Chapter 1. Introduction

1. Molecular Biology

The term molecular biology was first used in 1945 by William Astbury who was referring to the study of the chemical and physical structure of biological macromolecules. By that time, biochemists had discovered many fundamental intracellular chemical reactions. The importance of specific reactions and of protein structure in defining the numerous properties of cells was also appreciated. However, the development of molecular biology had to await the understanding that the most advantageous approaches would be made by studying “simple” systems such as bacteria and bacteriophages which yield information about the basic biological processes more readily than animal cells. In fact, the faith in the basic uniformity of life processes was an important factor in rapid growth of molecular biology. That is, it was believed that fundamental biological principles that govern the activity of simple organisms, such as bacteria and viruses, must apply to more complex cells; only the details should vary. This faith has been amply justified by experimental results.

The roots of molecular biology were established in 1953 when an Englishman, Francis Crick and a young American, James Watson working at Medical Research Council Unit, Cavendish Laboratory, Cambridge, proposed a double helical model for the structure of DNA (deoxyribonucleic acid) molecule which was well known as the chemical bearer of genetic informations of certain microorganisms (bacteria, bacteriophages, etc.) due to pioneer discoveries made by Grifith (1928), Avery, Macleod and McCarthy (1944) and Hershey and Chase (1952). This discovery was followed by a thorough search of occurrence of DNA as the genetic material in other microorganisms, plants and animals and also by investigations of the molecular and atomic nature of different reactions of living cells. From all these studies has emerged the realization that the basic chemical organization and the metabolic processes of all living things are remarkably similar despite their morphological diversity and that the physical and chemical principles governing living systems are similar to those governing non-living systems.

The present understanding of molecular biology is that in most organisms the phenotype or the body structure and function ultimately depend for their determination on the structural and functional (i.e., enzymatic) proteins or polypeptides. The synthesis of polypeptides is specified, directed and regulated by self-duplicating genes which are borne within molecules of DNA which is the universally accepted chemical bearer of genetic informations of most living organisms except certain viruses in which this function is carried by RNA, another nucleic acid. The genetic informations for polypeptide synthesis are initially dictated by the disposition of nitrogen bases in DNA molecule and are copied down by the process of transcription. During transcription stage copies (that is, transcripts) of an individual gene or genes are synthesized. These copies are molecules of RNA that include such familiar classes as ribosomal RNA, messenger RNA and transfer RNA. The biochemical interplay of these RNA copies which leads to the synthesis of a polypeptide chain, is called translation, meaning, literally, that the genetic message encoded in a messenger RNA molecule is translated into the linear sequence of amino acids in a polypeptide. The polypeptide in its turn determines the phenotype of the organism.

1. HISTORICAL BACKGROUND

The molecular biology is a very young biological discipline and has a very short history. Certain notable accomplishments of molecular biologists can be summarized as follows :

1928 F. Grifth discovered the phenomenon of transformation in bacteria.

1934 M. Schlesinger demonstrated that the bacteriophages are composed of DNA and protein.

1941 G.W. Beadle and E.L.Tatum published their classical study on the biochemical genetics of Neurospora.

1944 O.T. Avery, C.M. MacLeod and M. McCarthy recognized the DNA nature of transforming principle of pneumococcus bacteria. The fact suggested that it is DNA and not protein which is the hereditary chemical.

1948 A. Boivin, R.Vendrely and C. Vendrely showed that in the different cells of an organism the quantity of DNA for each haploid set of chromosome is constant.

1950 E. Chargaff demonstrated that in DNA the numbers of adenine and thymine groups are always equal and so are the numbers of guanine and cytosine groups.

1952 A.D. Hershey and M. Chase demonstrated that only the DNA of T2 bacteriophage enters the host, the bacterium Escherichia coli, whereas the protein remains behind.

1953 J. D. Watson and F.H.C. Crick proposed a model for DNA comprising of two helically intertwined chains tied together by hydrogen bonds between the purines and pyrimidines.

1956 A. Gierer and G. Schramm demonstrated that RNA is the genetic material of tobacco mosaic virus (TMV).

1957 H. Fraenkel-Contrat and B.Singer separated RNA from the protein of TMV viruses, produced hybrid RNA viruses and confirmed the view that RNA is the genetic material of some viruses. Mathew Meselson and Franklin W. Stahl performed a density-gradient experiment (using heavy isotope of nitrogen, 15N) in bacteria to confirm the Watson and Crick’s semiconservative theory of DNA replication.

1958 G. Beadle and E. Tatum received Nobel Prize for their contribution in biochemical genetics of fungus. J. Laderberg got Nobel Prize for the discovery of bacterial recombination.

1959 R.L. Sinsheimer isolated singlestranded DNA from a small virus φ-X- 174 which attacks Escherichia coli. S. Ochoa ; A. Kornberg received Nobel Prize for artificial synthesis of nucleic acids.

1961 M.W. Nirenberg and J.H. Matthaei cracked the messenger RNA code. F.H.C.Crick and his colleagues showed that the genetic language is made up of three-letter words (i.e., triplet codons). F.Jacob and J. Monod put forward the operon concept.

1962 J. Watson and F. Crick ; M. Wilkens got Nobel Prize for the discovery of molecular nature of DNA.

1963 J.P. Waller reported that nearly one-half of all proteins in E.coli cells have the amino acid methionine in the N-terminal position.

1964 K.A. Marcker and F. Sanger discovered a peculiar aminoacyltRNA in E.coli, called N-formyl- methionyl - tRNA and suggested that this molecule may play a role in the special mechanism of chain elongation. R.W. Holley and his colleagues gave detailed structure of alanyl tRNA (tRNA ala) from yeast. Holley died in 1993.

1965 F.H.C. Crick proposed the wobble hypothesis for anticodons of tRNA and explained how several codons meant for same amino acid are recognized by same tRNA. H.

Wallace and M.L. Birnstiel isolated ribosomal RNA genes in Xenopus. F. Jacob., A. Lwoff, and J. Monod received Nobel Prize for the discovery of protein synthesis mechanism in virus.

1968 R.W. Holley ; H.G. Khorana and M.W. Nirenberg got Nobel Prize for deciphering the genetic code.

1969 A.D. Hershey, M. Delbruck and S.E. Luria shared Nobel Prize in medicine for their contribution to replication and recombination in viruses (bacteriophages). Britten and Davidson proposed the gene-battery model for regulation of protein synthesis in eukaryotes.

1970 Howard Temin and David Baltimore demonstrated the synthesis of DNA on RNA template tumour viruses. Both were awarded Nobel Prize in 1975 for the discovery of an enzyme called RNA directed DNA polymerase (or reverse transcriptase) which is present in the core of virus particle (rous sarcoma virus). Biotechnology emerged as a new discipline due to marriage of biological science with technology (see Dubey, 1995). Knippers ; Kornberg and Gefter ; Moses and Richardson isolated DNA-polymerase-II enzyme.

1972 Mertz and Davis in 1972 demonstrated that cohesive termini of cleaved DNA molecule could be covalently sealed with *E.coli* DNA ligase and were able to produce recombinant DNA molecules. Cohen *et al*., for the first time reported the cloning of DNA by using plasmid as vector. R. Porter; G.M. Edelman received Nobel Prize (physiology and medicine) for the discovery of chemical structure of antibodies. C.B. Anfinsen; S. Moore and W.H. Stein got Nobel Prize (chemistry) for the discovery of chemical structure and activity of the enzyme ribonuclease.

1973 S.H. Kim suggested three dimensional structure, i.e., L-shaped model, of tRNA.

1975 E.M. Southern developed a method, called Southern blotting technique for analysing the related genes in a DNA restriction fragment. D. Pribnow discovered Pribnow box or minus ten sequence in E. coli genome.

1977 P.A. Sharp and R.J. Roberts discovered split genes of adenovirus. D.S. Hogness, I.B. David and N. Davidson studied split genes for 28 S rRNA in Drosophila. P. Chambon, P. Leder and R.A. Flavell studied split genes of B'globin, ovalbumin and tRNA. Itakura et al., first of all produced human insulin (humulin) by means of recombinant technology.

1978-79 W. Gilbert first of all used the terms exon and intron (for split genes).

1978 Hinnen et al., first of all described the transformation of yeast (Saccharomyces cervisae) by the help of plasmid of E.coli.

1979 Khorana reported completion of the total synthesis of a biologically functional gene. Alwine et al., developed northern blotting technique in which mRNA bands are blot transferred from the gel onto chemically reactive paper. Towbin et al., developed the western blotting technique to find out the newly encoded protein by a transformed cell.

1980 Fredrick Sanger got the second Nobel Prize for discovering complete sequence of 5400 nucleotides of single stranded DNA of φ × 174 bacteriophage.

1982 A. Klug was awarded Nobel Prize in chemistry for providing three-dimensional structure of tRNAs. Rubin and Spradling for the first time introduced Drosophila gene of xanthine dehydrogenase into a P-element (= parental element) which then was microinjected into embryo deficient for this gene. R.D. Palmiter and R.L. Brinter produced transgenic mice by genetic engineering.

1983 Marilyn Kozak proposed the scanning hypothesis for initiation of translation by eukaryotic ribosomes.

1984 Robert Tijan identified a DNA-binding protein called SP1 which is involved in eukaryotic gene regulation.

1985 Karry Mullis discovered polymerase chain reaction (PCR) which is widely exploited in gene cloning for genetic engineering. He made use of a thermostable enzyme (acts best on 720 C temperature), called Taq DNA polymerase, isolated from Thermus aquaticus.

1984,86 Alec Jeffreys discovered the technique of DNA fingerprinting.

1987 S. Tonegawa was awarded Nobel Prize for discovering the mode of rearrangements of DNA sequences of mammalian immunoglobulin genes to produce a large variety of antibodies. Stanford and coworkers developed the particle bombardment gun which shooted foreign DNA into plant cells or tissues at a very high speed.

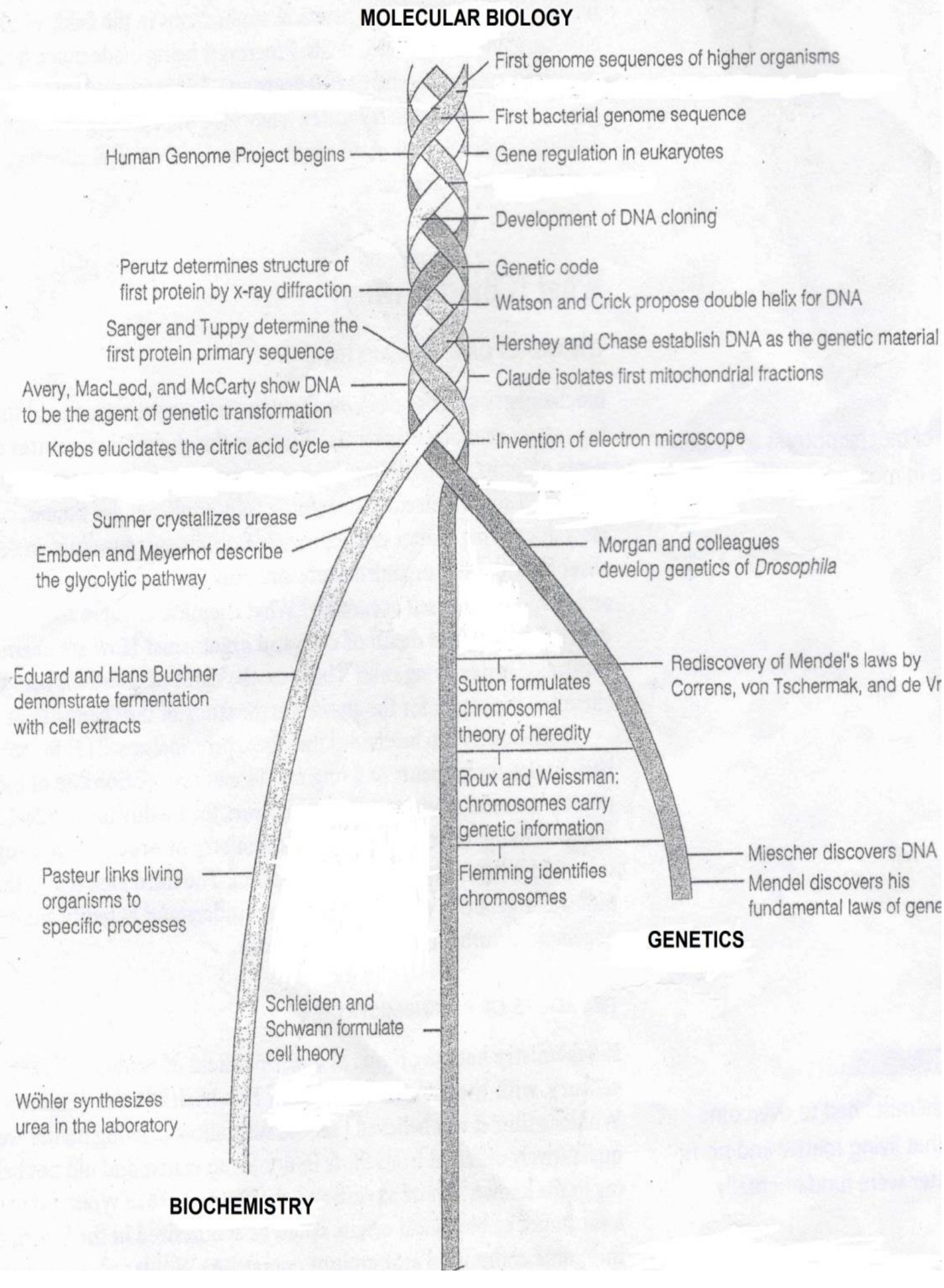
1988 J.W. Black, G.B. Elion and G.H. Hitchings were awarded Nobel Prize for formulating drugs such as 6- mercaptopurine and thioguanine, which lead to inhibition of DNA synthesis and of cell division. This prove effective in cancer chemotherapy. They also designed drugs for treating gout, malaria and viral infections such as herpes.

1989 T. Cech and S. Altman awarded Nobel Prize for showing enzymatic role of some RNA molecules, called ribozymes.

1991 Dr. Lalji Singh at CCMB, Hyderabad has developed a new technique of DNA fingerprinting by using BKM-DNA probe (BKM = banded krait minor satellite). He discovered this probe while he was working on sex determination in snake, the banded krait (Bungarus fasciatus) for his Ph.D. work.

1992 Edwin G. Krebs and Edmond H. Fisher were awarded Nobel Prize for the pioneering work on “reversible protein phosphorylation as a biological regulator mechanism.” Phosphorylation of proteins is shown to affect transcription, translation, cell division and many other cellular processes. Prof. Asis Dutta of JNU, New Delhi, was selected for the Birla Award for Science and Technology for cloning and characterization of two novel genes–gene for oxalate decarboxylase from Lathyrus sativus (in 1991) and gene for a seed specific nutritionally balanced protein from Amaranthus (in 1992).

1993 M.J. Chamberlain proposed the inchworm model for elongation of transcript of DNA template. This year’s Nobel Prize in chemistry was shared by Kary Mullis (for the discovery of PCR) with Michael Smith (for site directed mutagenesis).


MOLECULAR BIOLOGY
First genome sequences of higher organisms
First bacterial genome sequence
Human Genome Project begins
Gene regulation in eukaryotes
Development of DNA cloning
Perutz determines structure of first protein by x-ray diffraction
Genetic code
Watson and Crick propose double helix for DNA
Sanger and Tuppy determine the first protein primary sequence
Hershey and Chase establish DNA as the genetic material
Claude isolates first mitochondrial fractions
Avery, MacLeod, and McCarty show DNA to be the agent of genetic transformation
Krebs elucidates the citric acid cycle
Invention of electron microscope
Sumner crystallizes urease
Embden and Meyerhof describe the glycolytic pathway
Morgan and colleagues develop genetics of *Drosophila*
Eduard and Hans Buchner demonstrate fermentation with cell extracts
Rediscovery of Mendel's laws by Correns, von Tschermak, and de Vr
Sutton formulates chromosomal theory of heredity
Roux and Weissman: chromosomes carry genetic information
Miescher discovers DNA
Pasteur links living organisms to specific processes
Flemming identifies chromosomes
Mendel discovers his fundamental laws of gene
GENETICS
Schleiden and Schwann formulate cell theory
Wöhler synthesizes urea in the laboratory
BIOCHEMISTRY

2. Achievements of Molecular Biology

Asifa Akhtar. There have been a number of important concepts that have emerged. One that particularly jumps to mind is the importance of epigenetics in gene regulation. The field of epigenetics has flourished over the past 10 years. It is clear that chromatin provides an ideal platform for various posttranslational modifications on DNA and histones, which act as a signalling platform for various cellular processes. I also think that the discovery that a combination of four transcription factors can induce a pluripotent state was phenomenal and has stimulated a lot of research in the stem cell field[1]. Last, but not least, the involvement of non-coding RNAs in various cellular and nuclear processes is totally fascinating. The mechanisms by which long non-coding RNAs regulate gene expression await exciting discoveries in the coming years.

Elaine Fuchs. For the stem cell field, there is no question that the findings of Shinya Yamanaka and his co-worker Kazutoshi Takahashi were paradigm-shifting. Their work reported the creation of induced pluripotent stem (iPS) cells from mouse skin fibroblasts when cultured in embryonic stem cell (ESC) conditions[1]. It was remarkable that transient overexpression of a mere four transcription factors, OCT4, SOX2, MYC and Krüppel-like factor 4 (KLF4) — all naturally expressed by ESCs — could achieve this dramatic dedifferentiation of fibroblasts. This finding has allowed researchers to derive patient-tailored iPS cells to study the biology of a host of different human diseases — a first step, but a major one, for the future development of new drugs and treatments in medicine.

Tim Mitchison. Reaction–diffusion gradients specifying positional information inside cells. Gradients of signalling molecules were long known in developmental biology and paracrine physiology. But gradients inside cells being used as a spatial organizing system is a new concept. Bicoid, a classic developmental morphogen, diffuses inside a syncytium, but this is a special case. Gradients of RAN•GTP from mitotic chromatin and of Aurora B activity from chromatin in M phase and midzones in cytokinesis are classic cellular signals that we now know organize space inside cells using a reaction–diffusion mechanism. I attribute this concept to Eric Karsenti, who mooted the idea in the mid 1980s for signals diffusing away from DNA in eggs. However, it wasn't proven until the development of fluorescence resonance energy transfer (FRET)-based activity biosensors in the past decade[2,3]. In general, fluorescence sensors of biochemical activity are a very important development.

Reuben J. Shaw. One area close to our own work is the unexpected re-emergence of metabolism and its relationship to growth control and cancer. Advances in autophagy continue to amaze me in terms of how little basic information we actually have on how a cell works. Autophagy regulators are highly conserved proteins in a central cell biological process that is deregulated in common human diseases, yet much of the biochemical framework for this process has been decoded only recently. Other newly decoded central regulators and processes, ranging from cilia to sirtuins, microRNAs (miRNAs) and pathways such as those involving Hippo and mammalian target of rapamycin (mTOR), underlying so much biology, have changed half of what we know. These are very exciting times.

Daniel St Johnston. Several surprising concepts have emerged during the past decade: first, the amazing extent to which basic cell biological processes have been conserved during the evolution of eukaryotes; second, how much gene regulation is post-transcriptional, particularly through small non-coding RNAs; third, how basic cellular processes, such as endocytic trafficking, microtubule dynamics or mitochondrial behaviour are modulated during the course of normal development; and last, the wide range of cell biological and developmental events that are regulated in response to cellular stresses, such as DNA damage or nutrient deprivation, and how these are used as signals during normal development.

The most important technical advances have been high-throughput sequencing, which has provided the complete sequence of many genomes, and the use of RNA interference (RNAi) to knock down gene function.

Andreas Strasser. One important concept to emerge is the ability to reprogramme differentiated cells, such as fibroblasts or hepatocytes, to assume a pluripotent stem cell fate. Another key finding has been the discovery that signal transducers undergo complex processes of modification by different forms of ubiquitin linkages and that these regulate cellular responses to extracellular signals, such as ligands of the tumour necrosis factor (TNF) family. In addition, an important result has been the discovery that caspase 8 regulates both apoptosis and another cell death process, termed necroptosis. It will now be important to determine the roles of necroptosis in cell death processes that are thought to shape embryonic development but are not affected by mutations that block apoptosis. Moreover, the mechanism by which caspase 8 prevents receptor-interacting protein 1 (RIP1)- and RIP3-mediated necroptosis is now an area of immense interest.

Susan Taylor. Genomic science has transformed the way we think about biology and provides us with a new paradigm for asking biological questions and for thinking about evolution. Sequencing technology has advanced at an extraordinary pace, as has computing, so that sequencing whole genomes is becoming rapid and inexpensive. This has changed the face of biology. The human microbiome and our dramatic co-evolution with microbes is one of the most surprising discoveries to emerge from genome science. In parallel, and also of comparable magnitude, is the recognition of the role that small RNA molecules have and their enormous importance in regulating biology.

Claire E. Walczak. The past 10 years have been remarkable in terms of our understanding of genome organization, chromatin structure and gene expression, which provide the foundation for specifying individual cell function. This information has also provided the basis for many genome-wide studies looking at a multitude of biological processes and disease states. Such studies have provided fundamental new insights into epigenetics, have elucidated a molecular understanding of altered gene regulation in disease, and have enabled fundamental new discoveries, such as RNAi and the existence of miRNAs in the genome.

Marino Zerial. In the past decade, we have progressively shifted our view and approaches drastically towards a genomic perspective. Today, our research of biological processes is no longer focused on single genes or proteins but tends to widen to the complexes, pathways or even systems level. Owing to this change in dimensionality, we have ‘changed gear’ and routinely benefit from comparing species, interrogating genomes and manipulating cells and organisms. This was unthinkable in the 1990s. For example, consider how RNAi has changed our approach to exploring the function of genes. In general, the genomic revolution has disclosed a horizon of interesting problems, such as the role of both coding and non-coding RNAs, to name one.

Q *There has been increasing collaboration between different research communities both within, and outside, cell biology. Where do you think the most interesting interfaces in molecular cell biology reside, and what do you envisage the most fruitful collaborations will be in the future?*

A.A. Indeed, in this post-genomic era, the way we do research has changed dramatically. On average, papers have a more interdisciplinary and collaborative flavour, especially in combining biochemical and genomic analyses. In the future, I can foresee even more fruitful collaborations between cell and molecular biologists and bioinformaticians or even physicists. In fact, I think the next generation of scientists are already on the way who will perform both wet and dry laboratory research equally well.

E.F. I find the interface between human genetics, cell biology and the pharmaceutical and biotechnology industries to be the most exciting. The ability to rapidly sequence many human cancer samples has led to the identification of frequent mutations in particular types of cancer. The advances in small-molecule screening and design have led to fruitful collaborations between basic and pharmaceutical chemists, as they begin to design drugs that target only the mutant form of the protein and not its wild-type counterpart. An example is the recent development of inhibitors against the Val600Glu mutation in BRAF, a frequent mutation in melanomas[4,5]. While tumour resistance still makes eradication problematic, application of this approach to tumour resistance mutations should lead to drugs that can overcome the tumour cell resistance. With a bit of vision towards the future, we can begin to imagine drug cocktails that will make many more cancers treatable.

T.M. The development of new microscopy technology is currently extremely exciting, and has been for two decades or more. This includes new instrumentation, which typically involves physicists, and new probes, which often involves chemists. Super-resolution is one exciting direction (photo-activated localization microscopy (PALM), stochastic optical reconstruction microscopy (STORM) and stimulated emission depletion (STED))[6–8]. Activity biosensors is another[2,3]. Intravital imaging in mice and humans is yet another.

The success of mathematical modelling has been more mixed. I am optimistic that it will help us truly understand collective protein behaviour in the future but, so far, I think the impact has been modest. One big problem is groups saying, "Our model works, therefore we have solved the problem". This is rarely true, and questionable assumptions are often hidden. But we do know that human intuition alone cannot explain collective protein behaviour in complex systems, and there seems to be no alternative to formal modelling. But, we do have to get it better integrated and be more critical.

Finally, DNA sequencing is getting cheaper by the day, and this will have a huge impact. If you can apply this resource to your question, you will make rapid progress. It will also greatly enable work on non-traditional organisms, which opens up all of biology for molecular cell biology approaches.

R.J.S. There have been incredible breakthroughs by technology-driven laboratories with expertise in physics, microscopy and mass spectrometry, which have revolutionized the ability to detect and quantify protein abundance, modifications and interactions, as well as the concentrations of intracellular metabolites that would have been unimaginable 10 years ago. This has also been coupled with advances in RNAi technology, DNA sequencing, ChIP–seq (chromatin immunoprecipitation followed by sequencing) and other techniques that bring high-throughput approaches to cell biology laboratories. Collaborations among adept practitioners using these disparate technologies to decode tissue-specific regulators and rate-limiting pathway components will uncover much fundamental cell biology as well as new targets for many different disease states.

D. St J. Cell biological research is becoming increasingly quantitative, and this has stimulated exciting collaborations with mathematicians and physicists in diverse areas, for example to measure forces in biological systems, to model morphogenetic events and to automate image analysis. Input from the physical sciences has also played an important role in the development of super-resolution microscopy, which has the potential to revolutionize cell biology if and when it can be improved to image faster and deeper inside cells.

A.S. Interactions between bioinformaticians and cellular and molecular biologists have been highly productive, for example by allowing one to make sense out of large data sets, such as gene expression profiles. Interactions between structural biologists, medicinal chemists and cell biologists have allowed us to define complex interactions of proteins in cell signalling, such as the functions of the B cell lymphoma 2 (BCL-2) family members in apoptosis. Importantly, such interdisciplinary interactions have facilitated the development of small molecules to manipulate these processes in a therapeutic setting, such as the treatment of certain cancers.

S.T. Building bridges between computational science and experimental biology is one of our great opportunities and also one of our greatest challenges. There are enormous opportunities for bridging biology with theory and computer science, as well as with translational medicine. Advances in computing have allowed us to gather enormous amounts of data; however, the relevance of the data is compromised without a mechanistic understanding of biological systems. It is absolutely essential to bring these communities together so that we find ways to truly speak a common language. Only in this way can we achieve a comprehensive view of biological systems.

C.E.W. Even only 10 years ago, cell biology was largely a ‘fuzzy’ science, in which we looked at pretty pictures and described what we saw. Quantitative approaches, aided by advances in imaging, have transformed the field to a more quantitative science. The development of mathematical models for biological processes has grown increasingly more complex, offering new insight into protein function. In the future, we need to increase collaboration with chemists and physicists utilizing nanotechnology to visualize and perturb proteins on smaller and smaller scales. Computer scientists are essential to help us organize, process and analyse the large datasets being generated by high-throughput methods.

M.Z. Biological research today makes use of more quantitative approaches than before. For example, imaging and image analysis can be very quantitative, sensitive and precise, and thus are tremendously powerful for exploring biological mechanisms in time and space. This makes the collaboration with theoreticians particularly productive. Biologists need to work with mathematicians, physicists and engineers interested in biological problems because they can help us to understand mechanisms in a more precise and predictive fashion. Previously, our problem was the ability to identify some components that could give us clues into molecular mechanisms. Now that we can get to such components relatively easily (for example, with ‘omics’), our problem becomes how to understand the ways in which the structure and functional properties of biological systems emerge from the interplay of individual components. For this, we need the support of theory.

A.A. The dynamics and quantitative nature of how various pathways and macromolecular complexes function remain poorly understood. We are also beginning to appreciate that spatial and temporal control contribute important regulatory steps in gene regulation. The same molecule in different cellular compartments may have very different regulatory functions, which could be missed during biochemical analyses. If we can gear our research to go from qualitative to quantitative biology and understand the real dynamics of our favourite molecules *in vivo*, we will make a great leap in our understanding of various cellular pathways.

E.F. The most pressing questions in my field are in many ways no different than they were 20–30 years ago, but the answers are closer at hand. How do stem cells build tissues during normal homeostasis and wound repair, and how does this go awry in human diseases, including cancers? And how can we exploit this information to understand the bases of these different diseases and develop new and improved therapies for the treatment of these disorders? With the recombinant DNA technology revolution of the early 1980s and the human genome revolution at the turn of the century, the interface between basic science and medicine is closing at a pace we never imagined possible as students. The tools and technologies available to address fundamental biological questions are advancing at a ferocious rate. The challenge ahead will be to ask the right questions and creatively develop strategies that exploit these tools to bridge this gap and revolutionize medicine.

R.J.S. A big challenge going forward comes out of this explosion of data from different systems: bridging the omics studies (RNAi screens, ChIP–seq, phosphoproteomes and mass spectrometry interactomes) to define what the key rate-limiting proteins in any biological process are. The world still needs careful mechanistic dissection of individual proteins and functions, which sometimes gets lost amidst the push for larger and larger datasets. Taking the findings in cellular systems and then bridging that to the physiology and pathology of diseases in the intact higher organism also remains a key challenge.

D. St J. Most recent cell biology has focused on a relatively small number of cell types (most often, unpolarized, transformed tissue culture cells) and has largely overlooked the astonishing array of different cell types with specialized functions that occur *in vivo*. I think that one of the key challenges for the future is to develop better ways of performing*in vivo* cell biology to examine cellular behaviours in the context of organs and tissues. The ability to induce iPS cells to form organs in culture will be an enormous help for this type of work.

A.S. One challenge is elucidating the precise definition of how cellular differentiation and functional activation are controlled; that is, how the many transcriptional regulators, modifications to the genome (for example, through methylation) and posttranscriptional regulatory processes (for example, through the impact of miRNAs) interact to regulate stepwise changes towards a differentiated state. Another is defining the mechanisms that regulate non-apoptotic, but still genetically programmed, cell death pathways and the definition of their role in normal physiology (for example, during embryonic development and tissue homeostasis in adulthood).

S.T. The biggest challenge for biology is always to ask the right question, and this is even more important now as technologies advance so rapidly. In our frenzy to collect more and more data, we need to learn how to ask the right questions and how to extract useful information from that data. In parallel with systems biology, we must have a mechanistic understanding of biology. Without understanding the underlying biochemical principles, the data mean little. Just as we need classical physiology to understand how molecules work in whole animals, we need biochemistry to have a true mechanistic understanding of biological events.

C.E.W. While the genomic revolution has provided us with a wealth of potentially important molecules, the large-scale functional genomics screens only scratch the surface of understanding the mechanisms by which these proteins act. The challenge is to develop creative approaches to answer the most fundamental biological questions. For example, although proteomic approaches have identified all of the components of the mitotic spindle and genome-wide screens have identified an array of molecules that affect the mitotic spindle, we still do not understand the fundamental mechanism by which each chromosome moves to the spindle equator and then is partitioned to the daughter cells.

M.Z. Cell biology must move to tissues and organisms. An outstanding problem is bridging between scales. Understanding how cellular components form complexes, how these assemble into organelles and how organelles form cells, which build organs and organisms, poses enormous technical and conceptual challenges. The integration of biological processes is one of the most difficult problems we face. Solving these problems requires trespassing across the traditional borders between fields and developing new experimental and analytical methods. At present, we can explain only small parts of biological mechanisms: we see a few pieces of a puzzle, but for the whole picture we must draw in complexity. There are no current solutions at the modelling or computational level. This problem requires the development of new theories.

Q *What would you consider to be the main bottlenecks for the productivity of your research and what advice would you give to young researchers facing these challenges?*

A.A. Despite technical advances, such as high-throughput sequencing strategies, which have been tremendous in providing us with a global view relatively rapidly, the real bottleneck remains the in-depth analysis of data from these strategies and how to make biological sense out of such information. I also think the challenge remains to understand how various biological pathways work mechanistically and, even more importantly, how various pathways are interconnected. These are challenges for all generations of scientists. For young laboratories, one of the major challenges is to hire the right group of people and to focus on a particular biological question. My advice would be to use a multidisciplinary approach to address the questions of interest, as this provides one with the possibility to look at the question from different angles and may reveal unexpected and exciting findings.

E.F. In the United States, the main bottleneck is the precarious funding climate we face and the diminished emphasis our country places on higher education. Our country has spent decades investing in biomedical research and we are now poised to capitalize on this foundation and make major breakthroughs in the coming decades. It is paramount that we work harder to educate policy makers and the public about the time it takes to translate scientific discoveries into cures. I am optimistic that we can do so, and I would encourage young researchers to be optimistic as well, to pursue their passion for science but also to become involved in efforts to communicate with policy makers and the public who hold the strings to our future.

T.M. Funding is a major bottleneck everywhere, and it particularly affects young scientists. One huge challenge is proving our worth to society, which we must do if we expect to be funded by taxpayers' money. The huge progress in molecular cell biology in the past two decades has not translated into a whole lot of improvement in the human condition — for example, new ways of preventing and treating disease. I believe basic science is having, and will have, that impact, but over long time periods and, often, in unpredictable ways. Solving this problem requires that some bright young people (but not all — basic progress is as important as ever) take the road of: first, translating basic research into useful applications, which in my mind includes synthetic biology as well as more conventional ideas like drugs and stem cell-based organ replacements; and second, educating and energizing the public, which includes innovating at all stages of conventional education as well as public outreach, political action and internet-based science activity.

My advice to young scientists is to avoid the road well travelled. There is no surefire route to success in sciences. But if you take the common road of doing what your advisor and her colleagues already do, you are guaranteed to end up in a crowded field in which it is hard to compete for resources and gain independent recognition. You need to take risks — in approach, in system and in questions.

R.J.S. Ironically, in the face of the kinds of technologies that have been developed for cost-effective RNAi screens and full proteomic mass spectrometry analyses of all cellular proteins and metabolites, one unmet need that is still rate-limiting in nearly every area of cell biology research is having tools and reagents that can selectively visualize pools of a given protein with specific post-translational modifications (for example, acetylation, sumoylation and phosphorylation) in intact cells and organisms. My advice to young scientists would be to explore areas at those interfaces between fields and always be incorporating new techniques and ways of thinking about a biological problem.

D. St J. Apart from the usual ones of too much bureaucracy and too little funding, one of the major difficulties we face in the laboratory is how to examine the functions of proteins that have multiple roles in the same cell lineage. Although there are a few specific tricks that allow one to knock out the function of a protein at a specific place or time, most of these actually knock out the gene and not the protein, leaving the problem of perdurance. It would be a huge help to have a standard way of engineering conditional mutant forms of proteins that are either light or heat sensitive.

My advice to young scientists is to not assume that everything that has been published is correct and that it is better to tackle interesting questions that are hard than minor questions that are easy.

A.S. First, I believe that it is important to work on a problem that you are passionate about (that is, for which you really, really want to be the first to know the answers). Second, it is a great joy to work in an environment that is highly collaborative and collegial. Nobody can cover all areas of expertise that are necessary to tackle 'big issues'. Therefore, having ready access, as a cellular or molecular biologist, to structural biologists and bioinformaticians who are eager to collaborate is a great boost for your ability to answer important questions. Finally, when choosing your postdoctoral position, it is in my opinion a good idea to join a group that desperately needs your expertise and has projects and/or techniques on offer that you want to learn. This will create a mutually beneficial or 'symbiotic' relationship, whereas joining a research programme that already has all the expertise that you can offer is like 'shipping coals to Newcastle'.

S.T. Formulate your questions carefully and then delve deeply into your system. Do not be afraid of collaborations and of learning new technologies and new systems. Do not fear reaching out. A major bottleneck for all biological research in the United States now is funding of RO1 investigator-initiated grants. The US National Institutes of Health have nurtured an explosion of biological discoveries over the past few decades, and these discoveries are having an enormous, and often unanticipated, impact on our understanding of disease. We have, unquestionably, been the dominant player in the international community. Other countries, however, are now outpacing us in their funding and we will have trouble maintaining our eminence and dominance.

C.E.W. Carrying out studies that will have a sustained impact requires an ever-increasing amount of multidisciplinary resources, including people, expertise, equipment and funding. Our educational system is not keeping pace with the rapid development of new technologies and still maintains fairly traditional disciplines. This makes it challenging to recruit young scientists who are willing to break new ground to make the exciting new discoveries. My advice to young researchers is to take every opportunity available to learn and discover, and never forget to take time to just think and reflect about what is interesting and cool.

M.Z. My strong advice to young researchers is to think in a truly multidisciplinary way and to go to institutes which can support this approach. Addressing a problem from different sides is essential. Good funding is not everything: I would also encourage young scientists to choose institutes where they can get good mentoring and be stimulated and challenged by faculty members who demonstrate true interest in their work. Importantly, treasure the value of central facilities, available to everyone and capable of supporting research beyond what a single laboratory can do. This kind of support raises the level of ambition and productivity of a young starting group more than any seemingly rich 'start-up package'.

Chapter 2. DNA Structure

4. Nucleic Acids

Nucleic acids are biopolymers, or large biomolecules, essential for all known forms of life. Nucleic acids, which include DNA (deoxyribonucleic acid) and RNA (ribonucleic acid), are made from monomers known as nucleotides. Each nucleotide has three components: a 5-carbon sugar, a phosphate group, and a nitrogenous base. If the sugar is deoxyribose, the polymer is DNA. If the sugar is ribose, the polymer is RNA. When all three components are combined, they form a nucleotide. Nucleotides are also known as phosphate nucleotides.

Nucleic acids are among the most important biological macromolecules (others being amino acids/proteins, sugars/carbohydrates, and lipids/fats). They are found in abundance in all living things, where they function in encoding, transmitting and expressing genetic information in other words, information is conveyed through the nucleic acid sequence, or the order of nucleotides within a DNA or RNA molecule. Strings of nucleotides strung together in a specific sequence are the mechanism for storing and transmitting hereditary, or genetic information via protein synthesis.

Nucleic acids were discovered by Friedrich Miescher in 1869. Experimental studies of nucleic acids constitute a major part of modern biological and medical research, and form a foundation for genome and forensic science, as well as the biotechnology and pharmaceutical industries.

Deoxyribonucleic acid

Deoxyribonucleic acid (DNA) is a nucleic acid containing the genetic instructions used in the development and functioning of all known living organisms. The DNA segments carrying this genetic information are called genes. Likewise, other DNA sequences have structural purposes, or are involved in regulating the use of this genetic information. Along with RNA and proteins, DNA is one of the three major macromolecules that are essential for all known forms of life. DNA consists of two long polymers of simple units called nucleotides, with backbones made of sugars and phosphate groups joined by ester bonds. These two strands run in opposite directions to each other and are, therefore, anti-parallel. Attached to each sugar is one of four types of molecules called nucleobases (informally, bases). It is the sequence of these four nucleobases along the backbone that encodes information. This information is read using the genetic code, which specifies the sequence of the amino acids within proteins. The code is read by copying stretches of DNA into the related nucleic acid RNA in a process called transcription. Within cells DNA is organized into long structures called chromosomes. During cell division these chromosomes are duplicated in the process of DNA replication, providing each cell its own complete set of chromosomes. Eukaryotic organisms (animals, plants, fungi, and protists) store most of their DNA inside the cell nucleus and some of their DNA in organelles, such as mitochondria or chloroplasts. In contrast, prokaryotes (bacteria and archaea) store their DNA only in the cytoplasm. Within the chromosomes, chromatin proteins such as histones compact and organize DNA. These compact structures guide the interactions between DNA and other proteins, helping control which parts of the DNA are transcribed.

Ribonucleic acid

Ribonucleic acid (RNA) functions in converting genetic information from genes into the amino acid sequences of proteins. The three universal types of RNA include transfer RNA (tRNA), messenger RNA (mRNA), and ribosomal RNA (rRNA).Messenger RNA acts to carry genetic sequence information between DNA and ribosomes, directing protein synthesis.Ribosomal RNA is a major component of the ribosome, and catalyzes peptide bond formation. Transfer RNA serves as the carrier molecule for amino acids to be used in protein synthesis, and is responsible for decoding the mRNA. In addition, many other classes of RNA are now known.

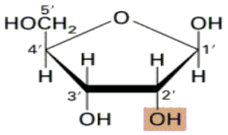
Artificial nucleic acid analogs

Artificial nucleic acid analogues have been designed and synthesized by chemists, and include peptide nucleic acid,morpholino- and locked nucleic acid, as well as glycol nucleic acid and threose nucleic acid. Each of these is distinguished from naturally occurring DNA or RNA by changes to the backbone of the molecule.

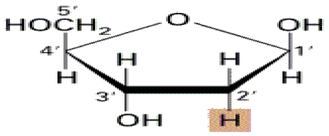
5. Chemical composition of DNA

The chemical DNA was first discovered in 1869, but its role in genetic inheritance was not demonstrated until 1943. In 1953James Watson and Francis Crick determined that the structure of DNA is a double-helix polymer, a spiral consisting of two DNA strands wound around each other. Each strand is composed of a long chain of monomer nucleotides. Thenucleotide of DNA consists of a deoxyribose sugar moleculeto which is attached a phosphate group and one of four nitrogenous bases: two purines (adenine and guanine) and two pyrimidines (cytosine and thymine). The nucleotides are joined together by covalent bonds between the phosphate of one nucleotide and the sugar of the next, forming a phosphate-sugar backbone from which the nitrogenous bases protrude. One strand is held to another by hydrogen bonds between the bases; the sequencing of this bonding is specific—i.e., adenine bonds only with thymine, and cytosineonly with guanine.

Deoxyribose (a pentose sugar derivative)

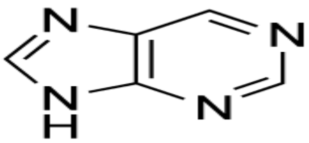


Ribose

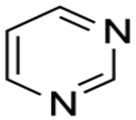


2-Deoxyribose

Nitrogenous Bases

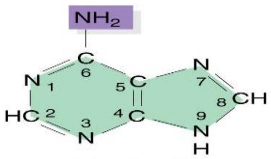


purine

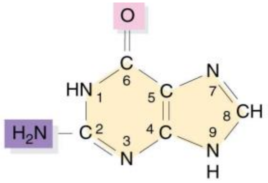


pyrimidine

Purines

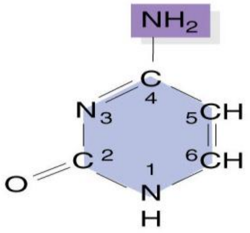


Adenine (A)

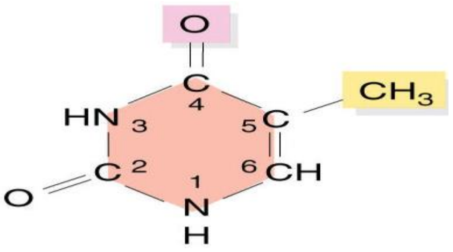


Guanine (G)

Pyrimidines

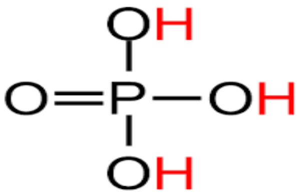


Cytosine (C)



Thymine (T)

Phosphoric acid



H_3PO_4

6. Nucleoside & Nucleotide

A nucleoside consists of a nitrogenous base covalently attached to a sugar (ribose or deoxyribose) but without the phosphate group. A nucleotide consists of a nitrogenous base, a sugar (ribose or deoxyribose) and one to three phosphate groups.

Nucleoside=Sugar+Base

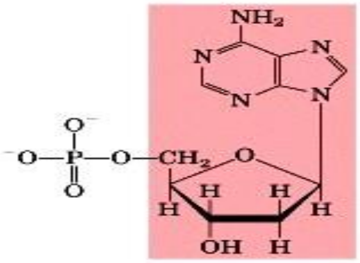
Nucleotide = Sugar + Base + Phosphate

Comparison chart

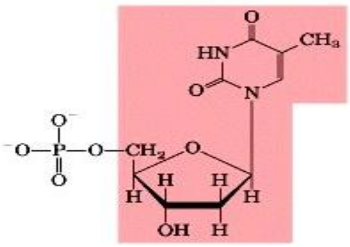
	Nucleoside	Nucleotide
Relevance in medicine	Several nucleoside analogues are used as antiviral or anticancer agents.	Malfunctioning nucleotides are one of the main causes of all cancers known of today.
Examples	Examples of nucleosides include cytidine, uridine, adenosine, guanosine, thymidine and inosine.	Nucleotides follow the same names as nucleosides, but with the indication of phosphate groups. For example, 5'-uridine monophosphate.
Chemical Composition	Sugar + Base. A nucleoside consists of a nitrogenous base covalently attached to a sugar (ribose or deoxyribose) but without the phosphate group. When phosphate group of nucleotide is removed by hydrolysis, the structure remaining is nucleoside.	Sugar + Base + Phosphate. A nucleotide consists of a nitrogenous base, a sugar (ribose or deoxyribose) and one to three phosphate groups.

7. Types of Deoxyribonucleotides

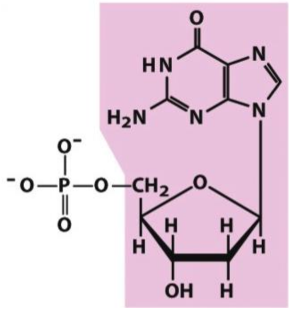
There are four types of Deoxy-ribonucleotides with respect to Nitrogen bases



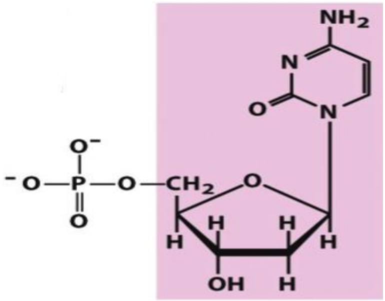
Deoxyadenylate



Deoxythymidylate



Deoxyadenylate



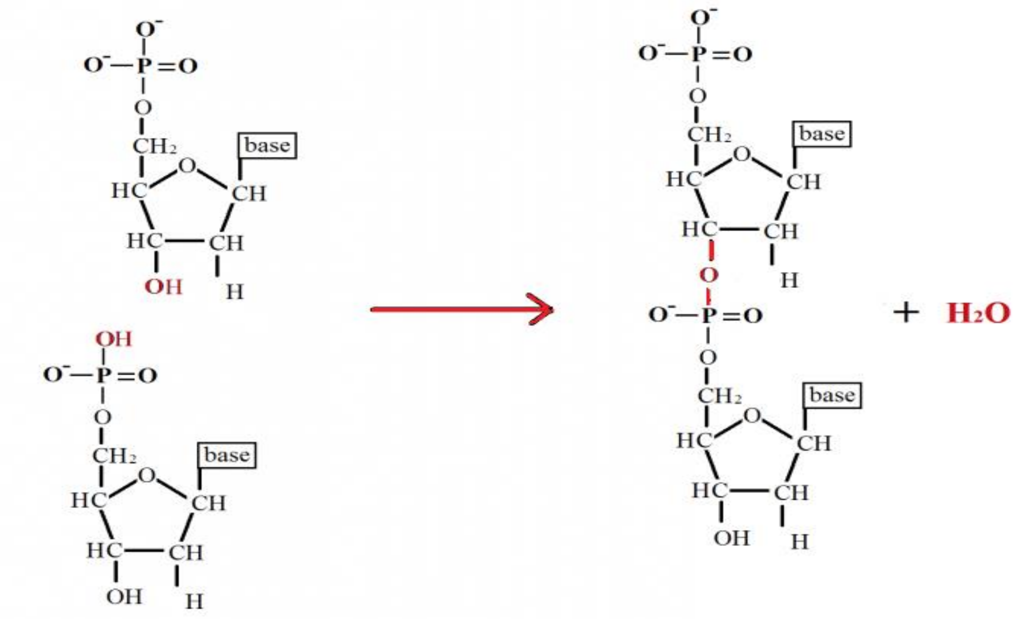
Deoxythymidylate

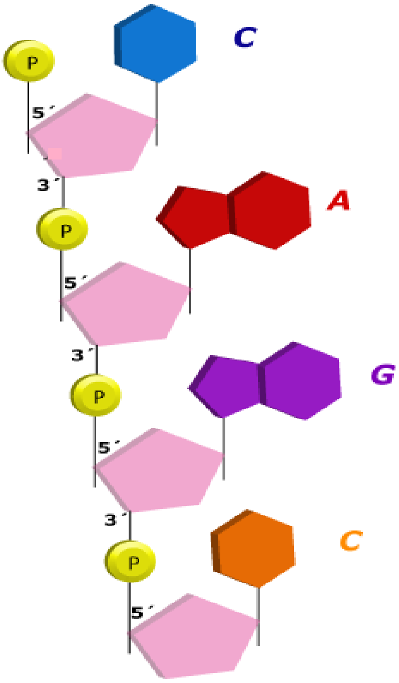
8. How do Deoxyribonucleotides Join?

A deoxyribonucleotide is the monomer, or single unit, of DNA, or deoxyribonucleic acid. Each deoxyribonucleotide comprises three parts: a nitrogenous base, a deoxyribose sugar, and one phosphate group. The nitrogenous base is always bonded to the 1' carbon of the deoxyribose, which is distinguished from ribose by the presence of a proton on the 2' carbon rather than an OH group. The phosphate groups bind to the 5' carbon of the sugar.

When deoxyribonucleotides polymerize to form DNA, the phosphate group from one nucleotide will bond to the 3' carbon on another nucleotide, forming a phosphodiester bond via dehydration synthesis. New nucleotides are always added to the 3' carbon of the last nucleotide, so synthesis always proceeds from 5' to 3'.

Joining of Deoxyribonucleotides





9. Structure of DNA

Miescher 's DNA structure

Although few people realize it, 1869 was a landmark year in genetic research, because it was the year in which Swiss physiological chemist Friedrich Miescher first identified what he called "nuclein" inside the nuclei of human white blood cells. (The term "nuclein" was later changed to "nucleic acid" and eventually to "deoxyribonucleic acid," or "DNA.") Miescher's plan was to isolate and characterize not the nuclein (which nobody at that time realized existed) but instead the protein components of leukocytes (white blood cells). Miescher thus made arrangements for a local surgical clinic to send him used, pus-coated patient bandages; once he received the bandages, he planned to wash them, filter out the leukocytes, and extract and identify the various proteins within the white blood cells. But when he came across a substance from the cell nuclei that had chemical properties unlike any protein, including a much higher phosphorous content and resistance to proteolysis (protein digestion), Miescher realized that he had discovered a new substance (Dahm, 2008). Sensing the importance of his findings, Miescher wrote, "It seems probable to me that a whole family of such slightly varying phosphorous-containing substances will appear, as a group of nucleins, equivalent to proteins" (Wolf, 2003).

More than 50 years passed before the significance of Miescher's discovery of nucleic acids was widely appreciated by the scientific community. For instance, in a 1971 essay on the history of nucleic acid research, Erwin Chargaff noted that in a 1961 historical account of nineteenth-century science, Charles Darwin was mentioned 31 times, Thomas Huxley 14 times, but Miescher not even once. This omission is all the more remarkable given that, as Chargaff also noted, Miescher's discovery of nucleic acids was unique among the discoveries of the four major cellular components (i.e., proteins, lipids, polysaccharides, and nucleic acids) in that it could be "dated precisely... [to] one man, one place, one date."

Levene Investigates the Structure of DNA

Meanwhile, even as Miescher's name fell into obscurity by the twentieth century, other scientists continued to investigate the chemical nature of the molecule formerly known as nuclein. One of these other scientists was Russian biochemist Phoebus Levene. A physician turned chemist, Levene was a prolific researcher, publishing more than 700 papers on the chemistry of biological molecules over the course of his career. Levene is credited with many firsts. For instance, he was the first to discover the order of the three major components of a single nucleotide (phosphate-sugar-base); the first to discover the carbohydrate component of RNA (ribose); the first to discover the carbohydrate component of DNA (deoxyribose); and the first to correctly identify the way RNA and DNA molecules are put together.

During the early years of Levene's career, neither Levene nor any other scientist of the time knew how the individual nucleotide components of DNA were arranged in space; discovery of the sugar-phosphate backbone of the DNA molecule was still years away. The large number of molecular groups made available for binding by each nucleotide component meant that there were numerous alternate ways that the components could combine. Several scientists put forth suggestions for how this might occur, but it was Levene's "polynucleotide" model that proved to be the correct one. Based upon years of work using hydrolysis to break down and analyze yeast nucleic acids, Levene proposed that nucleic acids were composed of a series of nucleotides, and that each nucleotide was in turn composed of just one of four nitrogen-containing bases, a sugar molecule, and a phosphate group. Levene made his initial proposal in 1919, discrediting other suggestions that had been put forth about the structure of nucleic acids. In Levene's own words, "New facts and new evidence may cause its alteration, but there is no doubt as to the polynucleotide structure of the yeast nucleic acid" (1919).

Indeed, many new facts and much new evidence soon emerged and caused alterations to Levene's proposal. One key discovery during this period involved the way in which nucleotides are ordered. Levene proposed what he called a tetranucleotide structure, in which the nucleotides were always linked in the same order (i.e., G-C-T-A-G-C-T-A and so on). However, scientists eventually realized that Levene's proposed tetranucleotide structure was overly simplistic and that the order of nucleotides along a stretch of DNA (or RNA) is, in fact, highly variable. Despite this realization, Levene's proposed polynucleotide structure was accurate in many regards. For example, we now know that DNA is in fact composed of a series of nucleotides and that each nucleotide has three components: a phosphate group; either a ribose (in the case of RNA) or a deoxyribose (in the case of DNA) sugar; and a single nitrogen-containing base. We also know that there are two basic categories of nitrogenous bases: the purines (adenine [A] and guanine [G]), each with two fused rings, and the pyrimidines (cytosine [C], thymine [T], and uracil [U]), each with a single ring. Furthermore, it is now widely accepted that RNA contains only A, G, C, and U (no T), whereas DNA contains only A, G, C, and T (no U) (Figure 1).

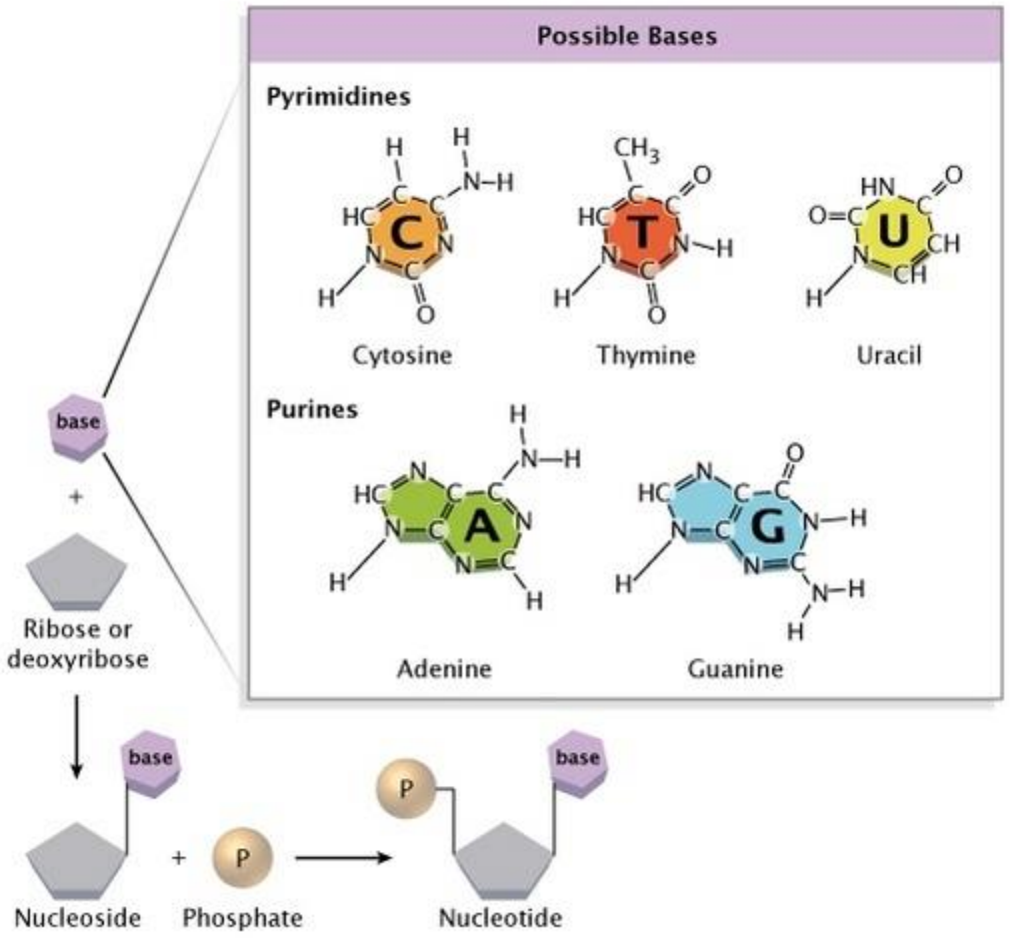


Figure : The chemical structure of a nucleotide.

