, J. H. Hoeijmakers, H. van Steeg & J. Vijg (2006). "Increased genomic instability is not a prerequisite for shortened lifespan in DNA repair deficient mice", Mutation Research 596(1-2): 22-35.

[137] Kobayashi, Y., I. Narumi, K. Satoh, T. Funayama, M. Kikuchi, S. Kitayama & H. Watanabe (2004). "Radiation response mechanisms of the extremely radioresistant bacterium Deinococcus radiodurans", Biological Sciences in Space 18(3): 134-135.

[138] Spindler, S. R. (2005). "Rapid and reversible induction of the longevity, anticancer and genomic effects of caloric restriction", Mechanisms of Ageing and Development 126(9): 960-966.

[139] Tissenbaum, H. A. & L. Guarente (2001). "Increased dosage of a sir-2 gene extends lifespan in Caenorhabditis elegans", Nature 410(6825): 227-230.

[140] Cohen, H. Y., C. Miller, K. J. Bitterman, N. R. Wall, B. Hekking, B. Kessler, K. T. Howitz, M. Gorospe, R. de Cabo & D. A. Sinclair (2004). "Calorie restriction promotes mammalian cell survival by inducing the SIRT1 deacetylase", Science 305(5682): 390-392.

[141] Cabelof, D. C., S. Yanamadala, J. J. Raffoul, Z. Guo, A. Soofi, A. R. Heydari (2003). "Caloric restriction promotes genomic stability by induction of base excision repair and reversal of its age-related decline", DNA Repair (Amst.) 2(3): 295-307.

[142] Stuart, J. A., B. Karahalil, B. A. Hogue, N. C. Souza-Pinto & V. A. Bohr. (2004). "Mitochondrial and nuclear DNA base excision repair are affected differently by caloric restriction", FASEB Journal 18(3): 595-597.

[143] Walker, D. W., G. McColl, N. L. Jenkins, J. Harris & G. J. Lithgow (2000). "Evolution of lifespan in C. elegans", Nature 405(6784): 296-297.

[144] Cromie, G. A., J. C. Connelly & D. R. Leach (2001). "Recombination at double-strand breaks and DNA ends: conserved mechanisms from phage to humans", Molecular Cell 8(6): 1163-1174.

[145] O'Brien, P. J. (2006). "Catalytic promiscuity and the divergent evolution of DNA repair enzymes", Chemical Reviews 106(2): 720-752.

[146] Maresca, B. & J. H. Schwartz (2006). "Sudden origins: A general mechanism of evolution based on stress protein concentration and rapid environmental change", Anat Rec B New Anat 289(1): 38-46.

[147] Pamela, C. C., A. H. Richard & R. F. Denise (2005). Lippincott's Illustrated Reviews: Biochemistry (3rd edition). Lippincott Williams & Wilkins.

[148] David, L. N. & M. C. Michael (2005). Lehninger Principles of Biochemistry (4th edition). W. H. Freeman.

[149] Ross, J. F. & M. Orlowski (1982). "Growth-rate-dependent adjustment of ribosome function in chemostat-grown cells of the fungus Mucor racemosus", Journal of Bacteriology 149(2): 650-653.