Chargaff Rules for DNA Structure

Erwin Chargaff was one of a handful of scientists who expanded on Levene's work by uncovering additional details of the structure of DNA, thus further paving the way for Watson and Crick. Chargaff, an Austrian biochemist, had read the famous 1944 paper by Oswald Avery and his colleagues at Rockefeller University, which demonstrated that hereditary units, or genes, are composed of DNA. This paper had a profound impact on Chargaff, inspiring him to launch a research program that revolved around the chemistry of nucleic acids. Of Avery's work, Chargaff (1971) wrote the following:

"This discovery, almost abruptly, appeared to foreshadow a chemistry of heredity and, moreover, made probable the nucleic acid character of the gene... *Avery gave us the first text of a new language, or rather he showed us where to look for it. I resolved to search for this text."*

As his first step in this search, Chargaff set out to see whether there were any differences in DNA among different species. After developing a new paper chromatography method for separating and identifying small amounts of organic material, Chargaff reached two major conclusions (Chargaff, 1950). First, he noted that the nucleotide composition of DNA varies among species. In other words, the same nucleotides do not repeat in the same order, as proposed by Levene. Second, Chargaff concluded that almost all DNA--no matter what organism or tissue type it comes from--maintains certain properties, even as its composition varies. In particular, the amount of adenine (A) is usually similar to the amount of thymine (T), and the amount of guanine (G) usually approximates the amount of cytosine (C). In other words, the total amount of purines (A + G) and the total amount of pyrimidines (C + T) are usually nearly equal. (This second major conclusion is now known as "Chargaff's rule.") Chargaff's research was vital to the later work of Watson and Crick, but Chargaff himself could not imagine the explanation of these relationships--specifically, that A bound to T and C bound to G within the molecular structure of DNA (Figure 2).

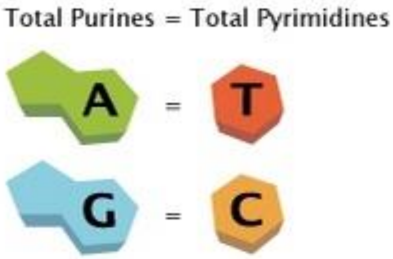


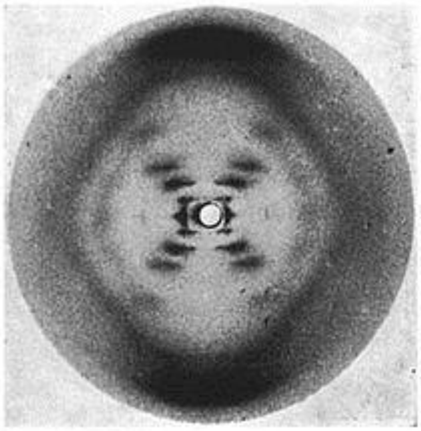
Figure : What is Chargaff's rule?

10. Wilkins and Franklin work on DNA

Maurice Wilkins and Rosalind Franklin, together with Ray Gosling, Alec Stokes and Herbert Wilson and other colleagues at the Randall Institute at King's, made crucial contributions to the discovery of DNA's structure in 1953.

Wilkins began using optical spectroscopy to study DNA in the late 1940s. In 1950 he and Gosling obtained the first clearly crystalline X-ray diffraction patterns from DNA fibres. Alec Stokes suggested that the patterns indicated that DNA was helical in structure.

The discovery of the structure of DNA in 1953 revealed the physical and chemical basis of how characteristics are passed down through the generations and how they are expressed in individual organisms.



X-ray diffraction pattern of DNA

11. Watson and Crick Model of DNA Structure

Chargaff's realization that A = T and C = G, combined with some crucially important X-ray crystallography work by English researchers Rosalind Franklin and Maurice Wilkins, contributed to Watson and Crick's derivation of the three-dimensional, double-helical model for the structure of DNA. Watson and Crick's discovery was also made possible by recent advances in model building, or the assembly of possible three-dimensional structures based upon known molecular distances and bond angles, a technique advanced by American biochemist Linus Pauling. In fact, Watson and Crick were worried that they would be "scooped" by Pauling, who proposed a different model for the three-dimensional structure of DNA just months before they did. In the end, however, Pauling's prediction was incorrect.

Using cardboard cutouts representing the individual chemical components of the four bases and other nucleotide subunits, Watson and Crick shifted molecules around on their desktops, as though putting together a puzzle. They were misled for a while by an erroneous understanding of how the different elements in thymine and guanine (specifically, the carbon, nitrogen, hydrogen, and oxygen rings) were configured. Only upon the suggestion of American scientist Jerry Donohue did Watson decide to make new cardboard cutouts of the two bases, to see if perhaps a different atomic configurationwould make a difference. It did. Not only did the complementary bases now fit together perfectly (i.e., A with T and C with G), with each pair held together by hydrogen bonds, but the structure also reflected Chargaff's rule (Figure 3).

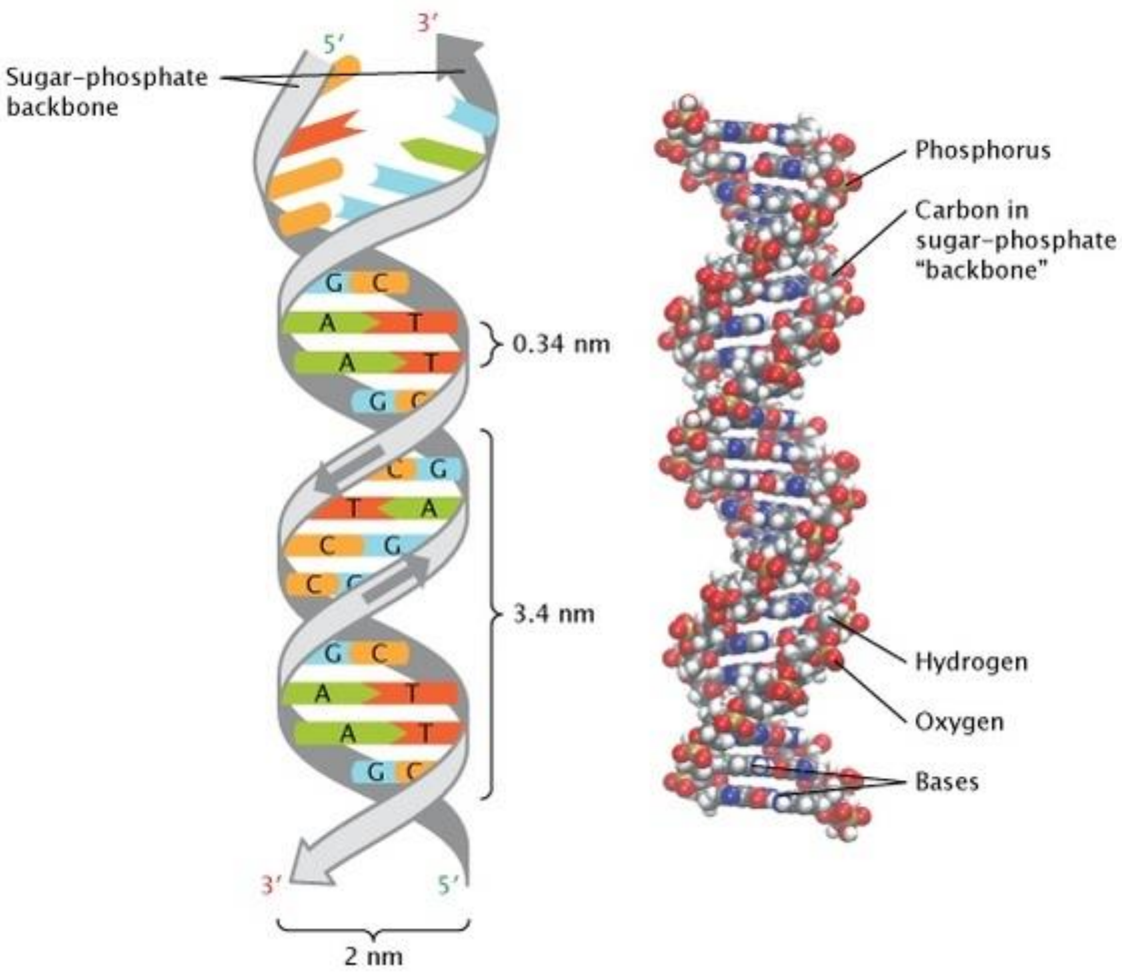


Figure : The double-helical structure of DNA.

Chapter 3. RNA Structure

12. Chemical Composition of RNA

Usually ribonucleic acid (RNA) is single stranded and made up of long, unbranched polynucleotide chain. The polynucleotide chain is formed by joining of ribonucleotides, with the help of 3' – 5' phosphodiester bonds in the same fashion as in case of DNA. But RNA is more stable than DNA because of intermolecular pairing.

Ribonucleotides = Pentose sugar (ribose) + N-base + phosphate group
Nitrogen bases are of two types

Purines

Adenine (A)
Guanine (G)

Pyrimidines

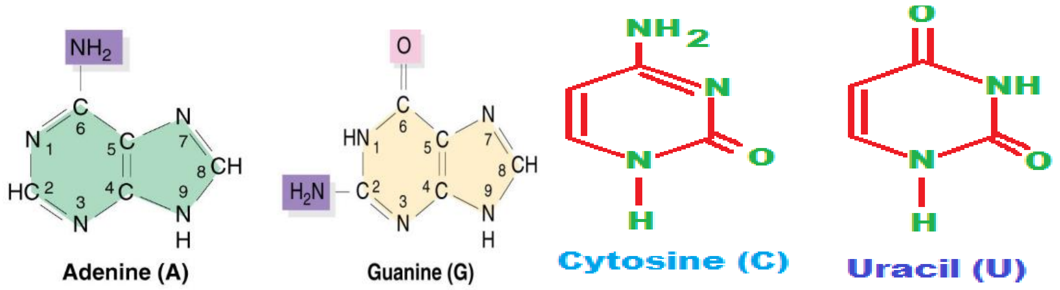
Cytosine (C)
Uracil (U)

Many RNAs possess number of minor bases in addition to above four bases, so there are unusual nucleotides like pseudouridine, inosine, dihydroxyuridine etc.

Following table summarizes N – bases, ribonucleosides and ribonucleotides.

N – base	Ribonucleoside	Ribonucleotide	Abbreviation
Adenine	Adenosine	Adenosine monophosphate (Adenylic acid)	AMP
Guanine	Guanosine	Guanosine monophosphate (Guanylic acid)	GMP
Cytosine	Cytidine	Cytidine monophosphate (Cytidylic acid)	CMP
Uracil	Uridine	Uridine monophosphate (Uridylic acid)	UMP

Only difference between DNA and RNA chemical composition is RNA has ribose sugar, has hydroxyl group (- OH) at the 2’ position and N-base thymine is replaced by uracil.



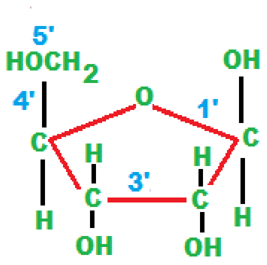
Adenine (A)

Guanine (G)

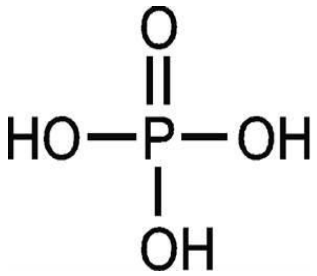
Cytosine (C)

Uracil (U)

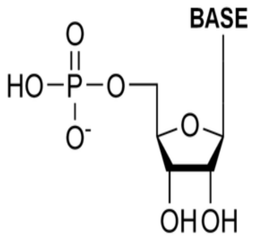
Nitrogenous Bases



Ribose (a pentose sugar)



Phosphoric acid

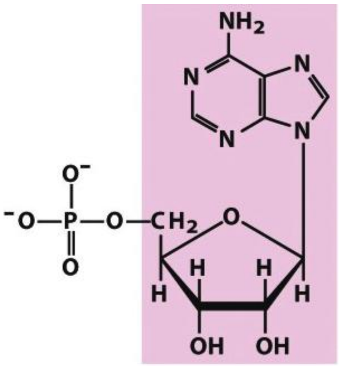


A Ribonucleotide

13. Types of Ribonucleotides

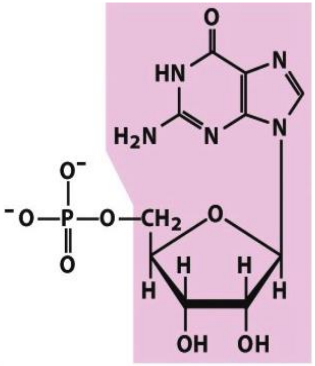
There are mainly four types of ribonucleotides depending upon the types of nitrogenous bases present in RNA.

Nucleotides	Symbols	Nucleoside
Adenylate (adenosine 5'-monophosphate)	A, AMP	Adenosine
Guanylate (guanosine 5'-monophosphate)	G, GMP	Guanosine
Uridylate (uridine 5'-monophosphate)	U, UMP	Uridine
Cytidylate (cytidine 5'-monophosphate)	C, CMP	Cytidine



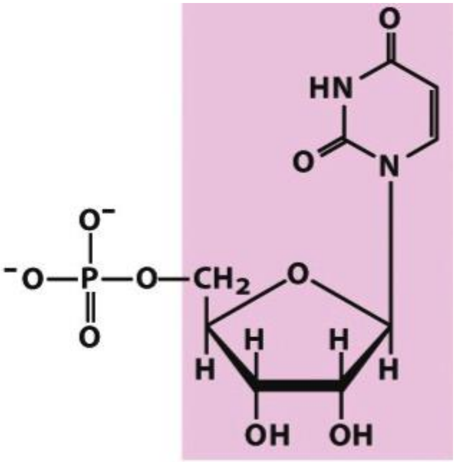
Adenylate (adenosine 5′-monophosphate)

Adenosine



Guanylate (guanosine 5′-monophosphate)

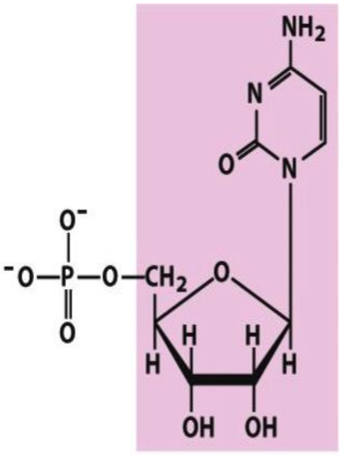
Guanosine



Uridylate (uridine 5′-monophosphate)

U, UMP

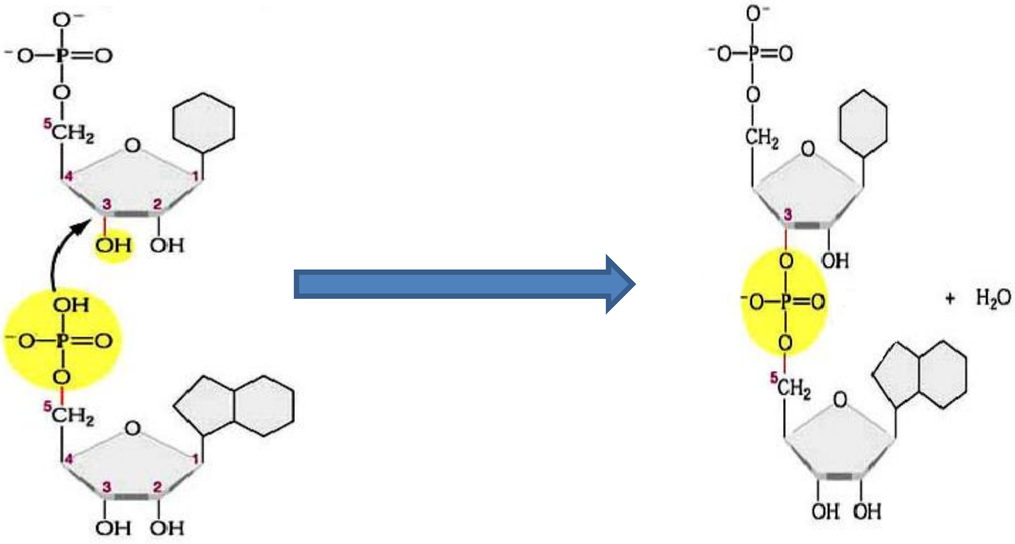
Uridine



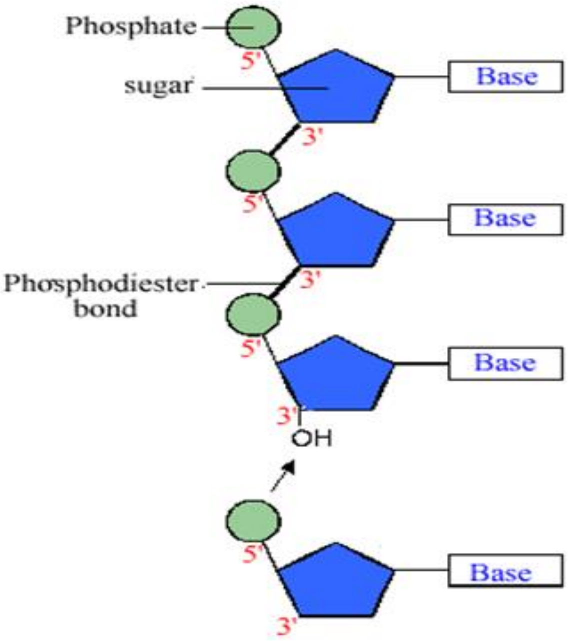
Cytidylate (cytidine 5′-monophosphate)

C, CMP

Cytidine



Joining of Ribonucleotides



A Poly-Ribonucleotide

14. Types of RNAs

There are actually several types of ribonucleic acids or RNAs, but mainly three types of Ribonucleic acids (RNAs) present in the cells of living organisms.

1. Messenger RNA (mRNA)
2. Transfer RNA (tRNA)
3. Ribosomal RNA (rRNA)

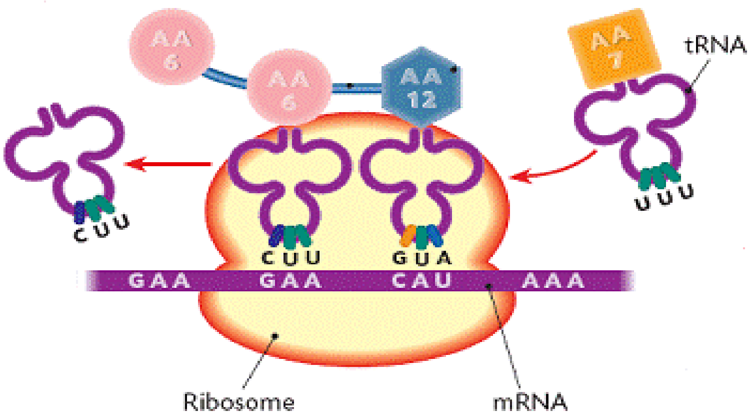
Messenger RNA (mRNA)

- It is the type of RNA that carries genetic information from DNA to the protein biosynthetic machinery of the ribosome.
- It provides the templates that specify amino acid sequences in polypeptide chains.
- The process of forming mRNA on a DNA template is known as transcription.
- It may be monocistronic or polycistronic.
- The length of mRNA molecules is variable and it depends on the length of gene.

Transfer RNA (tRNA)

- Transfer RNAs serve as adapter molecules in the process of protein synthesis.
- They are covalently linked to an amino acid at one end.

- They pair with the mRNA in such a way that amino acids are joined to a growing polypeptide in the correct sequence.

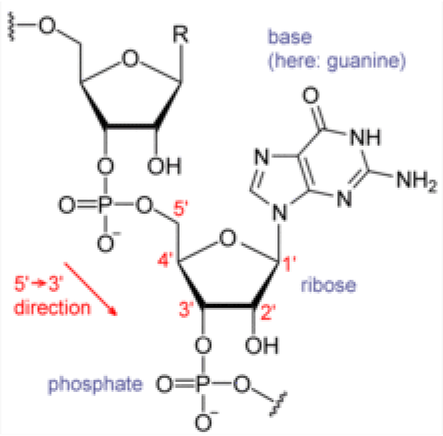


Ribosomal RNA (rRNA)

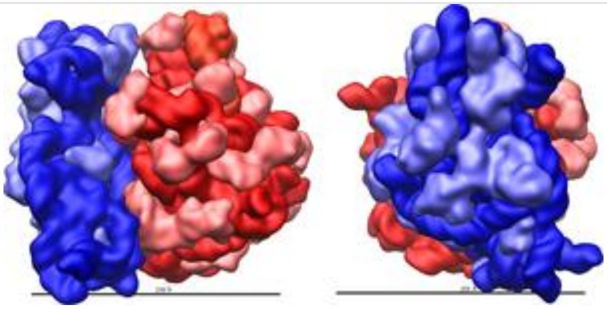
- Ribosomal RNAs are components of ribosomes.
- rRNA is a predominant material in the ribosomes constituting about 60% of its weight.
- It has a number of functions to perform in the ribosomes.

15. Structures of RNAs

Each nucleotide in RNA contains a ribose sugar, with carbons numbered 1' through 5'. A base is attached to the 1' position, in general, adenine (A), cytosine(C), guanine (G), or uracil (U). Adenine and guanine are purines, cytosine and uracil are pyrimidines. A phosphate group is attached to the 3' position of one ribose and the 5' position of the next. The phosphate groups do not have a negative charge each at physiological pH, making RNA a charged molecule (polyanion). The bases form hydrogen bonds between cytosine and guanine, between adenine and uracil and between guanine and uracil.[8] However, other interactions are possible, such as a group of adenine bases binding to each other in a bulge,[9] or the GNRA tetraloop that has a guanine–adenine base-pair.



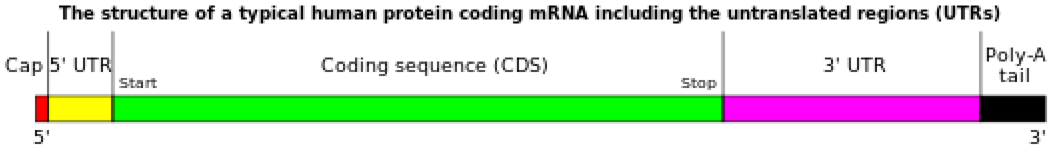
Ribosomal RNA



Ribosomal RNA (rRNA) is the RNA component of a ribosome

Ribosomes are non membranous organelles that participate in the translation of mRNA into a protein product. The ribosome structure is composed of 2 subunits. A small and a large subunit each of which primarily consists of rRNA of various size and a small quantity of proteins. rRNA constitutes about the 80% of the whole RNA present in an eukaryotic cell. The large subunit consists of rRNA of 5S, 5.8S and 28S sizes whereas the small subunit consists of rRNA of 18S size. (where S is the unit for rRNA size). These rRNAs are synthesized by transcription of the rRNA genes. However, the rRNA genes encode for all rRNAs except from the 5S rRNA, which is synthesized by the tRNA genes along with all the nuclear tRNAs. RNA polymerase type I is responsible for the transcription of rRNA genes by binding on the core element−CE, which overlaps the Transcription Start Site−TSS, along with the Transcription Factors inducing the so called Transcription Initiation Complex designated as TIC. The rate of the transcription is controlled by an Upstream Control Sequence−UCS located 100 base pairs upstream to the TSS. The transcription process begins and the genes are transcribed into pre−rRNA in the following order as situated on the gene: -18S - 5.8S - 28S-. The transcription comes to an end when the Transcription Complex reaches an area rich in Adenines found at about 600 base pairs downstream of the gene, indicating its end. The pre−rRNA formed includes all rRNAs on a single strand so that cleavage has to be performed so that different size rRNAs are separated. This task is untertaken by RNases that cleave the rRNA giving rise to the differential size rRNAs.

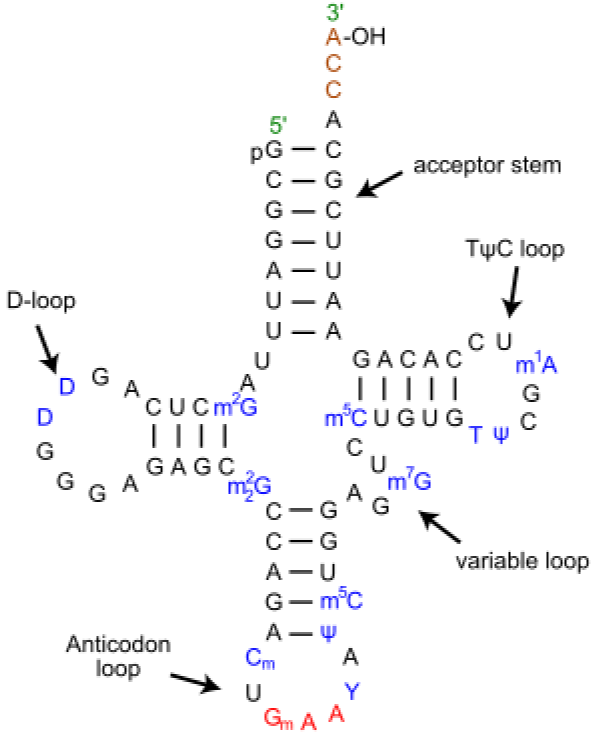
Messenger RNA


The structure of a typical human protein coding mRNA including the untranslated regions (UTRs)
Cap
5' UTR
Start
Coding sequence (CDS)
Stop
3' UTR
Poly-A tail
5'
3'

The structure of a mature eukaryotic mRNA. A fully processed mRNA includes a 5' cap, 5' UTR,coding region, 3' UTR, and poly(A) tail.

mRNA genes are the genes that encode only for proteins but this encoding has an RNA intermediate. The DNA is firstly transcribed into mRNA and subsequently translated into a protein product. So the mRNA genes are the genes that encode for mRNA in order to synthesize proteins. mRNA constitutes only the 5% of the total RNA. RNA polymerase II is the enzyme responsible for the transcription of the corresponding genes into mRNA. The polymerase binds on the TATA box which acts more or less as a promoter, located about 25 base pairs upstream the Transcription Start Site−TSS, along with the transcription factors giving rise to the Transcription Initiation Complex−TIC. In order for this complex to be functional a proper sequence of events of binding the TF and the polymerase on the promoter must occur as: TFII-D, TFII-A, TFII-B, RNA pol II, TFII-F, TFII-H TFII-E TFII-J. As soon as the TIC is formed then transcription begins giving rise to pre−mRNA which include both exons and introns. Transcription ends without recognition of an adenine rich area but rather by automatic disassembling of the Transcription Complex. The pre−mRNA is then submitted to processing that involves splicing - removal of introns and merging of the adjacent exons- and capping - addition of 7−methylguanosine on the 5' end of the mRNA so that mRNA cannot be cleaved by exonucleases. It also serves as a recognition site of the mRNA prior to translation for the small ribosomal subunit.

Transfer RNA



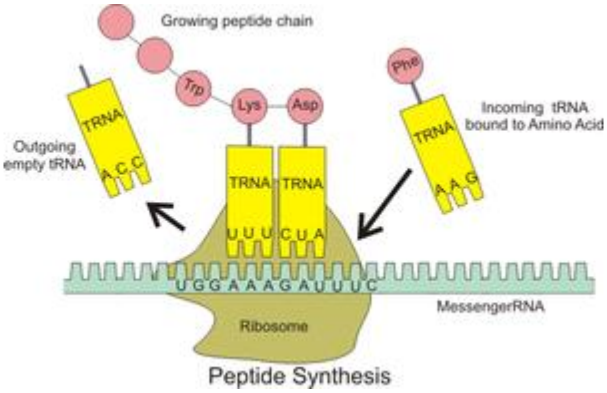
Secondary *cloverleaf structure* of $tRNA^{Phe}$ from yeast.

Transfer RNA is encoded by genes that also encode for the 5S size rRNA. RNA polymerase III is responsible for the transcription of these genes by binding on the promoter, situated about 100 base pairs downstream the Transcription Start Site -TSS, along with the Transcription Factors giving rise to the Transcription Initiation Complex. As soon as this complex is formed transcription process can begin and when the Transcription Complex faces an Adenine rich region transcription comes to an end as this area is an indication for the gene end. tRNA constitutes 15% of the total RNA and is directly involved in the translation of the mRNA. More specificaly tRNA binds onto a specific amino acid and brings it along the translation site so that it is bound on the newly synthesized peptide.

- tRNA binds to its specific amino acid recognized by its side R chain in presence of the aminoacyl tRNA synthetase enzyme. The synthetase binds the 5'-CCA-OH-3' acceptor arm with the —COOH group of the amino acid.
- When the small ribosomal subunit faces an AUG codon on the mRNA it indicates the commencing of the peptide formation. As soon as the AUG codon is recognized then the first tRNA binds on the small ribosomal subunit and on the mRNA through its anticodon arm, giving rise to the Translation Initiation Complex designated as $tRNA_i^{met}$. Eventually the large ribosomal

subunit binds on the complex indicating the initiation of the translation process. Translation always begins with the methionine amino acid on the newly synthesized peptide.

- After translocation of the Translation complex the $tRNA_i^{met}$ enter the Peptidyl site of the complex leaving the Aminoacyl site vacant for the next tRNA to enter, bringing together the two adjacent amino acids so that a peptide bond can be formed in presence of the peptidyl transferase enzyme. As soon as the peptide bond is formed, the tRNA is released from its amino acid in presence of the tRNA deacylase.
-



The interaction of tRNA and mRNA in protein synthesis

Chapter 4. DNA a genetic material

16. Nature of Genetic Material

After establishment of the fact that genes are the physical units located on the chromosomes. A major problem for the biologists was to find out the molecules responsible for carrying the hereditary information.

Characteristics of Genetic Material
Genetic material must contain complex information.
Genetic material must replicate faithfully.
Genetic material must encode phenotype.

Three sets of experiments provided a pivotal evidence that DNA rather than protein, is the hereditary material.

Griffith's Experiments (1928)
Avery's Experiments (1944)
Hershey-Chase experiments (1952)

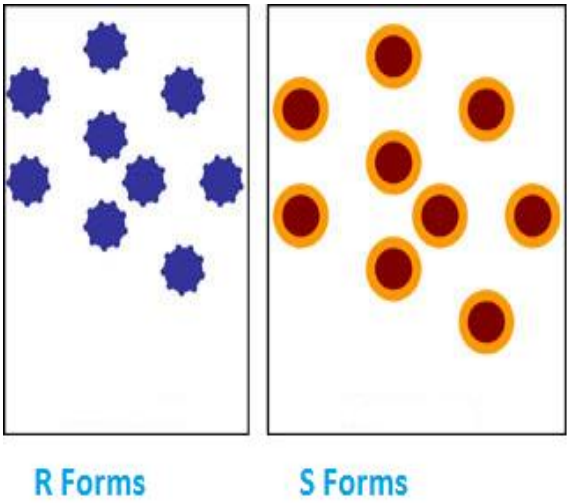
17. Griffith's Experiments (1928)

Griffith's experiment, reported in 1928 byFrederick Griffith,[1] was the first experiment suggesting that bacteria are capable of transferring genetic information through a process known as transformation. Griffith's findings were followed by research in the late 1930s and early 40s that isolated DNA as the material that communicated this genetic information.

Pneumonia was a serious cause of death in the wake of the post-WWI Spanish influenza pandemic, and Griffith was studying the possibility of creating a vaccine. Griffith used two strains of pneumococcus (*Streptococcus pneumoniae*) bacteria which infect mice – a type III-S (smooth) which was virulent, and a type II-R (rough) strain which was nonvirulent. The III-S strain covered itself with apolysaccharide capsule that protected it from the host's immune system, resulting in the death of the host, while the II-R strain did not have that protective capsule and was defeated by the host's immune system. A German bacteriologist, Fred Neufeld, had discovered the three pneumococcal types (Types I, II, and III) and discovered the Quellung reaction to identify them in vitro. Until Griffith's experiment, bacteriologists believed that the types were fixed and unchangeable, from one generation to another.

In this experiment, bacteria from the III-S strain were killed by heat, and their remains were added to II-R strain bacteria. While neither alone harmed the mice, the combination was able to kill its host. Griffith was also able to isolate both live II-R and live III-S strains of pneumococcus from the blood of these dead mice. Griffith concluded that the type II-R had been "transformed" into the lethal III-S strain by a "transforming principle" that was somehow part of the dead III-S strain bacteria.

Today, we know that the "transforming principle" Griffith observed was the DNA of the III-s strain bacteria. While the bacteria had been killed, the DNA had survived the heating process and was taken up by the II-R strain bacteria. The III-S strain DNA contains the genes that form the protective polysaccharide capsule. Equipped with this gene, the former II-R strain bacteria were now protected from the host's immune system and could kill the host. The exact nature of the transforming principle (DNA) was verified in the experiments done by Avery, McLeod and McCarty and by Hershey and Chase.



18. Griffith's experimental work

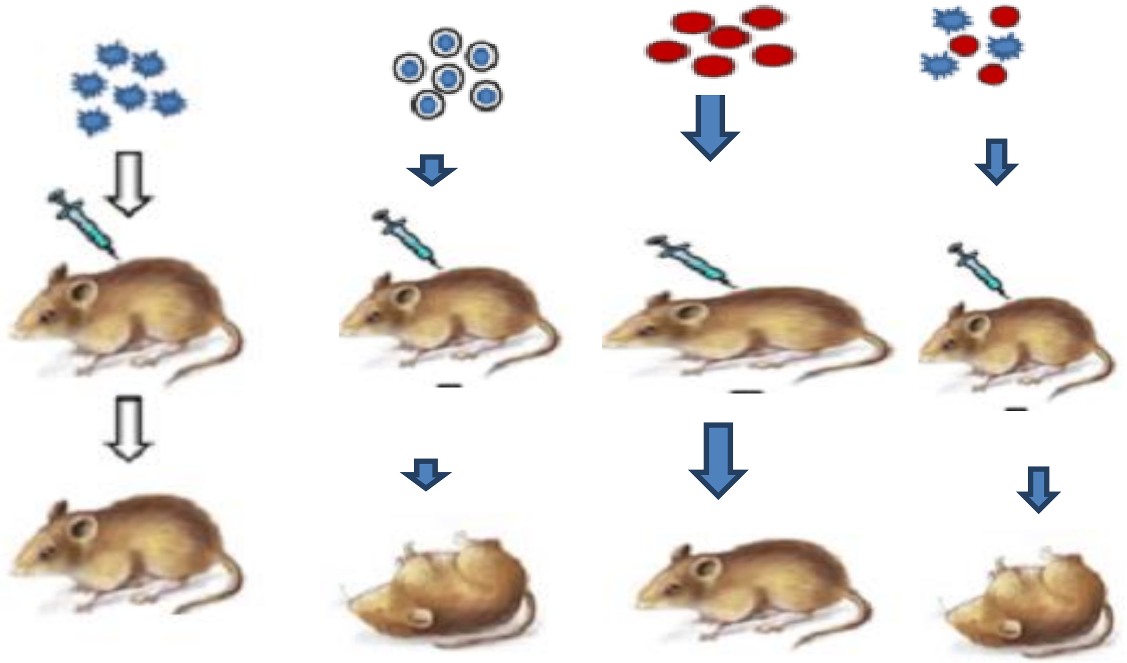


Fig. Griffith's experiment discovering the "transforming principle"

Possible Interpretations

It could have been the case that S- Type bacteria were not completely killed and a few live bacteria remained in the culture.

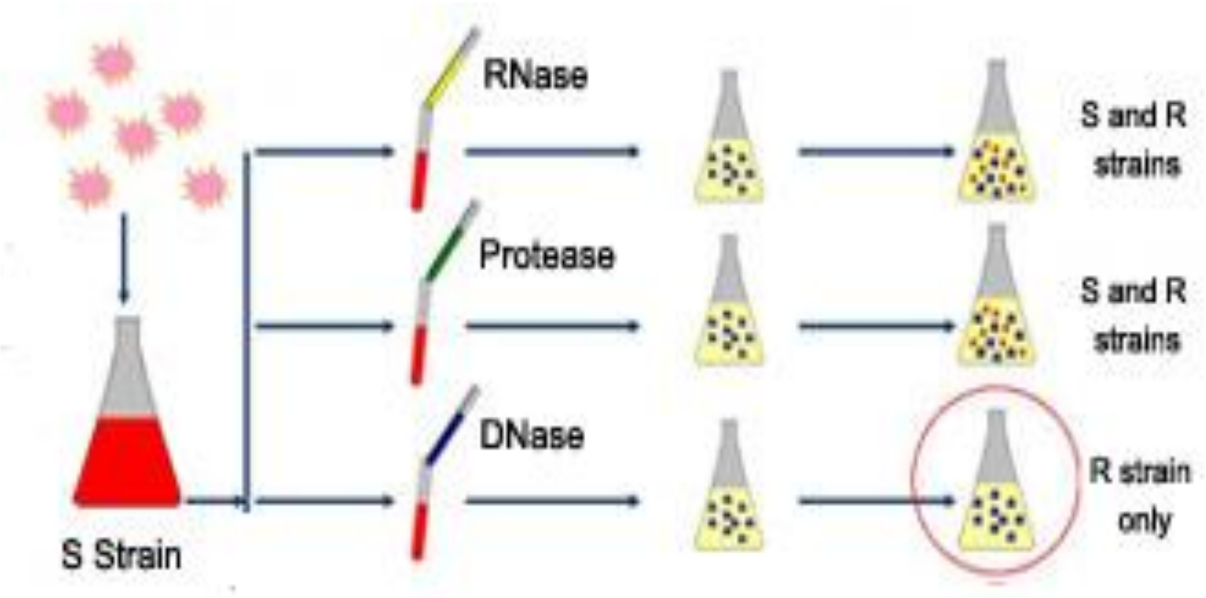
A second interpretation was that the live R-Type bacteria had mutated to the virulent S form.

Griffith finally concluded that R-Type bacteria had been *transformed.*

Griffith theorized that some substance of the dead bacteria might be responsible for that transformation . He called that substance as the transforming principle.

19. Avery's Experiments (1944)

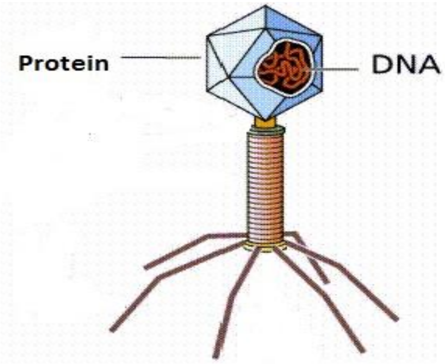
In 1944, experiments by Oswald T. Avery showed that DNA is the substance that causes bacterial transformation, in an era when it had been widely believed that it was proteins that served the function of carrying genetic information (with the very word *protein* itself coined to indicate a belief that its function was *primary*). It was the culmination of research in the 1930s and early 1940s at the Rockefeller Institute for Medical Research to purify and characterize the "transforming principle" responsible for the transformation phenomenon first described in Griffith's experiment of 1928: killed *Streptococcus pneumoniae* of the virulent strain type III-S, when injected along with living but non-virulent type II-R pneumococci, resulted in a deadly infection of type III-S pneumococci.Avery, MacLeod and McCarty succeeded in isolating and purifying the transforming substance in 1944. They showed that it had a chemical composition closely matching that of DNA and quite different from that of proteins. They showed that proteolytic enzymes had no effect on the transforming substance. Ribonuclease also had no effect on it. However, enzymes capable of destroying DNA, destroyed that substance.



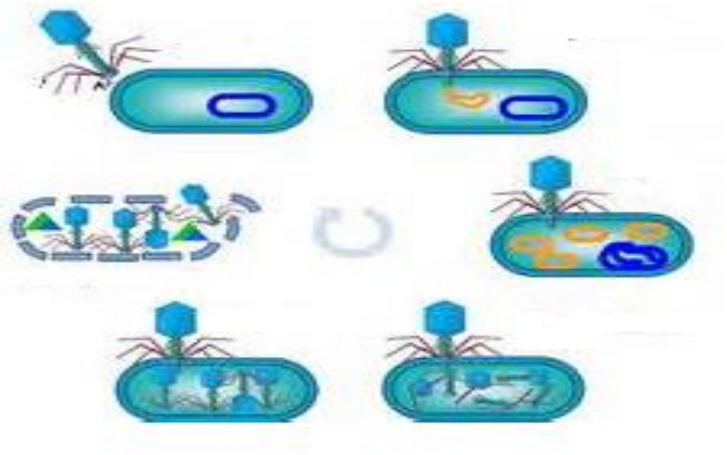
Avery, MacLeod, and McCarty further showed that the purified transforming substance precipitated at about the same rate as purified DNA. It absorbed ultraviolet light at the same wavelengths as does DNA. These findings provided compelling evidence that the transforming principle—and therefore the genetic information—resides in DNA.

20. Hershey-Chase experiments (1952)

The Hershey–Chase experiments were a series ofexperiments conducted in 1952 by Alfred Hershey andMartha Chase that helped to confirm that DNA is genetic material. While DNA had been known to biologists since 1869, many scientists still assumed at the time that proteinscarried the information for inheritance because DNA appeared simpler than proteins. In their experiments, Hershey and Chase showed that when bacteriophages, which are composed of DNA and protein, infect bacteria, their DNA enters the host bacterial cell, but most of their protein does not. Although the results were not conclusive, and Hershey and Chase were cautious in their interpretation, previous, contemporaneous and subsequent discoveries all served to prove that DNA is the hereditary material. Knowledge of DNA gained from these discoveries has applications in forensics,crime investigation and genealogy.



Bacteriophage

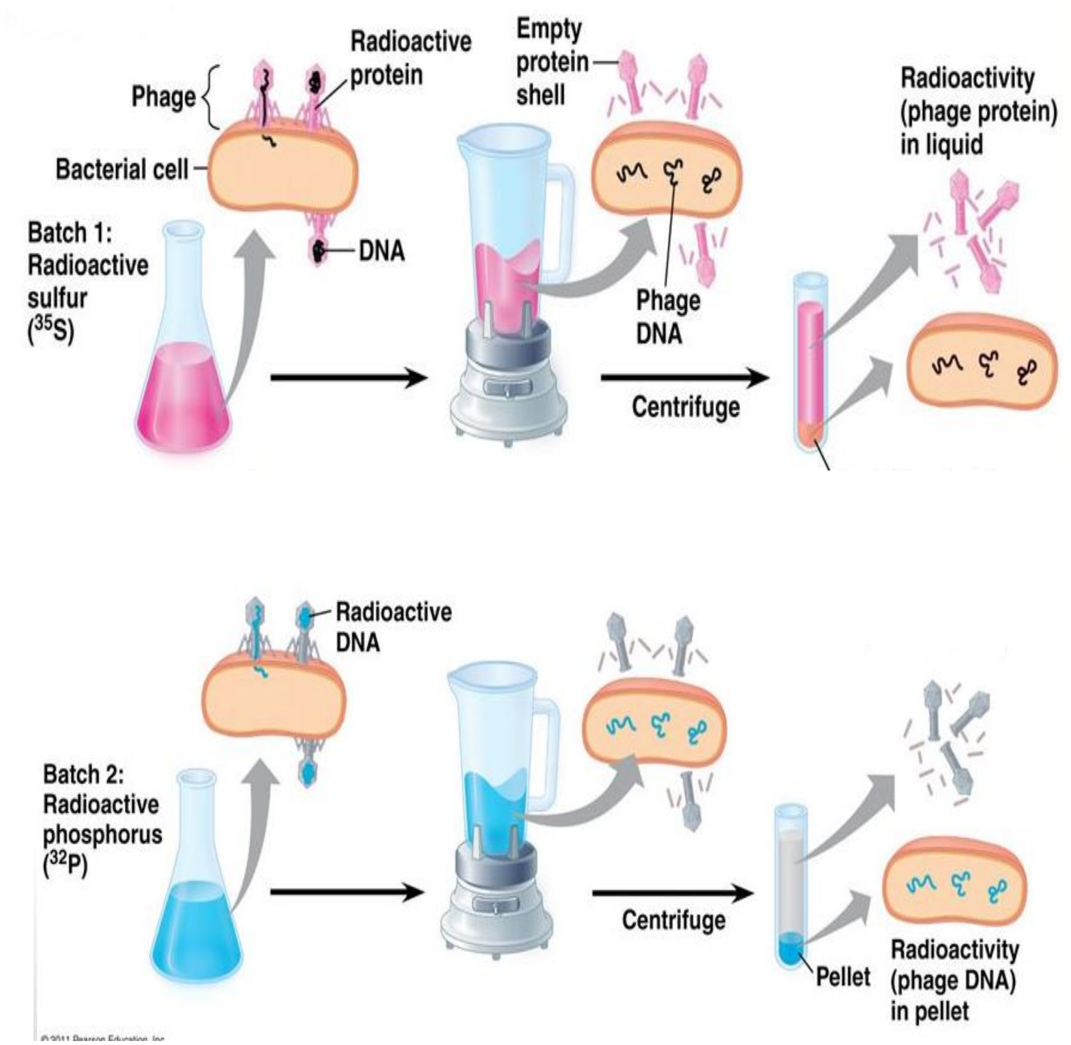


Life Cycle of a Bacteriophage

They used radioactive isotopes of phosphorus and sulfur. DNA contains P but not S; so Hershey and Chase used ^{32}P to follow phage DNA during reproduction. Protein contains sulfur but not phosphorus; so they used ^{35}S to follow the protein. Hershey and Chase first grew *E. coli* in a medium containing ^{32}P and infected the bacteria with T2 so that all the new phages would have DNA labeled with ^{32}P. They grew a second batch of *E. coli* in a medium containing ^{35}S and infected these bacteria with T2 so that all these new phages would have protein labelled with ^{35}S.

Hershey and Chase then infected separate batches of unlabeled *E. coli* with the ^{35}S- and ^{32}P-labeled phages. Then they placed the *E. coli* cells in a blender and sheared off the empty protein coats from the cell walls. They separated out the protein coats and cultured the infected bacterial cells. Eventually, the cells burst and new phage particles emerged.

When phages labeled with ^{35}S infected the bacteria, most of the radioactivity separated with the protein coats. When new phages emerged from the cell, they contained almost no radioactivity. When phages labelled with ^{32}P infected the bacteria, and removed the protein coat, radioactivity was present in the cells. When new phages emerged from the cell, they were also radioactive.

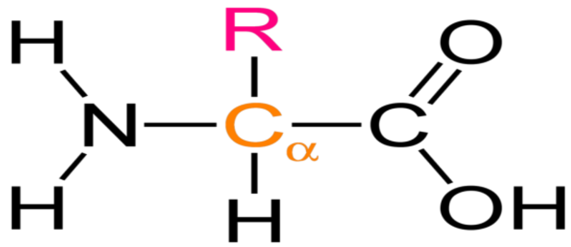

Phage
Radioactive protein
Empty protein shell
Radioactivity (phage protein) in liquid
Bacterial cell
Batch 1: Radioactive sulfur (^{35}S)
DNA
Phage DNA
Centrifuge
Radioactive DNA
Batch 2: Radioactive phosphorus (^{32}P)
Centrifuge
Pellet
Radioactivity (phage DNA) in pellet

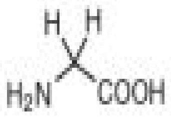
© 2011 Pearson Education, Inc.

Chapter 5. Protein structure

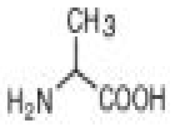
21. Chemical composition of protein

Proteins are polymers of amino acids.They range in size from small to very large. All the proteins are made up of twenty different types of amino acids. So these amino acids are called standard amino acids.

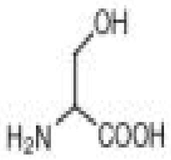




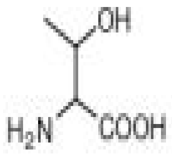
Glycine (Gly, G)



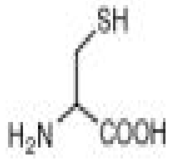
Alanine (Ala, A)



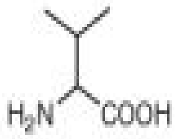
Serine (Ser, S)



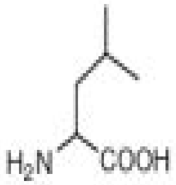
Threonine (Thr, T)



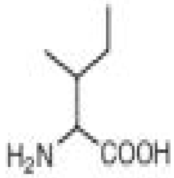
Cysteine (Cys, C)



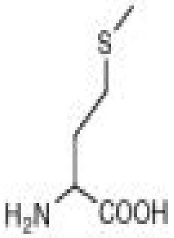
Valine (Val, V)



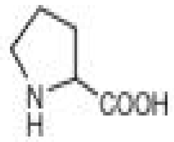
Leucine (Leu, L)



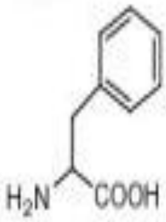
Isoleucine (Ile, I)



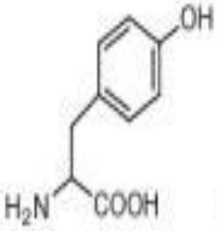
Methionine (Met, M)



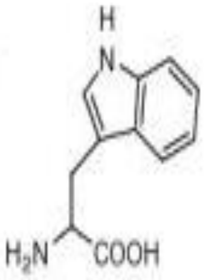
Proline (Pro, P)



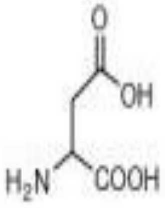
Phenylalanine (Phe, F)



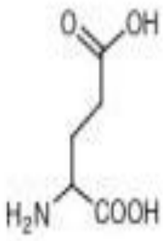
Tyrosine (Tyr, Y)



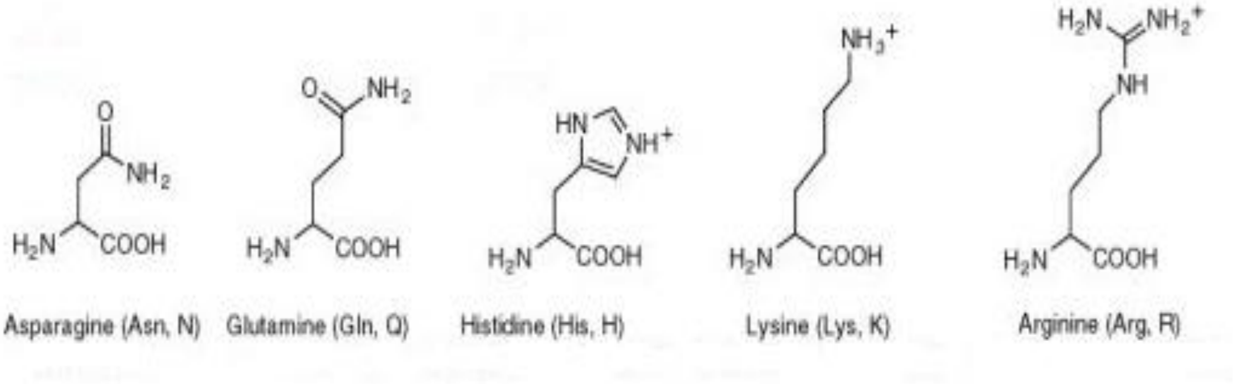
Tryptophan (Trp, W)



Aspartic Acid (Asp, D)

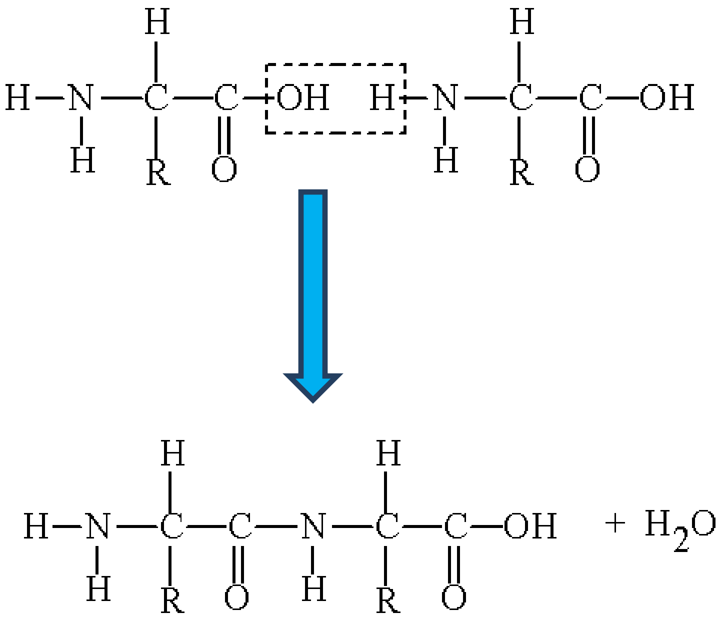


Glutamic Acid (Glu, E)



Asparagine (Asn, N) Glutamine (Gln, Q) Histidine (His, H) Lysine (Lys, K) Arginine (Arg, R)

In a protein molecule, each amino acid residue is joined to its neighbour by a specific type of covalent bond which is called Peptide Bond.

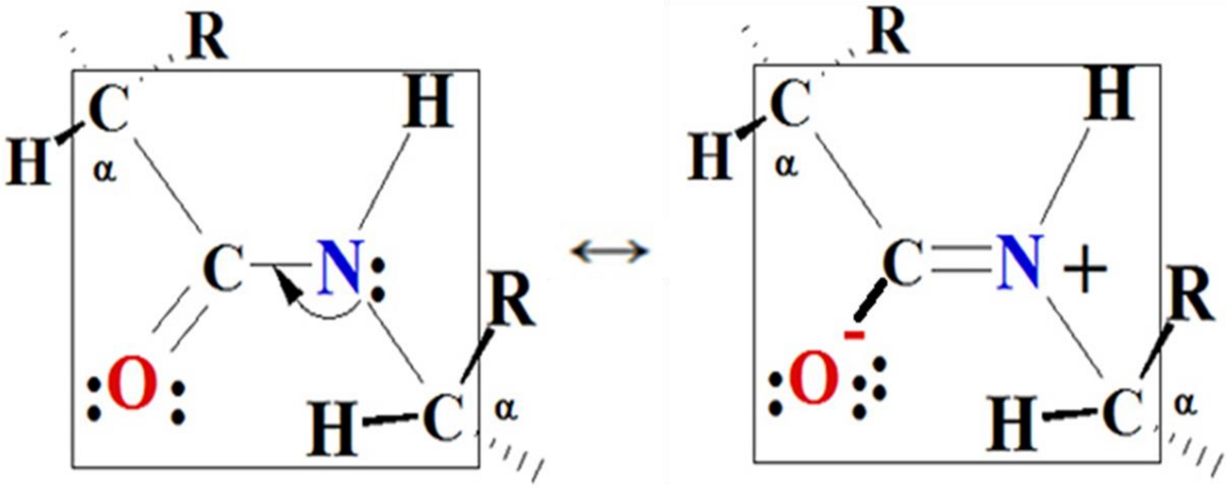


Amino acids can successively join to form dipeptides, tripeptides, tetrapeptides, oligo peptides and polypeptides.

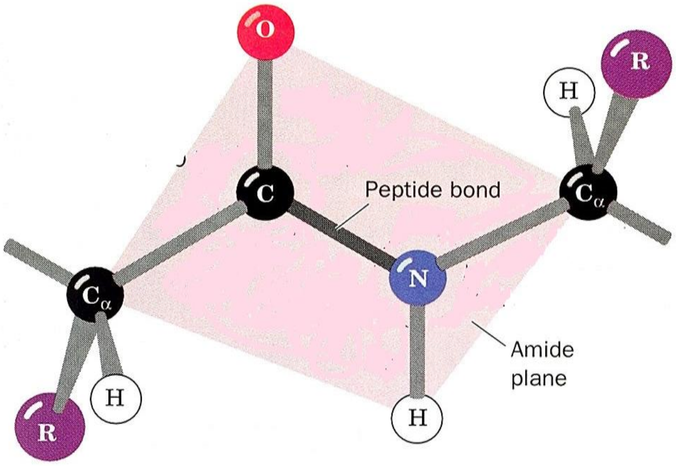
22. Primary structure of proteins

Primary structure or covalent structure of protein refers to the amino acid sequence of its polypeptide chain. Each type of protein has a unique amino acid sequence. In general, polypeptides are unbranched polymers, so their primary structure can often be specified by the sequence of amino acids along their backbone. However, proteins can become cross-linked, most commonly by disulfide bonds, and the primary structure also requires specifying the cross-linking atoms, e.g., specifying the cysteines involved in the protein's disulfide bonds. Other crosslinks include desmosine. The chiral centers of a polypeptide chain can undergo racemization. In particular, the L-amino acids normally found in proteins can spontaneously isomerize at the alpha carbon atom to form D-amino acids, which cannot be cleaved by most proteases.

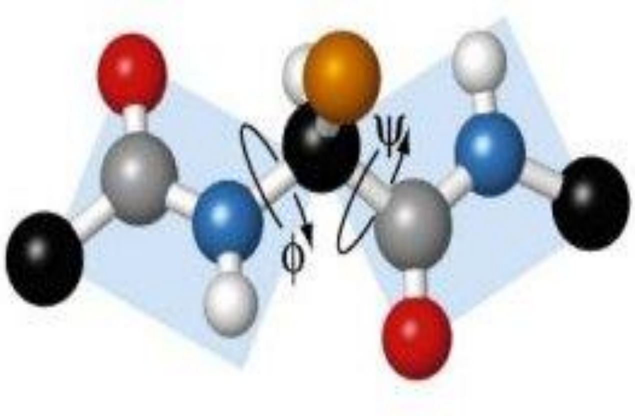
Peptide Bond Is Rigid and Planar.Linus Pauling and Robert Corey carefully analyzed the peptide bond. Their findings laid the foundation for our present understanding of protein structure. They demonstrated that the peptide C - N bond is somewhat shorter than the C - N bond in a simple amine.



The six atoms of the peptide group are co-planar i.e., lie in a single plane, with the oxygen atom of the carbonyl group and the hydrogen atom of the amide nitrogen trans to each other.



Pauling and Corey concluded that the peptide C - N bonds are unable to rotate freely because of their partial double-bond character.Rotation is permitted about the N - αC and the αC - C bonds.



The bond angles resulting from rotations at C are labelled ϕ (phi) for the N - αC bond and ψ(psi) for the αC - C bond.In principle, ϕ and ψ can have any value between +180 & -180.

23. Secondary structure of proteins

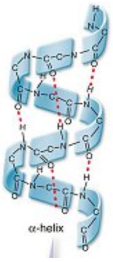
In biochemistry and structural biology, protein secondary structure is the general three-dimensional form of *local segments* of proteins. Secondary structure can be formally defined by the pattern of hydrogen bonds of the protein (such as alpha helices and beta sheets) that are observed in an atomic-resolution structure. More specifically, the secondary structure is defined by the patterns of hydrogen bonds formed between amine hydrogen and carbonyloxygen atoms contained in the backbone peptide bonds of the protein. The secondary structure may alternatively be defined based on the regular pattern of backbone dihedral angles in a particular region of the Ramachandran plot; thus, a segment of residues with such dihedral angles may be called a helix, regardless of whether it has the correct hydrogen bonds. The secondary structure may be provided by crystallographers in the corresponding PDB file.

Secondary structure does not describe the specific identity of amino acids in the protein which are defined as theprimary structure, nor the *global* atomic positions in three-dimensional space, which are considered to be tertiary structure. Other types of biopolymers such as nucleic acids also possess characteristic secondary structures.

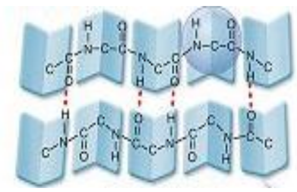
The concept of secondary structure was first introduced by Kaj Ulrik Linderstrøm-Lang at Stanford in 1952.

The most prominent are:-

α-helix

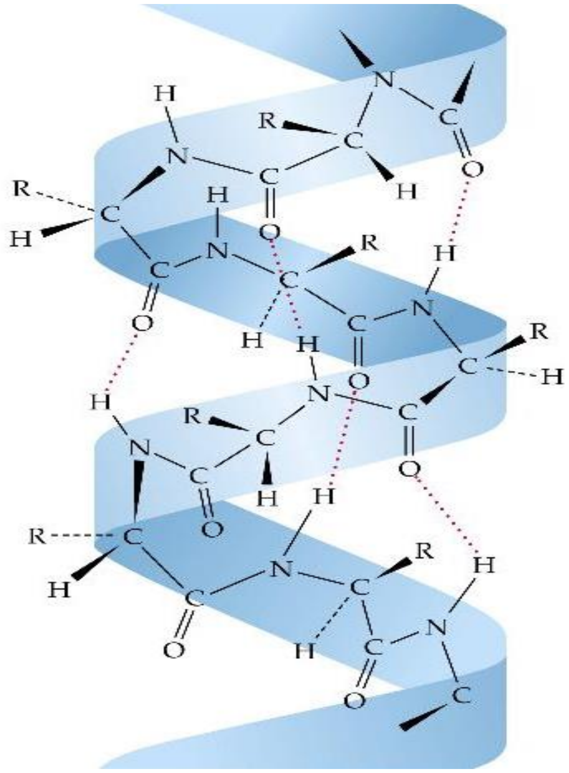


β- conformations.

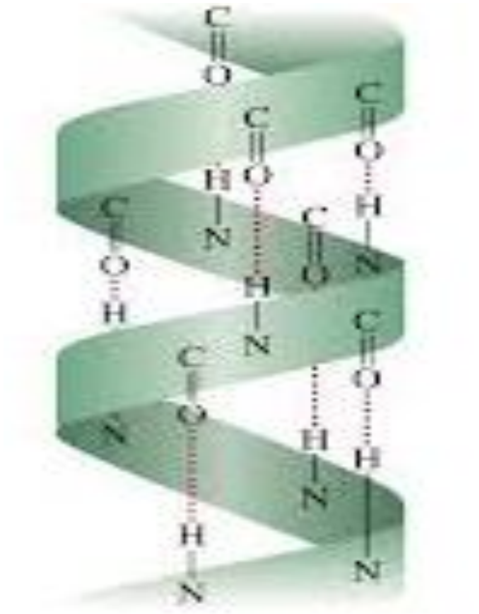


24. α- Helix

The alpha helix (α-helix) is a common secondary structure of proteins and is a righthand-coiled or spiral conformation (helix) in which every backbone N-H group donates a hydrogen bond to the backbone C=O group of the amino acid four residues earlier (i+4=i, ydrogen bonding). This secondary structure is also sometimes called a classic Pauling–Corey–Branson alpha helix (see below). The name 3.6_{13}-helix is also used for this type of helix, denoting the number of residues per helical turn, and 13 atoms being involved in the ring formed by the hydrogen bond. Among types of local structure in proteins, the α-helix is the most regular and the most predictable from sequence, as well as the most prevalent.



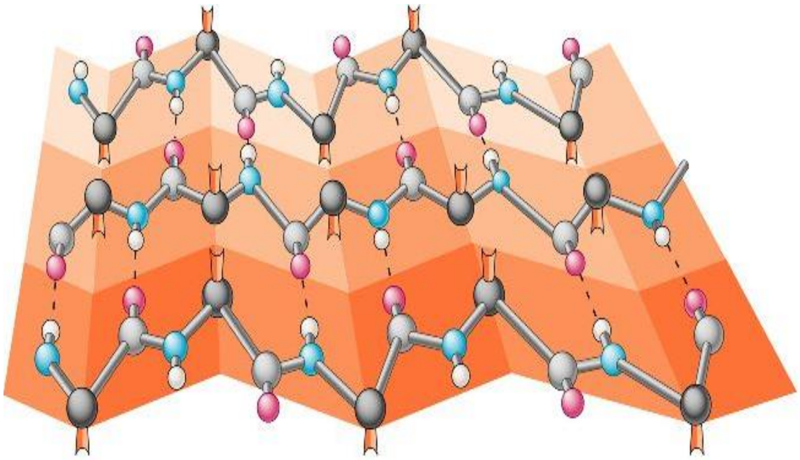
The helical twist of the α-helix found in all proteins is right-handed. The repeating unit is a single turn of the helix, which extends about 5.4 Å (includes 3.6 amino acid residues) along the long axis. The amino acid residues in an helix have conformations with psi = –45 to –50 and phi = – 60. An helix makes optimal use of internal hydrogen bonds.



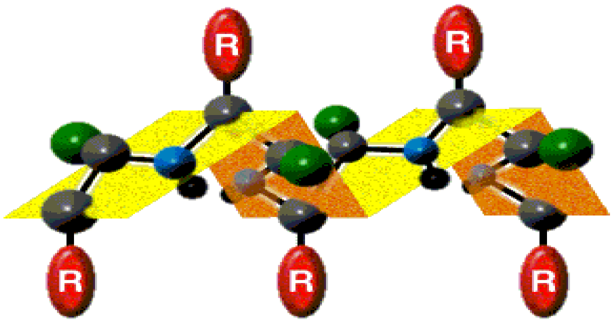
About one-fourth of all amino acid residues in polypeptides are found in α-helices while in some proteins it is the predominant structure.

25. β- Pleated Sheets

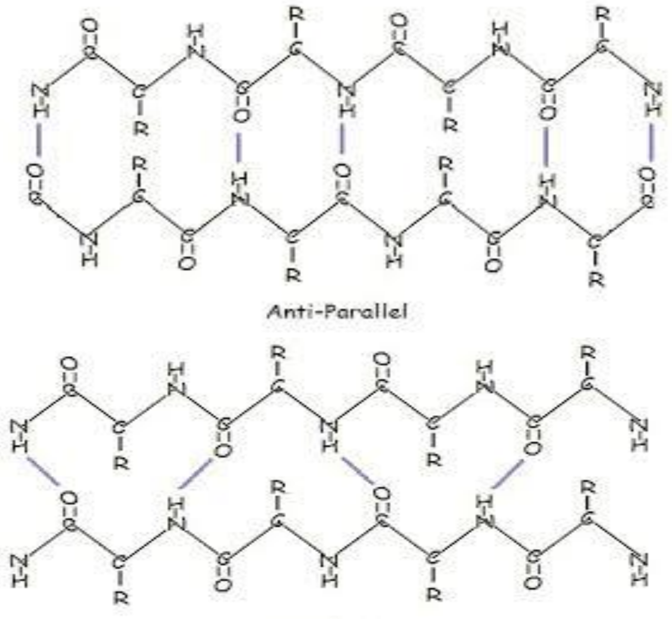
Pauling and Corey predicted a second type of secondary structure which they called **β-sheets.** This is a more extended conformation of polypeptide chains. The β sheet (also β-pleated sheet) is the second form of regular secondary structure in proteins. Beta sheets consist of beta strands connected laterally by at least two or three backbone hydrogen bonds, forming a generally twisted, pleated sheet. A beta strand (also β strand) is a stretch of polypeptide chain typically 3 to 10 amino acids long with backbone in an extended conformation. The higher-level association of β sheets has been implicated in formation of the protein aggregates and fibrils observed in many human diseases, notably theamyloidoses such as Alzheimer's disease.



The R groups of adjacent amino acids protrude from the zigzag structure in opposite directions.



Hydrogen bonds are formed between adjacent segments of polypeptide chain.The adjacent polypeptide chains in a sheet can be either parallel or antiparallel.



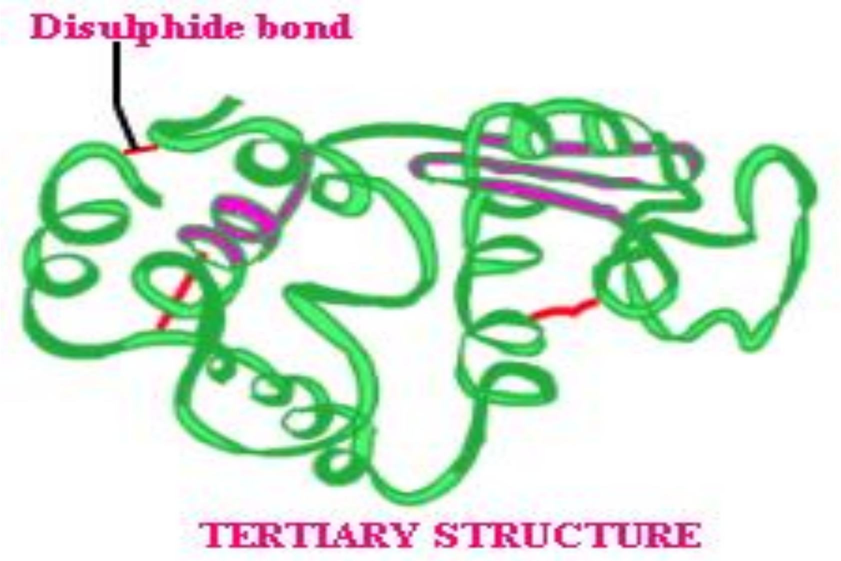
Anti-Parallel

Parallel

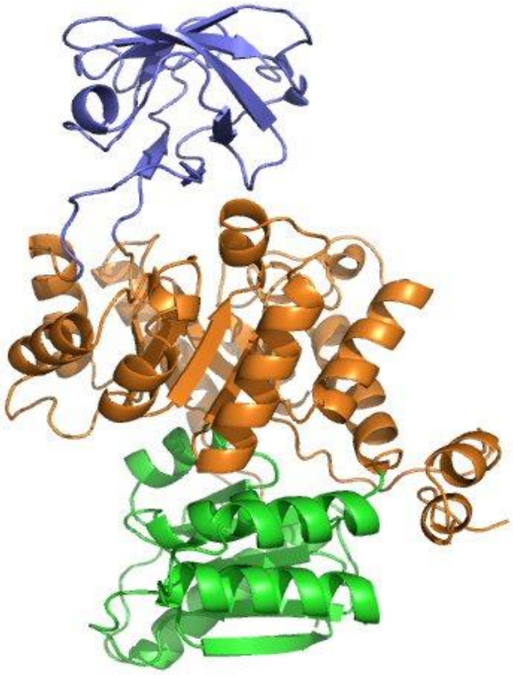
26. Tertiary Structure of Proteins

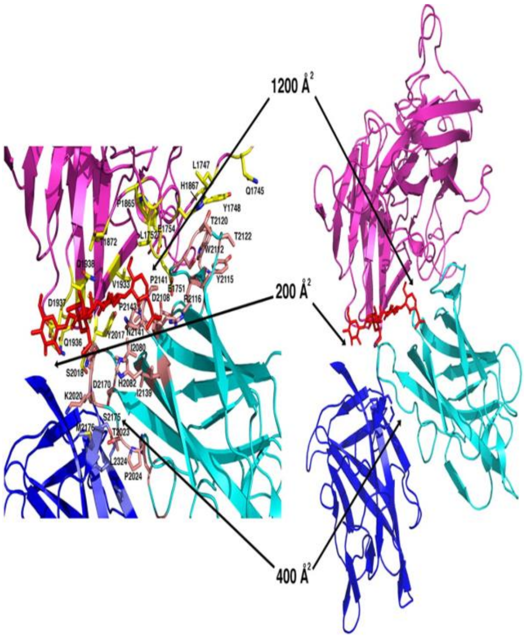
The term protein tertiary structure refers to a protein's geometric shape. The overall three-dimensional arrangement of all atoms in a protein is referred to as the protein’s tertiary structure.

The tertiary structure will have a single polypeptide chain "backbone" with one or more protein secondary structures, the protein domains. Amino acid side chains may interact and bond in a number of ways. The interactions and bonds of side chains within a particular protein determine its tertiary structure. The protein tertiary structure is defined by its atomiccoordinates. These coordinates may refer either to a protein domain or to the entire tertiary structure.[1][2] A number of tertiary structures may fold into a quaternary structure.



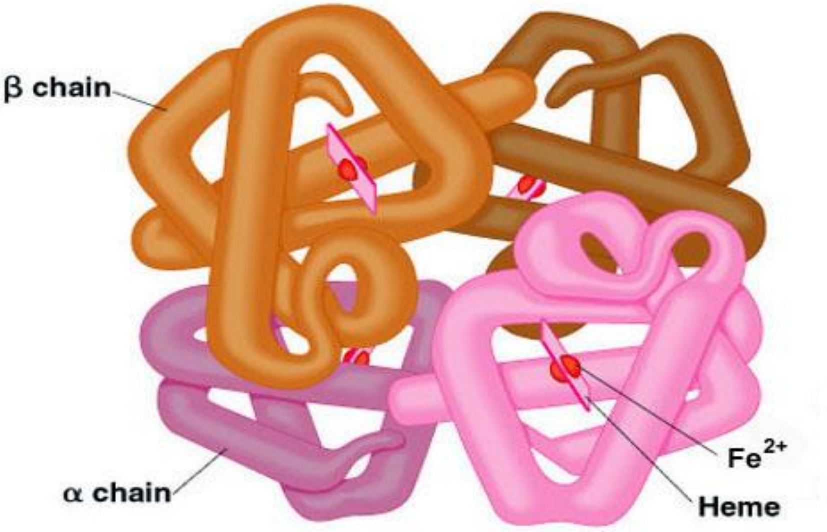
It includes longer-range aspects of amino acid sequence. Amino acids that are far apart in the polypeptide chain may interact within the completely folded structure of a protein. Interacting segments of polypeptide chains are held in their characteristic tertiary positions by different kinds of weak interactions (and sometimes by covalent bonds) between the segments. Large polypeptide chains usually fold into two or more globular clusters known as domains, which often give these proteins a bi- or multilobal appearance.



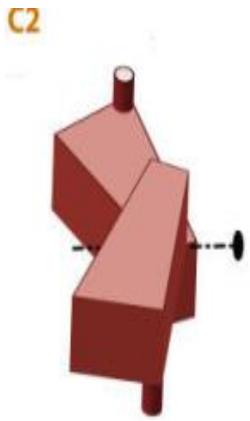


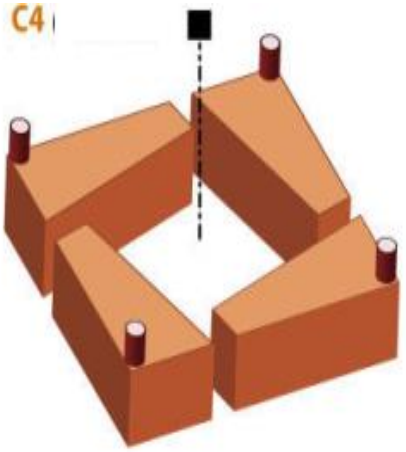
27. Quaternary Structure of Proteins

Quaternary structure is the number and arrangement of multiple folded protein subunits in a multi-subunit complex. It includes organisations from simple dimers to large homooligomers and complexes with defined or variable numbers of subunits. Some proteins contain two or more separate polypeptide chains or subunits, which may be identical or different. The spatial arrangement of these subunits is known as a protein's quaternary structure. A multi-subunit protein is also referred to as a multimer. A multimer with just a few subunits is called as oligomer and a single subunit or a group of subunits, is called a protomer.

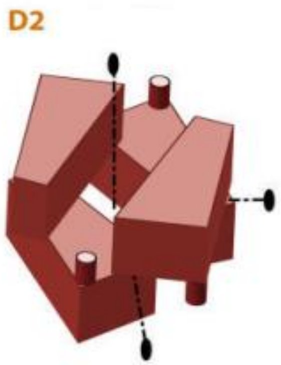


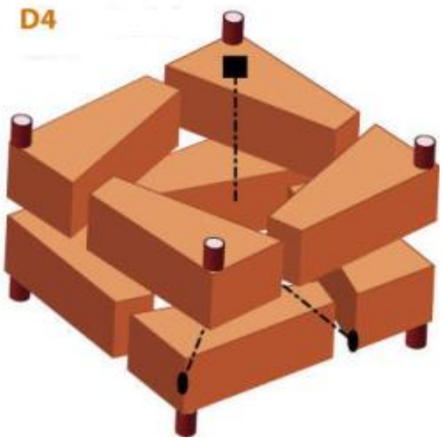
Identical subunits of multimeric proteins are generally arranged in a symmetric patterns. Oligomers can have either rotational symmetry or helical symmetry. There are several forms of rotational symmetry. The simplest is cyclic symmetry, involving rotation about a single axis.



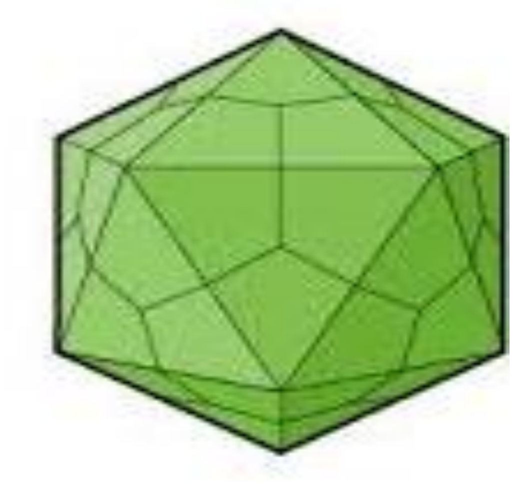


A somewhat more complicated rotational symmetry is dihedral symmetry, in which a twofold rotational axis is present.

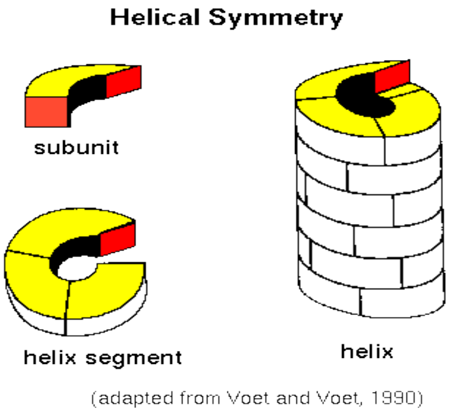




More complex rotational symmetries include icosahedral symmetry. An icosahedron is a regular 12-cornered polyhedron having 20 triangular faces.



The other major type of symmetry found in oligomers is helical symmetry.



Chapter 6. Organisation of Genetic Material

28. Genetic Materials in Viruses

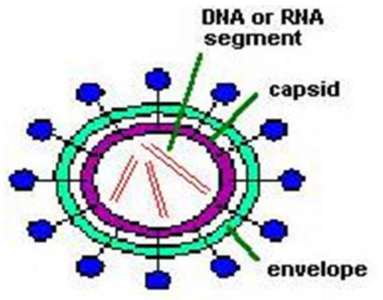
Viruses are exceptionally simple and extremely small microorganisms. An enormous variety of genomic structures can be seen among viral species; as a group, they contain more structural genomic diversity than plants, animals, archaea, or bacteria. There are millions of different types of viruses,[4] although only about 5,000 types have been described in detail.[3] As of September 2015, the NCBI Virus genome database has more than 75,000 complete genome sequences, but there are doubtlessly many more to be discovered.

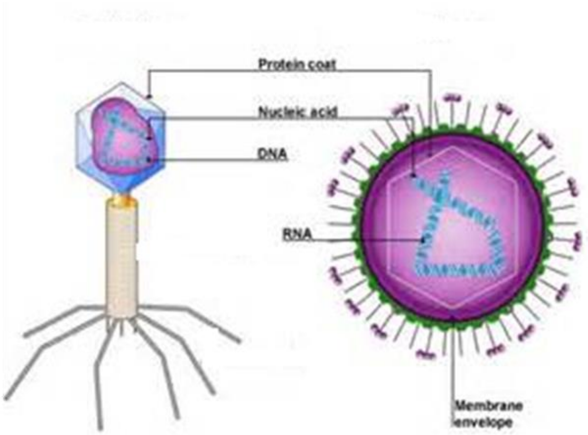
A virus has either a DNA or an RNA genome and is called a DNA virus or an RNA virus, respectively. The vast majority of viruses have RNA genomes. Plant viruses tend to have single-stranded RNA genomes and bacteriophages tend to have double-stranded DNA genomes.

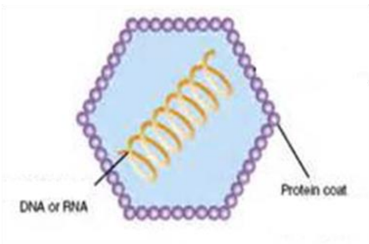
Viral genomes are circular, as in the polyomaviruses, or linear, as in the adenoviruses. The type of nucleic acid is irrelevant to the shape of the genome. Among RNA viruses and certain DNA viruses, the genome is often divided up into separate parts, in which case it is called segmented. For RNA viruses, each segment often codes for only one protein and they are usually found together in one capsid. However, all segments are not required to be in the same virion for the virus to be infectious, as demonstrated by brome mosaic virus and several other plant viruses.

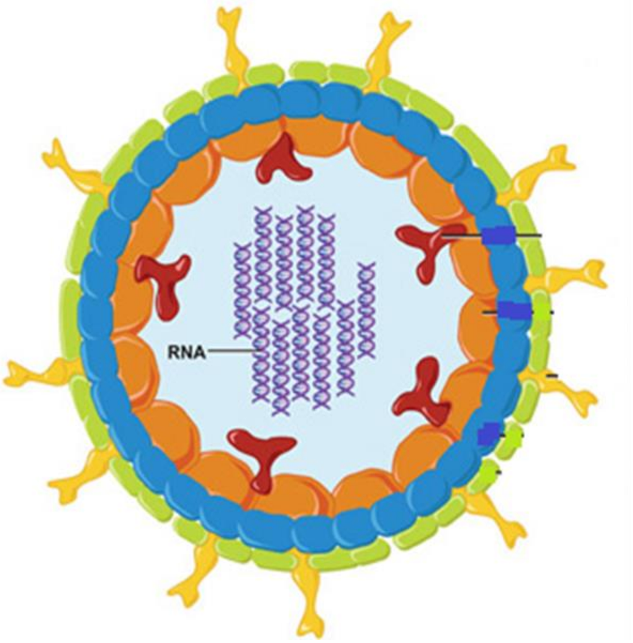
A viral genome, irrespective of nucleic acid type, is almost always either single-stranded or double-stranded. Single-stranded genomes consist of an unpaired nucleic acid, analogous to one-half of a ladder split down the middle. Double-stranded genomes consist of two complementary paired nucleic acids, analogous to a ladder. The virus particles of some virus families, such as those belonging to the Hepadnaviridae, contain a genome that is partially double-stranded and partially single-stranded. For most viruses with RNA genomes and some with single-stranded DNA genomes, the single strands are said to be eitherpositive-sense (called the plus-strand) or negative-sense (called the minus-strand), depending on if they are complementary to the viral messenger RNA (mRNA). Positive-sense viral RNA is in the same sense as viral mRNA and thus at least a part of it can be immediately translated by the host cell. Negative-sense viral RNA is complementary to mRNA and thus must be converted to positive-sense RNA by an RNA-dependent RNA polymerase before translation. DNA nomenclature for viruses with single-sense genomic ssDNA is similar to RNA nomenclature,

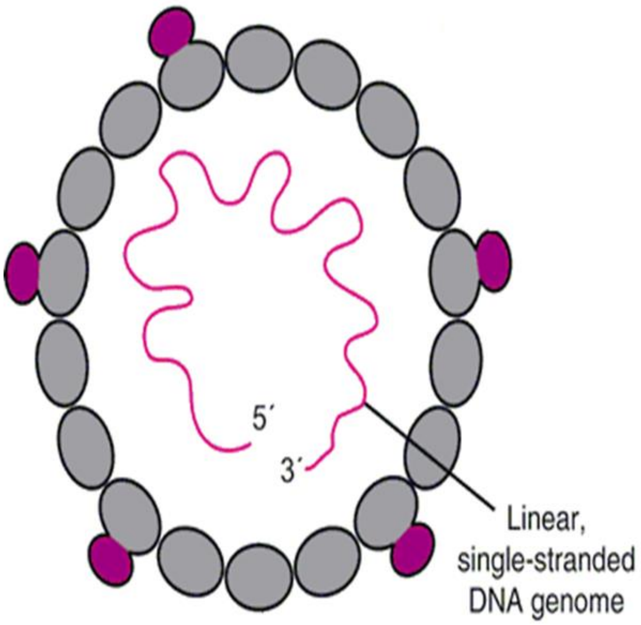
They have a very simple structural organization consisting of a molecule of nucleic acid and a protein coat.

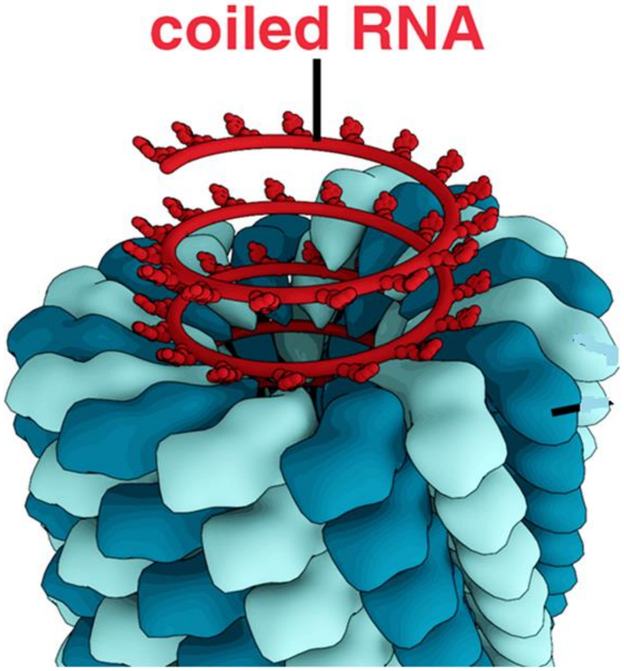


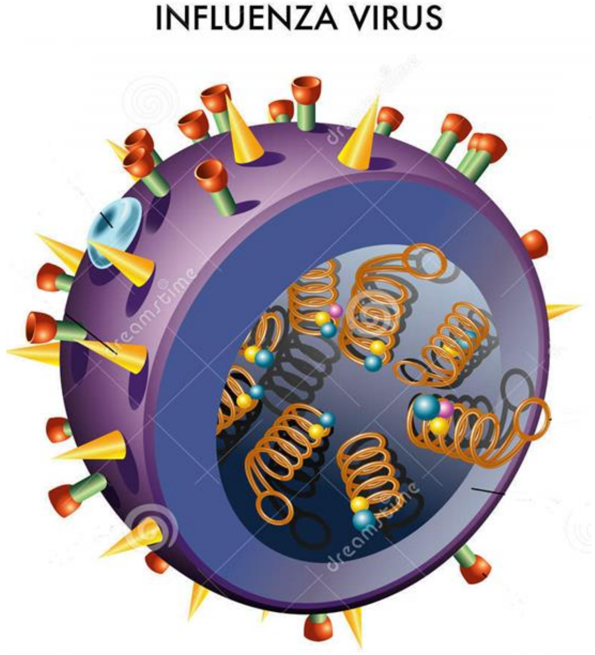








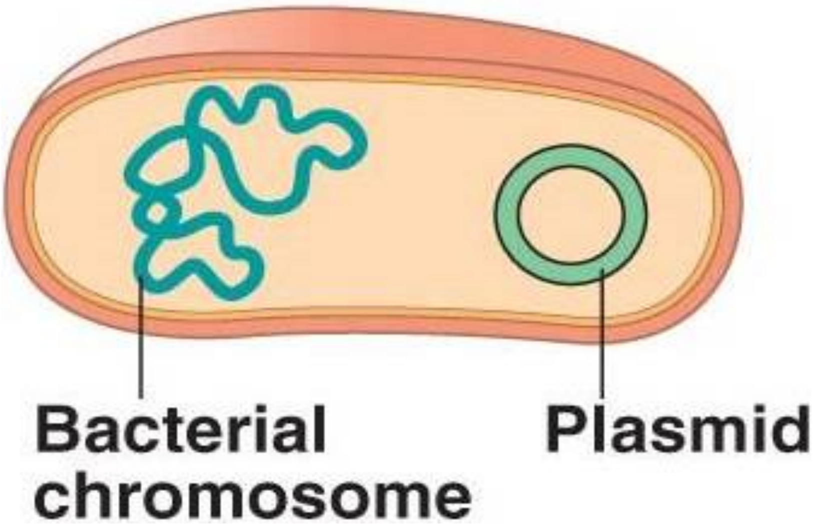




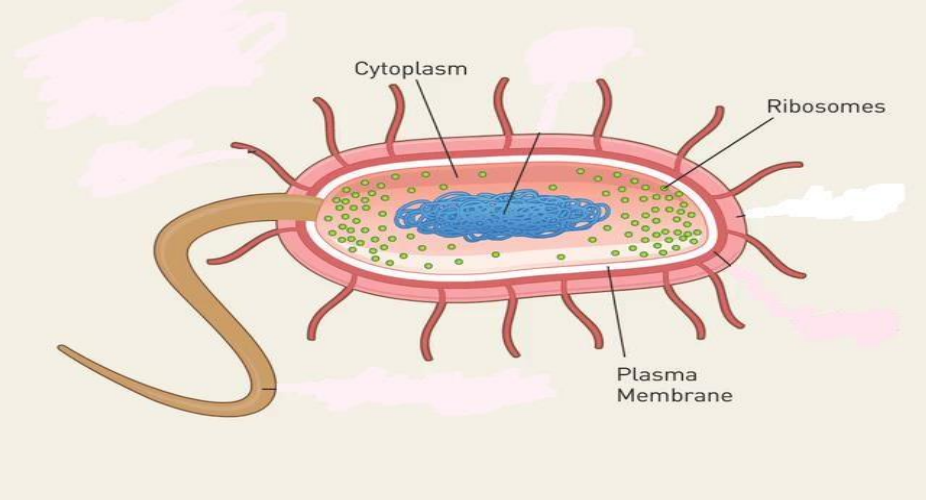
The percentage of nucleic acid in relation to protein is about 1% for the influenza virus and about 50% for some bacteriophages.The total amount of nucleic acid varies from a few thousand nucleotides to as many as 250,000 nucleotides. *E. coli's* chromosome consists of approx. 4 million nucleotide pairs.

29 30. Organization of Genetic Material in Bacteria

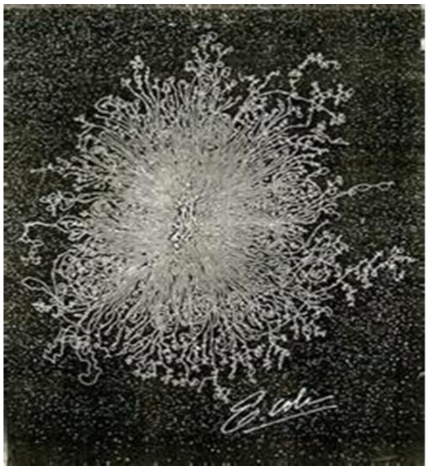
Bacterial genetics is the subfield of genetics devoted to the study of bacteria. Bacterial genetics are subtly different from eukaryotic genetics, however bacteria still serve as a good model for animal genetic studies. One of the major distinctions between bacterial and eukaryotic genetics stems from the bacteria's lack of membrane-bound organelles. Bacteria typically have a single circular chromosome consisting of a single circular molecule of DNA with associated proteins. The bacterial chromosome is a very long (up to 1mm). It is looped and folded and attached at one or several points to the plasma membrane.

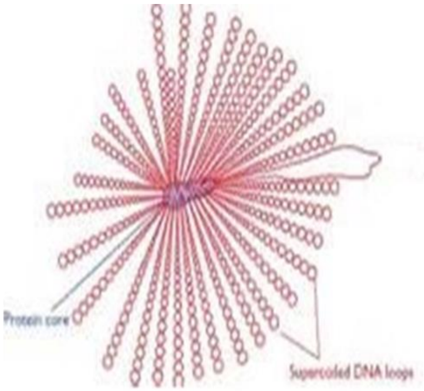


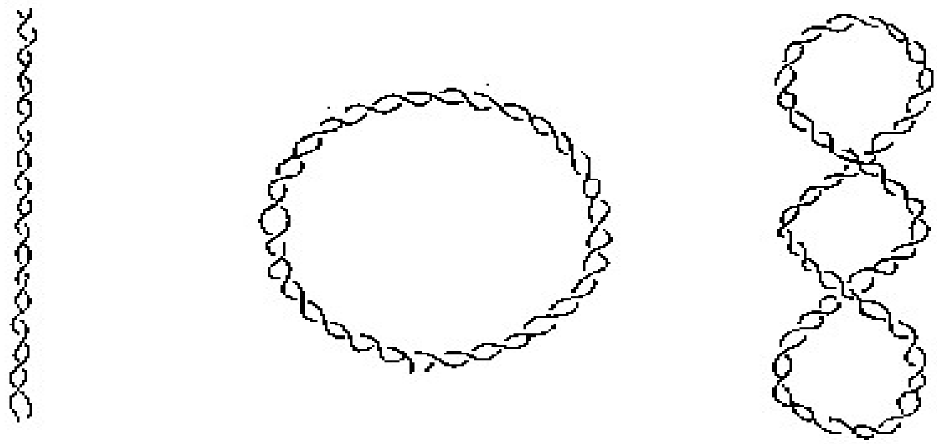
Specific proteins interact with the bacterial DNA to form a highly condensed nucleoprotein complex called the nucleoid.



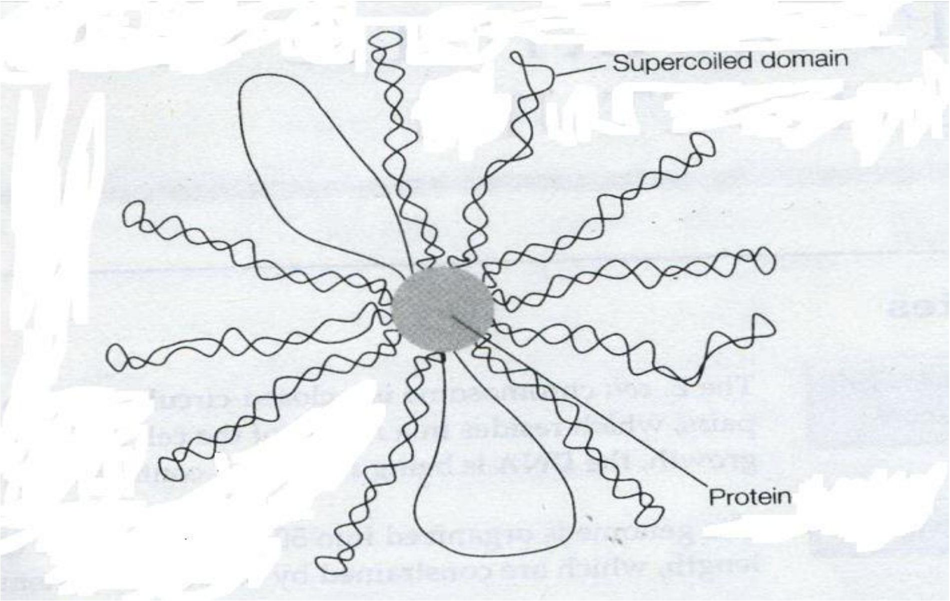
Bacterial chromatin can be released from the cell by gentle lysis of the cell. Electron micrograph of the chromatin reveals that it consists of multiple loops which emerge from a central region of the chromatin. Some of the loops are super-coiled while some are relaxed. Relaxed loops are formed as a result of a nick introduced into super-coiled loops by a cellular DNase.







If a super-coiled DNA molecule receives a nick, the strain of under-winding is immediately removed, and all the super-coiling is lost. Studies confirm that continued nuclease treatment increases number of relaxed loops.



The bacterial DNA is arranged in super-coiled loops that are fastened to a central protein matrix, so that each loop is topologically independent from all the others. So a nick that causes one super-coiled loop to relax would have no effect on other super-coiled loops. The super-coiled loops are dynamic structures which change during cell growth & division. An *E. coli* chromosome is estimated to have about 400 super-coiled loops. Each loop has an average length of about 10-20 kbp. The DNA compaction in a bacterial cell is contributed by super-coiling of loops, macromolecular crowding and DNA-binding proteins.

31. Organization of Genetic Material in Eukaryotes

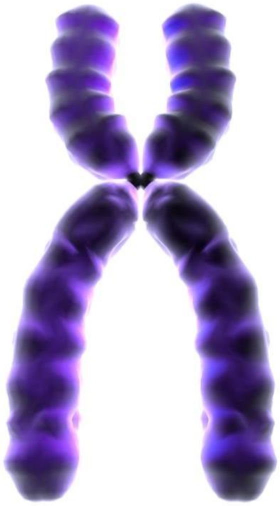
The genome is segmented in eukaryotes. It is made up of a number of linear chromosomes. A genome is the entire set of unique chromosomes that segregate to a gamete or to a haploid life stage in organisms with an alternation of generations like plants.

A diploid has two copies 2n = 2x of their genome while a tetraploid has 4 copies 2n = 4x of the basic monoploid genome X. A diploid then has homologous paired chromosomes.

The strands are organized by coiling the strand of DNA about an octomer of histones into nucleosomes. Each Nucleosome is an assembly of histone proteins with the DNA making two turns around the group and clamped by another histone, H1. This condenses the strand length. The strand is super coiled for further compaction.

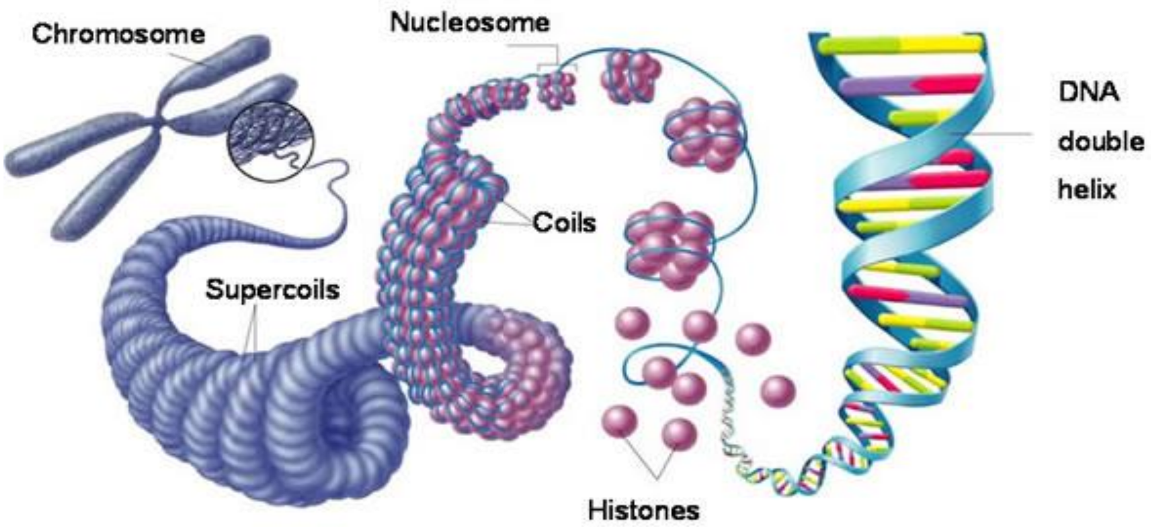
DNA packaging is increased during mitosis or meiosis with proteins like condensin & cohesin linking replicated chromatids dyads at the centromere and holding the supercoils together in each arm. The centromere permits spindle attachments to each pole so daughter cells each get a copy inmitosis.

Located along the DNA strand are sequence patterns the regulate gene expression as well as the open reading frames of the gene loci. Chromosomes contain hundreds to thousands of gene loci depending on the length of the chromosome.



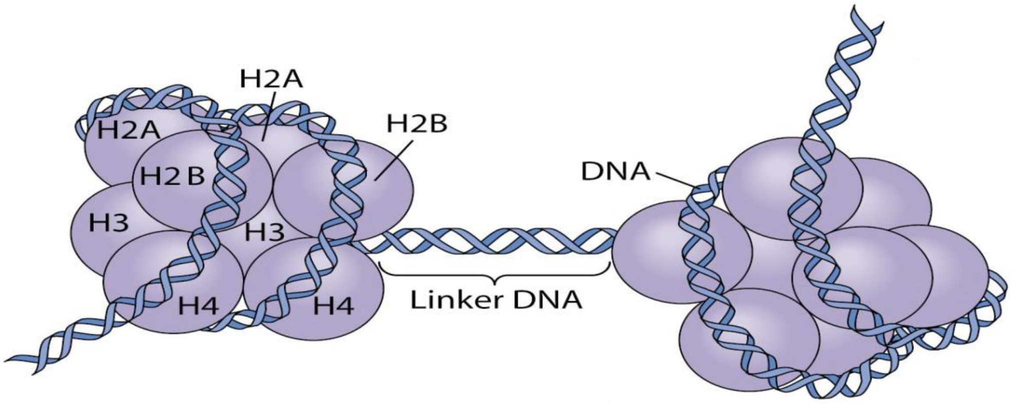
Chromosome

Each un-replicated chromosome consists of a single molecule of DNA. If stretched out, some human chromosomes would be several centimetres long. To package such a tremendous length of DNA into this small volume, each DNA molecule is coiled again and again and tightly packed around histone proteins. As eukaryotic chromosomes are not circular, so instead of super-coiling, the mechanism of packaging involves winding the DNA around special proteins, the histones. DNA with bound histones in the eukaryotes is called as chromatin. Chromatin consists of roughly spherical subunits, the nucleosomes, each containing approx. 200 bp of DNA and nine histones. A condensed mitotic chromosome is about 50,000 times shorter than fully extended DNA. Highly condensed chromatin is known as heterochromatin. The more extended form is known as euchromatin.



32. Histone Proteins

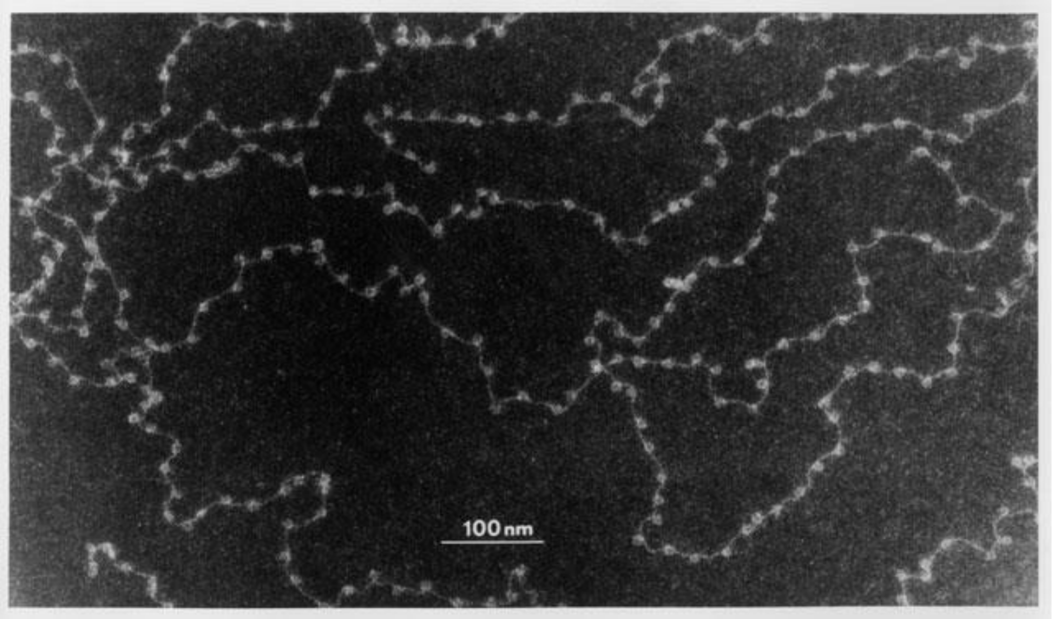
Most abundant proteins in the chromatin are histones. Histones are highly alkaline proteins found in eukaryotic cell nuclei that package and order the DNAinto structural units called nucleosomes. They are the chief protein components of chromatin, acting as spools around which DNA winds, and playing a role in gene regulation. Without histones, the unwound DNA in chromosomes would be very long (a length to width ratio of more than 10 million to 1 in human DNA). For example, each human cell has about 1.8 meters of DNA, (~6 ft) but wound on the histones it has about 90 micrometers (0.09 mm) of chromatin, which, when duplicated and condensed during mitosis, result in about 120 micrometers of chromosomes. There are nine types of histones including two each of H2A, H2B, H3 and H4 and one of H1. These histones fall in five major classes i.e., H1, H2A, H2B, H3 and H4.A typical human cell contains about 60 million copies of each kind of histone.



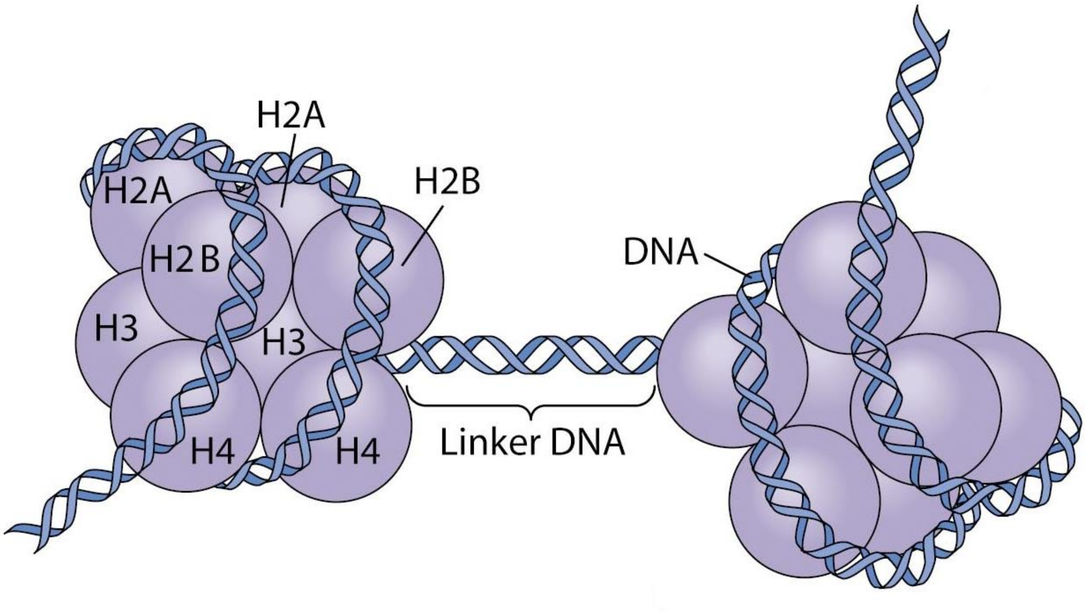
All histones have a high percentage of arginine and lysine but the lysine-to-arginine ratio differs in each type of histone. The positively charged side chains of lysine and arginine enable histones to bind to the negatively charged phosphate groups of the DNA. The electrostatic attraction is an important stabilizing force in the chromatin.

33. The Nucleosome

The uncondensed chromatin resembles beads on a string when viewed under the electron microscope.Each bead is a nucleoprotein complex called nucleosome. A nucleosome is a basic unit of DNA packaging in eukaryotes, consisting of a segment of DNA wound in sequence around eight histone protein cores.[2] This structure is often compared to thread wrapped around a spool.

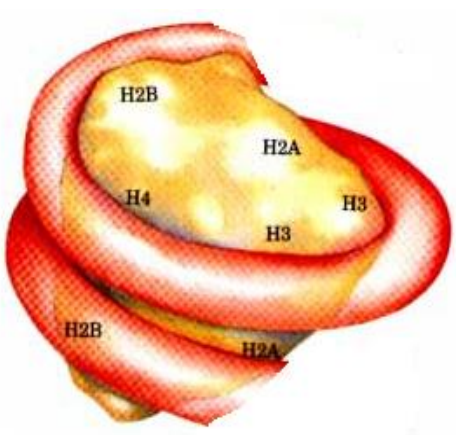


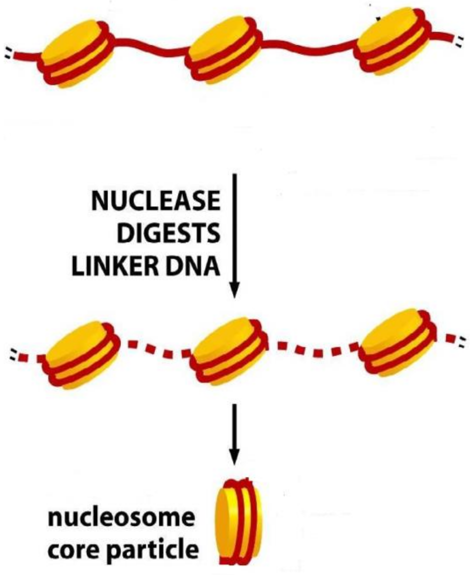
Nucleosomes form the fundamental repeating units of eukaryotic chromatin, which is used to pack the large eukaryotic genomes into the nucleus while still ensuring appropriate access to it (in mammalian cells approximately 2 m of linear DNA have to be packed into a nucleus of roughly 10 µm diameter). Nucleosomes are folded through a series of successively higher order structures to eventually form a chromosome; this both compacts DNA and creates an added layer of regulatory control, which ensures correct gene expression. Nucleosomes are thought to carry epigenetically inherited information in the form of covalent modifications of their core histones. Nucleosomes were observed as particles in the electron microscope by Don and Ada Olins and their existence and structure (as histone octamers surrounded by approximately 200 base pairs of DNA) were proposed by Roger Kornberg. The role of the nucleosome as a general gene repressor was demonstrated by Lorch et al. in vitro [8] and by Han and Grunstein in vivo.



The size of the linker DNA between the nucleosomes varies among different organisms and even different organs of the same organism. The length of DNA wrapped around nucleosomes also varies from one organism to the other ranging from about 170-240 bp. Prolonged nuclease digestion of chromatin cleaves additional nucleotides. The structure that remains is the nucleosome core particle. The nucleosome core particle consists of an octameric protein complex (two copies of each H2A, H2B, H3 & H4) with a 146 bp DNA fragment wound around it.

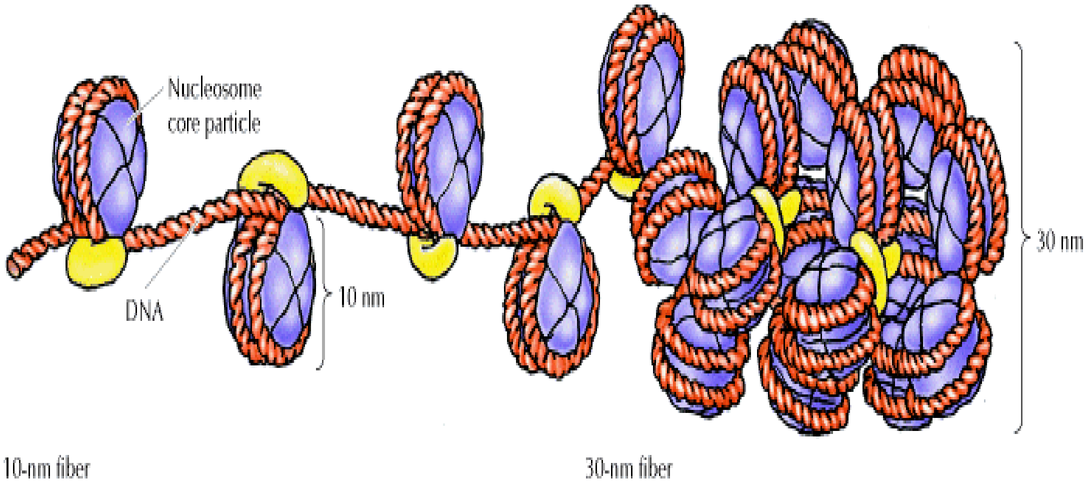
The crystal structure of the nucleosome core particle (PDB ID:1EQZ . Histones H2A,H2B, H3 and H4 are coloured, DNA is gray.The nucleosome core particle consists of approximately 147 base pairs of DNA wrapped in 1.67 left-handed superhelical turns around a histone octamer consisting of 2 copies each of the core histones H2A, H2B, H3, and H4. Core particles are connected by stretches of "linker DNA", which can be up to about 80 bp long. Technically, a nucleosome is defined as the core particle plus one of these linker regions; however the word is often synonymous with the core particle. Genome-wide nucleosome positioning maps are now available for many model organisms including mouse liver and brain.



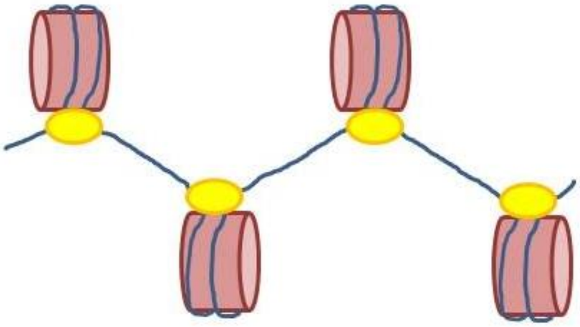


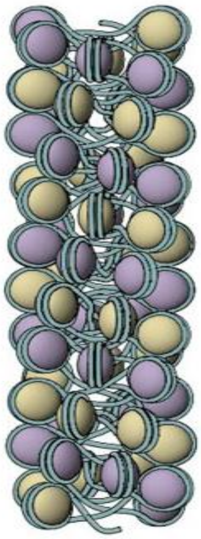
34 &35. The 30-nm Fiber

It is still unclear that how a chain of nucleosomes folds into higher order structures.The next level of organization is a 30-nm fiber.

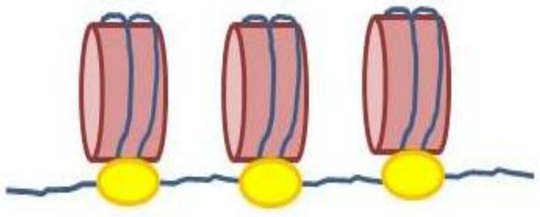


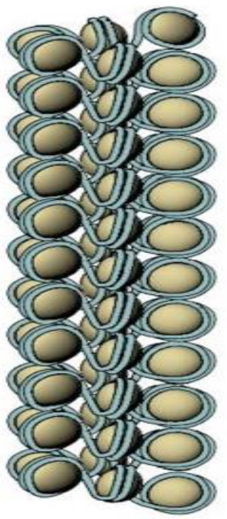
Various models have been proposed to explain how nucleosomes fold to form the 30-nm fiber. However, two models gained the most support:-
Zigzag Model
Solenoid Model
Zigzag model predicts that the linker DNA forms a straight path between successive nucleosomes.The nucleosomes lie on opposite sides of the fiber.





The solenoid model predicts that the nucleosome chain forms a helical structure with about 6 nucleosomes per turn.Linker DNA is bent to connect neighbouring nucleosomes. Some of the reconstitution experiments appear to support Zigzag models while some others support Solenoid model.





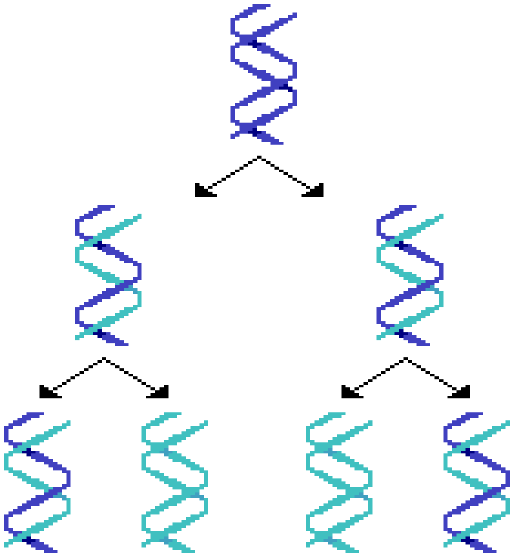
Chapter 7. DNA Replication

36. Replication of DNA

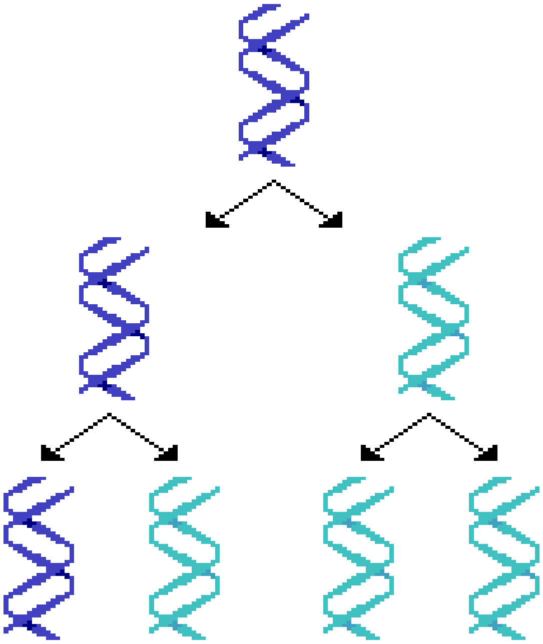
DNA replication is the process of producing two identical replicas from one original DNA molecule. This biological process occurs in all living organisms and is the basis for biological inheritance. DNA is made up of two strands and each strand of the original DNA molecule serves as a template for the production of the complementary strand. The double-helical model for DNA includes the concept that the two strands are complementary.

Thus, each strand can in principle serve as the template for making its own partner. A number of models were proposed to explain the mode of replication of DNA.

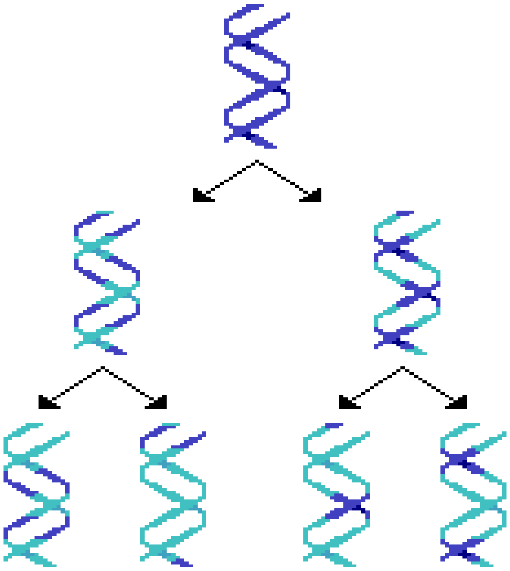
But the semiconservative model for DNA replication is the correct one. The Watson–Crick model for DNA replication proposed that the two parental strands separate and that each then serves as a template for a new progeny strand. This is called semiconservative replication because each daughter duplex has one parental strand and one new strand which means that one of the parental strands is “conserved” in each daughter duplex.



Another potential mechanism is conservative replication, in which the two parental strands stay together and somehow produce another daughter helix with two completely new strands.



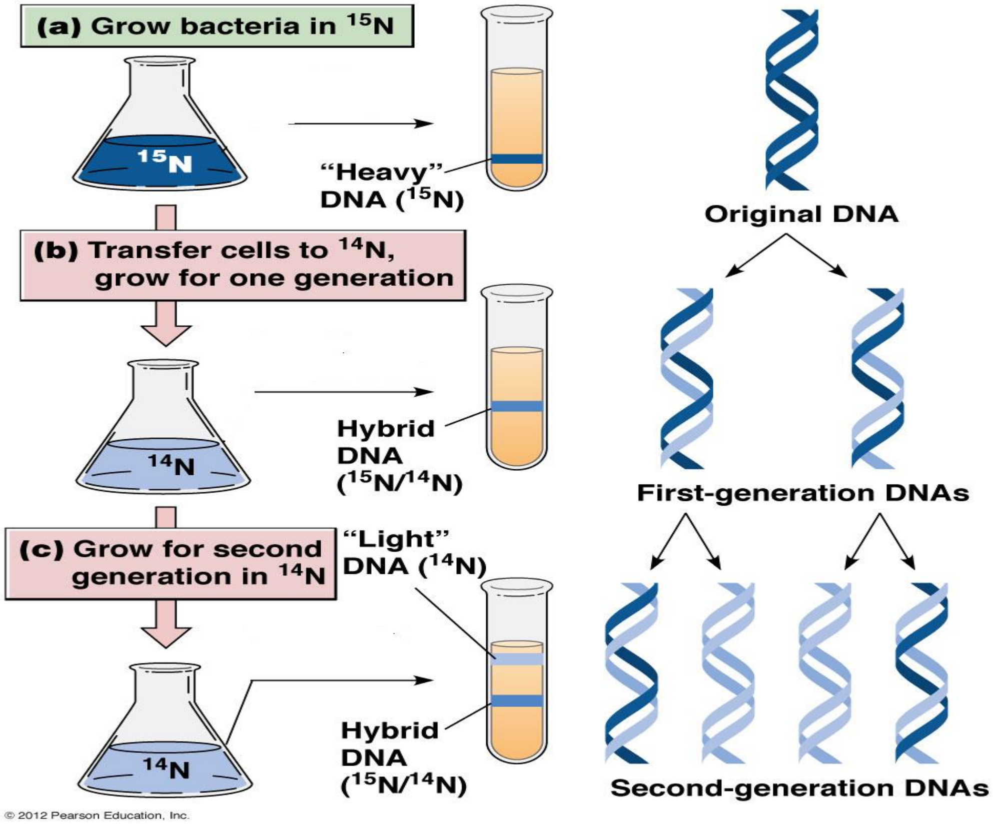
Yet another possibility is dispersive replication, in which the DNA becomes fragmented so that new and old DNAs coexist in the same strand after replication.



37. Experiment of Meselson & Stahl

The Meselson–Stahl experiment was an experiment by Matthew Meselson and Franklin Stahl in 1958 which supported the hypothesis that DNA replication wassemiconservative. In semiconservative replication, when the double stranded DNA helix is replicated, each of the two new double-stranded DNA helixes consisted of one strand from the original helix and one newly synthesized. It has been called "the most beautiful experiment in biology. Meselson and Stahl decided the best way to tag the parent DNA would be to change one of the atoms in the parent DNA molecule. Since nitrogen is found in the nitrogenous bases of each nucleotide, they decided to use an isotope of nitrogen to distinguish between parent and newly-copied DNA. The isotope of nitrogen had an extra neutron in the nucleus, which made it heavier. They labeled *E. coli* DNA with heavy nitrogen (15N) by growing cells in a medium enriched in this nitrogen isotope. This made the DNA denser than normal. Then they switched the cells to an ordinary medium containing primarily ^{14}N, for various lengths of time.

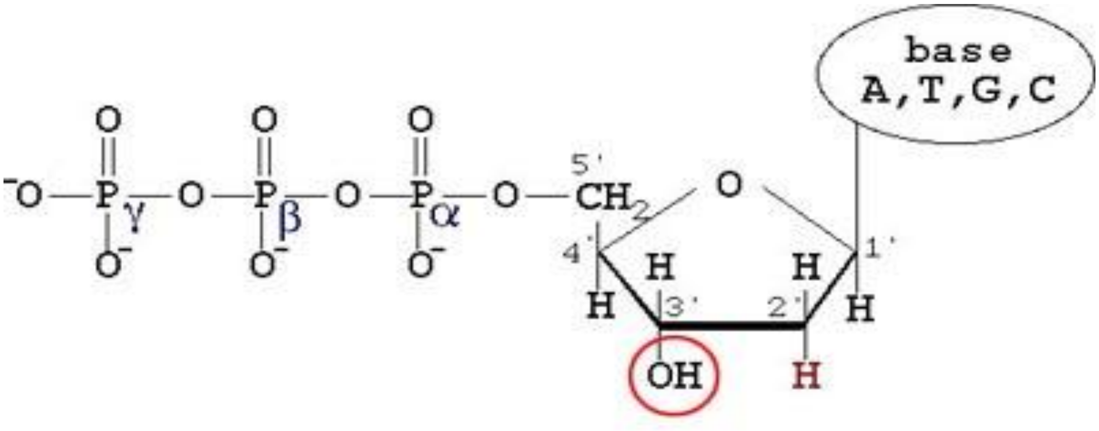
Finally, they subjected the DNA to density gradient centrifugation to determine the density of the DNA.


(a) Grow bacteria in ^{15}N
^{15}N
"Heavy" DNA (^{15}N)
(b) Transfer cells to ^{14}N, grow for one generation
^{14}N
Hybrid DNA ($^{15}N/^{14}N$)
(c) Grow for second generation in ^{14}N
"Light" DNA (^{14}N)
^{14}N
Hybrid DNA ($^{15}N/^{14}N$)
Original DNA
First-generation DNAs
Second-generation DNAs

38 & 39. Chemistry of DNA Synthesis

Oligonucleotide synthesis is the chemical synthesis of relatively short fragments of nucleic acids with defined chemical structure (sequence). Two key substrates are required for the synthesis of DNA to proceed.
Deoxynucleoside triphosphates
Primer:template junction

Four deoxynucleoside triphospahtes namely dGTP, dCTP, dATP & dTTP are required.
Nucleoside triphosphates have three phosphoryl groups attached to the 5' hydroxyl of deoxyribose. The innermost phosphoryl group is called the α-phosphate whereas the middle and outermost groups are called β- and ɣ- phosphates.



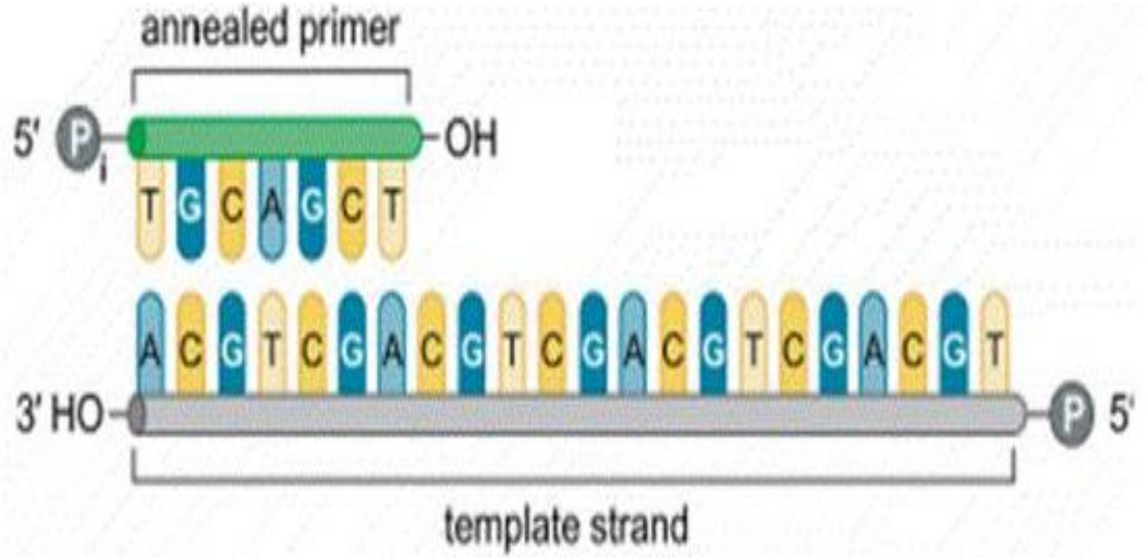
dNTP
deoxyribonucleotide triphosphate

The second important substrate for DNA synthesis is a particular arrangement of single stranded DNA (ssDNA) and double stranded DNA (dsDNA).This particular arrangement is called a primer:template junction.

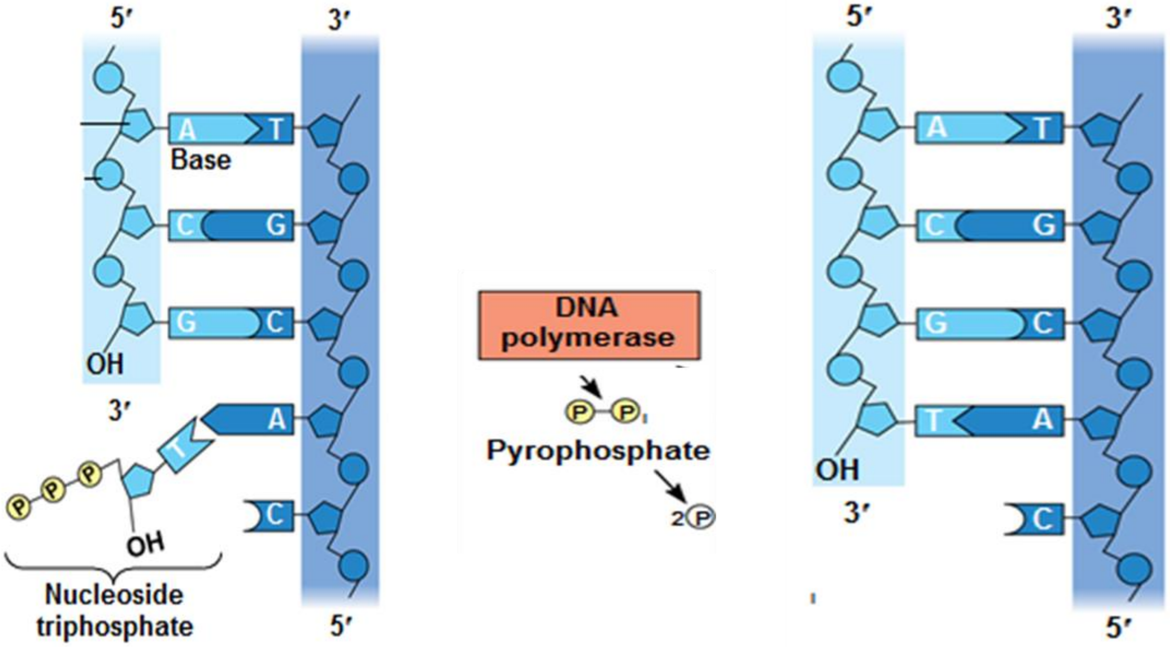
It has two components:-

The Template

The Primer



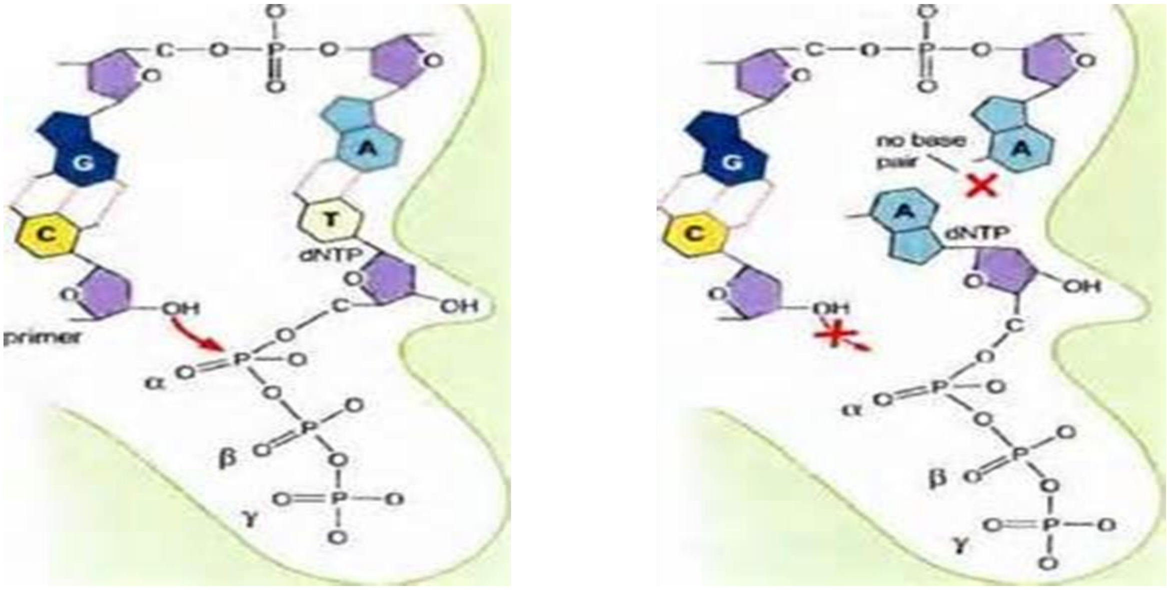
The new chain of DNA grows by extending the 3’ end of the primer. The phosphodiester bond is formed in an SN2 reaction. In this reaction, the hydroxyl group of the 3’ end of the primer attacks the α-phosphoryl group of the incoming nucleoside triphosphate. The leaving group of the reaction is pyrophosphate which arises from the release of β- and ɣ- phosphates of the nucleoside.



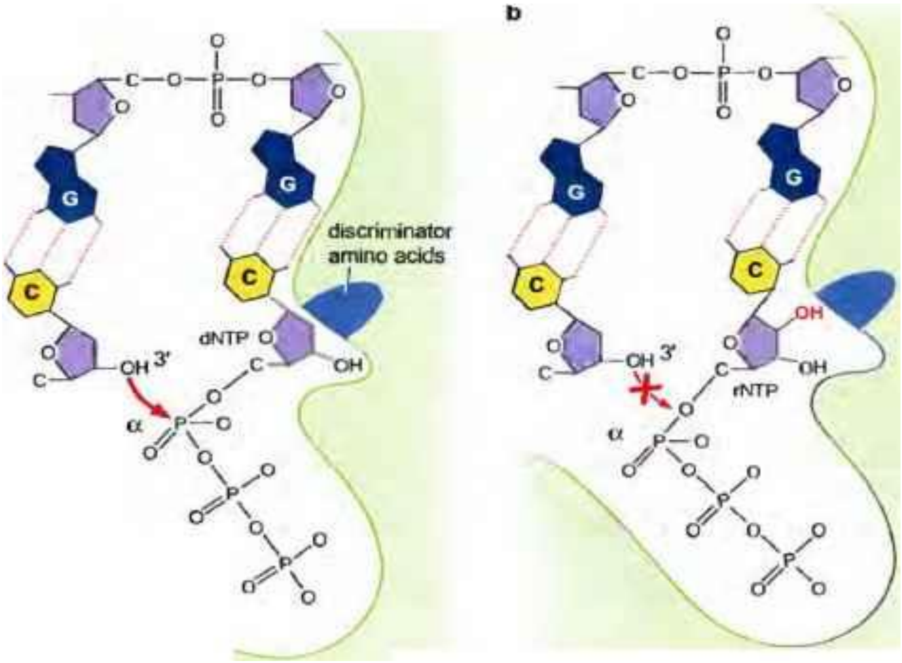
The template strand directs which of the four nucleoside triphosphates is added. The incoming nucleoside triphosphate base pairs with the template strand. The free energy for this reaction is provided by the rapid hydrolysis of the pyrophosphate into two phosphate groups by an enzyme known as pyrophosphatase. The net result of nucleotide addition and pyrophosphate hydrolysis is the simultaneous breaking of two high energy phosphate bonds. Therefore, DNA synthesis is a coupled process. This reaction is highly favourable with high value of $K_{\mathbf{eq}}$ which means that its an irreversible reaction.

40. Mechanism of DNA Polymerase

The DNA polymerases are enzymes that create DNA molecules by assembling nucleotides, the building blocks of DNA. These enzymes are essential to DNA replication and usually work in pairs to create two identical DNA strands from a single original DNA molecule. Every time a cell divides, DNA polymerase is required to help duplicate the cell's DNA, so that a copy of the original DNA molecule can be passed to each of the daughter cells. It uses a single active site to catalyze the addition of any of four deoxynucleoside triphosphates. DNA polymerase monitors the ability of the incoming nucleotide to form an A:T or G:C base pair, rather than detecting the exact nucleotide that enters the active site. Only when a correct nucleotide comes, the 3'-OH of the primer and the α-phosphate of the nucleotide align in optimum position for catalysis to take place.

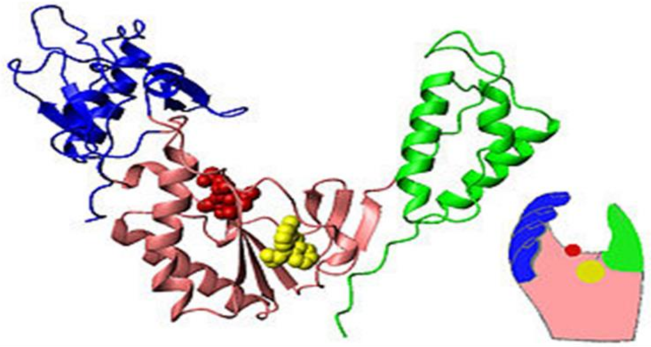


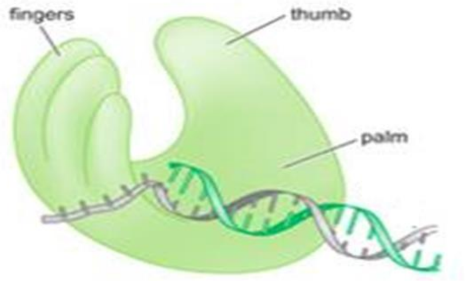
Incorrect base pairing leads to dramatically lower rate of nucleotide addition as a result of catalytically unfavourable alignment of these substrates. DNA polymerase shows an impressive ability to distinguish between ribonucleoside (rNTPs) and deoxyribonucleoside triphosphates (dNTPs). Although rNTPs are present at approx. ten-fold higher concentration in the cell, yet their incorporation rate is 1000-folds lower than dNTPs. This discrimination is mediated by the steric exclusion of rNTPs from the active site of DNA polymerase. In DNA polymerase, the nucleotide-binding pocket cannot accommodate a 2’-OH on the in-coming nucleotide. This space is occupied by two amino acids that make van der Waals contacts with the deoxyribose ring. These amino acids are called discriminator amino acids.



41 & 42. DNA Polymerases Resemble a Hand

The structural studies on DNA polymerases reveal that the DNA substrate sits in a large cleft that resembles a partially closed right hand. Based on the hand analogy, the three domains of the DNA polymerase are called the thumb, fingers and palm.





The palm domain is composed of a β-sheet and contains the primary elements of the catalytic site.

This region of DNA polymerase binds two divalent metal ions (Mg^{2+} or Zn^{2+}). One metal ion reduces the affinity of the 3'-OH for its hydrogen.

This generates a $3'\text{-}O_2$ that is primed for the nucleophilic attack of the α-phosphate of the incoming dNTP. The second metal ion coordinates the negative charges of the β- and γ-phosphates of the dNTP and stabilizes the pyrophosphate produced by joining the primer and the incoming nucleotide. In addition to its role in catalysis, the palm domain also monitors the base pairing of the most recently added nucleotides. The fingers of the polymerase are also important for catalysis.

Several residues located within the fingers bind to the incoming dNTP. More importantly, once a correct base pair is formed between the incoming dNTP and the template, the finger domain moves to enclose the dNTP.

This closed form of the polymerase "hand" stimulates catalysis. In contrast to the fingers and the palm, the thumb domain is not intimately involved in catalysis.

Instead, it interacts with the DNA that has been most recently synthesized. This serves two purposes:-

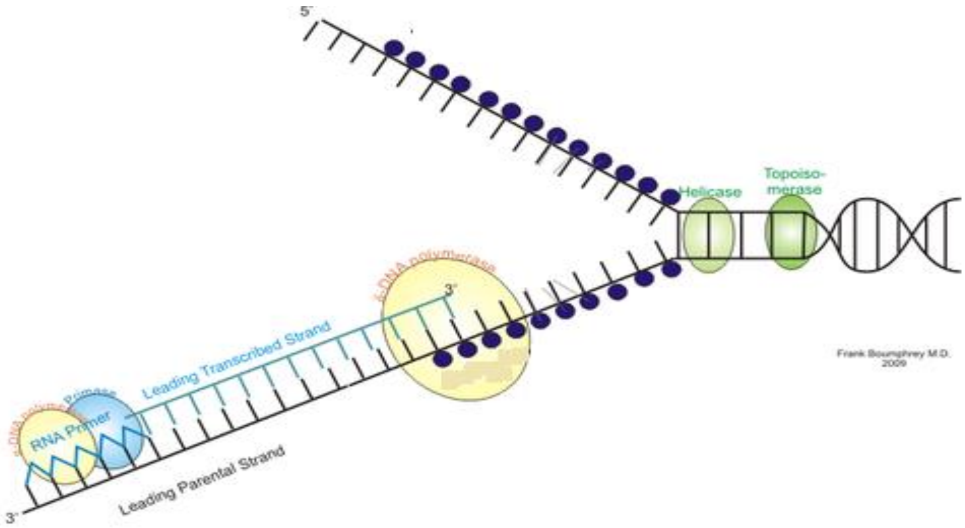
First, it maintains the correct position of the primer and the active site. Second, the thumb helps to maintain a strong association between the DNA polymerase and its substrate.

This association contributes to the ability of the DNA polymerase to add many dNTPs.

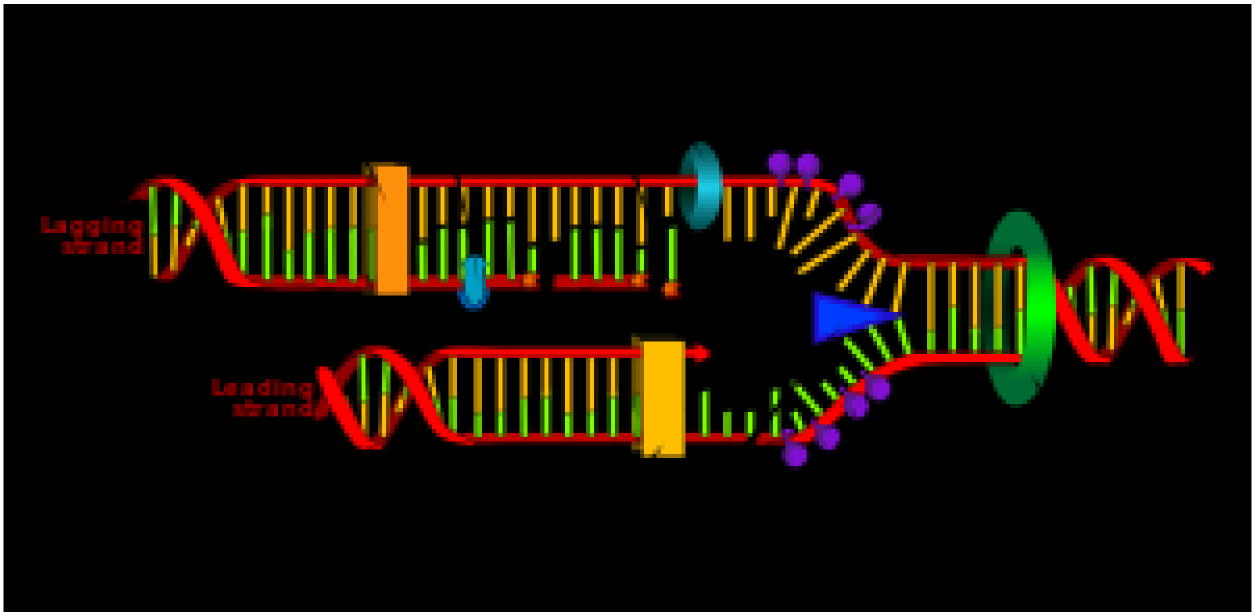
43. THE REPLICATION FORK

In the cell, both strands of the DNA duplex are replicated at the same time. So it requires separation of the two strands of the double helix to create two template DNAs. The junction between the newly separated template strands and the unreplicated duplex DNA is known as the Replication Fork.

The replication fork moves continuously towards the duplex region of unreplicated DNA. As the fork moves, it creates two ssDNA templates that each directs the synthesis of a complementary DNA strand. The antiparallel nature of DNA creates a complication for the simultaneous replication of the two exposed templates at the replication fork. Because DNA is synthesized only by elongating a 3’ end, only one of the two exposed templates can be replicated continuously as the replication fork moves. The newly synthesized DNA strand directed by this template is known as the leading strand.



Synthesis of the newDNA strand directed by the other ssDNA template is more complicated.
This template directs the DNA polymerase to move in the opposite direction of the replication fork. The new DNA strand directed by this template is known as the lagging strand.
This strand of DNA must be synthesized in a discontinuous fashion. Synthesis of the lagging strand must wait for movement of the replication fork to expose a substantial length of template before it can be replicated. Each time a substantial length of the template is exposed, DNA synthesis is initiated and continues until it reaches the 5' end of the previous newly synthesized fragment of lagging strand DNA. The resulting short fragments of new DNA formed on the lagging strand are called Okazaki fragments. They vary in length from 1000 to 2000 nucleotides in bacteria and from 100 to 400 nucleotides in eukaryotes.

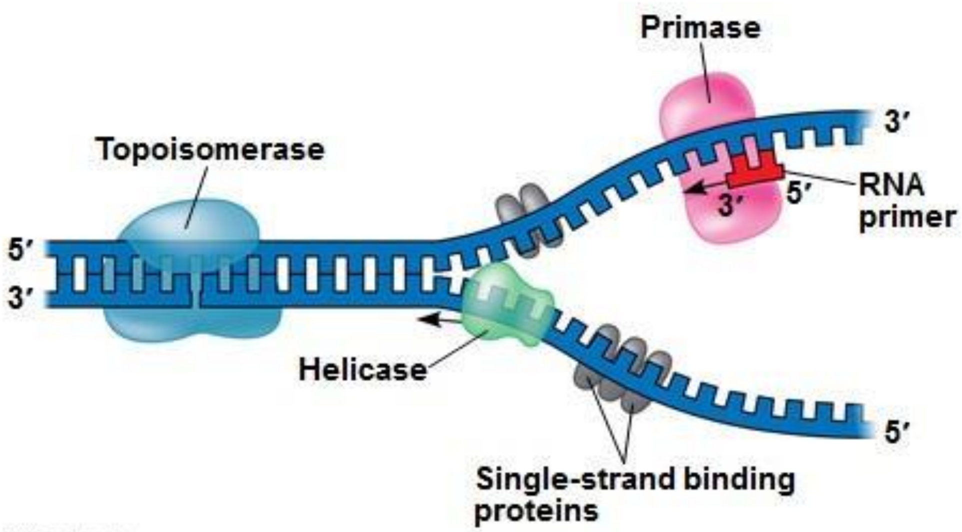


Shortly after being synthesized, Okazaki fragments are covalently joined together to generate a continuous, intact strand of new DNA. Okazaki fragments are therefore transient intermediates in DNA replication.

44. THE RNA PRIMER

All DNA polymerases require a primer with a free 3'-OH. They cannot initiate the synthesis of new DNA strand de novo. To accomplish this, the cell takes advantage of the ability of RNA polymerases to do what DNA polymerases cannot: start new RNA chains de novo. Primase is a specialized RNA polymerase dedicated to making short RNA primers (5–10 nucleotides long) on a ssDNA template. These primers are then extended by DNA polymerase.

Figure 16.13



© 2011 Pearson Education, Inc.

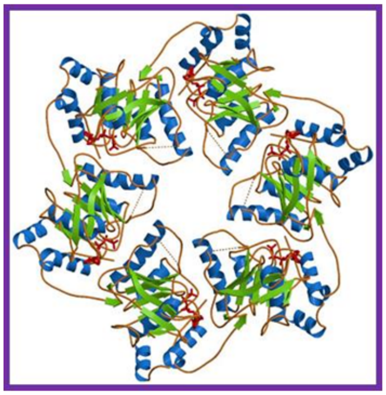
●

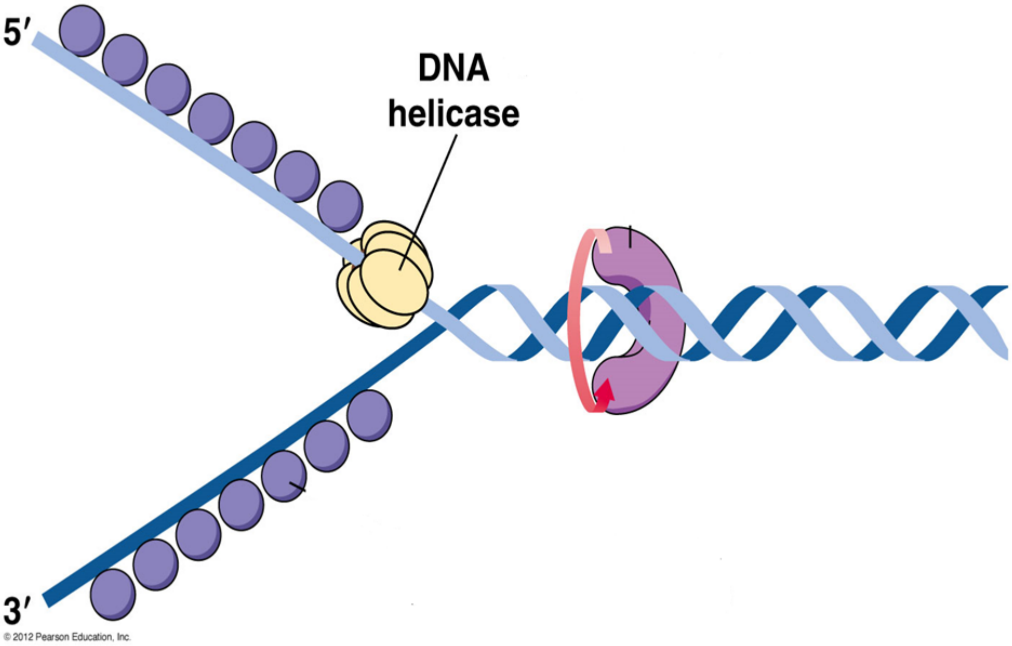
Both the leading and lagging strands require primase to initiate DNA synthesis. Each leading strand requires only a single RNA primer. The discontinuous synthesis of the lagging strand means that new primers are needed for each Okazaki fragment.

Synthesis of the lagging strand can require hundreds of Okazaki fragments and their associated RNA primers. Primase activity is dramatically increased when it associates with another protein that acts at the replication fork called DNA Helicase. This protein unwinds the DNA at the replication fork, creating an ssDNA template that can be acted on by primase.

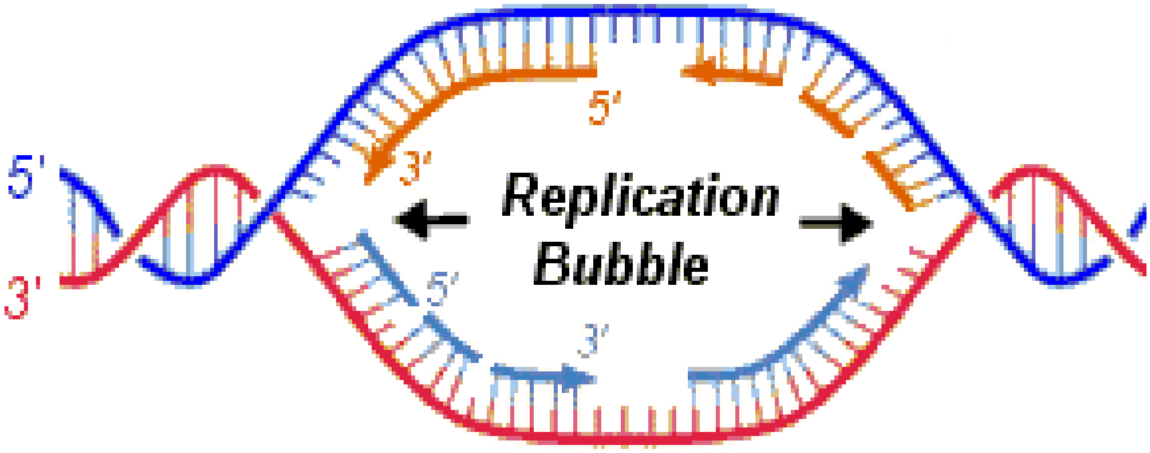
45. THE DNA HELICASE

DNA polymerases are unable to separate the two strands of duplex DNA. Therefore a third class of enzymes, called DNA Helicases catalyze the separation of the two strands of duplex DNA at the replication fork. DNA helicases are hexameric proteins that assume the shape of a ring. This ring encircles one of the two single strands at the replication fork adjacent to the single-stranded:double-stranded junction.





DNA helicases found at replication forks exhibit high processivity because they encircle the DNA. They associate with the DNA and unwind multiple base pairs of DNA. Release of the helicase from the DNA therefore requires the opening of the hexameric protein ring, which is a rare event. However, the helicase can dissociate when it reaches the end of the DNA strand. Of course, this arrangement of enzyme and DNA poses problems for the binding of the DNA helicase to the DNA strand in the first place. Thus, there are specialized mechanisms that open the DNA helicase (hexameric) ring and place it around the DNA before re-forming the ring. Each DNA helicase moves along ssDNA in a defined direction This property is referred to as the polarity of the DNA helicase. DNA helicases can have a polarity of either 5’ 3’ or 3’ 5’. This direction is always defined according to the strand of DNA bound rather than the strand that is displaced.



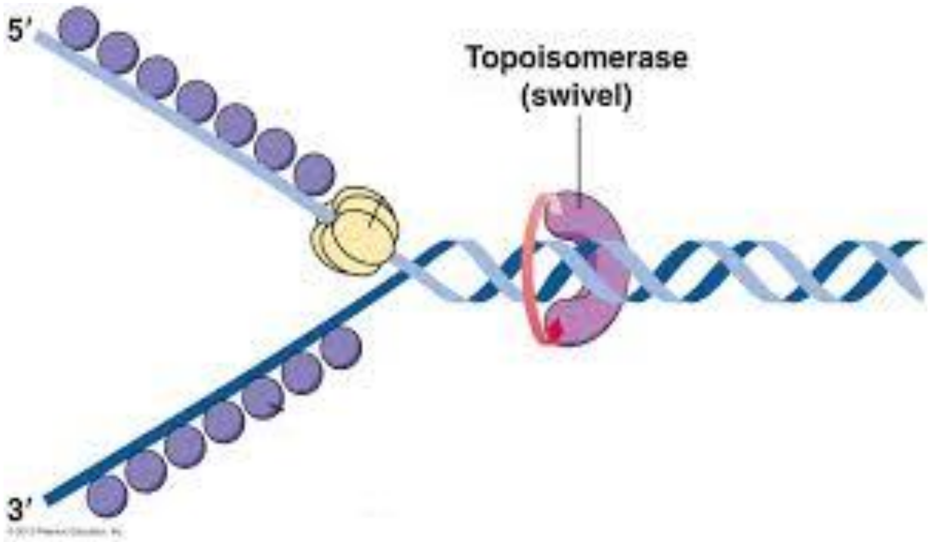
(c) 2000 Chemis

46. TOPOISOMERASES

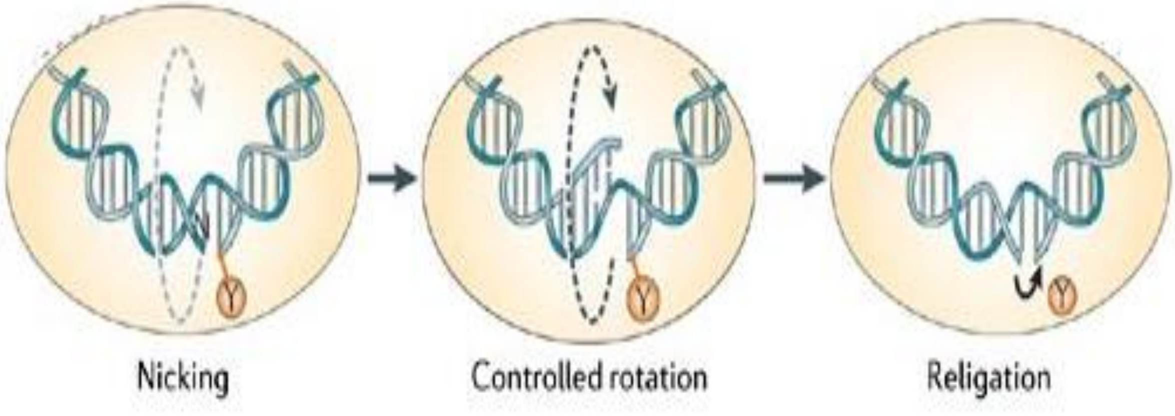
Topoisomerases are enzymes that regulate the overwinding or underwinding of DNA. The winding problem of DNA arises due to the intertwined nature of its double-helical structure. During DNA replication and transcription, DNA becomes overwound ahead of a replication fork. If left unabated, this tension would eventually stop the ability of RNA & DNA polymerase involved in these processes to continue down the DNA strand.

In order to prevent and correct these types of topological problems caused by the double helix, topoisomerases bind to either single-stranded or double-stranded DNA and cut the phosphate backbone of the DNA. This intermediate break allows the DNA to be untangled or unwound, and, at the end of these processes, the DNA backbone is resealed again. Since the overall chemical composition and connectivity of the DNA do not change, the tangled and untangled DNAs are chemical isomers, differing only in their global topology, thus their name. Topoisomerases are isomerase enzymes that act on thetopology of DNA.

Bacterial topoisomerase and human topoisomerase proceed via the same mechanism for replication and transcription.



These enzymes do this by breaking either one or both strands of the DNA without letting go of the DNA and passing the same number of DNA strands through the break.



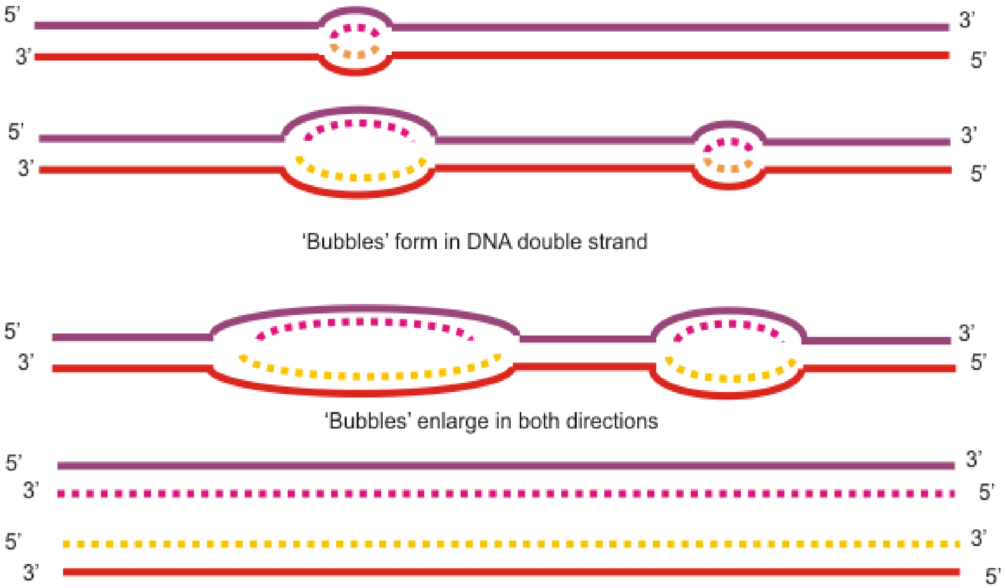
This action relieves the accumulation of supercoils. In this way, topoisomerases act as a “swivelase” that prevents the accumulation of supercoils ahead of the replication fork.

47. INITIATION OF REPLICATION

For a cell to divide, it must first replicate its DNA. This process is initiated at particular points in the DNA, known as "origins", which are targeted by initiator proteins. In *E. coli* this protein is DnaA; in yeast, this is the origin recognition complex. Sequences used by initiator proteins tend to be "AT-rich" (rich in adenine and thymine bases), because A-T base pairs have two hydrogen bonds (rather than the three formed in a C-G pair) which are easier to unzip. Once the origin has been located, these initiators recruit other proteins and form the pre-replication complex, which unzips the double-stranded DNA.

The initial formation of a replication fork requires the separation of the two strands of the DNA duplex to provide the ssDNA. ssDNA is required for DNA helicase binding and to act as a template for the synthesis of both the RNA primer and new DNA. Although DNA strand separation (DNA unwinding) is most easily accomplished at chromosome ends, but DNA synthesis generally initiates at internal regions. As the circular chromosomes lack the chromosome ends so it makes internal DNA unwinding essential for the replication initiation. The specific sites at which DNA unwinding and initiation of replication occur are called Origins of Replication. Depending on the organism, there may be as few as one or as many as thousands of origins per chromosome.

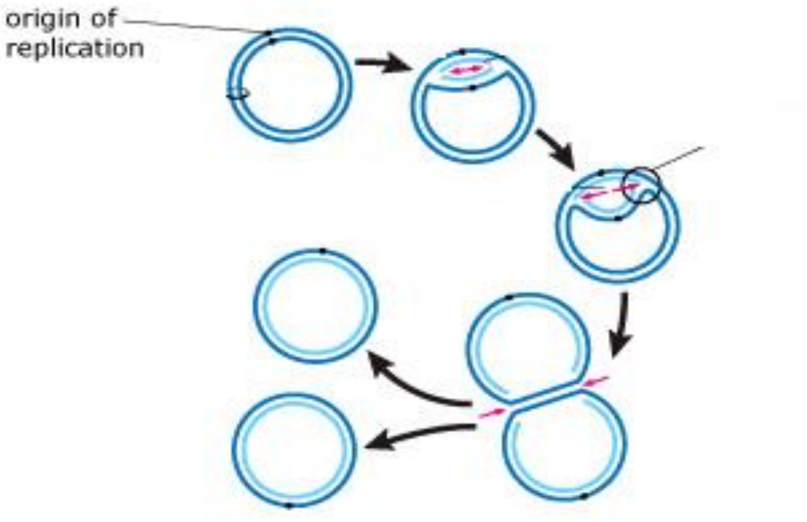
Dotted line= duplicating DNA strands



'Bubbles' form in DNA double strand

'Bubbles' enlarge in both directions

'Bubbles' join up forming Duplicate strands



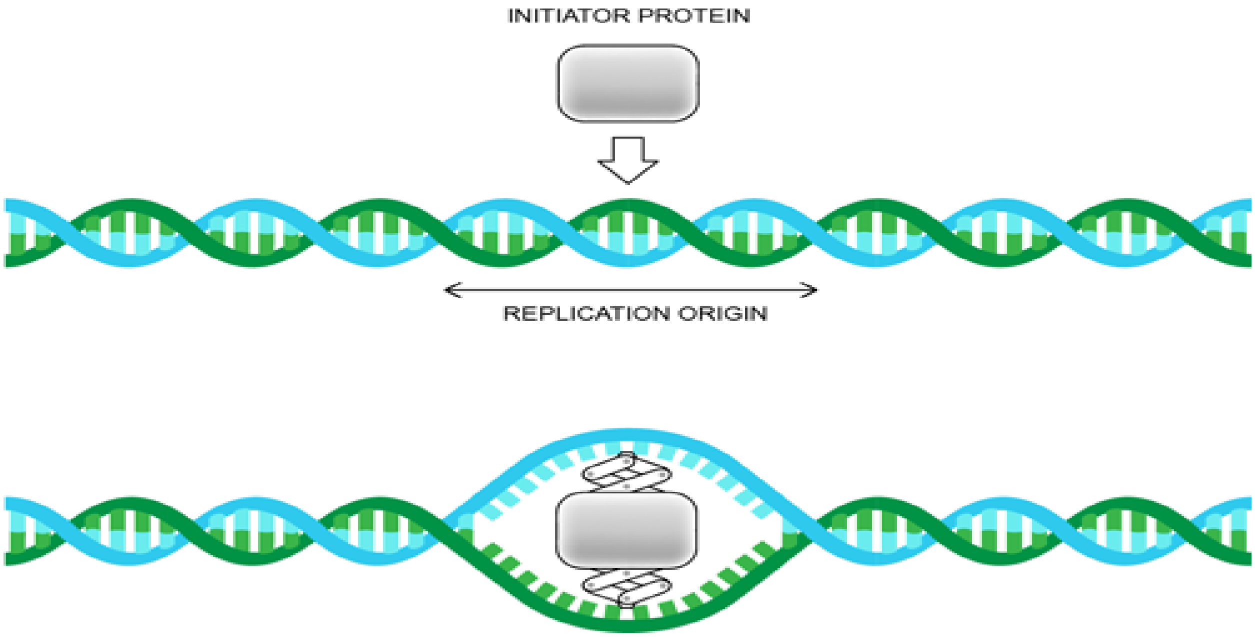
From *Biology* by Campbell and Reece
© 2008 Pearson Education, Inc.

48. THE REPLICON MODEL

Initiation of replication was explained by Francois Jacob, Sydney Brenner and Jacques Cuzin in 1963, which is called as The Replicon Model of replication Initiation. They defined all of the DNA replicated from a particular origin of replication as a replicon. As the single chromosome found in *E. coli* cells has only one origin of replication, the entire chromosome is a single replicon. In contrast, the presence of multiple origins of replication divides each eukaryotic chromosome into multiple replicons; one for each origin of replication.

The replicon model proposed two components that controlled the initiation of replication; the replicator and the initiator. The replicator is defined as the cis-acting DNA sequences that are sufficient to direct the initiation of DNA replication. This is in contrast to the origin of replication, which is the physical site on the DNA where the DNA is unwound and DNA synthesis initiates. Although the origin of replication is always part of the replicator, sometimes the origin of replication is only a fraction of the DNA sequences required to direct the initiation of replication (the replicator).

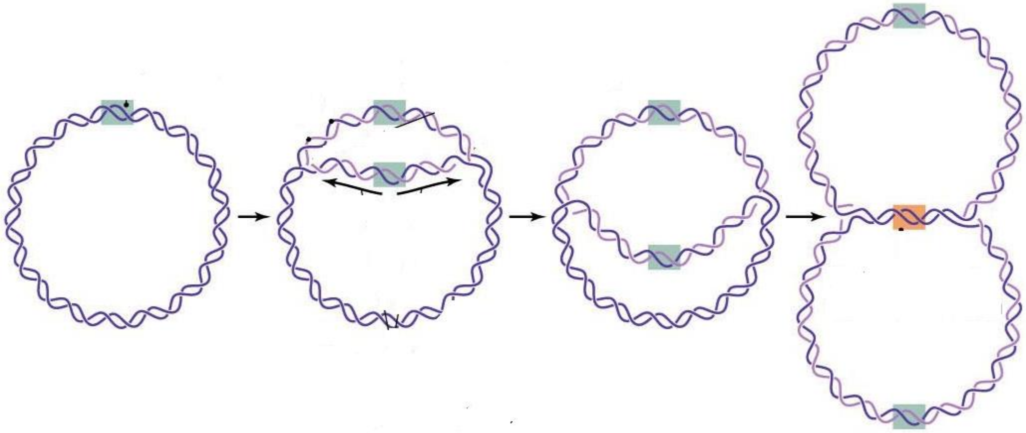
The second component of the replicon model is the initiator protein. This protein specifically recognizes a DNA element in the replicator and activates the initiation of replication. All initiator proteins select the sites that will become origins of replication. The initiator protein is the only sequence-specific DNA binding protein involved in the initiation of replication.



All the remaining proteins other than initiator protein, required for replication initiation do not bind to a DNA sequence specifically.

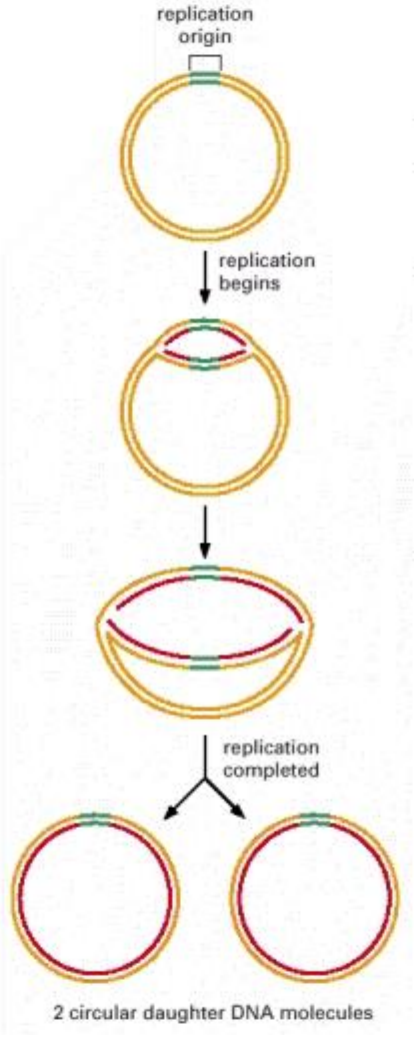
49. FINISHING REPLICATION

Completion of DNA replication requires a set of specific events. These events are different for circular chromosomes and linear chromosomes. In case of circular chromosome, the conventional replication fork machinery replicates the entire molecule, but the resulting daughter molecules are topologically linked to each other.



While in case of linear chromosome, the replication fork machinery cannot complete replication of the very ends of linear chromosomes. Therefore, organisms containing linear chromosomes have developed novel strategies to replicate their chromosome ends. After replication of a circular chromosome is complete, the resulting daughter DNA molecules remain linked together as catenanes.

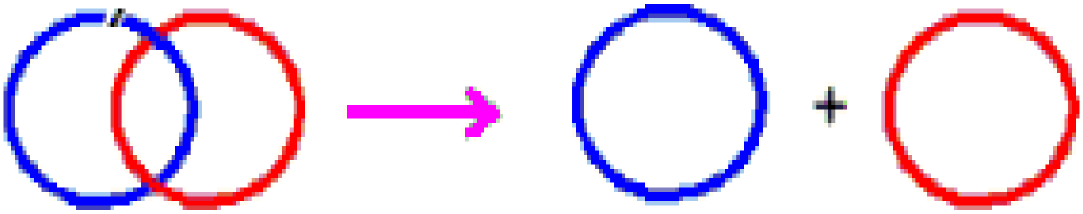
While in case of linear chromosome, the replication fork machinery cannot complete replication of the very ends of linear chromosomes.



50. TYPE II TOPOISOMERASES

Type IIA topoisomerases are essential in the separation of entangled daughter strands during replication. This function is believed to be performed by topoisomerase II in eukaryotes and by topoisomerase IV in prokaryotes. Failure to separate these strands leads to cell death. Type IIA topoisomerases have the special ability to relax DNA to a state below that of thermodynamic equilibrium, a feature unlike type IA, IB, and IIB topoisomerases. This ability, known as topology simplification, was first identified by Rybenkov et al. (Science 1997). The hydrolysis of ATP drives this simplification, but a clear molecular mechanism for this simplification is still lacking.

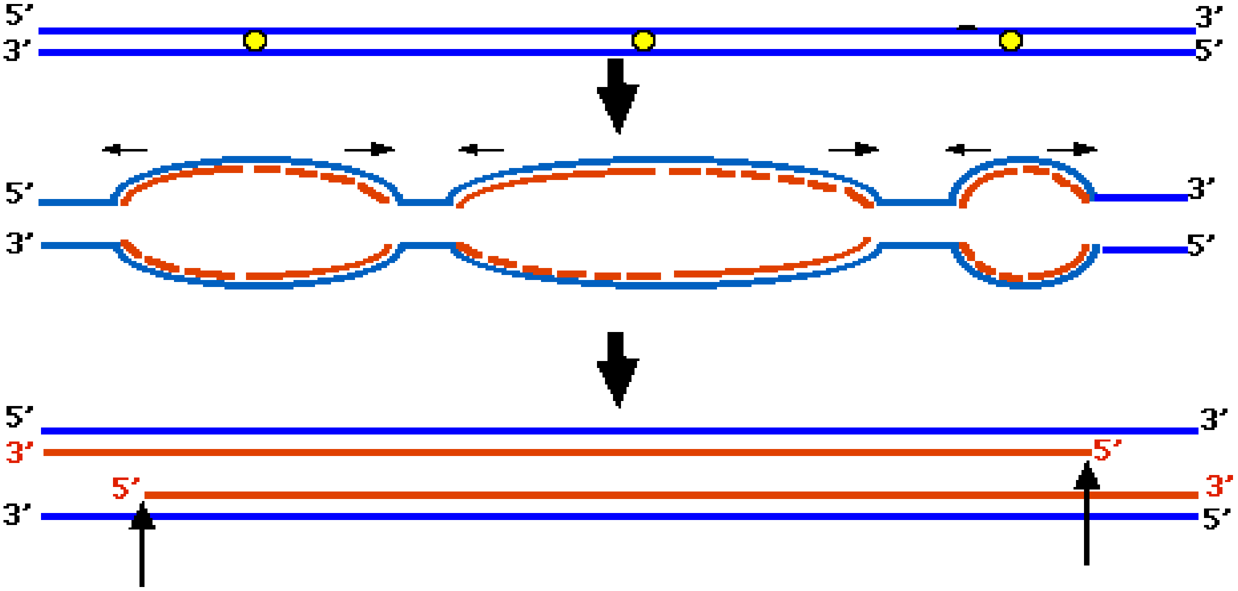
Type II Topoisomerases are the enzymes which have the ability to break a dsDNA molecule and pass a second dsDNA molecule through this break.



So this reaction can easily decatenate the two circular daughter chromosomes and allow their segregation into separate cells. The activity of type II topoisomerases is also critical to the segregation of large linear molecules. Although there is no inherent topological linkage after the replication of a linear molecule, the large sized chromosomes necessitates the intricate folding of the DNA into loops which are attached to a protein scaffold. These attachments lead to many of the same problems that circular chromosomes have after replication.
So type II topoisomerases allow these linked DNAs to be separated. So as in the case of circular chromosomes, type II topoisomerases also allow these linked DNAs to be separated.

51. TELOMERASE

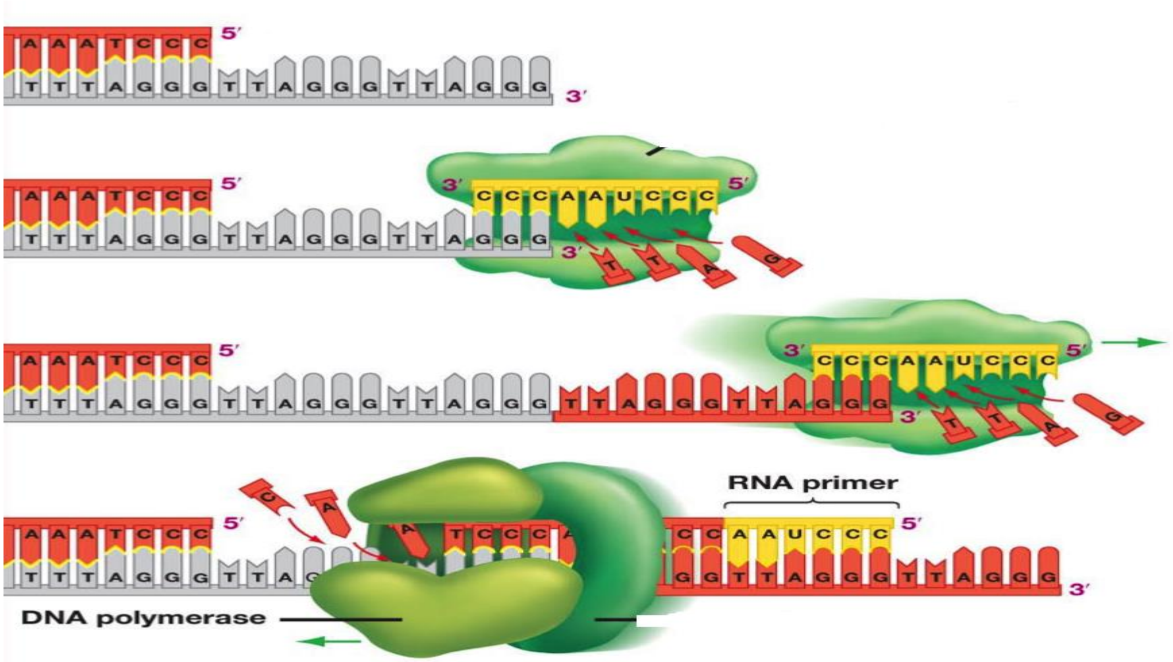
The requirement for an RNA primer to initiate all new DNA synthesis creates a dilemma for the replication of the ends of linear chromosomes. This is called as the end replication problem. This difficulty is not observed during the duplication of the leading-strand template. Because it requires only one RNA primer which completes the DNA synthesis up to extreme terminus of the strand. In contrast, the requirement for multiple primers to complete lagging-strand synthesis means that a complete copy of its template cannot be made. Even if the end of the last RNA primer for Okazaki fragment synthesis anneals to the final base of the lagging-strand template. Once this RNA molecule is removed, there will remain a short region (the size of the RNA primer) of un-replicated ssDNA at the end of the chromosome.



Telomerase, also called telomere terminal transferase,[1] is aribonucleoprotein that adds the polynucleotide "TTAGGG" to the 3' end oftelomeres, which are found at the ends of eukaryotic chromosomes. A telomere is a region of repetitive sequences at each end of a chromatid, which protects the end of the chromosome from deterioration or from fusion with neighbouring chromosomes.

Telomerase is a reverse transcriptase enzyme that carries its own RNAmolecule (with the pattern of "CCCAAUCCC" in vertebrates), which is used as a template when it elongates telomeres.

Like all other DNA polymerases, telomerase acts to extend the 3’ end of its DNA substrate. But unlike most DNA polymerases, telomerase does not need an exogenous DNA template to direct the addition of new dNTPs. Instead, the RNA component of telomerase serves as the template for adding the telomeric sequence to the 3’ terminus at the end of the chromosome. Telomerase specifically elongates the 3’-OH of telomeric ssDNA sequences using its own RNA as a template.
As a result, the newly synthesized DNA is single-stranded. So when telomerase acts on the 3’ end of the telomere, it extends only one of the two strands of the chromosome. This is accomplished by the lagging-strand DNA replication machinery.
By providing an extended 3’ end, telomerase provides additional template for the lagging-strand replication machinery. By synthesizing and extending RNA primers using the telomerase extended 3’ end as a template, the cell can effectively increase the length of the 5’ end of the chromosome as well.



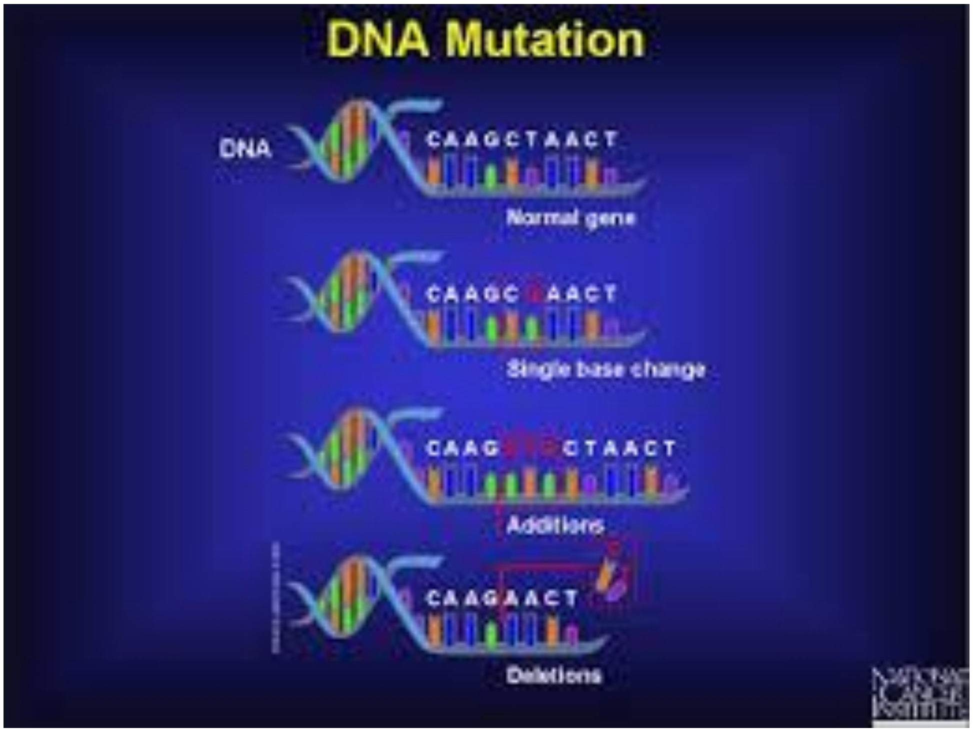
Chapter 8. DNA Mutation

52. DNA MUTATIONS

Mutation is a permanent alteration of the nucleotide sequence of the genome of an organism, virus, or extrachromosomal DNA or other genetic elements. Mutations result from damage to DNA which is not repaired, errors in the process of replication, or from the insertion or deletion of segments of DNA by mobile genetic elements. Mutations may or may not produce discernible changes in the observable characteristics (phenotype) of an organism. Mutations play a part in both normal and abnormal biological processes including: evolution, cancer, and the development of the immune system, including junctional diversity.

Mutation can result in many different types of change in sequences. Mutations in genes can either have no effect, alter the product of a gene, or prevent the gene from functioning properly or completely. Mutations can also occur in nongenic regions.

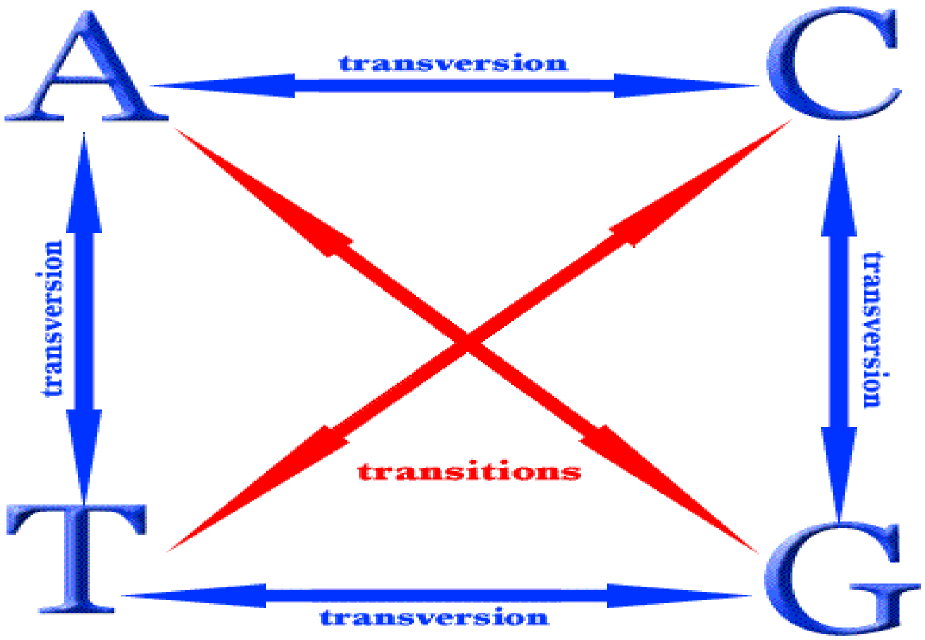
DNA can be easily damaged even under normal physiological conditions. Many different kinds of chemical and physical agents can damage DNA. Some of these agents are endogenous which are produced inside the cells as a result of normal metabolic pathways. While some others are exogenous agents which come from the surrounding environment. On one hand, DNA stability is required to ensure that the genetic information may pass accurately from one generation to the next. It is also require for the correct functioning of thousands of genes. On the other hand the genetic variation is needed to drive evolution. If this variation would be lacking, the new species, including humans, would have not arisen. So the life and biodiversity depend on a happy balance between DNA damage (mutation) and its repair.



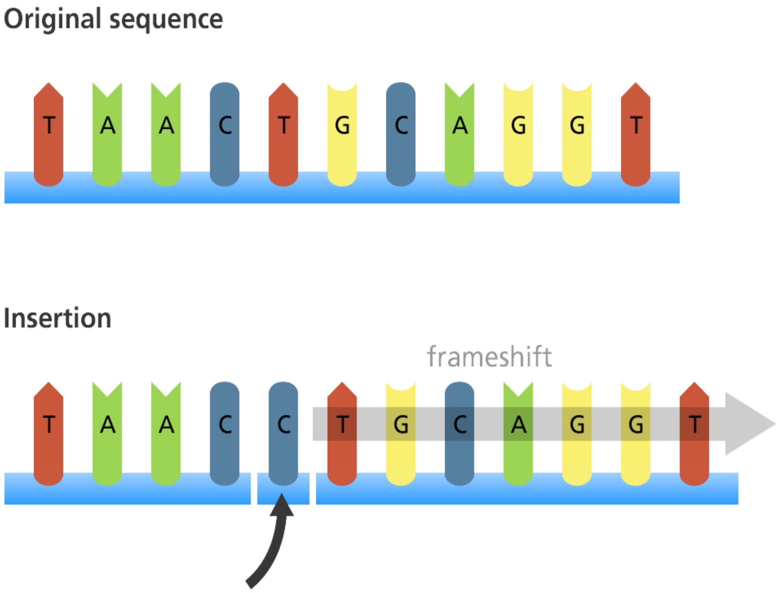
53. NATURE OF MUTATIONS

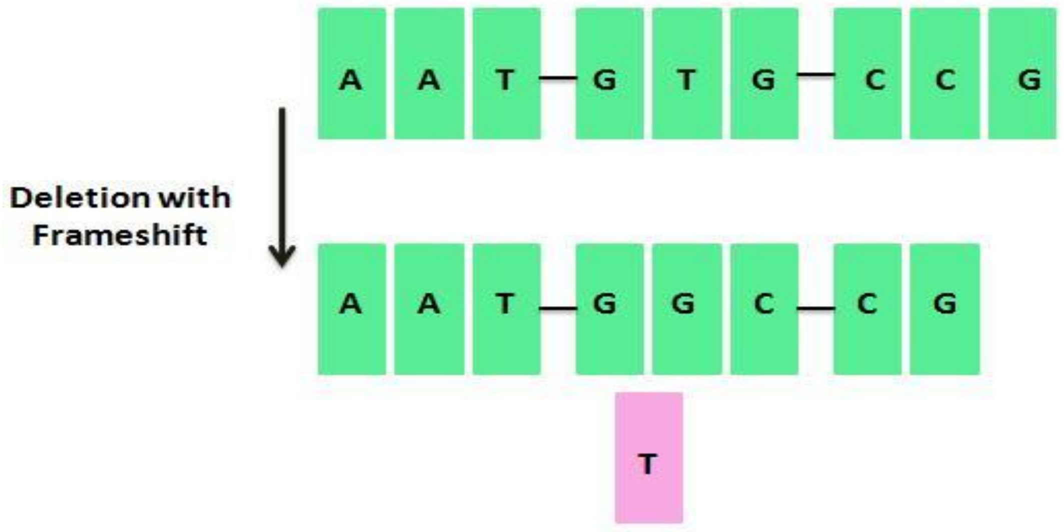
DNA mutations may be very simple (single base change) or very complex and including several thousands of nucleotides. The simplest mutations are switches of one base for another.
There are two kinds of such mutations which include:-
Transitions
Transversions

Transitions are pyrimidine-to-pyrimidine and purine-to-purine substitutions, such as thymine (T) to cytosine (C) and adenine (A) to guanine (G). Transversions are pyrimidine-to-purine and purine-to-pyrimidine substitutions, such as T to G or A and A to C or T.

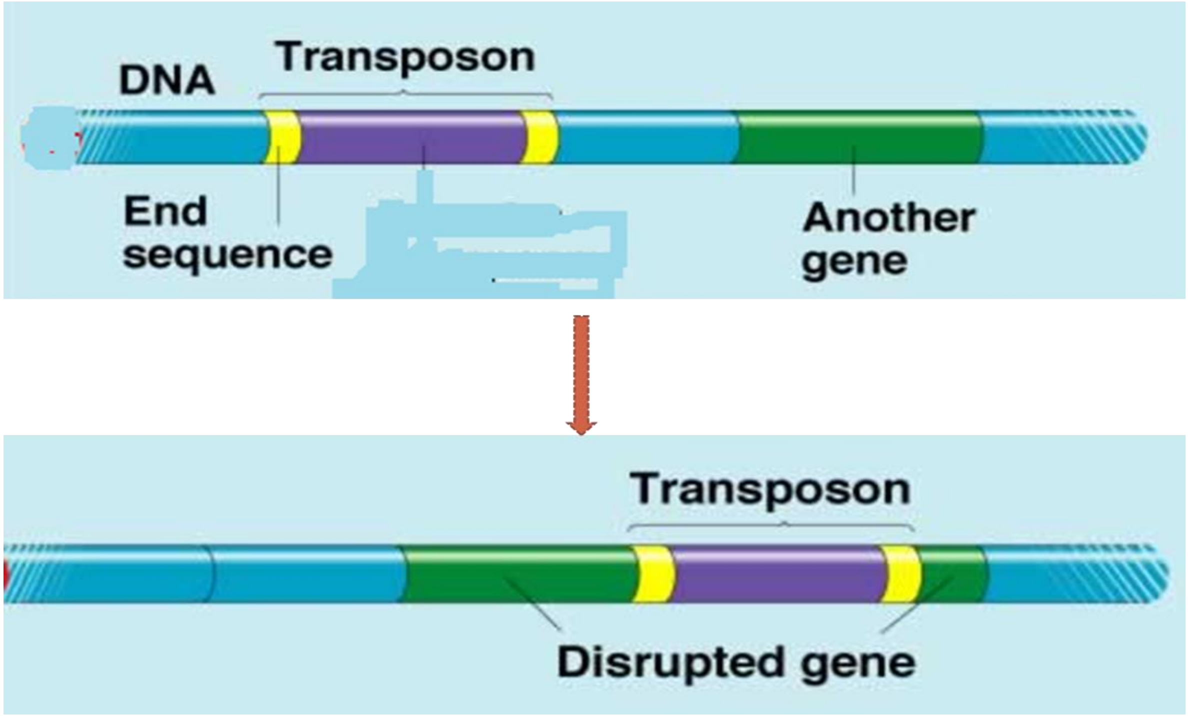


Other simple mutations are insertions or deletions of a nucleotide or a small number of nucleotides.

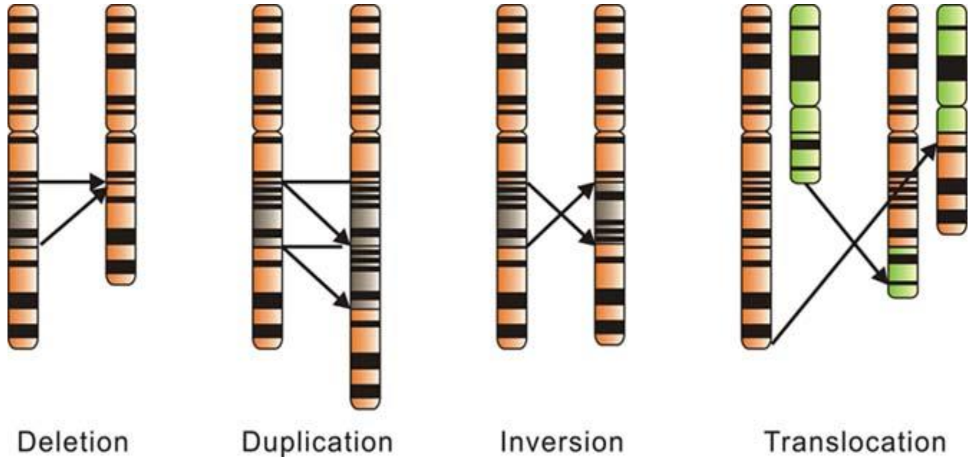




All such mutations that alter a single nucleotide are called point mutations. Other kinds of mutations cause more drastic changes in DNA, such as extensive insertions and deletions and gross rearrangements of chromosome structure. Such changes might be caused, for example, by the insertion of a transposon, which typically places many thousands of nucleotides of foreign DNA in the coding or regulatory sequences of a gene.



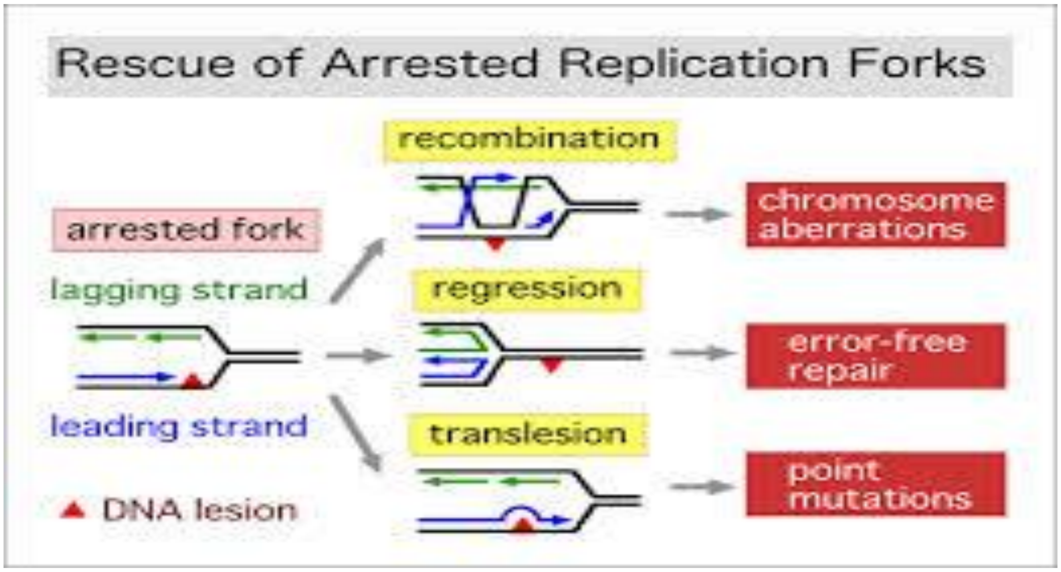
Another type of mutations which are more drastic occur at chromosomal levels. These are changes in appearance of the individual chromosomes through mutation-induced rearrangements.

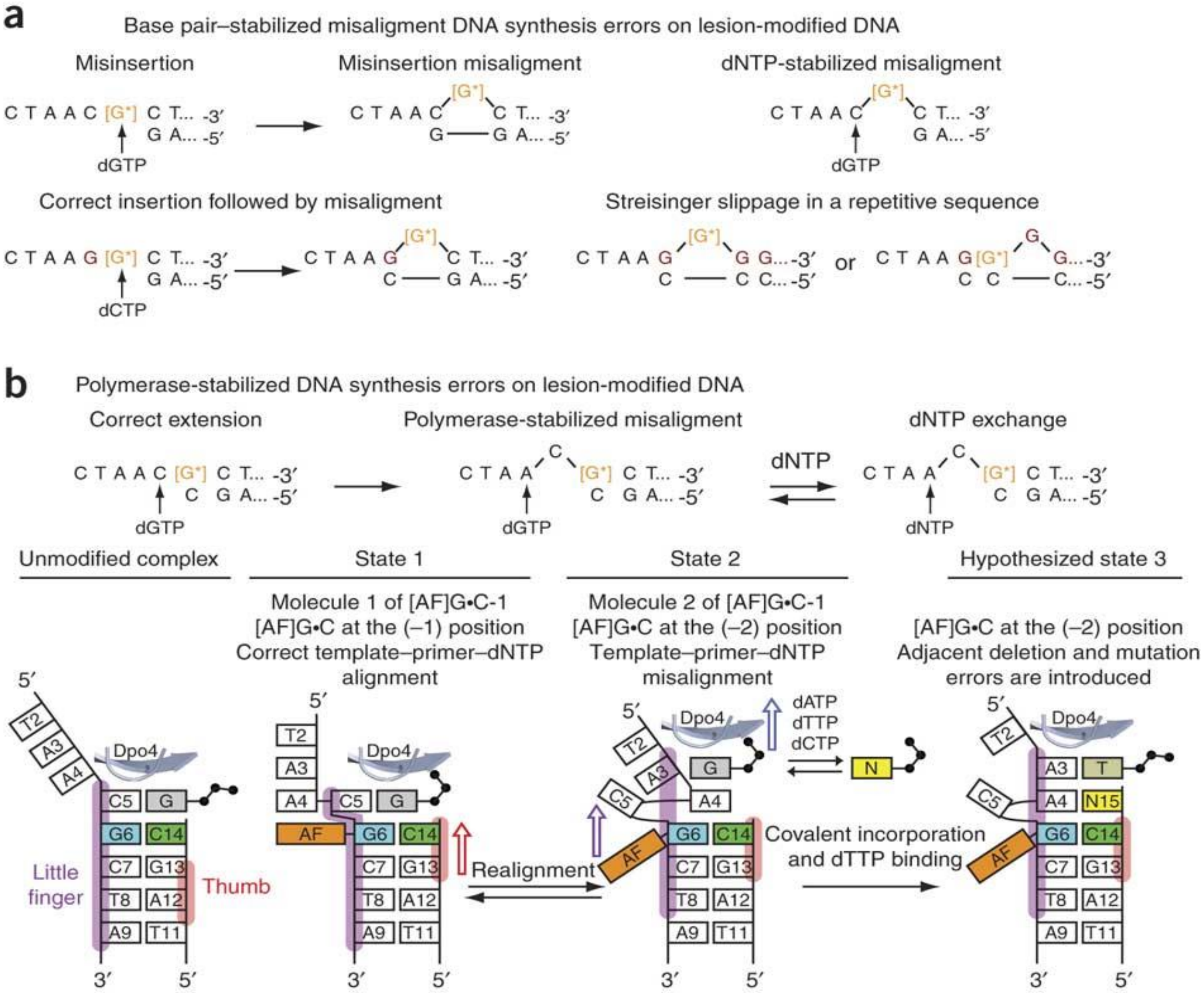


54. REPLICATION ERRORS

The replication machinery achieves a remarkably high degree of accuracy using a proofreading mechanism, which removes wrongly incorporated nucleotides. However, this proofreading is not foolproof.

Some mis-incorporated nucleotides escape detection and become a mismatch between the newly synthesized strand and the template strand. If the misincorporated nucleotide is not subsequently detected and replaced, the sequence change will become permanent in the genome. During a second round of replication, the mis-incorporated nucleotide will direct the incorporation of its complementary nucleotide into the newly synthesized strand. At this point, the mismatch will no longer exist; instead, it will have resulted in a permanent change (a mutation) in the DNA sequence.


Rescue of Arrested Replication Forks
arrested fork
lagging strand
leading strand
▲ DNA lesion
recombination
regression
translesion
chromosome aberrations
error-free repair
point mutations

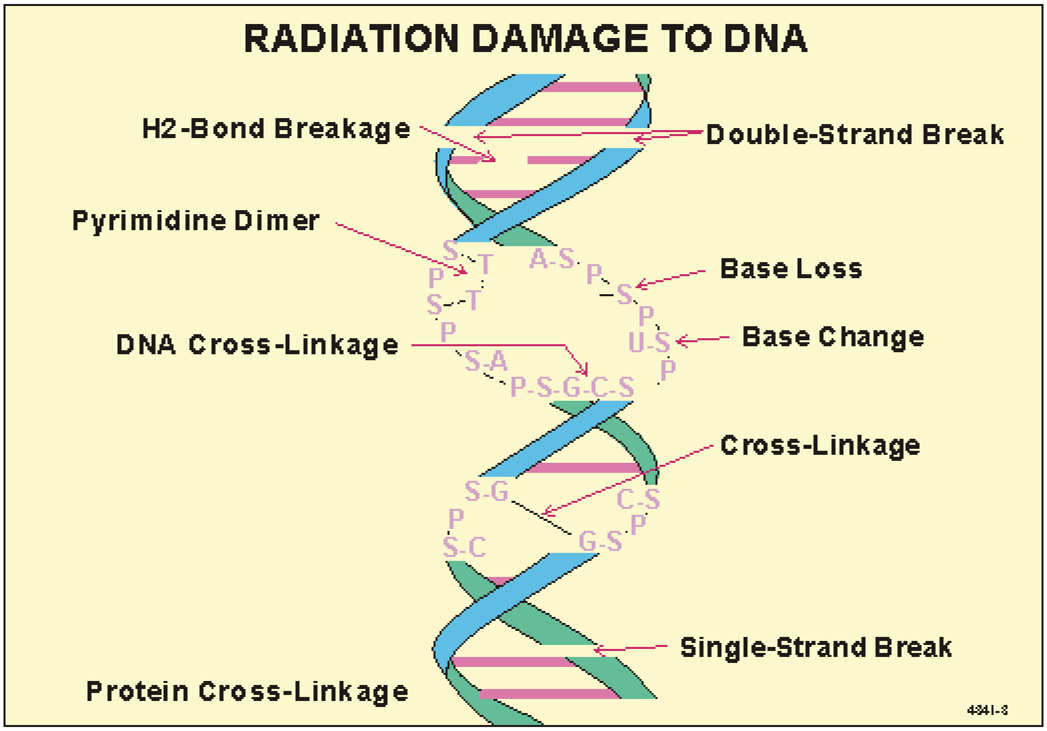

a
Base pair–stabilized misaligment DNA synthesis errors on lesion-modified DNA
Misinsertion
Misinsertion misaligment
dNTP-stabilized misaligment
C T A A C [G*] C T... -3′
G A... -5′
dGTP
Correct insertion followed by misaligment
Streisinger slippage in a repetitive sequence
or
dCTP
b
Polymerase-stabilized DNA synthesis errors on lesion-modified DNA
Correct extension
Polymerase-stabilized misaligment
dNTP exchange
dNTP
Unmodified complex
State 1
State 2
Hypothesized state 3
Molecule 1 of [AF]G•C-1
[AF]G•C at the (–1) position
Correct template–primer–dNTP alignment
Molecule 2 of [AF]G•C-1
[AF]G•C at the (–2) position
Template–primer–dNTP misalignment
[AF]G•C at the (–2) position
Adjacent deletion and mutation errors are introduced
Dpo4
Little finger
Thumb
Realignment
dATP
dTTP
dCTP
Covalent incorporation and dTTP binding
AF
3′
5′

55. RADIATION DAMAGE

Cells are exposed to three types of high energy electromagnetic radiations which can damage their DNA.
These radiations include:-
Ultraviolet light (wavelength 100-400nm)
X-rays (wavelength 0.01-100nm)
Gamma rays (wavelength <0.01nm)
Later two are ionizing radiations.
Ultraviolet light is divided into three bands:

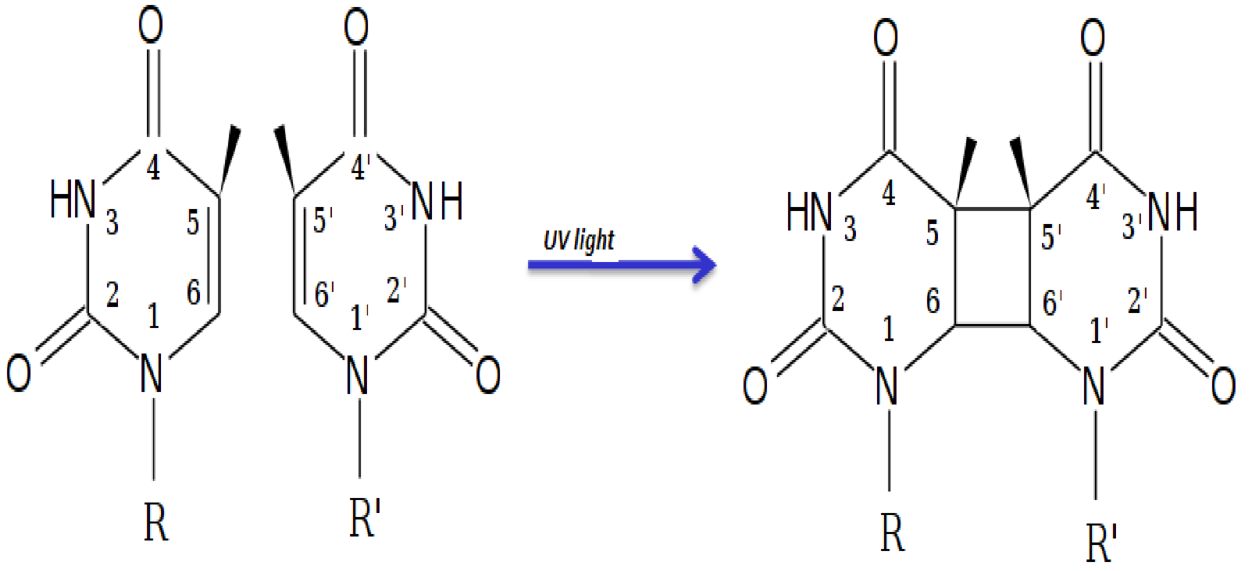
- UV –A (321-400nm)
- UV –B (296-320nm)
- UV –C (100-295nm)

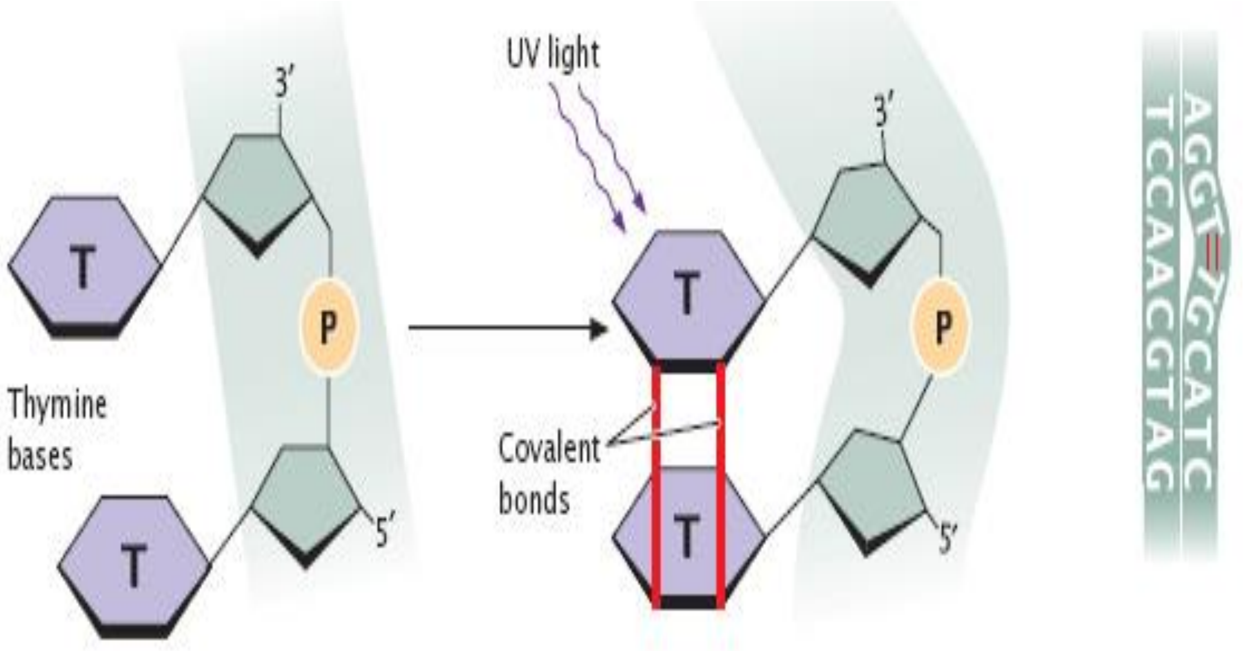
The majority of UV light reaching on earth is UV – A. This is least energetic band and so does little damage to DNA. UV – B accounts for about 10% of the UV radiation reaching the earth's surface. It is responsible for most of the DNA damage in the skin. UV – C includes the wavelength of maximum DNA absorbance (260nm). So it would cause a great deal of DNA damage to exposed organisms if it were able to penetrate the earth's surface. Fortunately, very little UV - C reaches the earth's surface because the ozone layer prevents it from penetration. However, during lab studies, the germicidal lamps that produce UV - C light are used.



56. Cyclobutane pyrimidine Dimer

Pyrimidine dimers are molecular lesions formed from thymine orcytosine bases in DNA via photochemical reactions.[1][2] Ultraviolet light induces the formation of covalent linkages by reactions localized on the C=C double bonds. Two major photoproducts account for nearly all of the UV - induced DNA damage. Which involve the dimer formation between adjacent pyrimidine bases on the same DNA strand. The first pathway includes the formation of cyclobutane pyrimidine dimer (CPD) which accounts for about 75% of all the UV - induced damage. The cyclobutane ring is generated by forming one bond between C-5 atoms and another between C-6 atoms on adjacent pyrimidine rings.





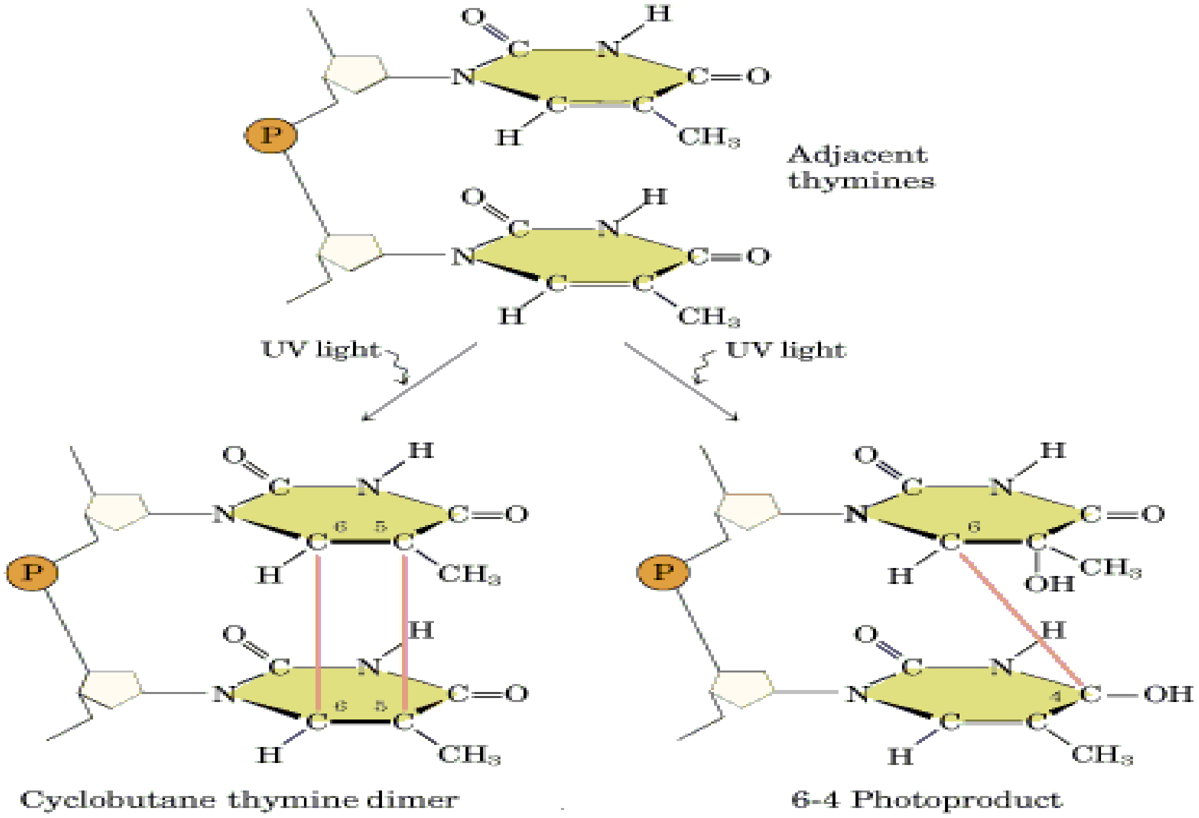
The most common cyclobutane pyrimidine dimer is the thymine-thymine (T< >T) dimer. Cytosine-thymine (C< >T) and cytosine- cytosine (C< >C) dimers are also form but at slower rates. Structural studies show that:-

- 1) B-DNA can accommodate a single T< >T dimer forcing the helical axis to bend by about 30˚ towards the major groove.
- 2) The dimer's 3'-thymine can form a normal base pairing with its adenine partner on the complementary strand.
- 3) The interaction between the 5'-thymine and its complementary adenine partner will be weaker than normal because a single hydrogen bond can be formed here.

Thymine-thymine dimers are often used to study DNA repair systems because they are stable, easy to form and easy to detect.

57. (6-4) Photoproducts

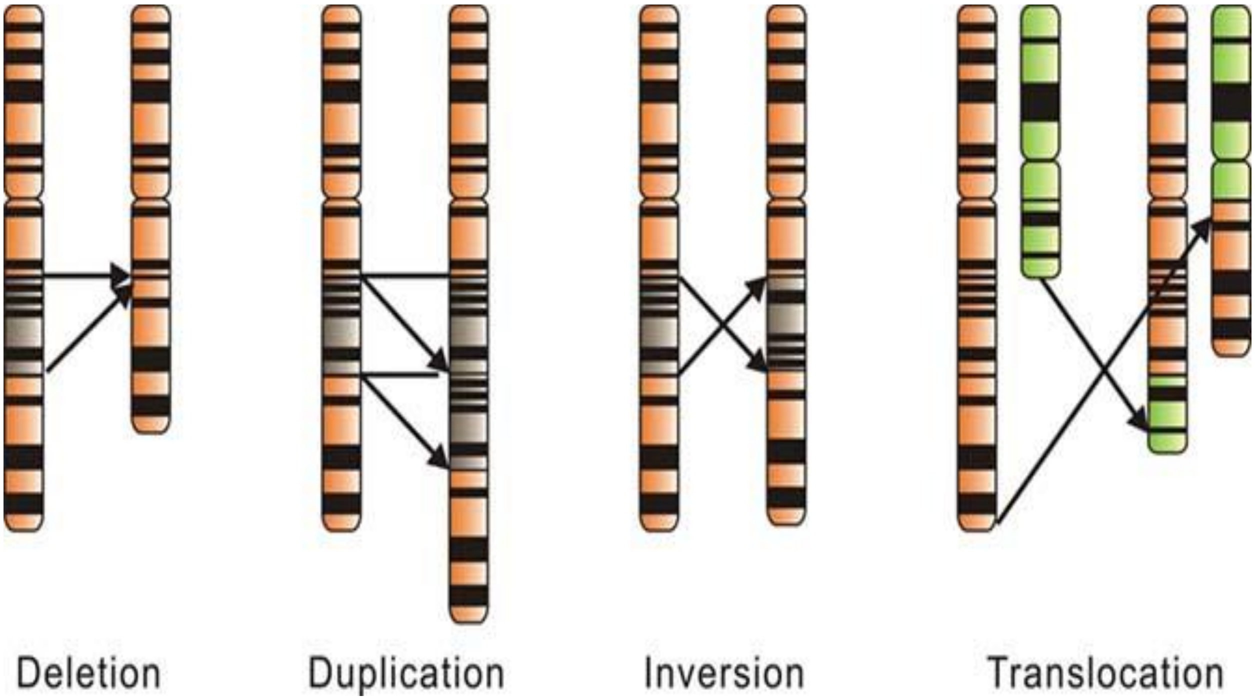
The second pathway, which accounts for most of the remaining UV-induced DNA damage, produces (6-4) photoproducts. A bond is formed between the C-6 atom of the 3'-pyrimidine (either thymine or cytosine) and the C-4 atom of the 5'-pyrimidine (usually cytosine).



The (6-4) photoproduct causes a major distortion in B-DNA because the two pyrimidine rings are perpendicular to each other. If not removed, a pyrimidine dimer or (6-4) photoproduct can interfere with the normal operation of the replication and transcription machinery. This interference results in mutations and cell death. Even if the lesion is removed, the result can be a mutation.

58. X-rays & gamma rays damage

Gamma radiations and X-rays (ionizing radiation) are particularly hazardous because they cause double-strand breaks in the DNA, which are difficult to repair. Ionizing radiations directly or indirectly generate many different kinds of DNA lesions.
Direct damage takes place when DNA or water bound to it absorbs the radiation. Indirect damage takes place when water molecule or any other molecule surrounding the DNA absorb the radiation and form reactive species that then damage the DNA. The DNA lesions caused due to these radiations may be isolated or clustered.
Many lesions within a few helical turns are called clustered lesions. One type of clustered lesion, the double-strand break, is generally thought to be the primary reason that ionization radiation is so lethal to cells. Double-strand breaks are also responsible for various chromosomal aberrations such as deletions, duplications, inversions and translocations.



About 65% of the DNA damage caused by these radiations is due to indirect effects, primarily due to transfer of photons to water. The photon transfer activates the water and causing it to undergo two types of primary reactions.
In the first of these reactions, the water molecule is ionized.

$$H2O \qquad H2O^{\bullet +} + e^{-}$$

The $H2O^{\bullet +}$ that is formed readily dissociates and release a proton and a hydroxyl radical ($^{\bullet}OH$).

$$H2O^{\bullet +} \qquad H^{+} + {}^{\bullet}OH$$

The electron generated by the first reaction can combine with molecular oxygen to form a superoxide radical ($^{\bullet}O2^{-}$).

$$e^{-} + O2 \qquad {}^{\bullet}O2^{-}$$

In the second type of primary reaction, excited water molecule (H2O*) splits into a hydrogen atom and a hydroxyl radical.

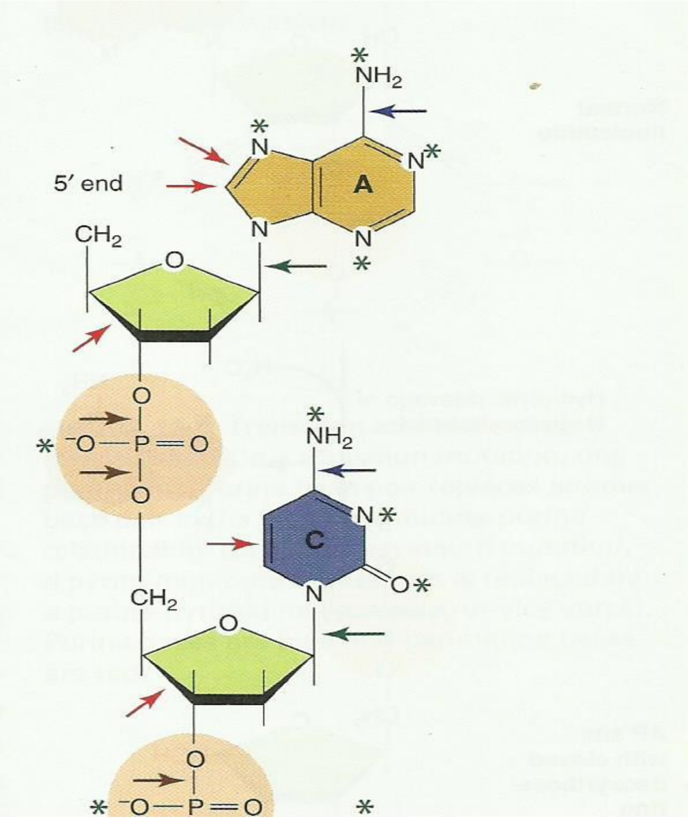
$$H2O^{*} \qquad {}^{\bullet}H + {}^{\bullet}OH$$

So the three highly reactive chemical species i,e., H, OH and $O2^{-}$ are produced by the two primary pathways. Each of these attacks and damages whatever biomolecule they encounter. A wide variety of changes take place when that molecule happens to be DNA.

59. DNA INSTABILITY IN WATER

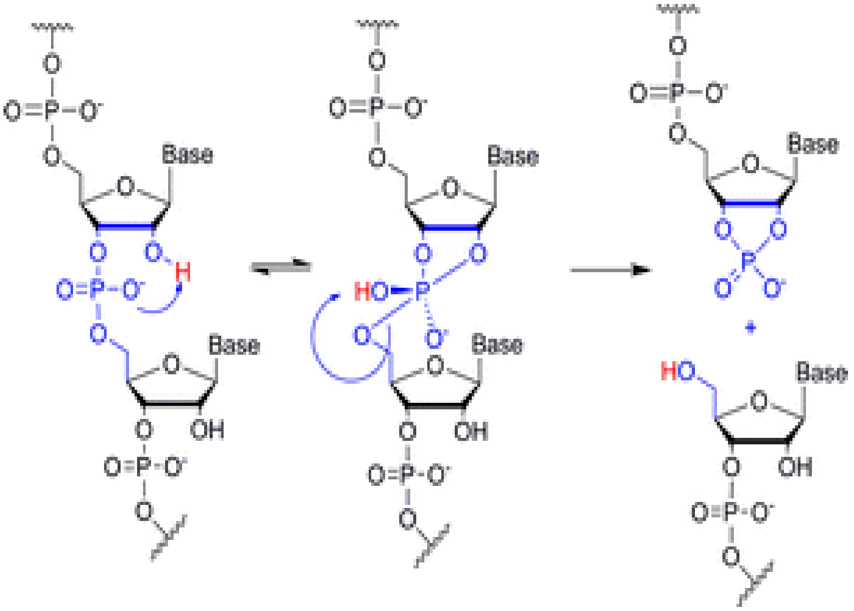
DNA has three kinds of bonds with the potential for hydrolytic cleavage, namely:

- ○ Phosphodiester bond
- ○ N-glycosyl bond
- ○ Bonds linking exocyclic amine groups to bases



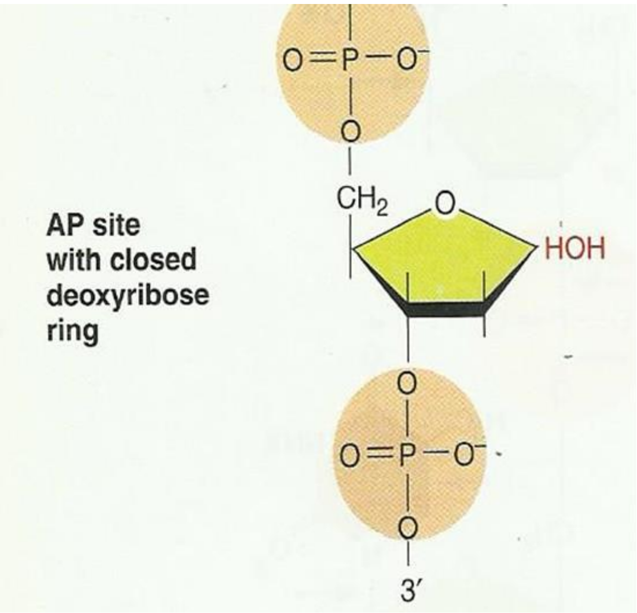
Hydrolytic bond Cleavage

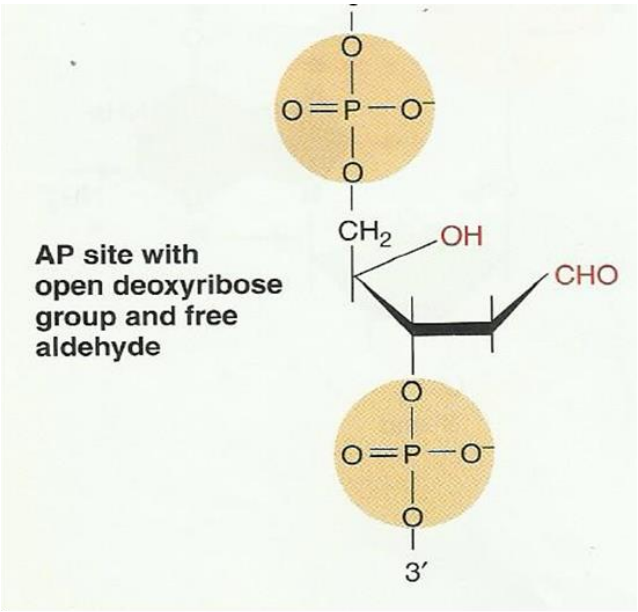
Spontaneous phosphodiester bond cleavage, which introduces a nick in the DNA strand, is a very rare occurrence and doesn't make a significant contribution to DNA damage.



N-glycosyl bond cleavage leads to the formation of an abasic site, which is also known as an AP site (AP for apurinic & apyrimidinic).

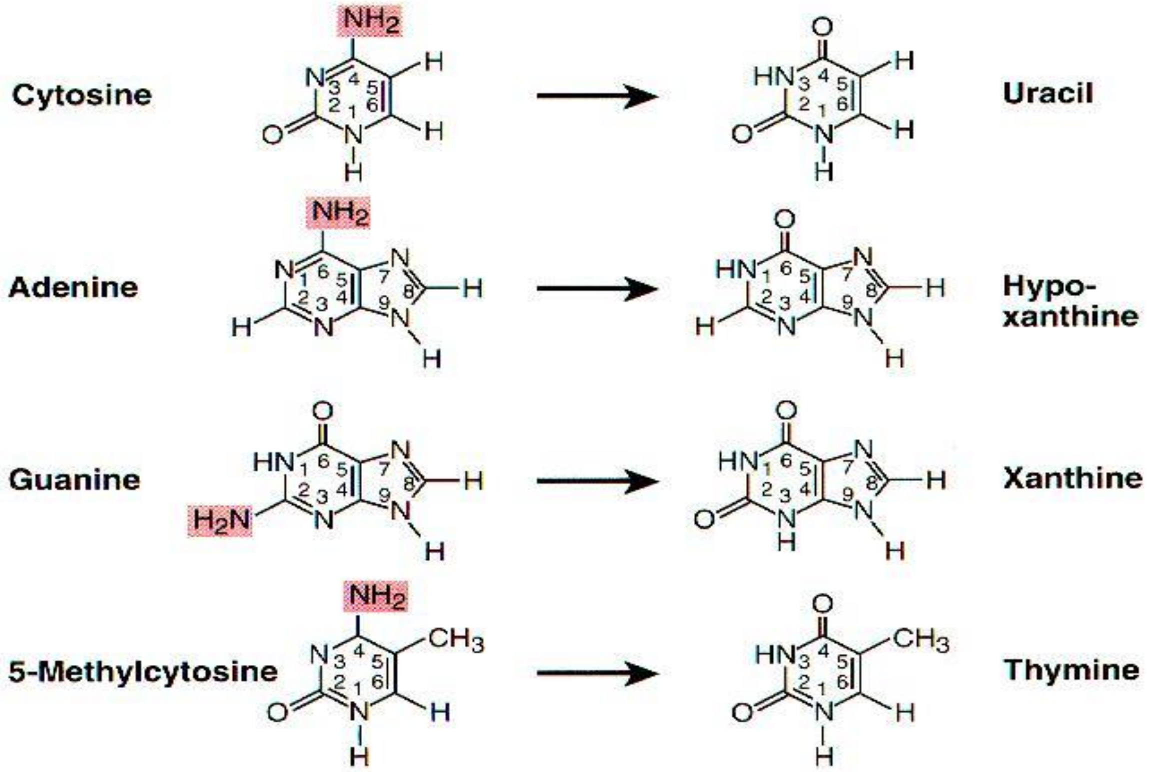
According to current etimates, about 10,000 purine and 500 pyrimidine bases are lost from DNA in a mammalian cell nucleus each day. Experiments are also showing that purine N-glycosyl bonds are more easily hydrolyzed than pyrimidine N-glycosyl bonds. AP site formation sensitizes the neighbouring 3’-phosphodiester bond to cleavage which can be attributed to the formation of a free aldehyde group. A DNA strand with one or more AP sites makes a poor template because it lacks the information required to direct accurate replication and transcription.



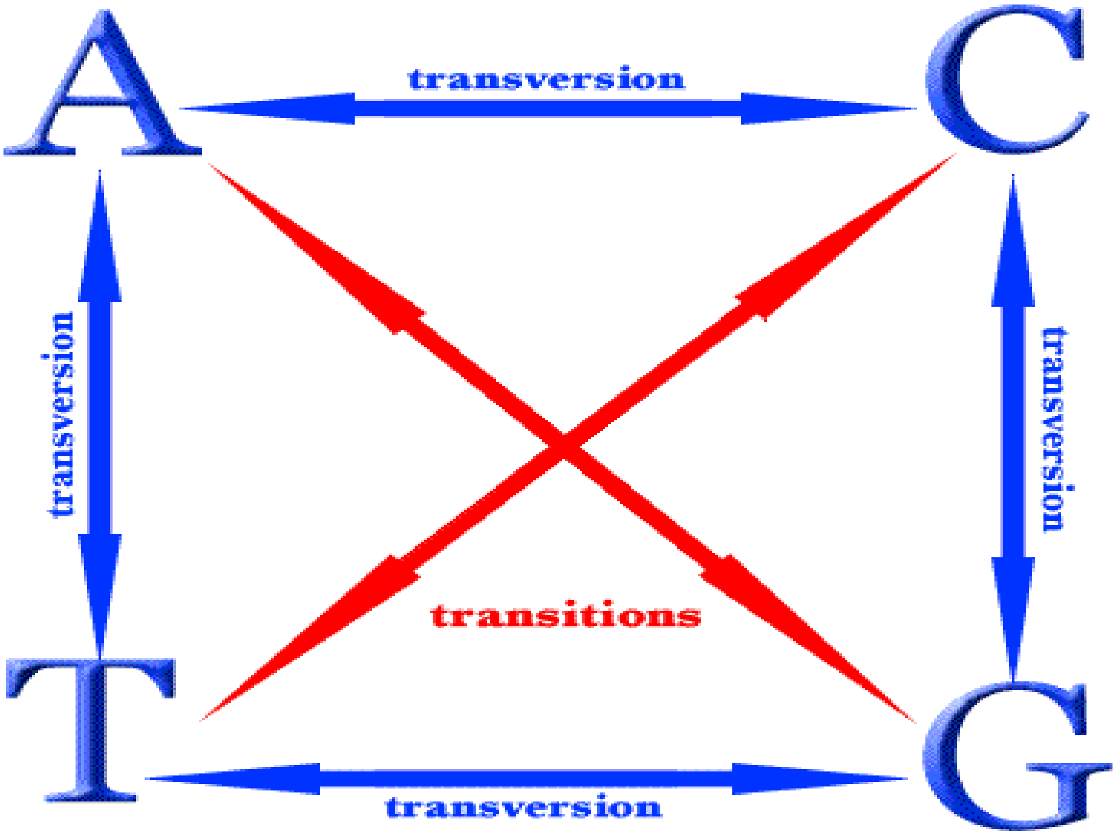


60 &61. Water-mediated Deamination

Water-mediated deamination converts cytosine, guanine and adenine to uracil, xanthine and hypoxanthine, respectively. Hydrolytic deamination of cytosine is estimated to take place about 100 to 500 times a day in a mammalian cell. Whereas, combined guanine and adenine deaminations are estimated to occur at about 1 or 2% of that of cytosine deamination.



The conversion of guanine to xanthine may result in mutations or arrested DNA synthesis because xanthine does not stable base pairs with either cytosine or thymine. While the conversion of adenine to hypoxanthine will cause a T – A base pair to be replaced by C – G base pair. Likewise, an uncorrected deamination that converts cytosine to uracil will cause a C – G base pair to be replaced with a T – A base pair. Mutations of this type in which a pyrimidine on one strand is replaced by a different pyrimidine and a purine on the other strand is replaced by a different purine are called transitions. Another type of replacement mutation which is termed as transversion mutation involves replacing a pyrimidine on one strand with purine and a purine on the other strand with a pyrimidine.



A few cytosine bases in eukaryotic DNA are converted into modified base 5-methylcytosine. This modified base is concentrated in so called CpG islands. CpG islands are small segments of DNA often present in regulatory elements called promoters that are located just before the transcription unit they regulate. The frequency of spontaneous deamination of 5-methylcytosine bases in CpG islands is even greater than those of cytosine. The product of this deamination is thymine and not uracil.

So it results in the conversion of a C – G base pair to a T – A base pair. Nitrous acid (HNO2) is formed from nitrites used as preservatives in processed meat.

It reacts with amine groups attached to the ring structure in C, A and G and greatly increase their rate of deamination. Bisulfite (HSO3) used as an additive in fruit juices and dry fruits also greatly increases the rate of cytosine deamination but does’nt affect purine or 5-methylcytosine deamination.

62 & 63. Oxidative Damage to DNA

The reactive oxygen species damage DNA.

So the reactive oxygen species produced during cellular respiration (ETC) may have the potential to damage DNA. But it is unlikely that they do so because:-

1) the respiratory chain doesn’t normally release these reactive oxygen species.

2) cells contain superoxide dismutase to convert superoxide radicals into molecular oxygen and hydrogen peroxide. Then catalase to convert hydrogen peroxide to oxygen and water.

3) superoxides and hydroxyl radicals are so reactive that, if released from respiratory chain, they would react with the nearby biomolecules before they had chance to reach nuclear DNA.

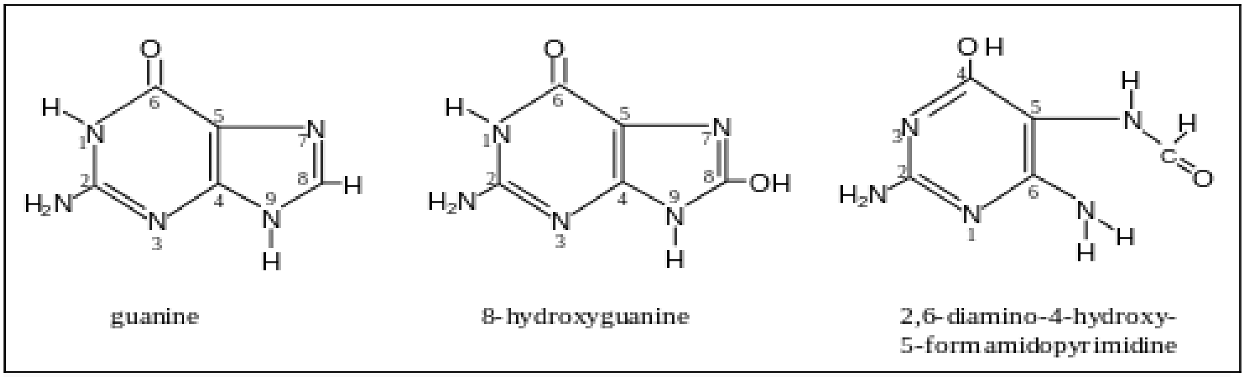
So the reactive oxygen species produced in this way do not damage the nuclear DNA under normal physiological conditions. The primary culprit appears to be the hydroxyl radical, which is produced by ionizing radiations. Hydroxyl radicals can also be produced chemically from hydrogen peroxide. Hydrogen peroxide is not as reactive as superoxide and hydroxyl radicals, so it has a much longer half-life in the cell, provided it escapes catalase and peroxidase.

If it does escape, hydrogen peroxide can be converted into hydroxide radical by the following reaction:-

$$\mathrm{Fe}^{2+} + \mathrm{H2O2} \xrightarrow{} \mathrm{Fe}^{3+} + \bullet\mathrm{OH} + \mathrm{OH}^{-}$$

This reaction is called the Fenton reaction and in this reaction copper, manganese and cobalt can replace iron. Hydroxyl radicals generated by any means are known to cause more than 80 different kinds of base damage.

Two of the oxidized base products are 8-oxoguanine (oxoG) and thymine glycol. 8-Oxoguanine can base pair with adenine or cytosine and if uncorrected, this 8-oxoG – A base pair will be replicated to form a T – A base pair, thus causing a transversion mutation. On the other hand, thymine glycol inhibits DNA replication and is therefore cytotoxic. Hydroxyl radicals produced by the Fenton reaction are tend to be more widely dispersed than those produced by ionizing radiations and therefore, much less likely to produce double-stranded breaks. Cells can repair single-stranded breaks much more easily than they can repair double-stranded breaks. Reactive oxygen species can also convert various biomolecules into reactive species that can then damage DNA. For example, polyunsaturated fatty acid oxidation produces two aldehyde products, malondialdehyde and 4- hydroxynonenal, which contribute to base damage.



oxidative damage of DNA, two of the most common changes guanine to 8-hydroxyguanine or to 2,6-diamino-4-hydroxy-5-formamidopyrimidine

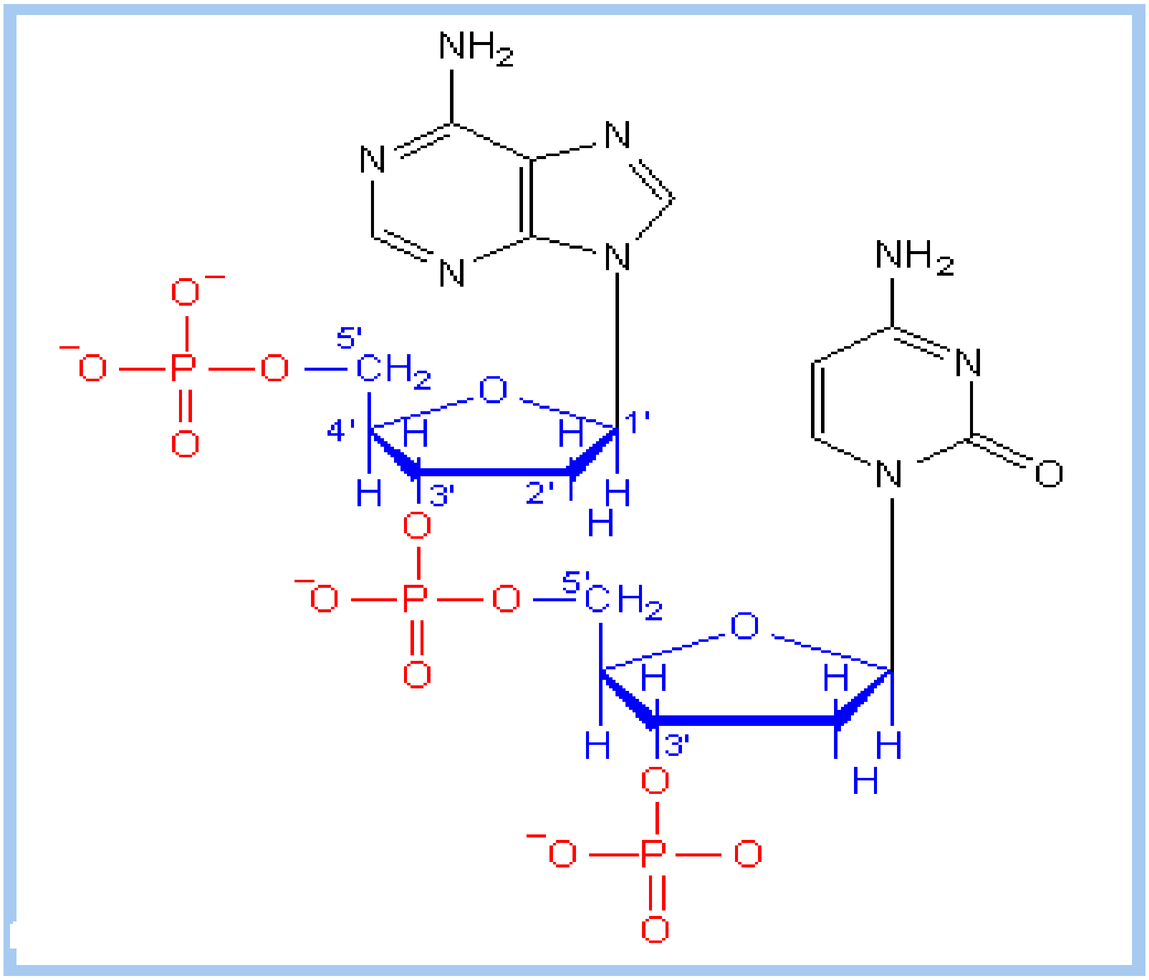
64 & 65. Alkylation Damage to DNA

DNA has electron-rich atoms that are readily attacked by electron-seeking chemicals called electrophiles or electrophilic agents. Alkylation is the transfer of an alkyl group from one molecule to another. The alkyl group may be transferred as an alkylcarbocation, a free radical, a carbanion or a carbene (or their equivalents).[1] An alkyl group is a piece of a molecule with the general formula CnH2n+1, where n is the integer depicting the number of carbons linked together. For example, a methyl group (CH3) is a fragment of a methane molecule (CH4); n = 1 in this case. Alkylating agents utilize selective alkylation by adding the desired aliphatic carbon chain to the previously chosen starting molecule. This is one of many known chemical syntheses. Alkylating agents are a highly reactive group of electrophiles which transfer methyl, ethyl, or alkyl groups to the electron-rich atoms in the DNA and damage it. Alkylation in DNA takes place at:-

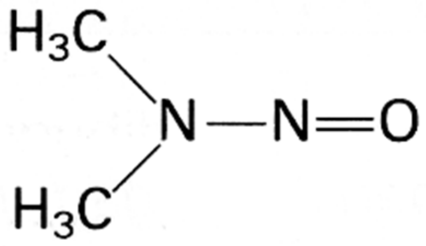
A) nitrogen and oxygen atoms external to the base ring systems;

B) non-bridging oxygen atoms in phosphate groups. C) nitrogen atoms in the base ring systems except those linked to deoxyribose.

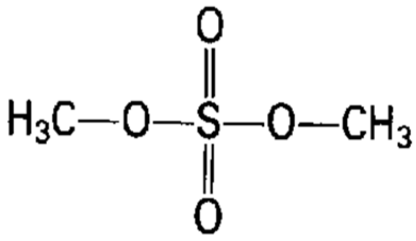
Many different kinds of naturally occurring and synthetic chemical agents are known to transfer alkylating agents to DNA.



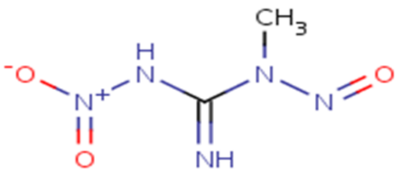
The product formed by attaching a chemical group to DNA is called an adduct. If the chemical group attaches to a single site on the DNA then the product is termed as monoadduct. The exposure of DNA to dimethylnitrosamine leads to the production of a monoadduct in which a single methyl group attaches to DNA.



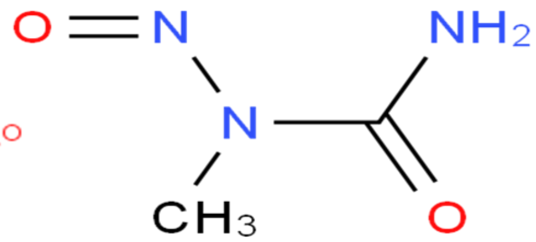
Dimethylnitrosamine



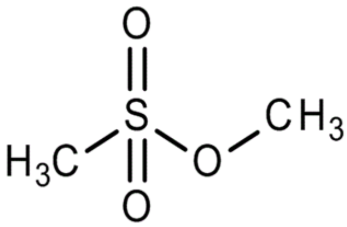
Dimethyl sulfate (DMS)



NTG



MNU

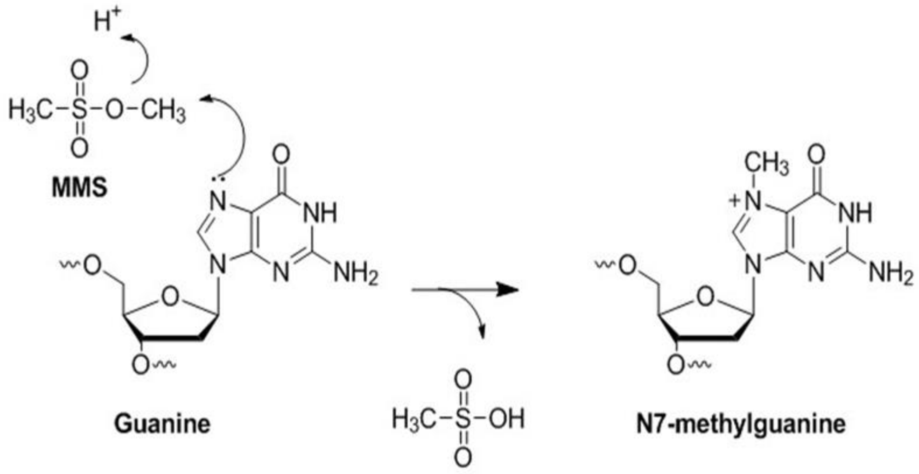


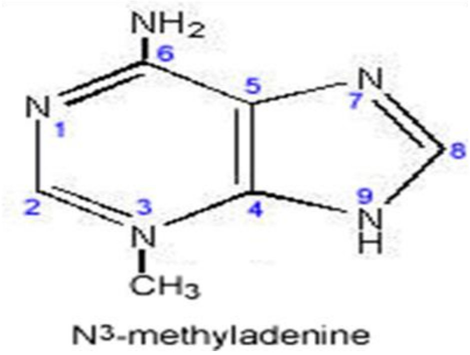
MMS

When DNA is exposed to methyl methane sulfonate (MMS) or N-methyl-N-nitrosourea (MNU), methylation takes place most frequently at:-

i) N-7 position in guanine and

ii) next most frequently at N-3 in adenine.



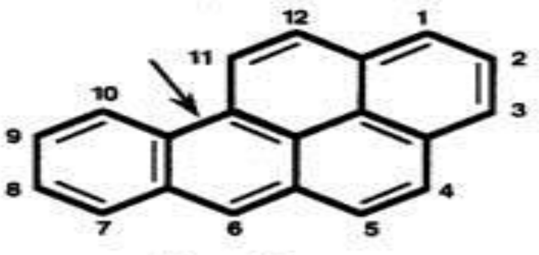


N-methylguanine forms a base pair with cytosine, but it is readily removed from the DNA with the resultant formation of an abasic site. Methylation at N-3 in adenine is of great practical significance because N-3 methyl adenine formation blocks DNA replication but does not appear to lead to mutations. Therefore, a methylating agent that can transfer methyl group exclusively to N-3 position in adenine would have the potential to kill cancer cells without causing mutations. Methylation at O-6 in guanine and O-4 in thymine are much less frequent events than either of the above described methylations. O^6-methylguanine and O^4-methylthymine formation are quite important because the methylated bases mispair during DNA replication, resulting in transition mutations. The phosphate groups in DNA backbone can also be methylated.
The resulting neutral phosphodiester is easily cleaved by water to produce single strand breaks.

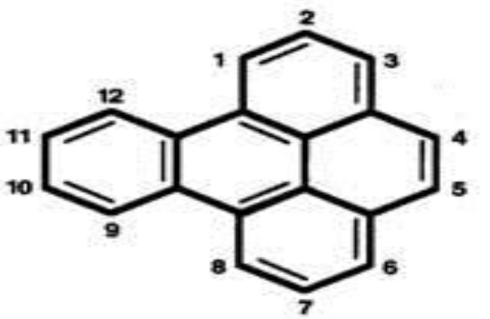
66. DNA Damage by PAHs

Many environmental agents become active alkylating agents only after they are metabolized in the cells. One such agent is a mixture of polycyclic aromatic hydrocarbons (PAHs) formed by the

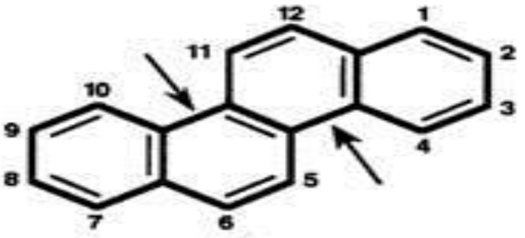
incomplete combustion of the burning wood or coal used as fuel. Similar type of polycyclic aromatic hydrocarbons are also present in tobacco smoke and charbroiled meats.
There are more than 100 different types of PAH compounds.



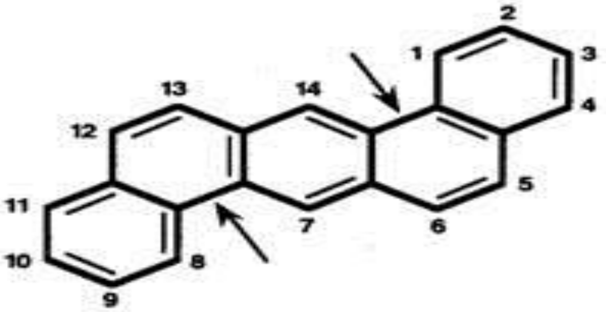
Benzo[a]pyrene



Benzo[e]pyrene

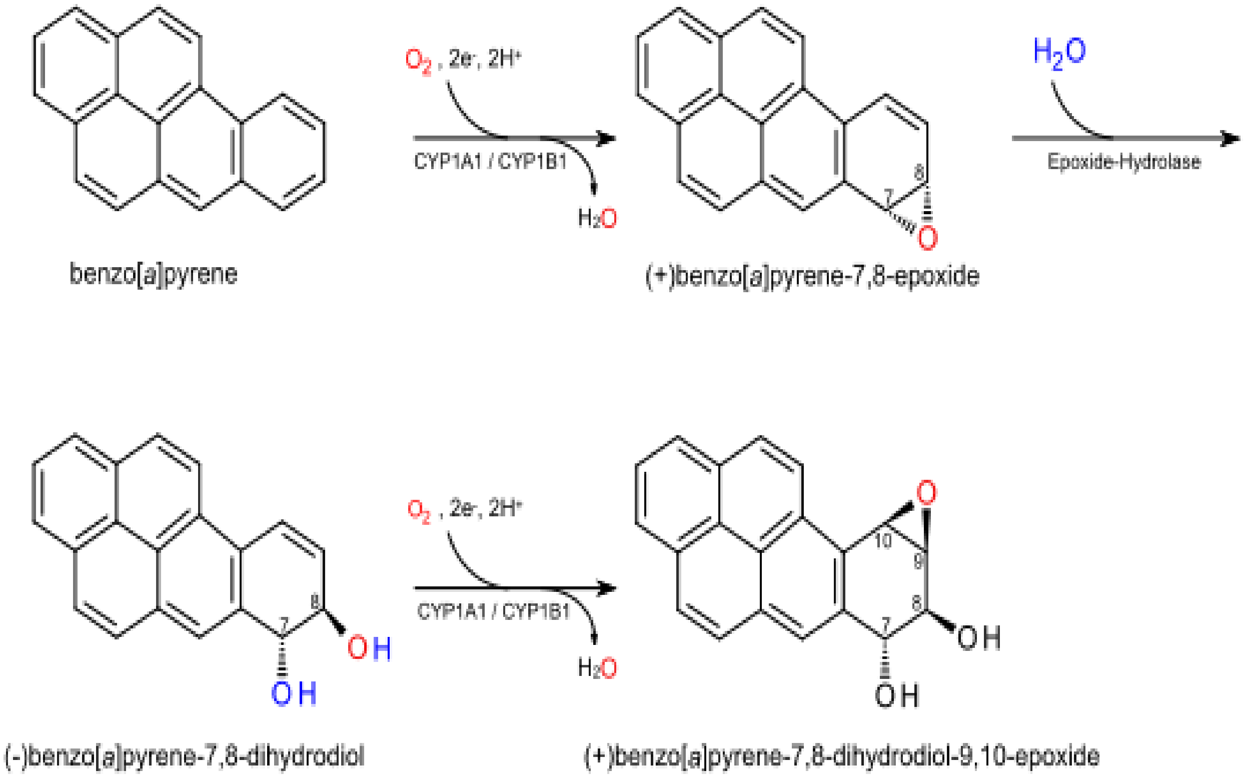


Chrysene

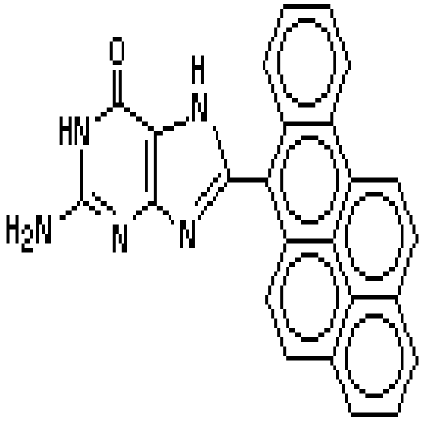


Dibenz[a,h]anthracene

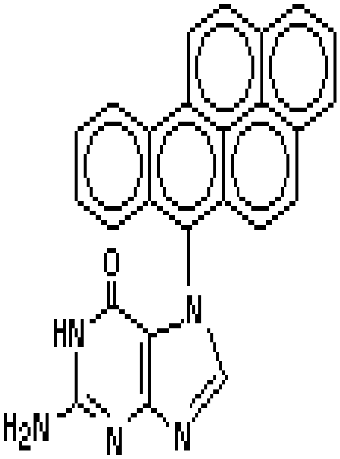
The common structural feature in all PAHs is two or more fused aromatic rings. These are not able to damage DNA unless they are metabolically activated in the cell. One of the polycyclic aromatic hydrocarbon Benzo[a]pyrene is converted into an active epoxide alkylating agent through a metabolic pathway.


O2 , 2e-, 2H+
CYP1A1 / CYP1B1
H2O
H2O
Epoxide-Hydrolase
O2 , 2e-, 2H+
CYP1A1 / CYP1B1
H2O
7
8
O
OH
OH
10
9
benzo[a]pyrene
(+)benzo[a]pyrene-7,8-epoxide
(-)benzo[a]pyrene-7,8-dihydrodiol
(+)benzo[a]pyrene-7,8-dihydrodiol-9,10-epoxide

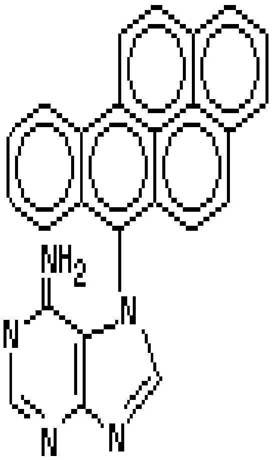
Cytochrome P450 enzymes being not highly specific can act on PAHs such as benzo[a]pyrene and add oxygen atoms to form reactive three- membered epoxide rings. These epoxides then alkylate DNA, causing replication errors that result in mutations, which ultimately convert a normal cell into a cancer cell.



Bap-8-C8GuA



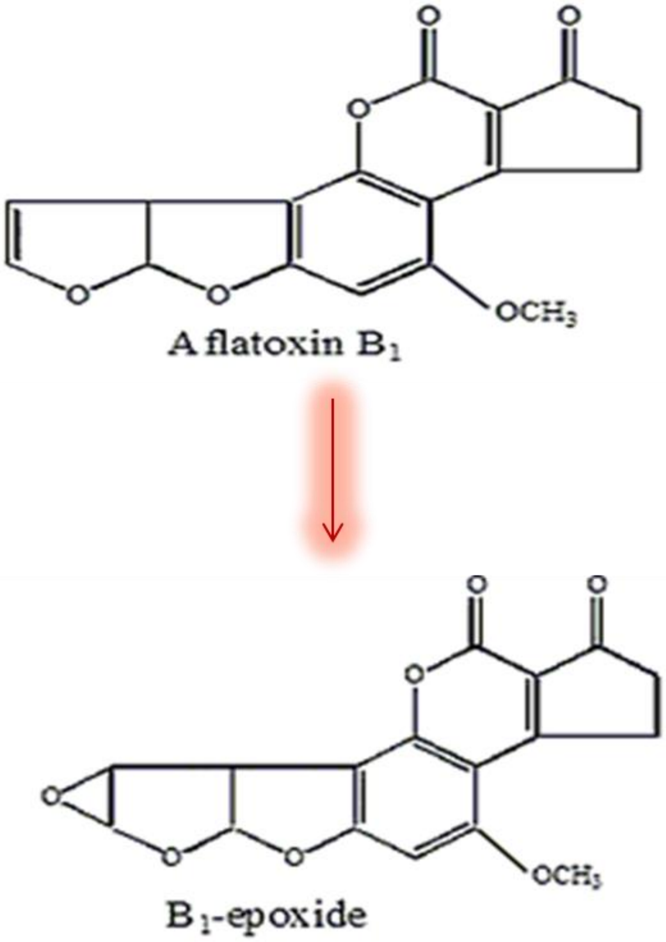
BuP-6-N7Gua



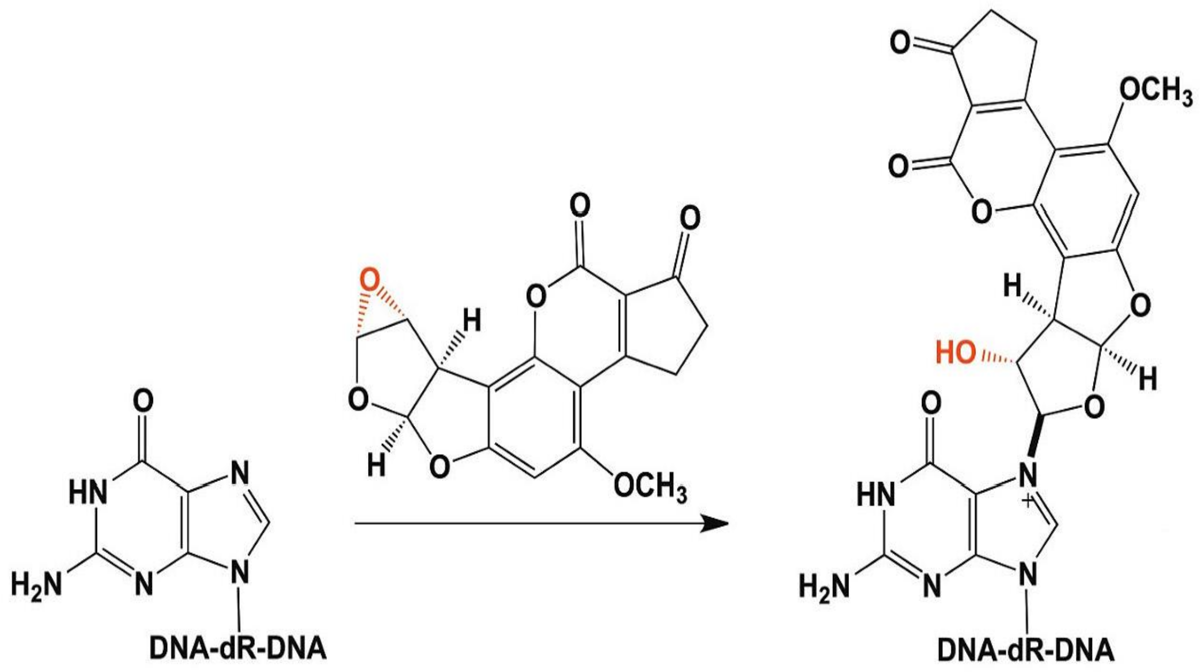
BuP-6-37Ade

67. DNA Damage by Aflatoxins

Another class of chemical carcinogens that must be activated before damaging DNA are called aflatoxins. They are produced from *Aspergillus flavus* and *A. parasiticus,* fungi that grow on peanut and other grains such as rice and corn. Animals feeding on contaminated peanuts or grains containing aflatoxins exhibit markedly increased rates of liver diseases including liver cancer. Aflatoxin B1, the most potent toxin produced by *A. flavus* presents a particularly serious health threat in the United States.
Cytochrome P450 converts it into an epoxide derivative that damages DNA.



Under ideal conditions, a small tripeptide glutathione will attack the epoxide ring, making the aflatoxin derivative soluble so that it can be excreted in the urine. If the reactive epoxide derivatives escape the attack of glutathione, they are free to attack guanine rings in DNA.



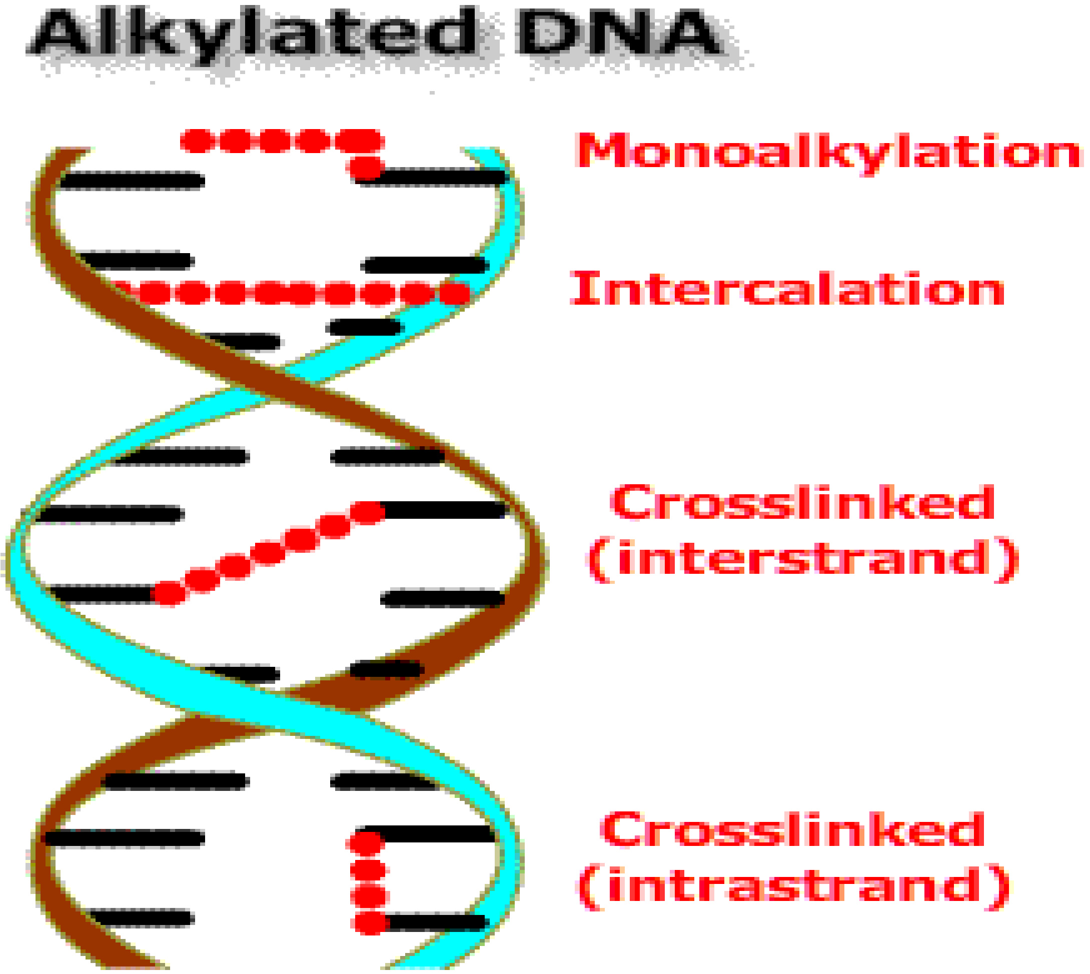
Guanine Base in DNA (dG)

AFB$_1$-N7-dG-Adduct

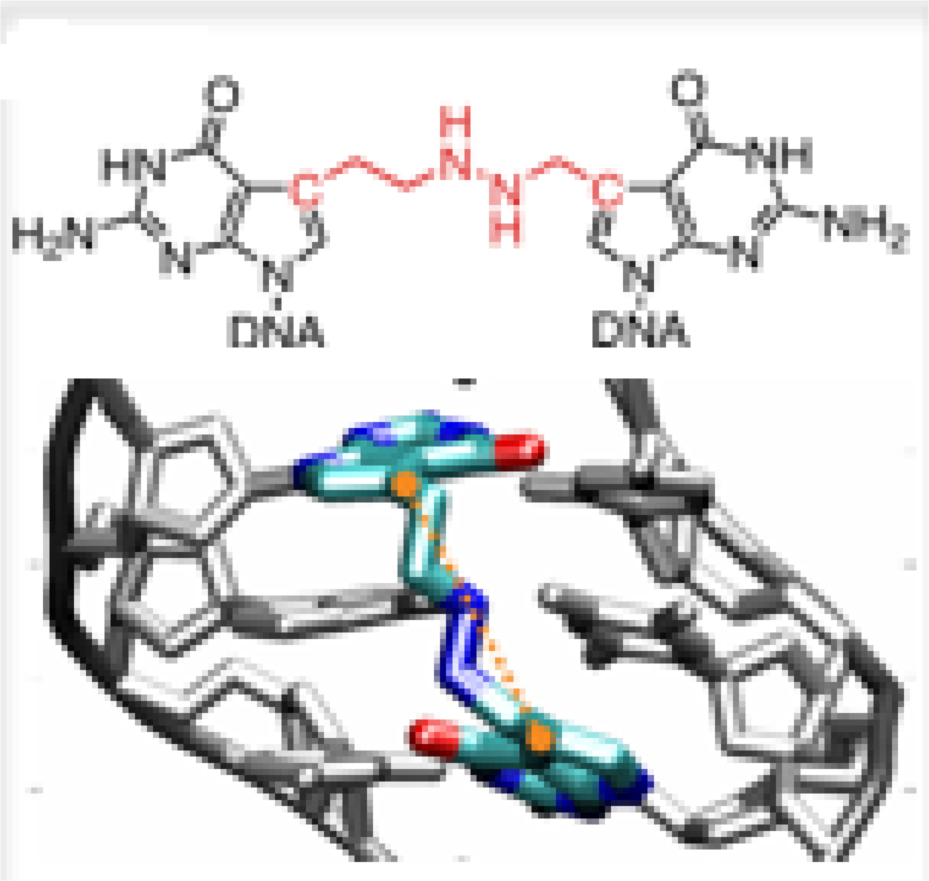
The flat aflatoxin ring system inserts between DNA bases causing helical distortion that in turn leads to replication errors. The flat aflatoxin ring system inserts between DNA bases causing helical distortion that in turn leads to replication errors.

68. Chemical Cross-Linking Agents

Many alkylating agents, having two reactive sites can form intra-strand or inter-strand cross-links in addition to forming monoadducts. Inter-strand cross-links are of special interest because they prevent strand separation an, if not corrected, are lethal.



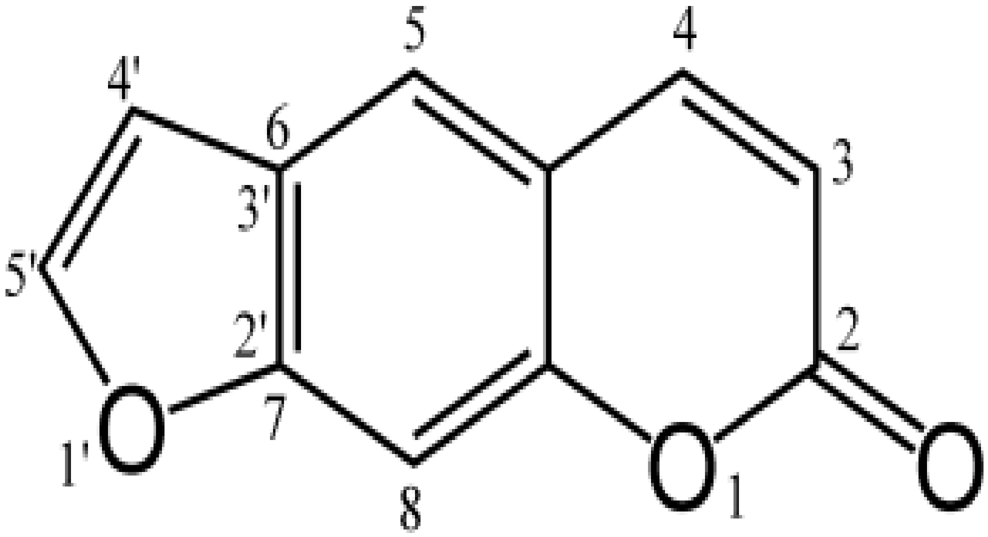
One of the simplest cross-linking agents, nitrogen mustard gas (bis[2-chloroethyl] methylamine damages DNA by forming inter-strand cross-links. It does so by attacking N-7 on two guanines, which are on opposite strands of DNA double helix.



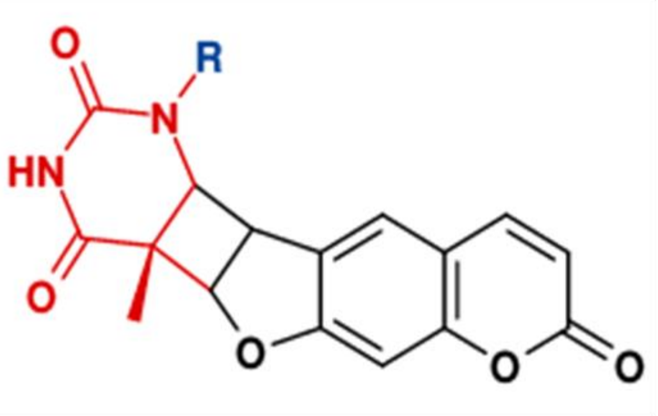
Although a very toxic substance, nitrogen mustard gas has found clinical application as a chemotherapeutic agent for treating certain forms of cancers.

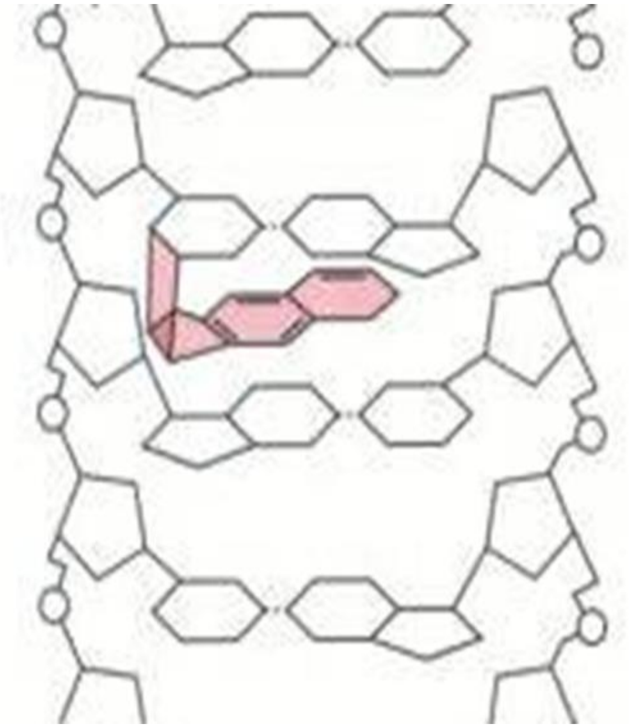
69. DNA damage by Psoralen

Psoralen can form mono-adducts or cross-links. It is a naturally occurring substance that can alkylate DNA if photoactivated. The planar psoralen molecule, which consist of a furan ring fused to a heterobicyclic ring system called coumarin, intercalates into the DNA molecule.

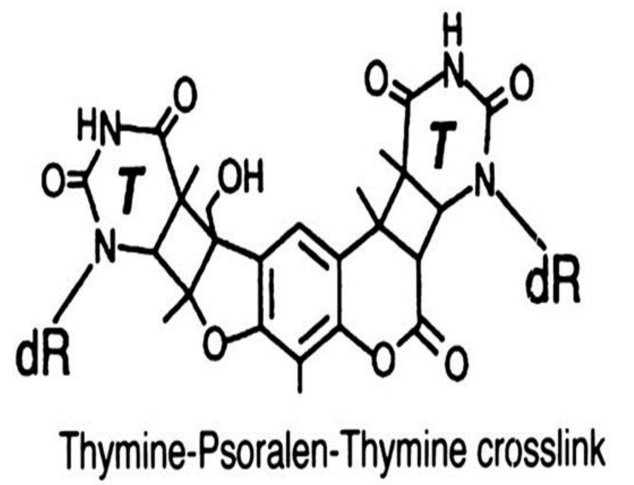


Upon exposure to light with a wavelength of 400 to 450nm, the furan ring in psoralen becomes activated and adds across the 5,6 double bond in a pyrimidine base to form the 4',5', monoadduct.

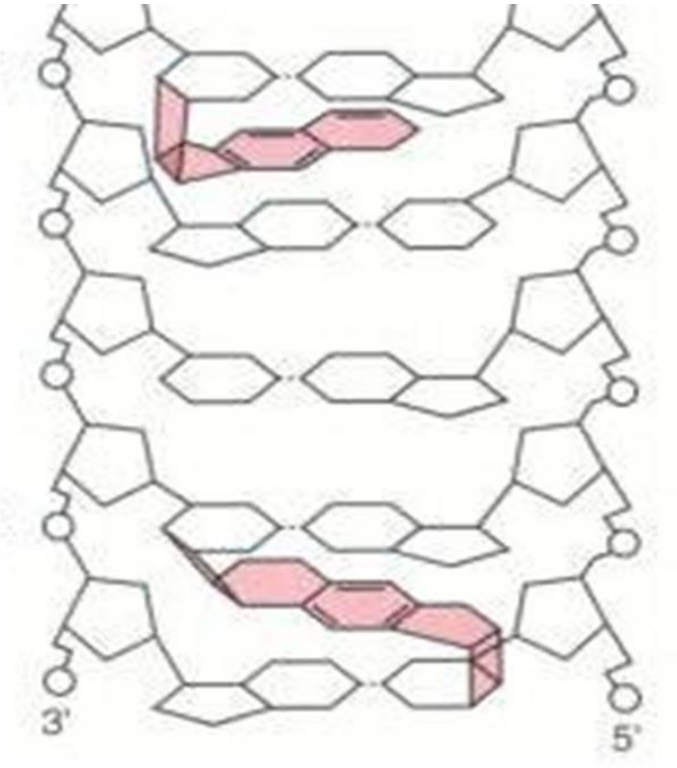




The planar tricyclic psoralen derivative in the monoadduct is in position to combine with a second pyrimidine base on the opposite DNA strand. But to do so, it must be activated by light with the wavelength of 320 to 400nm. The resulting photoproduct contain a cross-link between pyrimidine bases on opposite strands of the DNA duplex.



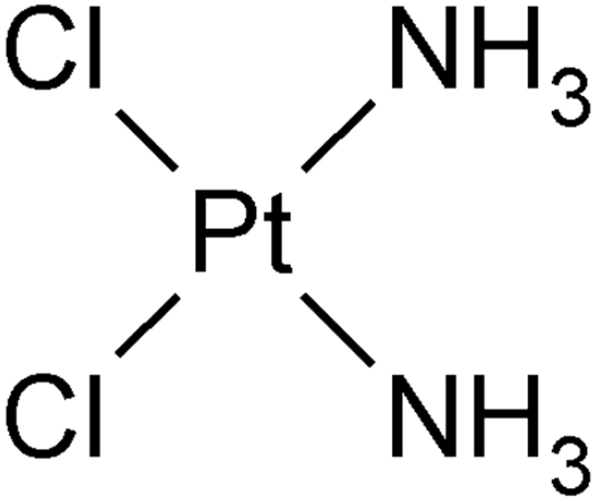
Thymine-Psoralen-Thymine crosslink



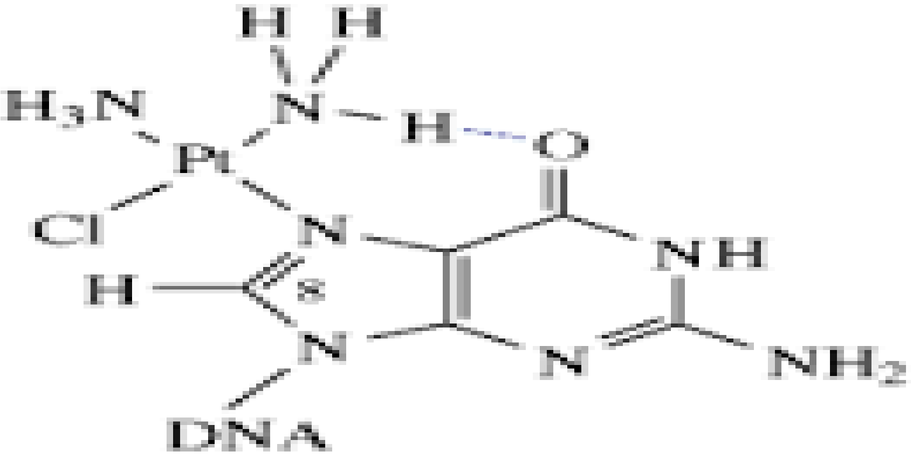
If not properly repaired, psoralen damage causes mutations and is lethal to cells.

70. DNA damage by Cisplatin

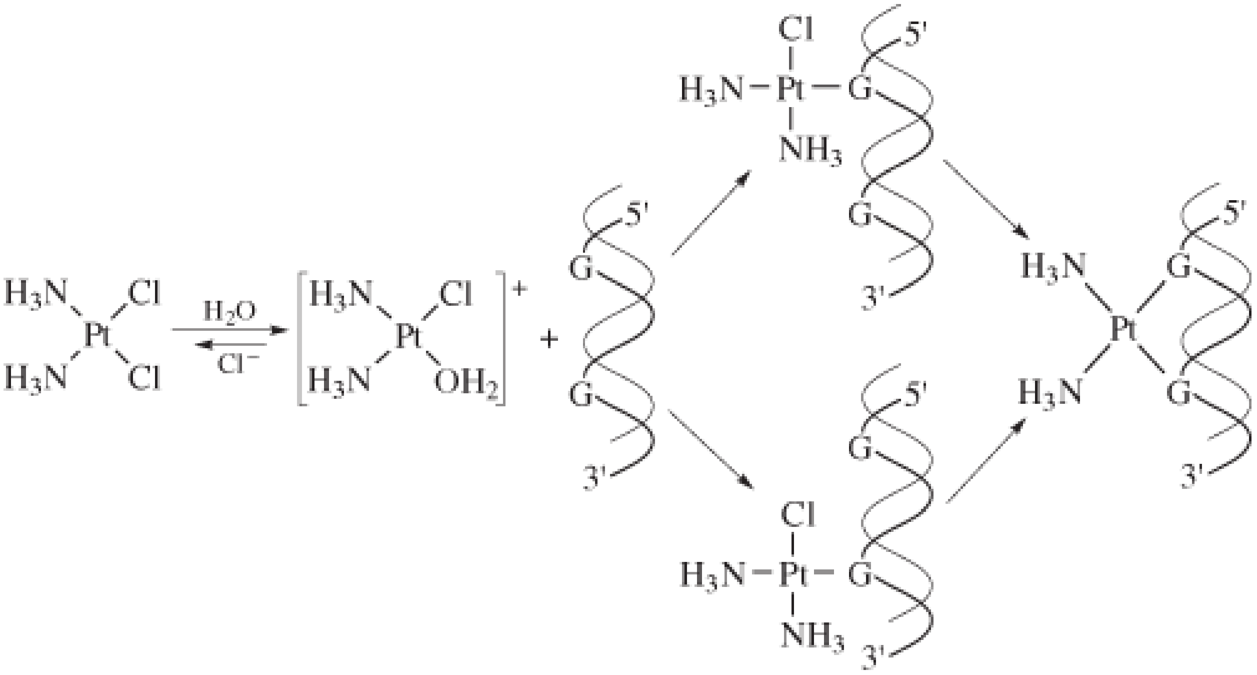
While examining the effects of electric currents on *E. coli*, a new compound viz; *cis*-diamminedichloroplatinum, better known as Cisplatin was found to block cell division in *E. coli*. Efforts were made to see if cisplatin would also inhibit cell division in other kinds of cells. It was revealed that cisplatin blocks the division of tumor cells.



After cisplatin enters the cell by passive diffusion or active transport, it undergoes hydrolysis to produce a highly reactive and charged complex (Pt(NH3)2ClH2O^{+}). This complex coordinates to the N-7 atom of either a guanine or adenine base in DNA.



Then the remaining chloride ligand is displaced by hydrolysis , allowing the platinum to coordinate to a second purine base on the same or opposite strand of the double stranded DNA.



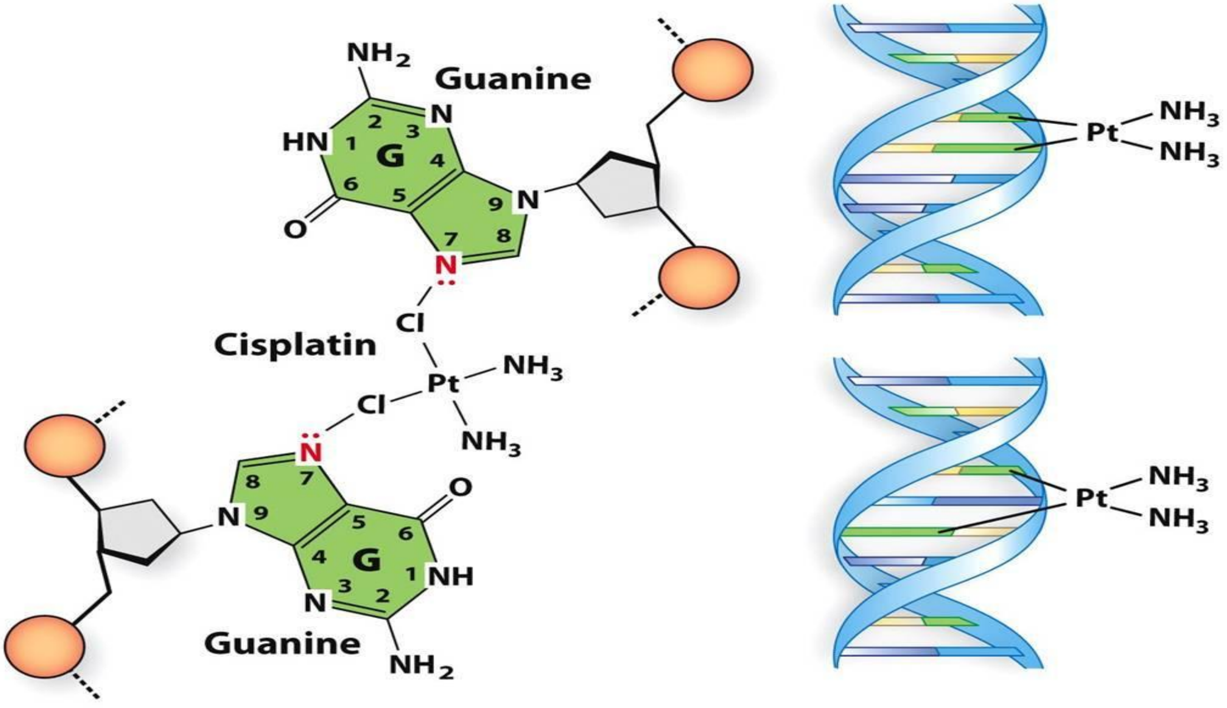
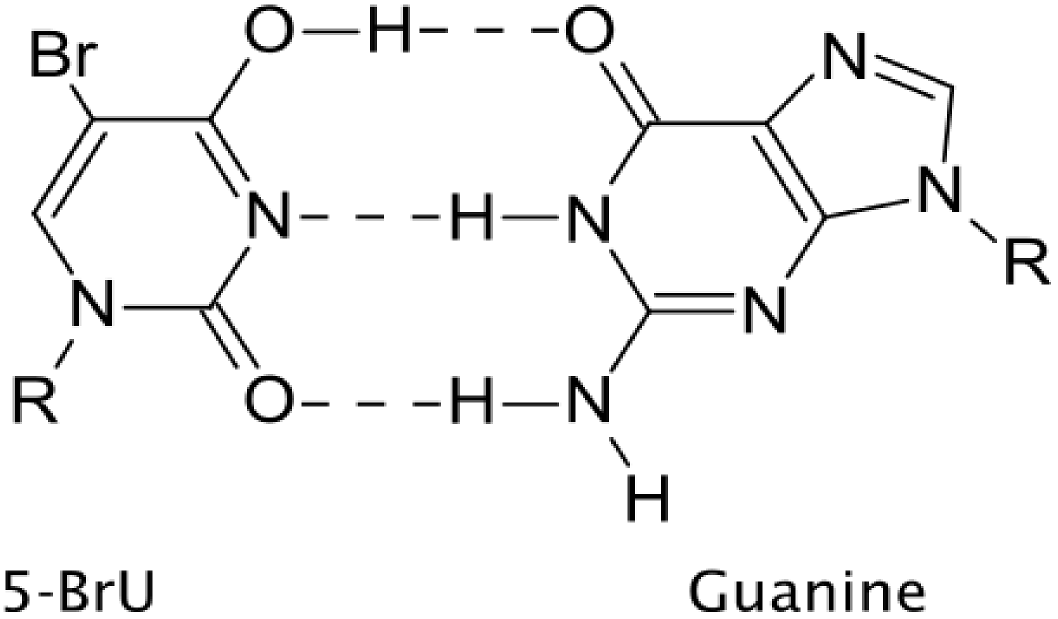


Figure 12-14a
Molecular Biology: Principles and Practice
© 2012 W. H. Freeman and Company

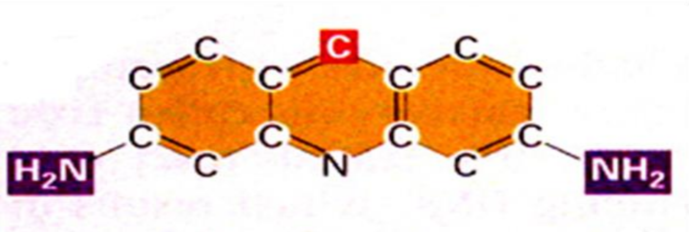
Cisplatin's cytotoxic effects appear to be due to inter-strand cross-linking which blocks the replication and transcription machinery. Cisplatin is a very effective chemotherapeutic agent for treating cancers of the bladder, ovaries and testicles, but has a number of side effects.

71. Base Analogs and Intercalating Agents

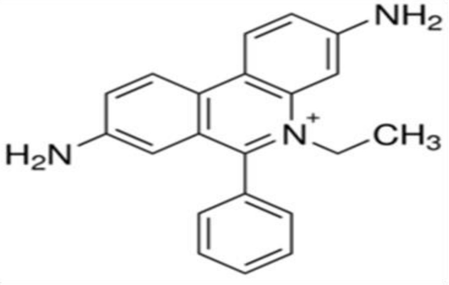
Mutations are also caused by compounds that substitute for normal bases called Base analogs or or the compounds that slip between the bases called Intercalating agents and cause errors in replication. Base analogs are structurally similar to proper bases but differ in ways that make them treacherous to the cell. Thus, base analogs are similar enough to the proper bases to get taken up by cells, converted into nucleoside triphosphates, and incorporated into DNA during replication. But, because of the structural differences from the proper bases, the analogs base-pair inaccurately, leading to frequent mistakes during the replication process. One of the most mutagenic base analogs is 5-bromouracil, an analog of thymine. The presence of the bromo substituent allows the base to mispair with guanine.



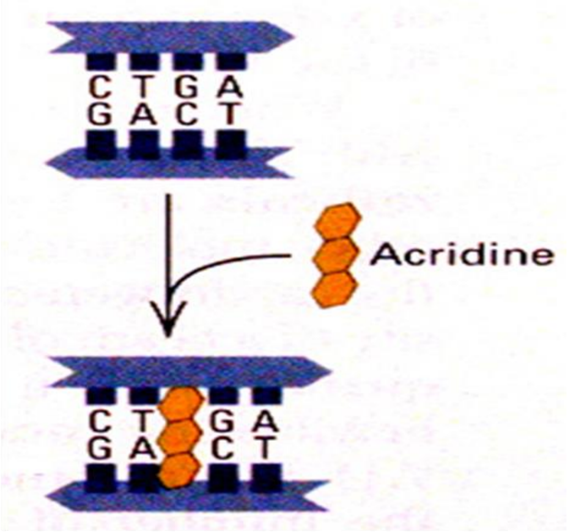
Intercalating agents are flat molecules containing several polycyclic rings that bind to the purine or pyrimidine bases of DNA, just as the bases bind or stack with each other in the double helix.



Proflavin



Ethidium



Intercalating agents, such as proflavin, acridine, and ethidium, cause the deletion or addition of a base pair or even a few base pairs. One possibility in the case of insertions is that, by slipping between the bases in the template strand, these mutagens cause the DNA polymerase to insert an extra nucleotide opposite the intercalated molecule. Conversely, in the case of deletions, the distortion to the template caused by the presence of an intercalated molecule might cause the polymerase to skip a nucleotide.

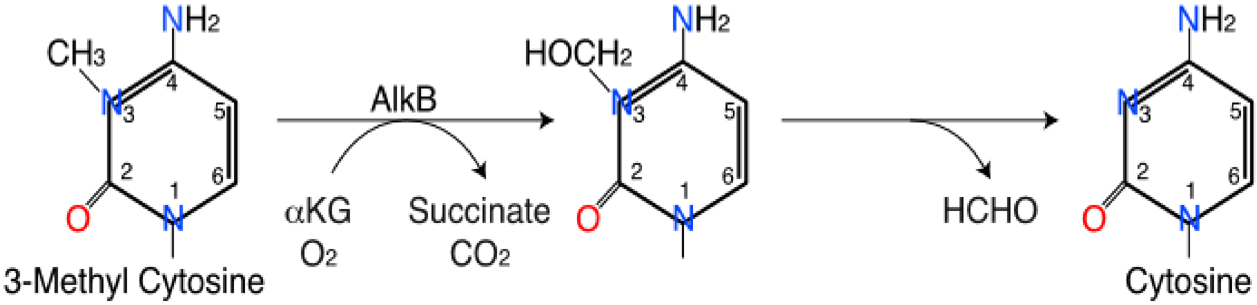
Chapter 9. DNA Repair

72. Direct Reversal of DNA Damage

The first clue to the existence of an enzyme that catalyzes the direct reversal of DNA damage was reported by Albert Kelner in 1949. In his experiments, Kelner first irradiated bacteria with UV light at doses that killed most of the bacteria. The he tested the survivors to isolate the desired mutants. Even though Kelner was very careful in his experimentation, he noticed a great deal of variation in the number of survivors from one experiment to the other. Finally, he found that cells placed in dark after UV treatment had a very low survival rate whereas those placed in light had a high survival rate. Exposure to light thus reversed the UV light's bactericidal effects. Similar phenomenon was also observed by Renato Dulbecco while studying UV-irradiated phage T2. Dulbecco prepared multiple plates each containing the same number of UV-irradiated phages and sensitive bacteria, and placed the plates in a stack under a florescent bulb in the lab. Each of the stacked plates should have about the same number of plaques.

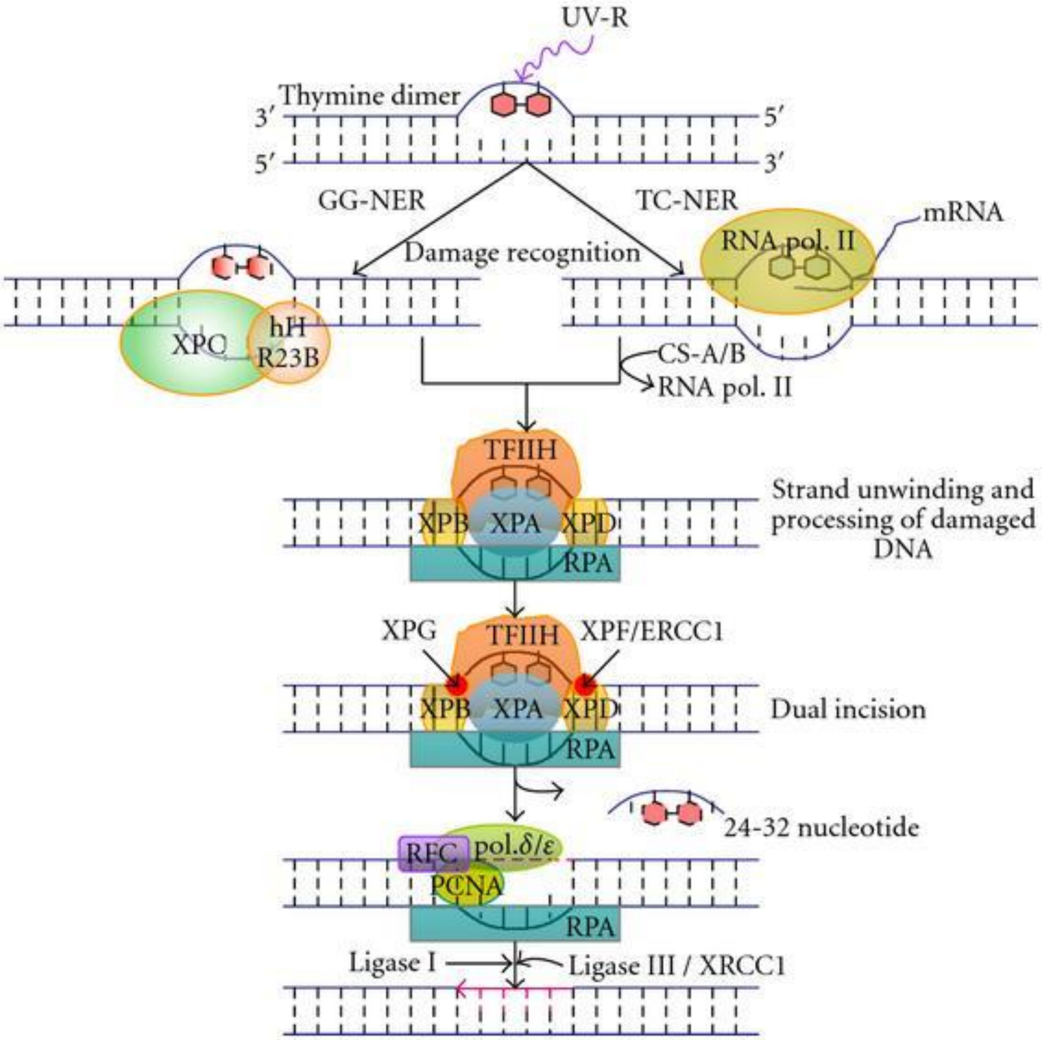
However, the plaque number decreased dramatically, going from top to bottom of the stack. Dulbecco explained this by proposing that the plates on the top of the stack were exposed to more light from the bulb as compared to the plates on the bottom of the stack. He tested this hypothesis by exposing some of the plates to more to fluorescent light while keeping others in dark. As expected, the number of plaques on plates exposed to light was much higher than those left in the dark. The bacteria were somehow using the visible light to repair the UV damaged DNA in the phage. The chemical basis for this light-dependant phenomenon, which Dulbecco called photoreactivation, remained to be elucidated.

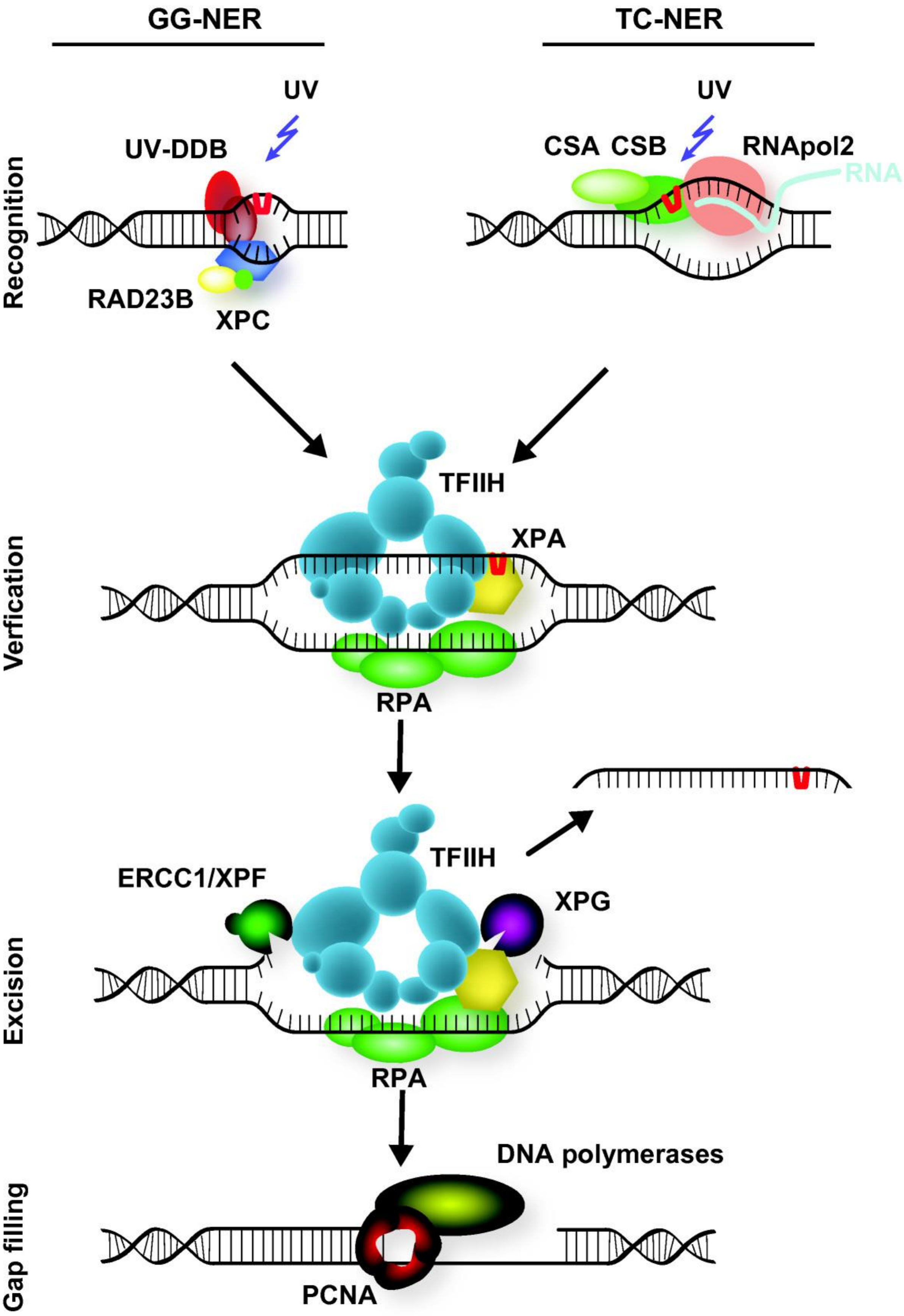
Direct DNA damage reversal is also provided by AlkB, a protein that is found in most, perhaps all, living organisms (including humans). This protein is a member of the class of enzymes called alpha-keto-glutarate-dependent and iron-dependent oxygenases (aKG-Fe(II)-oxygenases), which use iron-oxo intermediates to oxidize chemically inert compounds. In the process, alpha-keto-glutarate is converted to succinate and CO_2, which is similar to one of the steps in the tricarboxylic acid cycle. AlkB is capable of reversing methylation at the 1 position of adenine and the structurally similar 3 position of cytosine, as shown in this diagram:



73. Photoreactivation

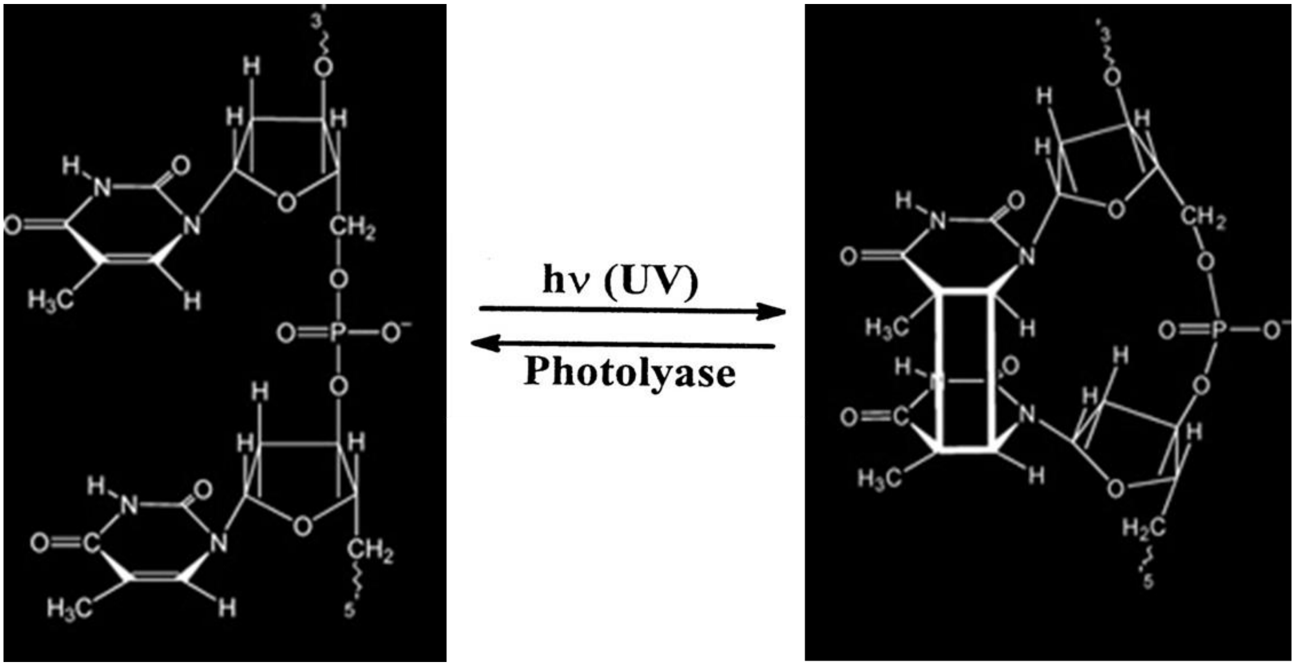
Claud S. Rupert and co-workers devised an in vitro photoreactivation system in 1957, taking a major step toward determining the chemical mechanism of photoreactivation. They used a straightforward approach in which they isolated DNA from the gram-negative bacteria *Haemophilus influenzae*. They irradiated this DNA with UV light to inactivate its transforming ability. Then they demonstrated that a cell-free *E. coli* extract restored the transforming activity in the presence of light. Although, there study had the potential to open the way for the purification and characterization of photoreactivation enzyme, investigators still needed to establish the chemical nature of this repair.


UV-R
Thymine dimer
3′
5′
5′
3′
GG-NER
TC-NER
Damage recognition
mRNA
RNA pol. II
XPC
hH R23B
CS-A/B
RNA pol. II
TFIIH
XPB
XPA
XPD
RPA
Strand unwinding and processing of damaged DNA
XPG
TFIIH
XPF/ERCC1
XPB
XPA
XPD
RPA
Dual incision
24-32 nucleotide
RFC
pol.δ/ε
PCNA
RPA
Ligase I
Ligase III / XRCC1


GG-NER
TC-NER
UV
UV
UV-DDB
CSA CSB
RNApol2
RNA
RAD23B
XPC
Recognition
TFIIH
XPA
Verfication
RPA
TFIIH
ERCC1/XPF
XPG
Excision
RPA
DNA polymerases
Gap filling
PCNA

74. CPD Photolyase

The problem of the chemical nature of photoreactivation was resolved over next few years when investigators came to know that UV irradiation induces the formation of cyclobutan pyrimidine dimer in DNA. Further studies showed that the photoreactivation enzyme reverses the UV induced damage by using the energy provided by blue light (350-450nm) to drive cyclobutane ring disruption. When it was recognized that the photoreactivation enzyme catalyzes the disruption of carbon – carbon bonds, it was given a more descriptive name of CPD photolyase.

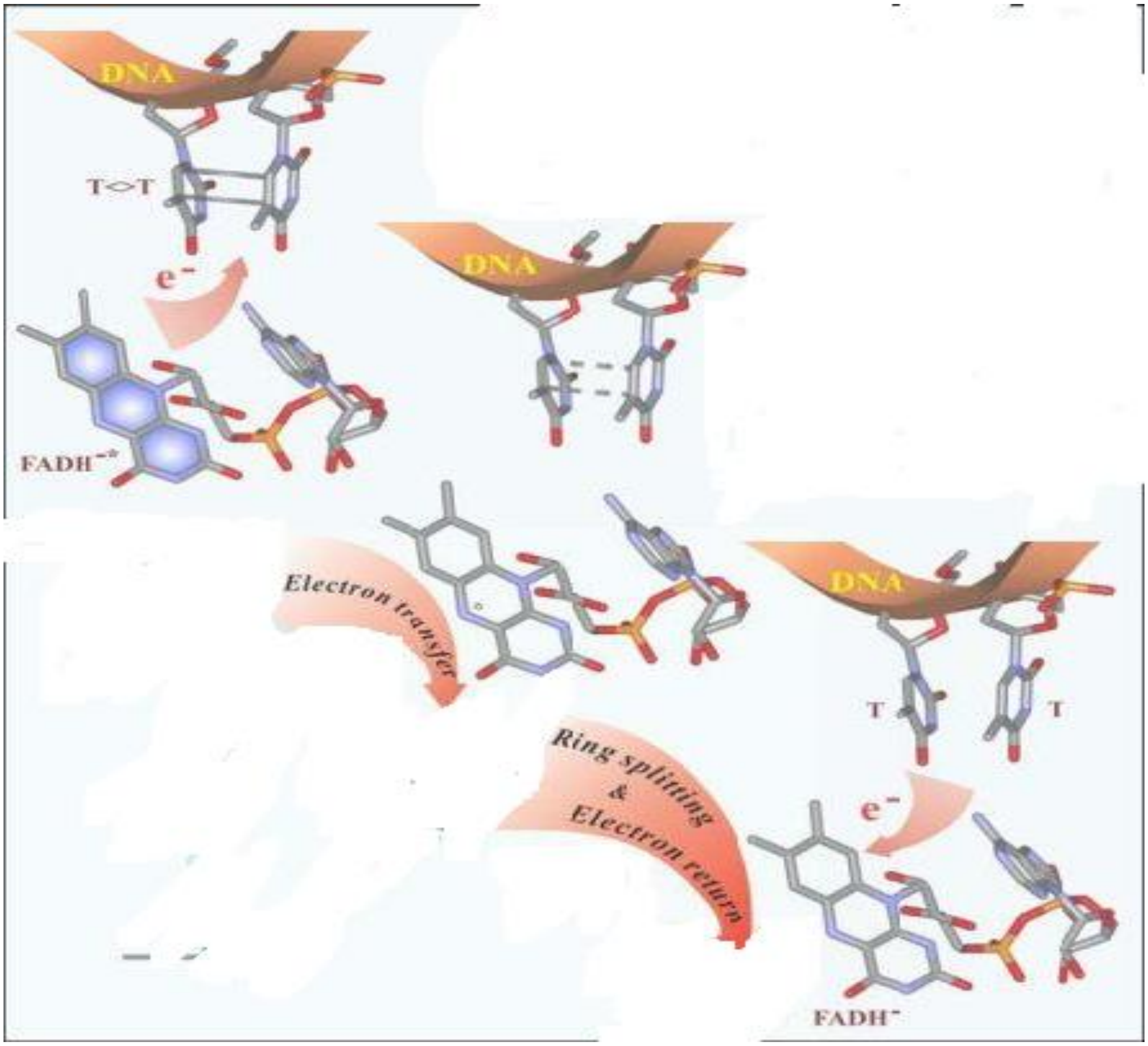


The bacteria that lack CPD photolyase can't repaire cyclobutane pyrimidine lesions through photoreaction thus can't reverse the UV induced damage in DNA. CPD photolyases are present in a wide variety of organisms including bacteria, archaea, plants, and animals but not in humans and other placental mammals. These are monomeric proteins ranging in size from about 450-550 amino acid residues.

All CPD photolyases have two domains, designated as N – and C – terminal domains. A light absorbing pigment, or chromophore pigment binds to each domain through non-covalent bonding. This chromophore factor acts as a photoantenna to capture light with wavelengths that would not otherwise be available.

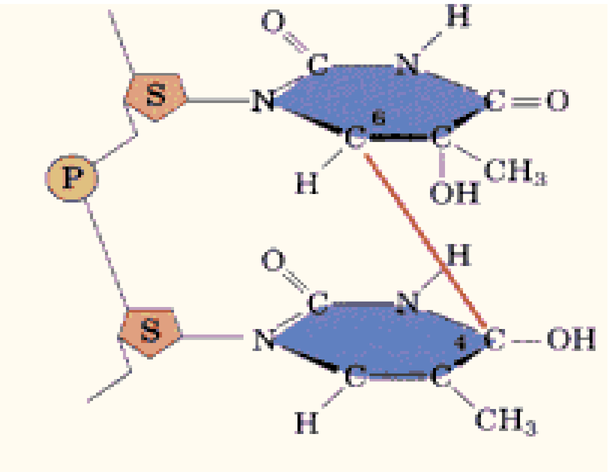
75. Mechanism of CPD Photolyase

CPD photolyase can bind to DNA in the dark by recognizing the altered DNA structure caused by a CPD formation rather than a specific nucleotide sequence. This binding is about 10^5 tighter when a DNA segment contains a CPD than when it does not. Half of the binding energy appears to come from interactions between enzyme and the DNA back bone. The other half of energy comes from interactions between the FADH at the active site and the CPD. However, the light harvesting antenna pigment does not influence binding. Once the enzyme – DNA complex is formed, the CPD is flipped out of the DNA double helix and into the enzyme's active site. After the CPD flips into the enzyme's active site, the energy of an absorbed photon is transferred from the light harvesting antenna pigment to the FADH . FADH then transfers an electron to the CPD to induce cyclobutane ring cleavage. The catalytic cycle is completed when the electron is transferred back from the repaired thymine to the FADH cofactor.

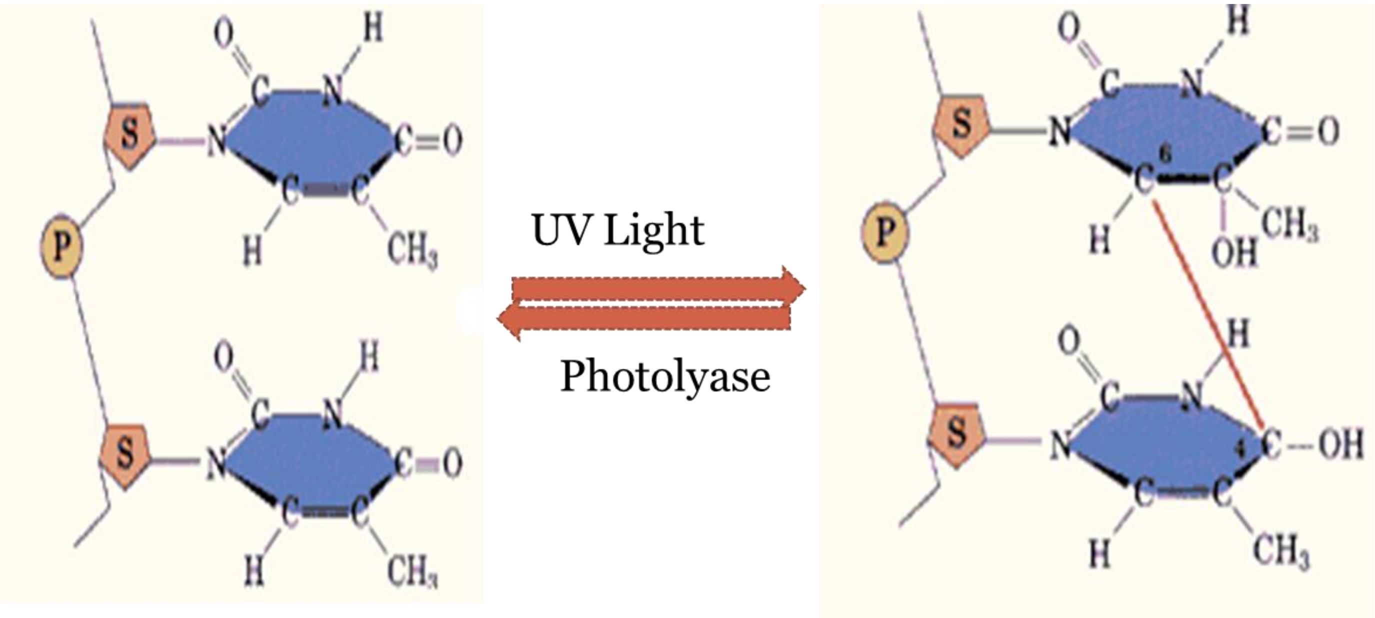


76. (6-4) Photolyase

UV irradiation also induces the formation of another type of pyrimidine dimer, the (6-4) photoproduct.



Takeshi Todo, Taisei Nomura and coworkrs reported in 1993 that *Drosophila melanogaster* has a photolyase that reverses (6-4) photoproduct lesions in the DNA. This photolyase was designated as (6 – 4) photolyase. It is widely distributed in plants and animals, but has not been detected in bacteria and mammals.



Considerably less information is known about the mechanism of action of (6-4) photolyase as compared to CPD photolyase; however, the two enzymes seem to work in a similar way. The (6-4) photolyase binds to damaged DNA, causing the (6-4) photoproduct to flip out of the DNA and into the active site of the enzyme. In the active site of the enzyme, it undergoes a rearrangement to produce a product that receives an electron from an excited FADH molecule. The final outcome is that the organisms containing (6-4) photolyase can use light energy to convert (6-4) photoproducts back to normal pyrimidine rings. Organisms can also repair dimer lesions introduced by UV light by excising damaged nucleotides and replacing them with normal nucleotides. This type of excision repair is the major pathay for repairing UV-induced damage to DNA in organisms such as humans that lack both the types of photolyases.

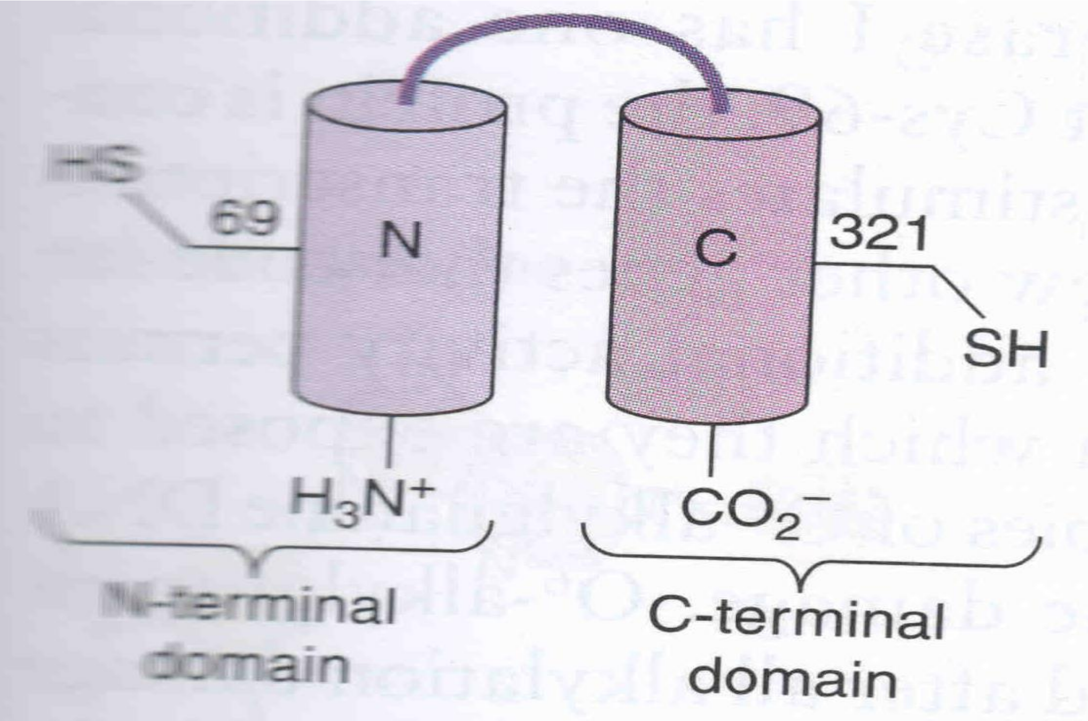
77. Damage Reversal by Dealkylation

Another means of direct reversal of DNA damage is by dealkylation. Direct dealkylation reactions have probably been most extensively studied in *E. coli*. Three different proteins can catalyze the direct removal of alkyl groups attached to oxygen atoms in DNA.
These include:-

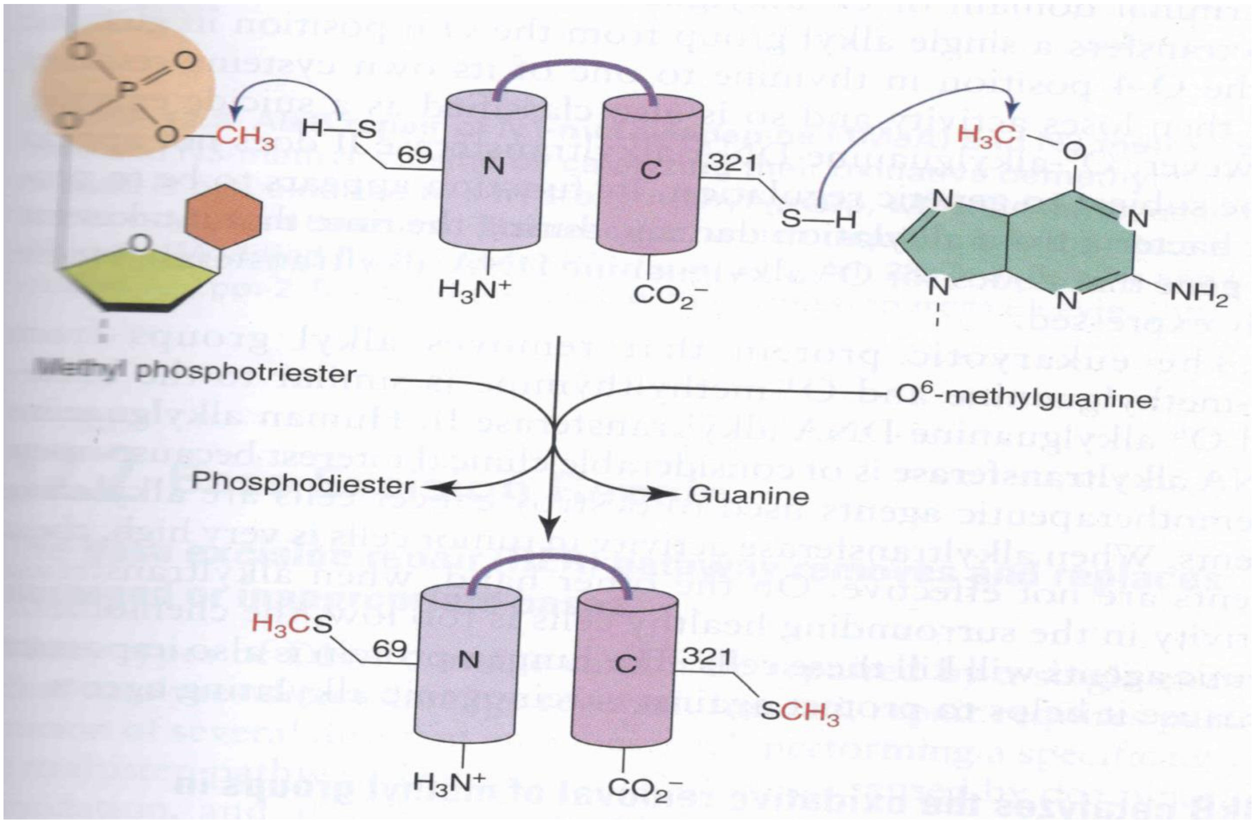
1) O^6-alkylguanine DNA alkyltransferase I
2) O^6-alkylguanine DNA alkyltransferase II
3) Alkylguanine DNA alkyltransferase

O^6-alkylguanine DNA alkyltransferase I can remove methyl & other alkyl groups attached to O-6 in guanine. It can also remove alkyl groups attached to O-4 of thymine and to phosphotriesters.

This enzyme is a monomer that has a flexible linker connecting its N- and C- terminal domains. Each domain has an active site that performs a specific function.



The N- terminal domain transfers an alkyl group from an Sp phosphotriester to one of its own cysteine residues, Cys-69.



The C- terminal domain transfers an alkyl group from either O^6-alkylguanine or O^4-alkylthymine to one of its own cysteine residues, Cys-321. Once alkylated, the protein can't be regenerated and therefore behaves more like an alkyl transferring agent than an enzyme. Such proteins as O^6-alkylguanine DNA alkyltransferase I, which lose their activity after acting one time, are called suicide enzymes. O^6-alkylguanine DNA alkyltransferase I, has one additional remarkable function. After methylation at Cys-69, this protein is converted to a transcriptional activator. This activator stimulates the transcription of the gene that codes for it as well as some other genes that code for the proteins that repair DNA damage. This additional activity permits the bacteria to adapt to environments in which they are exposed to alkylating agents by synthesizing more copies of these enzymes which can repair damaged DNA. O^6-alkylguanine DNA alkyltransferase I that is synthesized after all the alkylation damage has been repaired will remain unmethylated. This unmethylated form of the enzyme blocks transcription of the same genes that were activated by the methylated proteins.

78. Dealkylation Enzymes

O^6-alkylguanine DNA alkyltransferase II

O^6-alkylguanine DNA alkyltransferase II is another E. coli's alkyl transferring enzyme. It has properties very similar to that of C-terminal domain of O^6-alkylguanine DNA alkyltransferase I. It also transfers a single alkyl group from the O – 6 position in guanine or the O-4 position in thymine to one of its own cysteine residues. After this transfer, the enzyme loses its activity and so is also classified as a suicide enzyme. However, this enzyme doesn't appear to be subjected to genetic regulation. Its function appears to be to protect bacteria from alkylation damage during the time that it takes for the gene of O^6-alkylguanine DNA alkyltransferase I to be fully expresed.

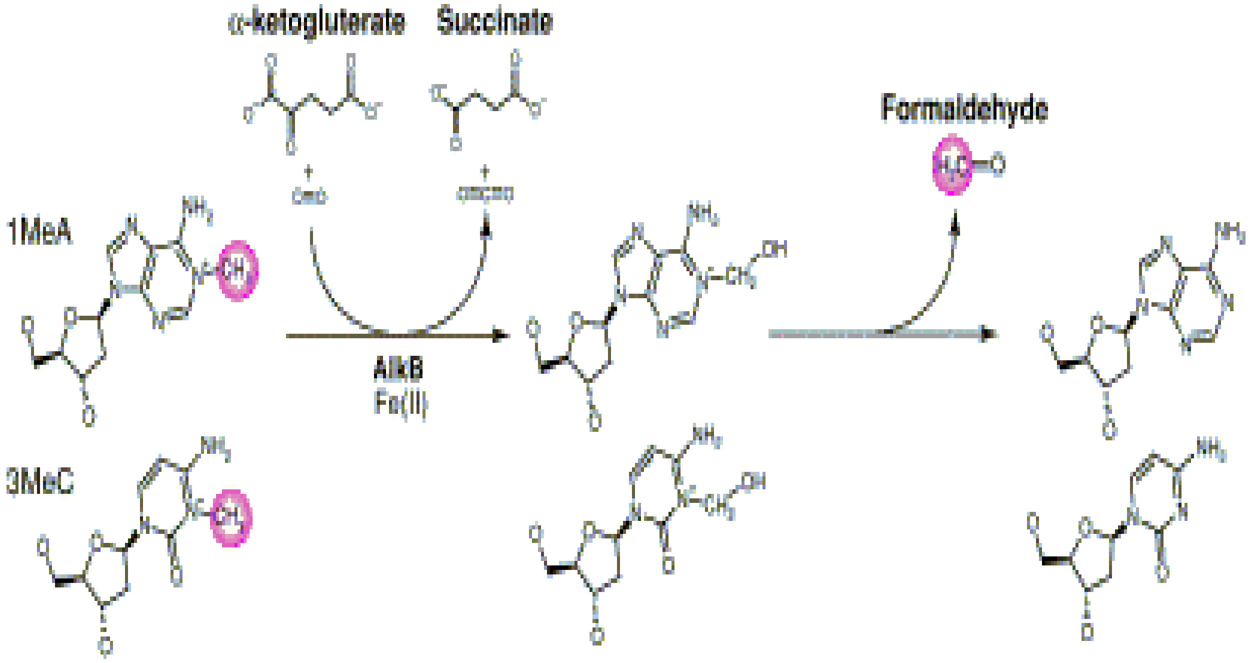
Alkylguanine DNA alkyltransferase

The eukaryotic protein that removes alkyl group from O^6-methylguanine or O^4- methylthymine is similar to the bacterial O^6-alkylguanine DNA alkyltransferase II. Human alkylguanine DNA alkyltransferase is of considerable clinical interest because many chemotherapeutic agents used to destroy cancer cells are alkylating agents. When alkyltransferase activity is very high in the tumor cells, these agents are not effective. On the other hand, when alkyltransferase activity in the surrounding healthy cells is too low, the chemotherapeutic agents will kill these cells. The human protein/enzyme is also important because it helps to protect the cells against carcinogenic alkylating agents.

AlkB

An entirely different type of alkylation damage repair activity was found in *E. Coli* in 2002 which is carried out by a protein AlkB. AlkB catalyzes the direct conversion of 1-methyladenine, 1-methylguanine, 3-methylcytosine and 3-methylthymine to adenine, guanine, cytosine and thymine, respectively. The AlkB catalyzed reaction requires, $Fe2^+$, molecular oxygen, and α -ketoglutarate.

A similar type of enzyme has been found in other organisms including human & mammals.

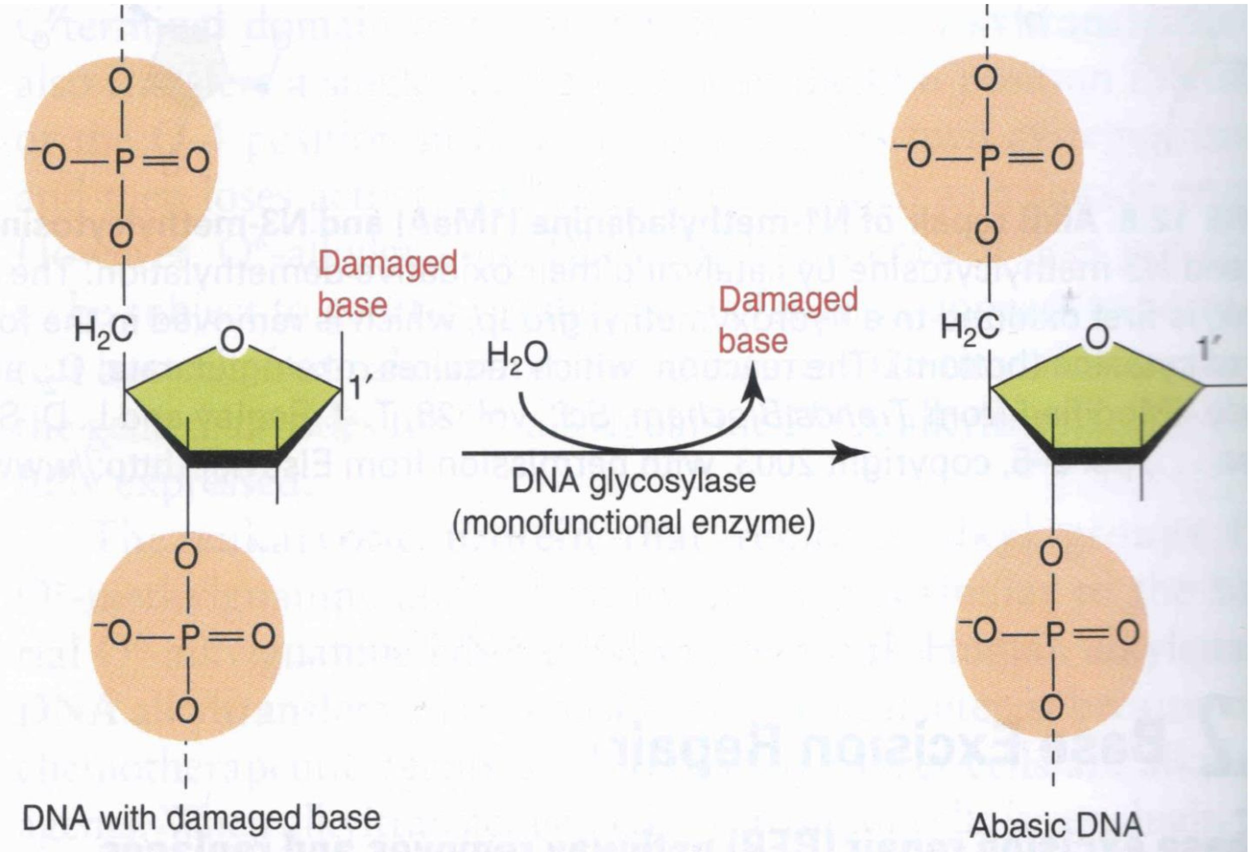


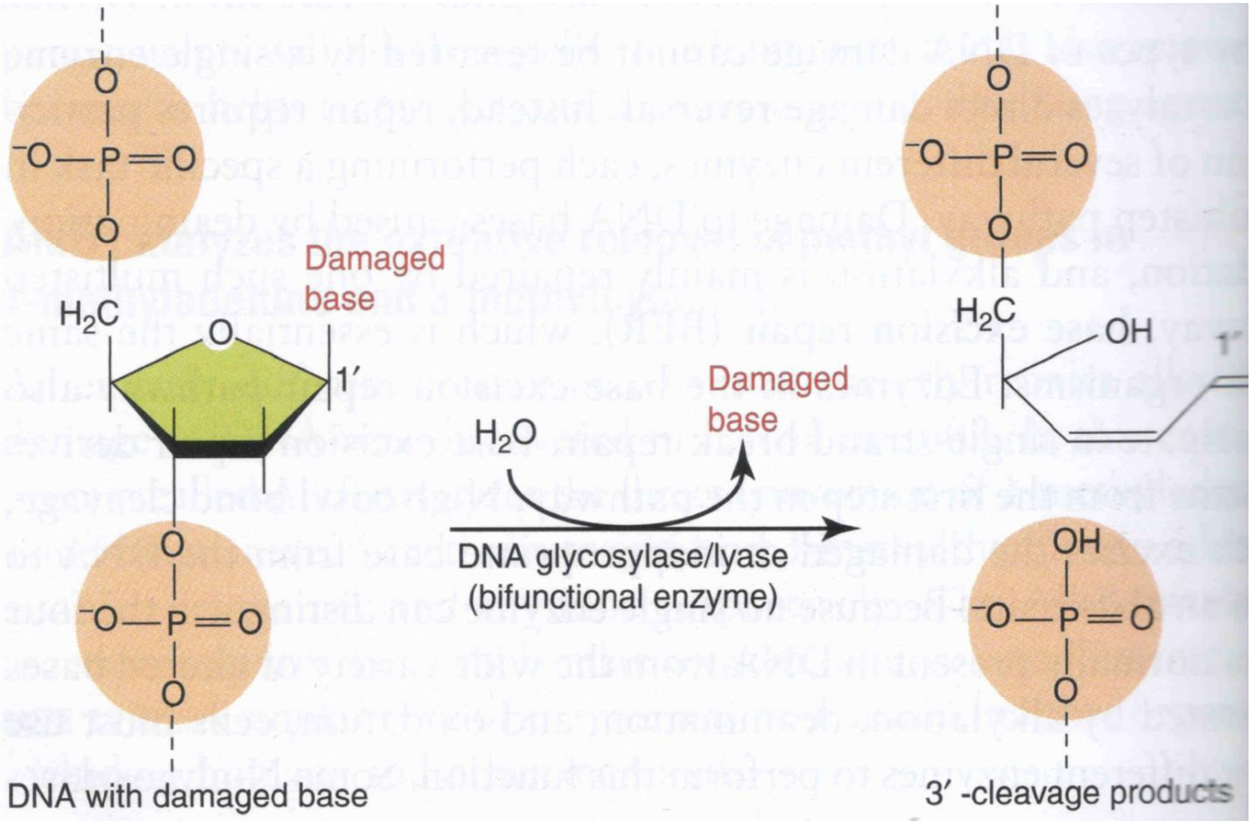
79. Base Excision Repair

Many types of DNA damages can’t be repaired by a single enzyme that catalyzes direct damage reversal. Instead, repair requires the participation of several different enzymes, each performing a specific task in multistep pathway. Damage to DNA bases caused by deamination, oxidation, and alkylation is mainly repaired by one such multistep pathway which is called, Base Excision Repair (BER). BER pathway is same in all organisms. Enzymes involved in the base excision repair pathway also participate in the repair of single-strand break in DNA. Base excision repair derives its name from the first step in the pathway, N-glycosyl bond cleavage.

This cleavage excises the damaged or inappropriate base from the DNA to form an abasic site. Because no single enzyme can distinguish the four bases present in DNA from a wide variety of altered bases, cells must use many different enzymes to perform this function. Some N-glycosylases are monofunctional enzymes with only the ability to excise a damaged base.

Such enzymes are called DNA glycosylases. Other N-glycosylases have an AP lyase activity that cleaves the bond between sugar and the phosphate 3’ to the damaged site. This enzymes is designated as DNA glycosylase/lyase.

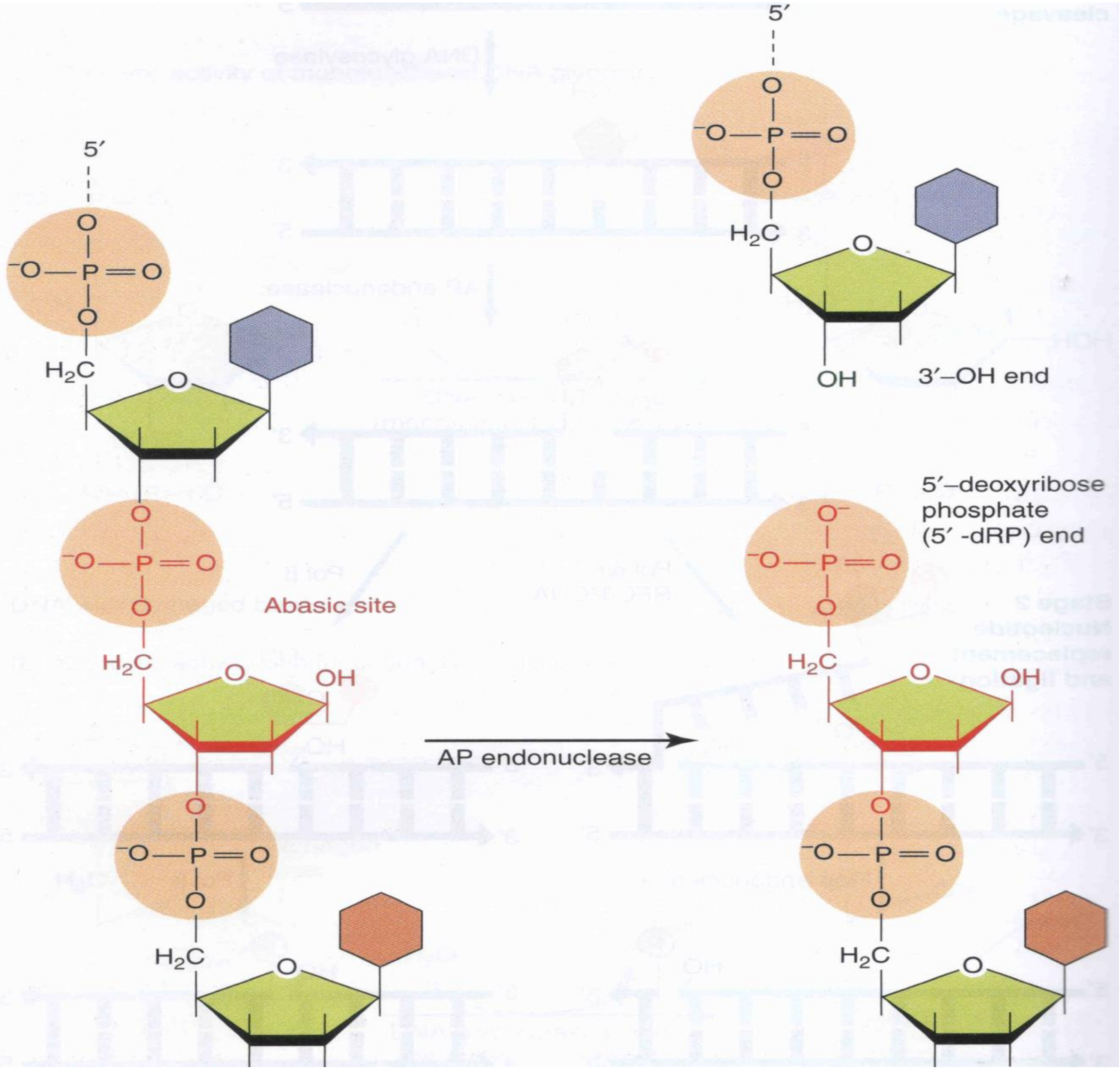

O
⁻O—P=O
O
Damaged base
H2C
O
1′
H2O
Damaged base
DNA glycosylase
(monofunctional enzyme)
O
⁻O—P=O
O
DNA with damaged base
O
⁻O—P=O
O
H2C
O
1′
O
⁻O—P=O
O
Abasic DNA



Both types of enzymes detect the damaged base, flip it out of the DNA helix into an active site pocket, and then cleave the N-glycosyl bond. Although some glycosylases excise a specific base, most have a somewhat broader specificity. The *E.coli* enzyme Uracil N-glycosylase (Ung) is specific for uracil. Whereas another *E. coli* N-glycosylase, 3-methyladenine DNA glycosylase (AlkA), acts on 3- or 7-methyl purines, 3- or 7-ethylpurine, ethenoadenine, and O^{2} -methyl pyrimidine. A null mutation in a gene that codes for a DNA glycosylase or DNA glycosylase/lyase is not lethal which probably reflects overlapping functional abilities among the glycosylases.

80. Base Excision Repair Pathway

The DNA glycosylase catalyzes base excision to produce an AP (apurinic and apyrimidininc) site. The next enzyme in the pathway, AP endonuclease, hydrolyzes the phosphodiester bond 5’ to the AP site to generate a nick.



E.coli has two well-characterized AP endonucleases:-

- Exonuclease III (Xth)
- Endonuclease IV (Nfo)

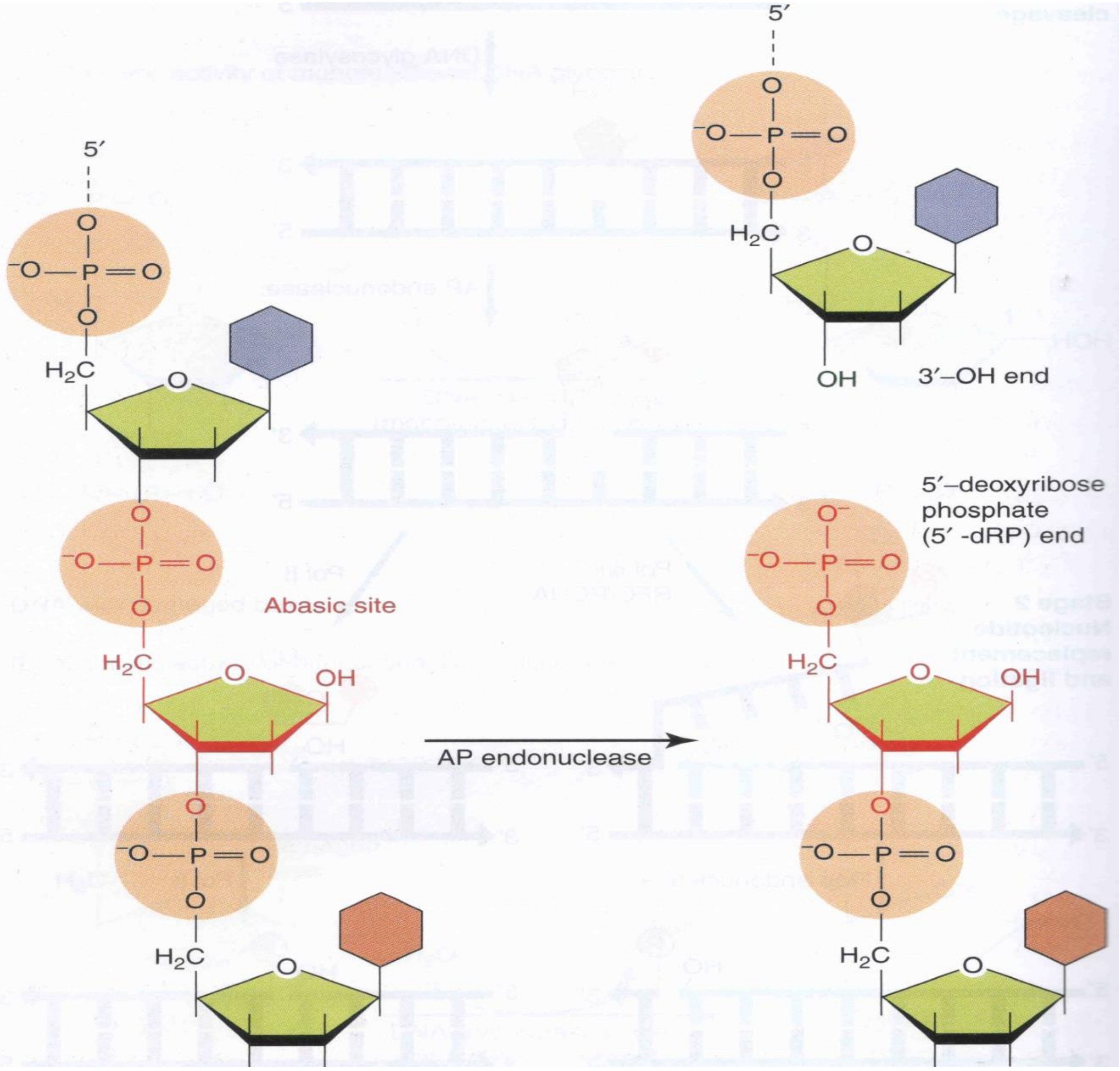
Exonuclease III (Xth) despite its name accounts for most of the bacterial AP endonuclease activity. Both are multifunctional enzymes that have 3’-phosphate and 3’-repair phosphodiesterase activities.

The former activity removes phosphate groups from the 3’ end of a DNA strand. The latter activity i.e., 3’-repair phosphodiesterase activity removes the 3’-unsaturated aldehydic group produced by DNA glycosylase/lyase action. These activities are important because DNA polymerase can’t attach new nucleotides to a blocked 3’-end.

The mammalian AP endonuclease, APE1 is homologous to *E. coli* exonuclease III.

81. Short Patch Repair

Base excision by DNA glycosylase and strand cleavage by AP endonuclease introduces a gap with a 5'-deoxyribose phosphate (5'-dRP) on one side and a 3'-OH on the other.



Additional enzymes are required to fill in the gap and remove the 5'-deoxyribose phosphate. Cells can repair the damage by two different pathways viz. short patch repair & long patch repair.

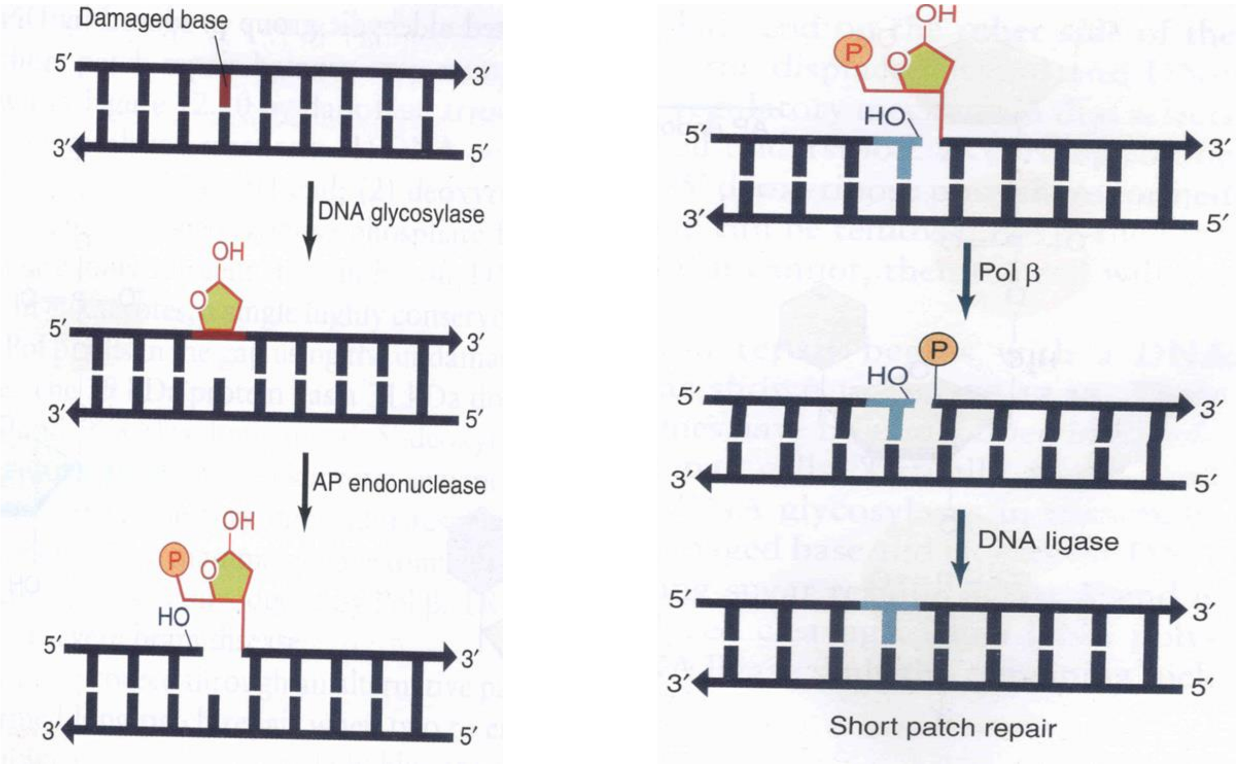
In the short patch repair, only a single nucleotide is replaced. This repair pathway involves the following enzyme catalyzed reactions:-

(1) DNA polymerase adds a deoxyribonucleotide to the 3'-OH end.

(2) Deoxyribose phosphate lyase (dRPase) removes 5'-deoxyribose phosphate from the 5'-end; and

(3) DNA ligase joins the adjacent ends.

In *E. coli*, DNA polymerase I fills in the gap. While in eukaryotes, a single highly conserved enzyme, DNA polymerase β (Pol β) fills in the gap using the undamaged DNA strand as the template.

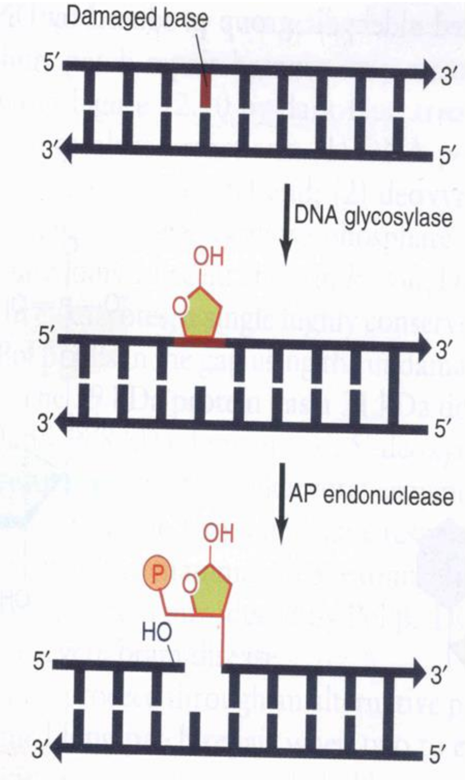


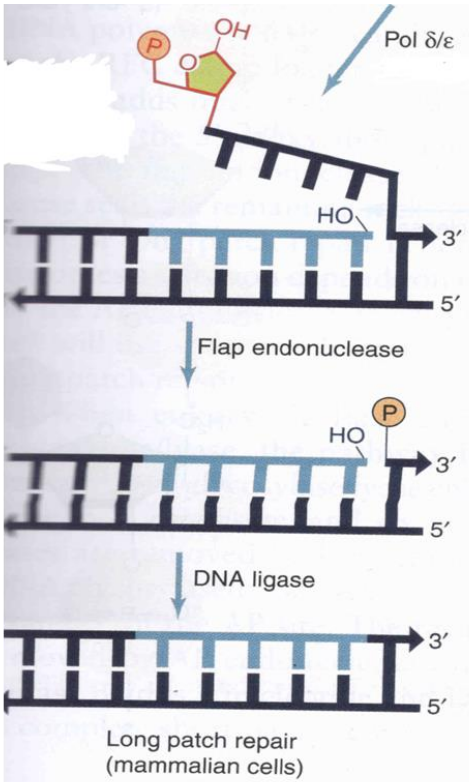
The eukaryotic enzyme differs from its bacterial counterpart in one very important respect, i.e., it lacks 3’ → 5’ proofreading activity._Two mammalian 3’ exonucleases, TREX1 and TREX2 may correct the errors introduced by Pol β.

82. Long Patch Repair

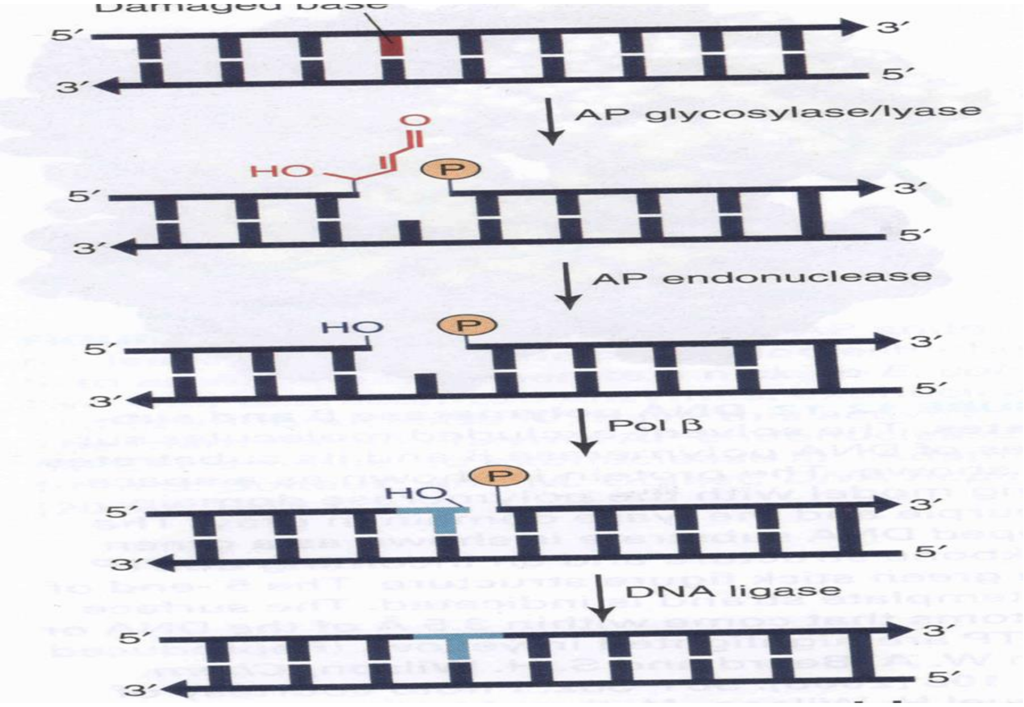
The alternative pathway to repair the damage in mammalian cells is termed as long patch repair. In this pathway, 2 – 8 nucleotides are replaced. In this case, DNA polymerase δ□ or Ɛ catalyzes chain extension with the assistance of the RFC (replication factor C) clamp loader and the PCNA sliding clamp. As the polymerase adds nucleotides to the 3’-OH end on one side of the gap, it displaces the 5’- deoxyribose phosphate end on the other side of the gap. The flap endonuclease cleaves the displaced strand and DNA ligase seals the remaining nick.

The regulatory mechanism that selects long or short patch repair is not well understood.

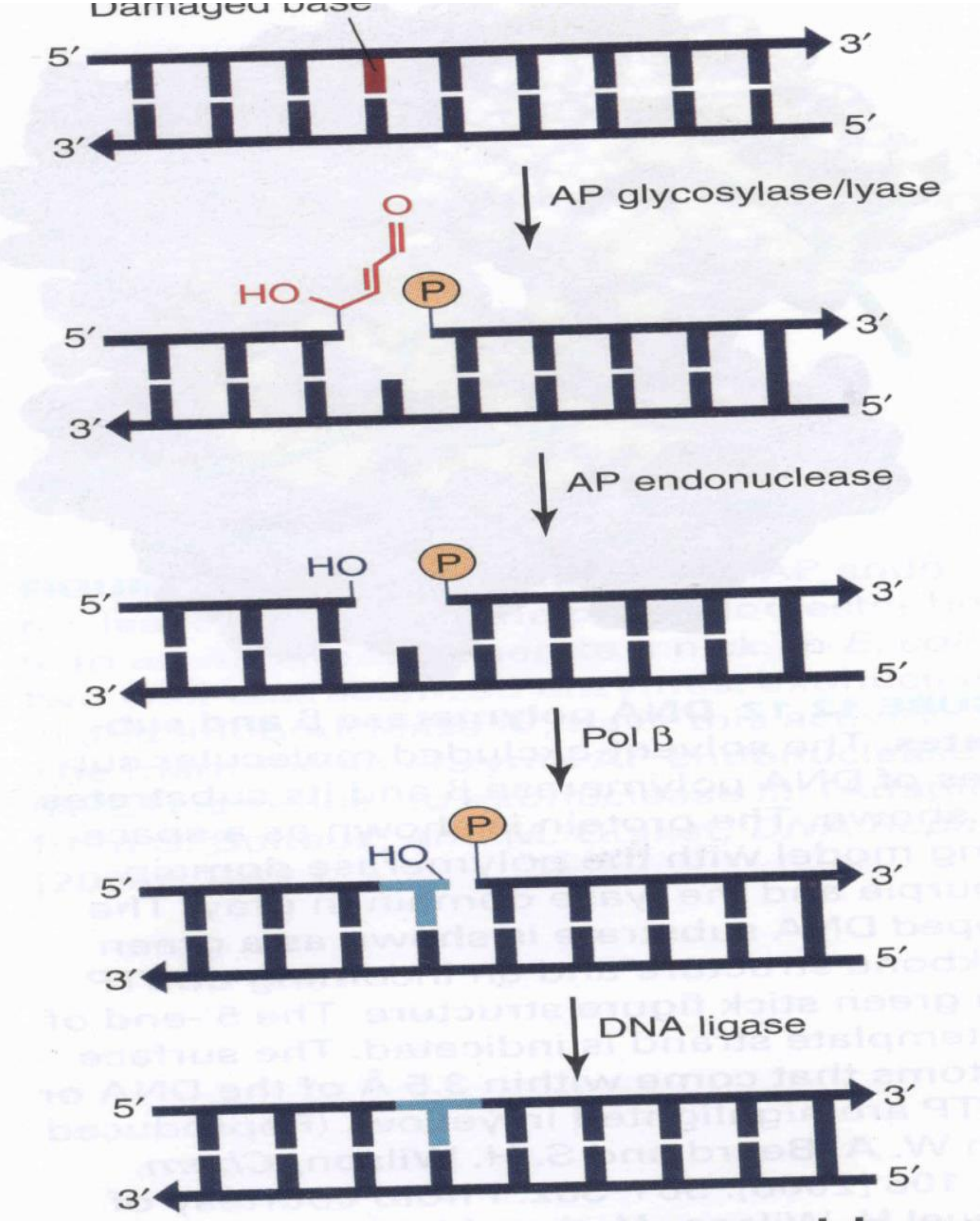




When eukaryotic base excision repair begins with a DNA glycosylase/lyase, the pathway is shown below:-



Three distinct DNA glycosylase/lyase enzymes have been identified in *E. coli*, three in *S. cerevisiae*, and six in human cells. Virtually all oxidized bases are removed by bifunctional DNA glycosylases in mammals. DNA glycosylase/ lyase excises the damaged base and cleaves the DNA strand 3’ of the AP site. The resulting sugar residue at the 3’-end is removed by the AP endonuclease catalyzed cleavage. Then DNA polymerase β adds a nucleotide and DNA ligase seals the remaining nick to complete short patch repair.


Damaged base
5′
3′
3′
5′
AP glycosylase/lyase
O
HO
P
5′
3′
3′
5′
AP endonuclease
HO
P
5′
3′
3′
5′
Pol β
P
HO
5′
3′
3′
5′
DNA ligase
5′
3′
3′
5′

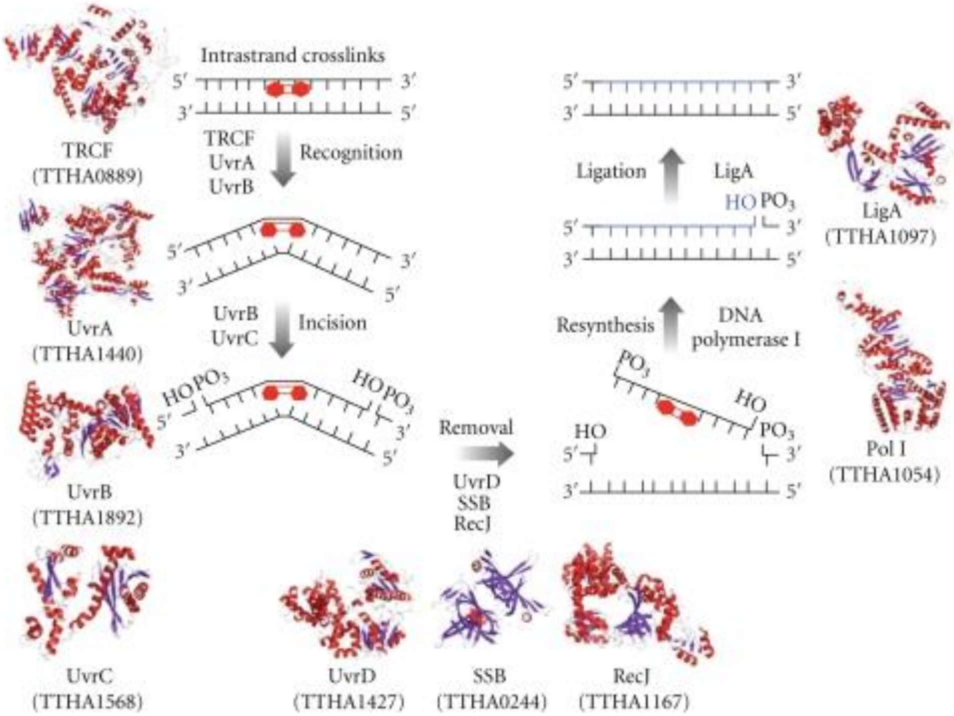
83. Nucleotide Excision Repair

Nucleotide excision repair (NER) pathway removes bulky adducts from DNA by excising an oligonucleotide bearing the lesion and replacing it with new DNA. This repair mechanism excises UV-induced cyclobutane pyrimidine dimers, (6-4) photoproducts, damaged bases formed by alkylating agents and certain types of cross-links. The efficiency of repair for different kinds of lesions can vary a lot.

In general, there is a direct relationship between the amount of helical distortions produced by lesion and the efficiency of this repair. The basic nucleotide excision repair pathway is same in all the organisms.It involves the following steps:-

1. damage recognition
2. an incision in the damaged DNA strand on each side of the lesion,
3. excision of the oligonucleotide created by the incision,
4. synthesis of new DNA to replace the excised DNA segment using the undamaged DNA strand as a template, and
5. ligation of the remaining nick.

Although the basic nucleotide excision repair pathway is similar in all the organisms, there are considerable differences in the proteins that carry out the various steps.



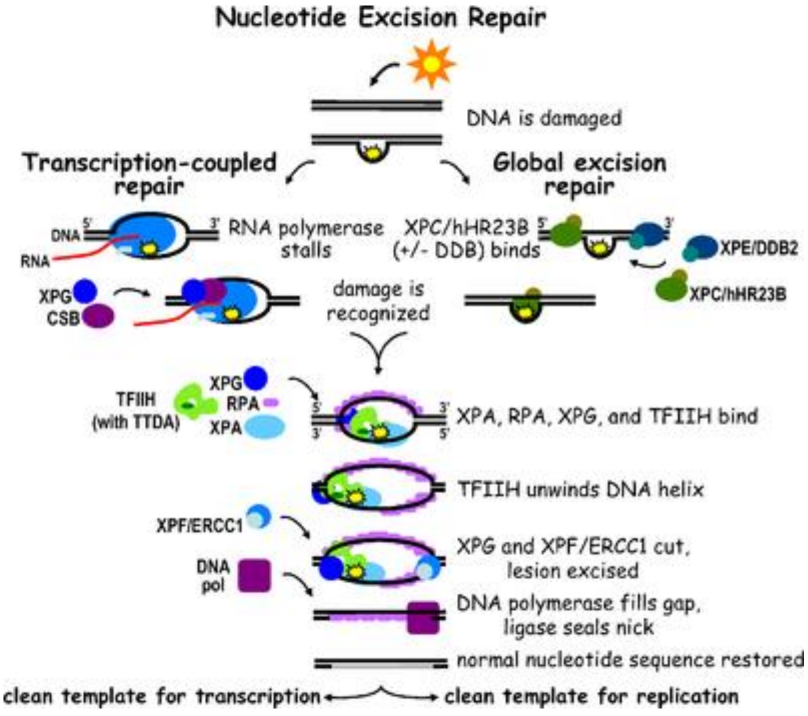
84. Nucleotide Excision Repair of UV-induced Damage

UV-irradiated *E. coli* can regain their ability to survive after incubation in dark. However, they recover more slowly than when incubated in the light. This observation suggests that the bacteria use some process other than photoreactivation to repair the UV-induced DNA damage. Richard Setlow & William Carrier and, working independantly, Richard Boyce & Paul Howard-Flanders, used a similar approach to investigate this alternative process in 1964. Both groups cultured *E. coli* in the presence of $[^{3}H]$thymine to label the DNA and then irradiated the cells with UV light to induce the formation of cyclobutane thymine dimer.

Then they:-

1) incubated the UV-irradiated cells in the dark so that the photoreactivation could not take place;
2) removed samples after various incubation times and added TCA to them;
3) separated acid-insoluble DNA from acid soluble nucleotides;
4) digested the DNA and oligonucleotides to release intact thymine cyclobutane dimers; and
5) detected the released dimers by chromatography.

The experiments revealed that as the incubation time in the dark increases, Cyclobutane thymine dimers disappear from the acid-insoluble DNA and appear in the acid soluble oligonucleotide fraction.These results were correctly interpreted to mean that bacteria can excise an oligonucleotide containing a lesion and replace the excised oligonucleotides with newly synthesized DNA. Subsequent studies showed that eukaryotes and the archaea also have nucleotide excision repair pathways.


Nucleotide Excision Repair
DNA is damaged
Transcription-coupled repair
Global excision repair
DNA
RNA
5'
3'
RNA polymerase stalls
XPC/hHR23B (+/- DDB) binds
XPE/DDB2
XPC/hHR23B
XPG
CSB
damage is recognized
TFIIH (with TTDA)
XPG
RPA
XPA
XPA, RPA, XPG, and TFIIH bind
TFIIH unwinds DNA helix
XPF/ERCC1
XPG and XPF/ERCC1 cut, lesion excised
DNA pol
DNA polymerase fills gap, ligase seals nick
normal nucleotide sequence restored
clean template for transcription
clean template for replication

85. UvrA, UvrB, and UvrC Proteins

UvrABC endonuclease is a multienzyme complex in *Escherichia coli* involved in DNA repair by nucleotide excision repair, and it is, therefore, sometimes called an excinuclease. This UvrABC repair process, sometimes called the short-patch process, involves the removal of twelve nucleotides where a genetic mutation has occurred followed by a DNA polymerase, replacing these aberrant nucleotides with the correct nucleotides and completing the DNA repair. The subunits for this enzyme are encoded in the *uvrA*, *uvrB*, and *uvrC* genes. This enzyme complex is able to repair many different types of damage, including cyclobutyl dimer formation.

Genetic studies revealed that three *E. coli* genes viz., *uvrA*, *uvrB*, and *uvrC*, code for proteins that are essential for damage recognition, incision, and excision. All three genes have been cloned and the proteins that they encode (UvrA, UvrB, and UvrC) have been purified and characterized. Under normal physiological conditions, *E. coli* has about 25 molecules of UvrA, 250 molecules of UvrB, and 10 molecules of UvrC. After UV damage of DNA, UvrA and UvrB levels increase ten- and four folds, but the UvrC level remains the same. Although UvrA, UvrB, and UvrC do not combine to form a stable ternary complex, the polypeptides nevertheless are said to be part of a UvrABC damage-specific endonuclease. The UvrABC damage-specific endonuclease is called as UvrABC endonuclease in short but some investigators preferably term them as UvrABC excinuclease. UvrABC excinuclease is termed so because the proteins participate in incision and excision reactions. The three polypeptides work in the order suggested by their names, that is their order of action is UvrA, UvrB, and the UvrC.

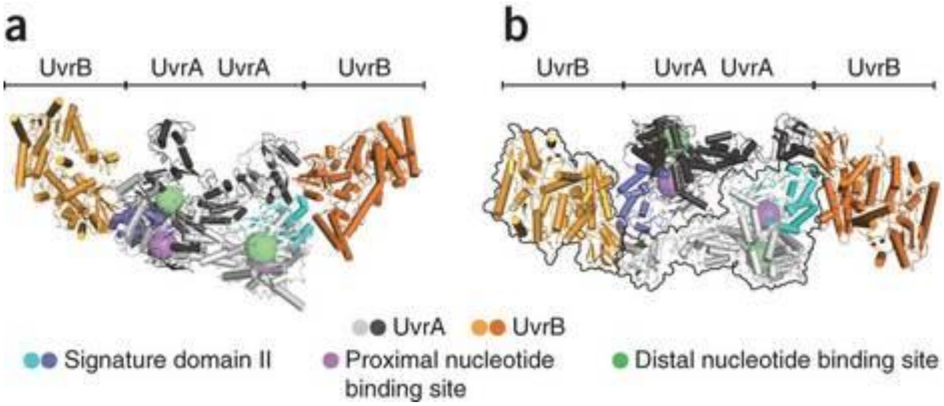


Fig. Architecture of the UvrA–UvrB DNA damage sensor.close

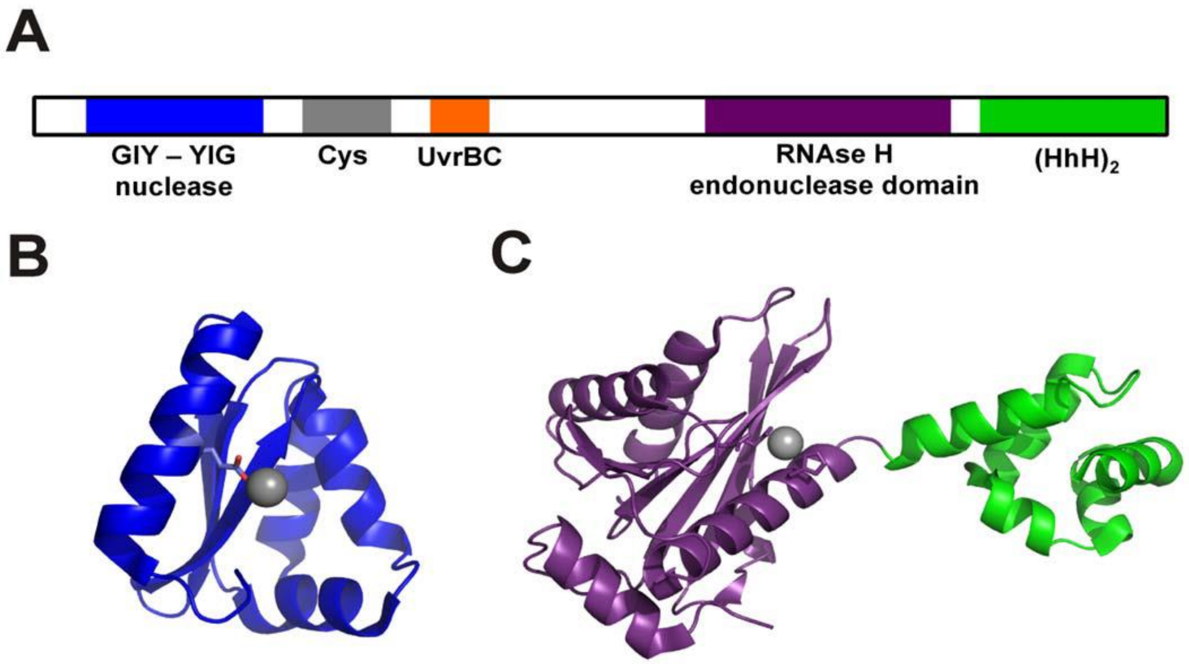


Fig. (A) UvrC consists of two endonuclease domains, particularly an N-terminal GIY-YIG nuclease domain (blue) and a C-terminal RNAse H-like endonuclease domain (purple). UvrC also comprises a Cys-rich region (gray), an UvrB-interacting domain (orange) and a C-terminal tandem Helix-hairpin-helix motif (green). (B) Crystal structure of the N-terminal GIY-YIG endonuclease domain of T. maritima UvrC (PDB: 1YD1) (Truglio et al, 2005). The residue E76 coordinating the Mg2+-ion (gray sphere) is shown as stick model. (C) Crystal structure of the C-terminal RNAse H-like endonuclease domain (purple) and the (HhH)2-domain (green) of T. maritima UvrC (PDB entry: 2NRZ) (Karakas et al, 2007). The catalytic triade DDH consisting of residues D367, D429 and H488 which coordinates one Mn2+-ion (gray sphere) is represented as stick model.

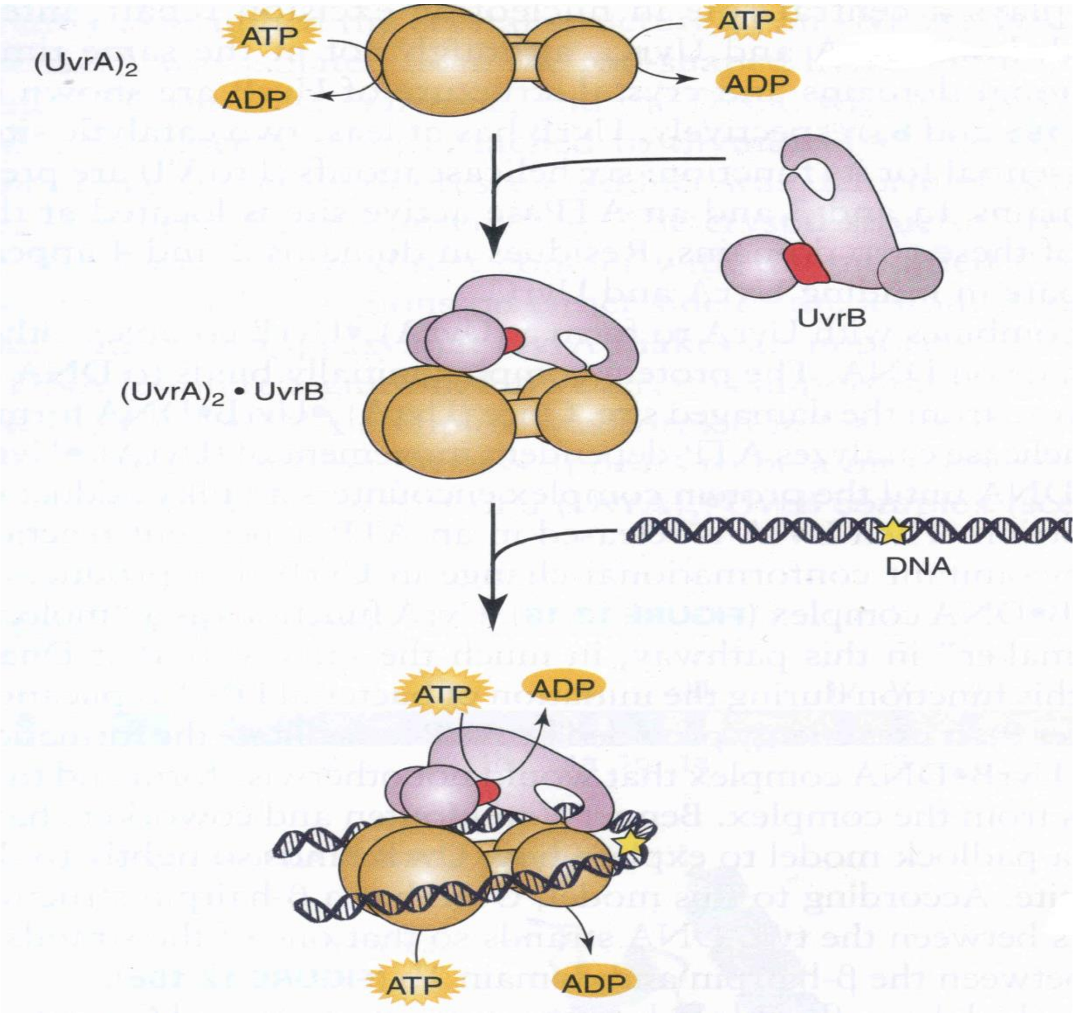
86. The NER Pathway

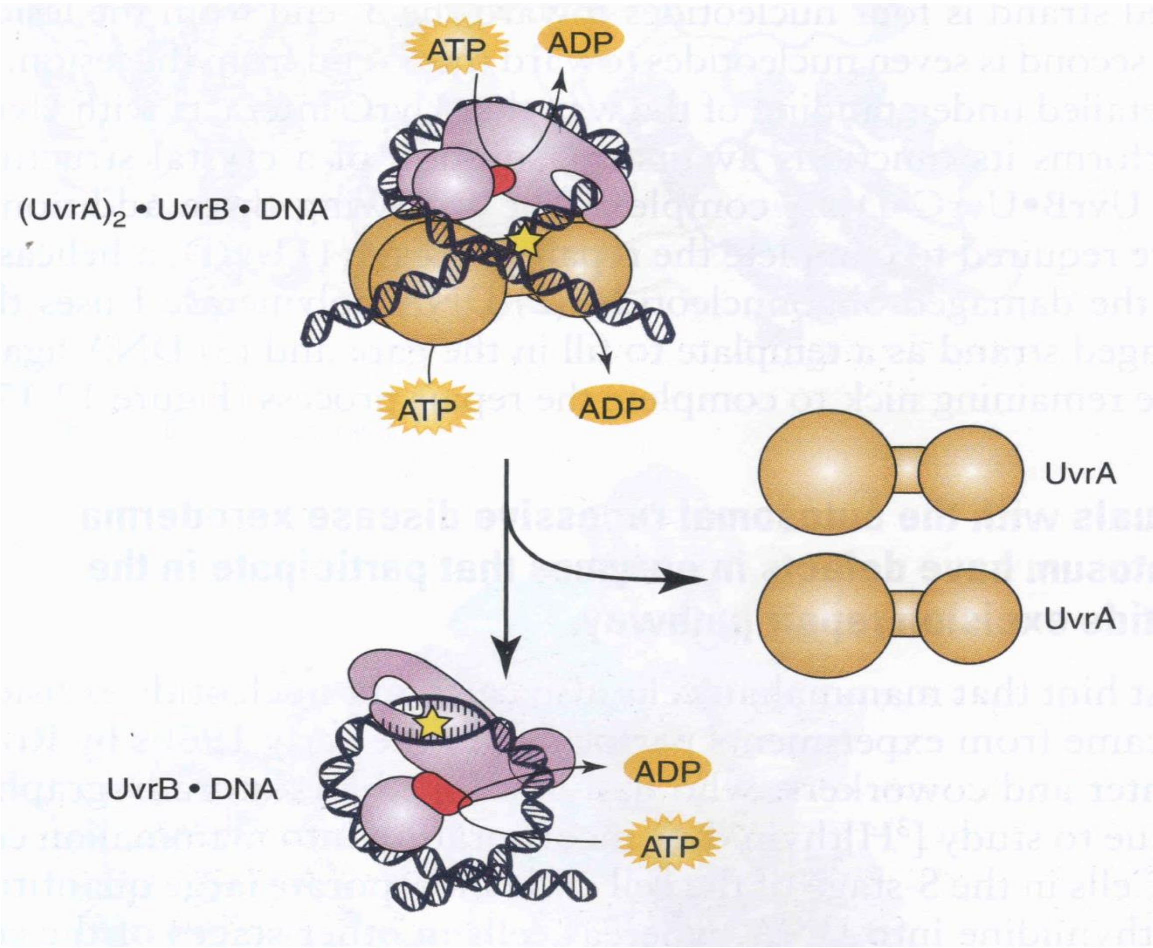
The crystal structure of UvrA•DNA complex reveals that UvrA does not make direct contact with the modified thymine but does bind to DNA regions on either side of the lesion. Based on this information, it appears that UvrA makes an important contribution towards recognition of DNA damaged lesion.

This UvrA•DNA complex is formed in vitro in the absence of UvrB. The recognition process appears to be more complicated in vivo where UvrA is a part of a (UvrA)2•UvrB complex. UvrB plays a central role in nucleotide excision repair by interacting with both the UvrA and UvrB, although not at the same time.

UvrB has at least two catalytic sites that are essential for its function. UvrB combines with UvrA to form a (UvrA)2•UvrB complex either in solution or on DNA.

Initially, this protein complex binds to DNA at some distance from the damaged site. Once (UvrA)2•UvrB • DNA complex is formed, the UvrB helicase catalyzes ATP-dependant movement of (UvrA)2•UvrB along the DNA until the protein complex encounters a bulky adduct or helix distortion. The UvrA is released in an ATP-dependant reaction with a concomitant conformational change in the UvrB that produces a stable UvrB•DNA complex.

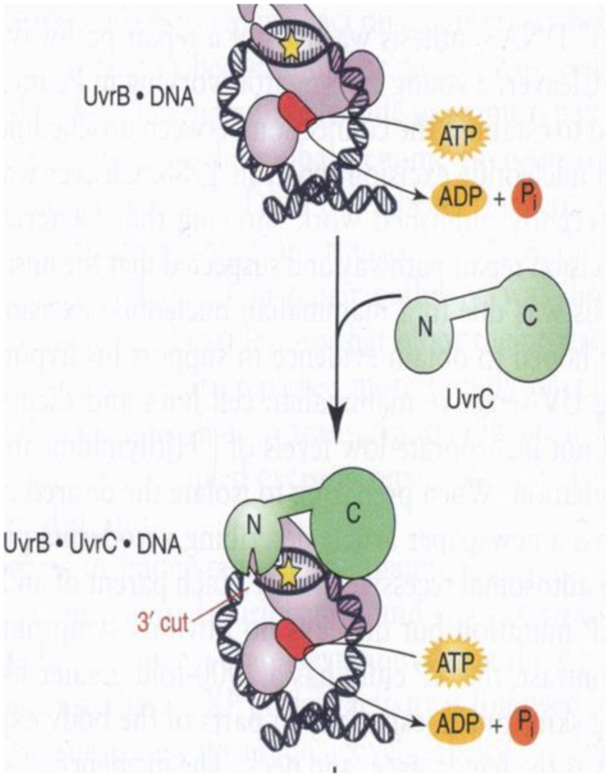

ATP
ATP
(UvrA)$_2$
ADP
ADP
UvrB
(UvrA)$_2$ • UvrB
DNA
ATP
ADP
ATP
ADP

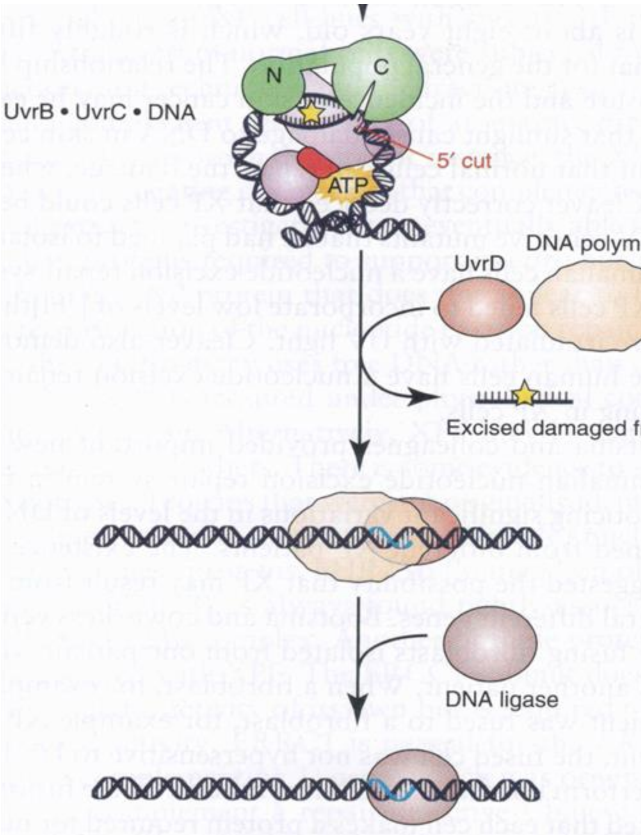


UvrA functions as a “molecular match maker” in this pathway. It uses energy of ATP to facilitate the formation of UvrB•DNA complex and then dissociates from the complex. UvrC, which has a flexible linker that connects its N- and C-terminal domains, binds to the UvrB•DNA complex and makes two incisions, one on each side of the lesion. The first incision of the damaged strand is four nucleotides towards the 3’-end from the lesion and the second is seven nucleotides toward the 5’-end from the lesion. How UvrC interacts with UvrB and performs its function is still to be described.

Following three more steps are required to complete the repair process:-

- 1). UvrD, a helicase, excises the damaged oligonucleotide;
- 2). DNA polymerase I uses the undamaged strand as a template to fill the gap;
- 3). DNA ligase seals the remaining nick to complete the repair process.



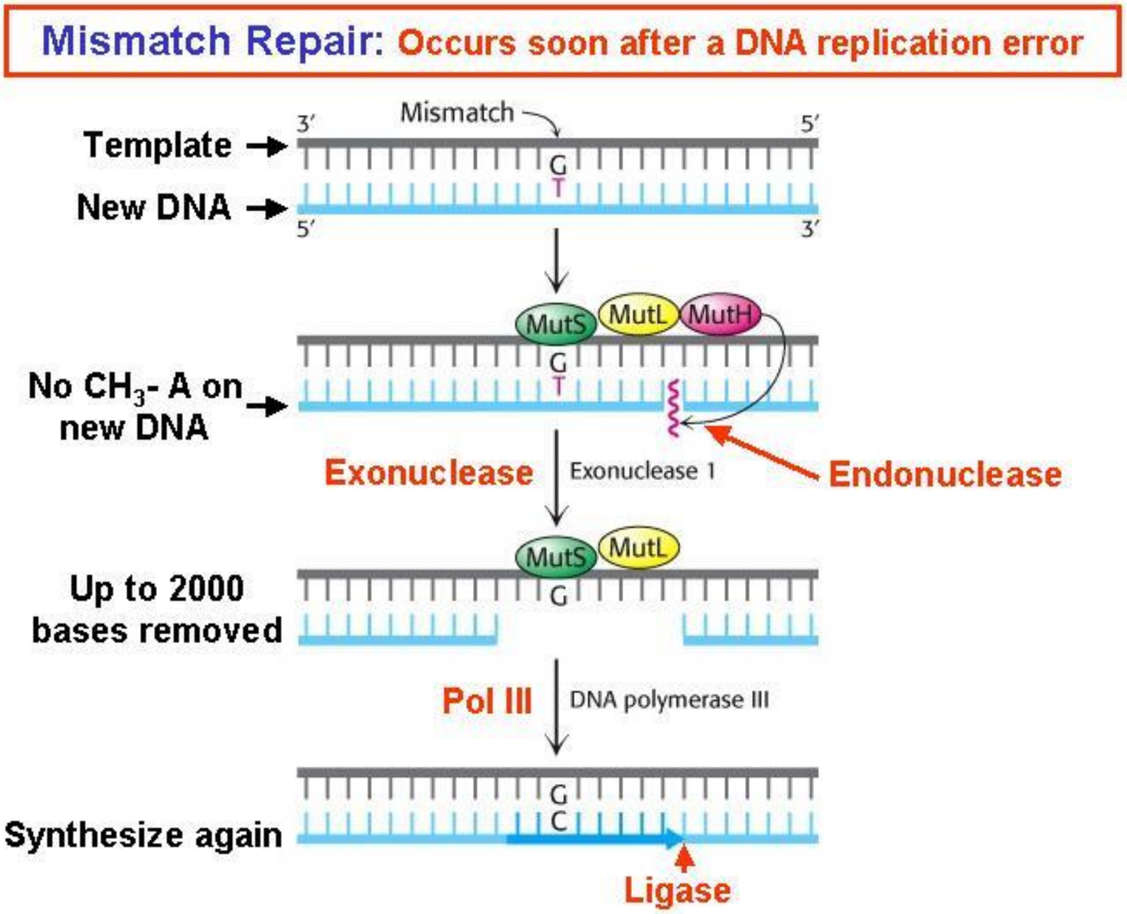


87. Mismatch Repair

The mismatch repair system corrects rare base pair mismatches and short insertions or deletions that appear in DNA following replication. DNA polymerases introduce about one mispaired nucleotide per 10^5 nucleotides.

However, the 3’ → 5’ proofreading exonuclease increases replication fidelity by 100- fold by removing mispaired nucleotides. Although an error frequency of one nucleotide in 10^7 may seem extremely low, it would result in a high mutation rate. The second type of error that occur during replication is the short insertions and deletions, which result from the fact that repeated-sequence motifs sometimes dissociate and then re-anneal incorrectly. As a result, the newly synthesized strand will have a different number of repeats than the template strand.

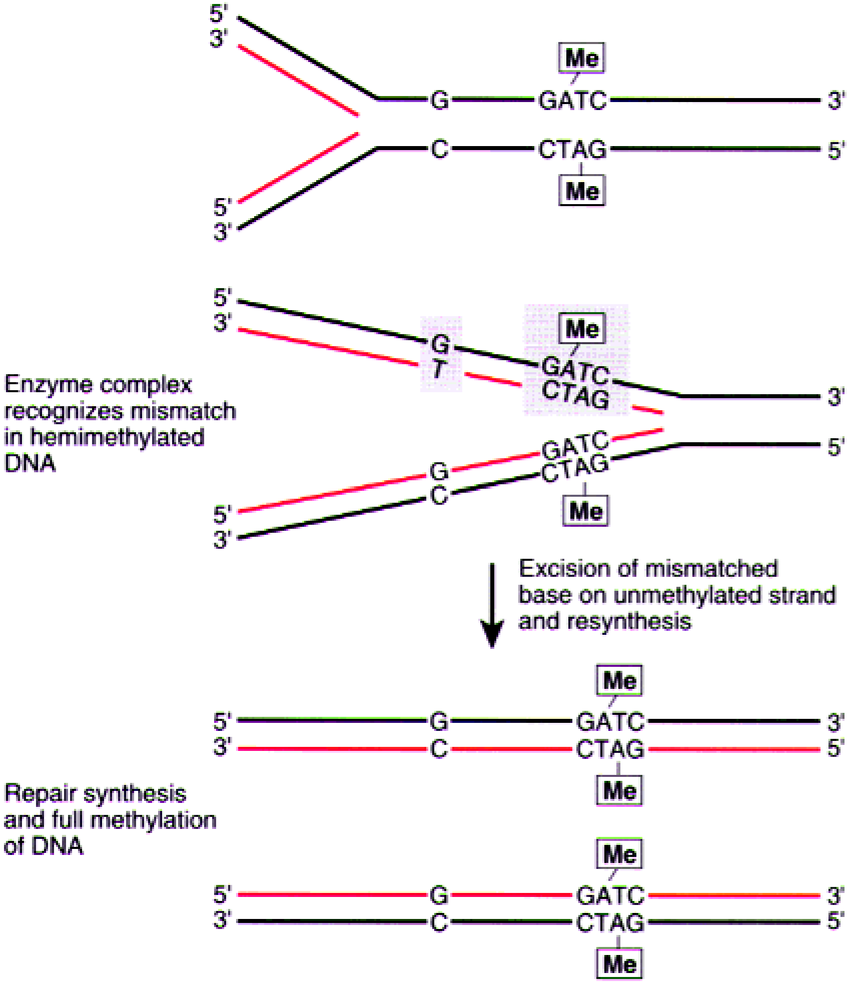
Introduction of an insertion or deletion into the newly synthesized DNA is likely to produce a mutation. Cells with a non-functional mismatch repair system have a high rate of mutation due to their inability to efficiently repair base pair mismatches, short insertions or deletions that arise during replication.



88. Mismatch Repair System in *E. coli*

Let us begin the examination of mismatch repair by considering the *E. coli* mismatch repair system because this system has been the most extensively studied. Although this system provides valueable informations for studying mismatch repair in other organisms, it differs from the mismatch repair systems of gram-positive bacteria & eukaryotes in one important respect. The *E. coli* mismatch repair system can distinguish a newly synthesized strand from a parental strand because only the latter has methyl groups attached to sites with the sequence GATC. *E. coli* has a deoxy- adenosine methylase that transfers methyl groups from S-adenosylmethionine molecules to deoxyadenosines in GATC sequences. The time of methylation by deoxyadenosine methylase, however, lags behind that of nucleotide addition at the replication fork about two minutes, so the newly synthesized strand is transiently unmethylated. The *E. coli* mismatch repair system exploits this period of transient unmethylation to identify and cut GATC sites in a newly synthesized strand with a mismatch.

GATC sequences in DNA are normally methylated (Me) at the 6 position of adenine. During semiconservative DNA synthesis, a G-T mismatch arises in one of the sister DNA duplexes. The enzymatic mechanism for repairing this lesion depends on discrimination between the newly synthesized (red) and parental (black) strands. This is achieved by recognition of the temporary lack of methylation of the newly synthesized strand before postreplicative DNA methylation takes place. The nonmethylated daughter strand containing the incorrect base is enzymatically attacked by mismatch correction enzymes, and the misincorporated base is removed. Repair synthesis and daughter-strand methylation at GATC sequences restore the sister DNA duplexes to their native state.


5'
3'
Me
G
GATC
3'
C
CTAG
5'
Me
5'
3'
5'
3'
G
T
Me
GATC
CTAG
3'
Enzyme complex recognizes mismatch in hemimethylated DNA
GATC
CTAG
5'
G
C
Me
5'
3'
Excision of mismatched base on unmethylated strand and resynthesis
Me
5'
G
GATC
3'
3'
C
CTAG
5'
Me
Repair synthesis and full methylation of DNA
Me
5'
G
GATC
3'
3'
C
CTAG
5'
Me

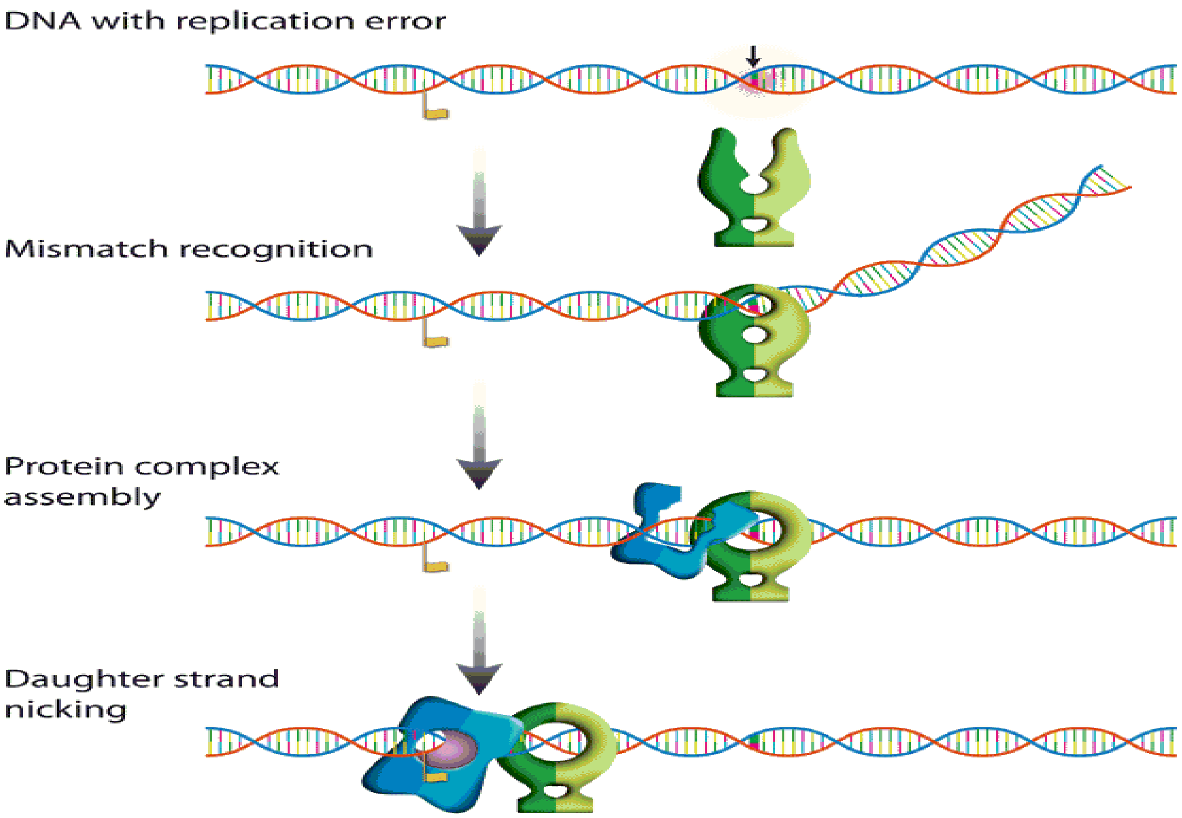
89. MutS, MutL, & MutH Proteins

Genetic and biochemical studies have demonstrated that three *E. coli* proteins viz., MutS, MutL, and MutH are dedicated to mismatch repair. Although these proteins are essential for mismatch repair, they are not sufficient. Several additional enzymes and protein factors also make important contributions. Among these enzymes and protein factors are:-

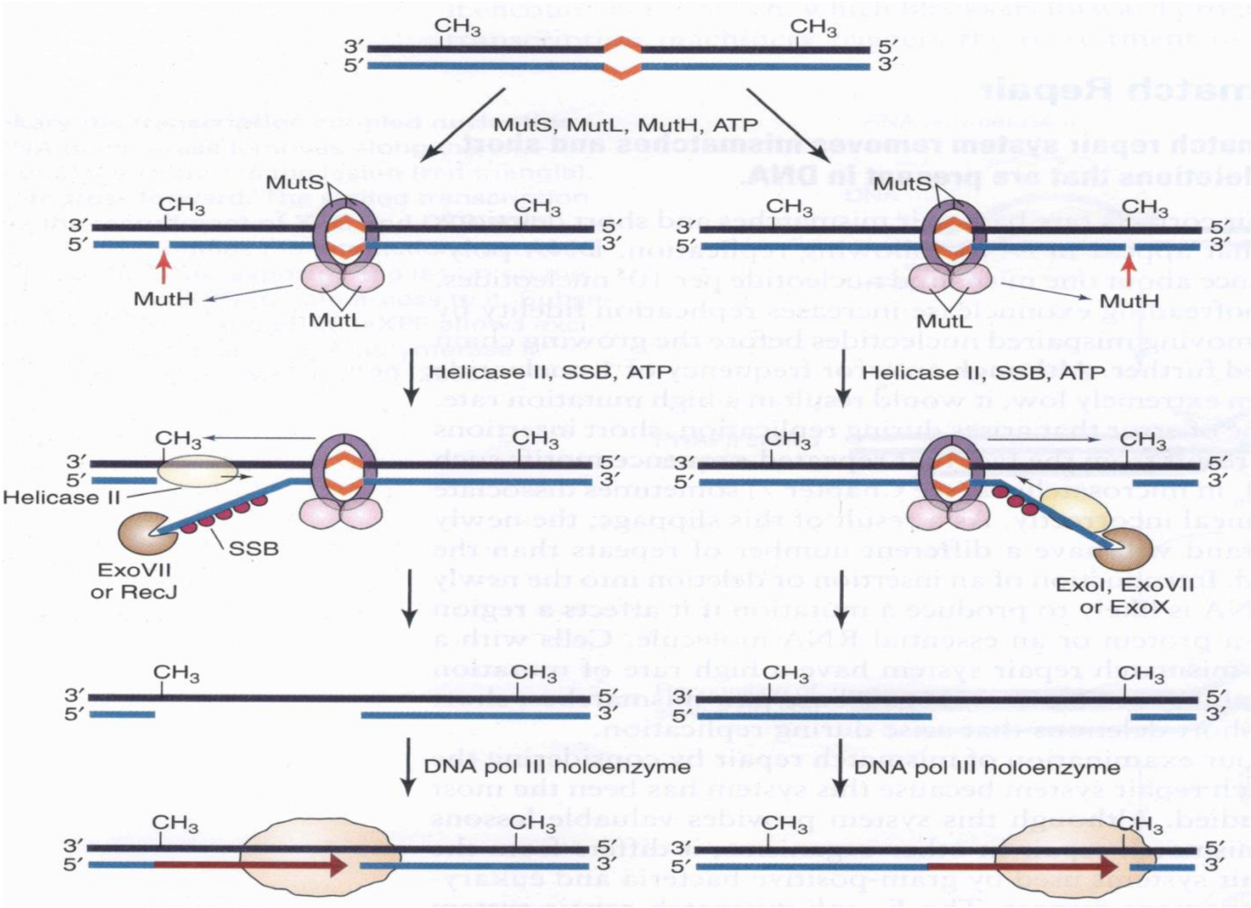
- i. DNA helicase II (UvrD),
- ii. Single-stranded DNA binding protein (SSB)

- iii. 5’→3’ exonucleases (ExoVII or RecJ)
- iv. 3’→5’ exonucleases (ExoI, ExoVII, or ExoX),
- v. DNA polymerase III holoenzyme,
- vi. DNA ligase, and
- vii. Deoxyadenosine methylase.

The process of repair begins when MutS (a homodimer or homotatramer) binds to the mismatch. MutS recruits MutL (a homodimer) in an ATP-dependant fashion. Then the MutS•MutL complex activates MutH, which makes an incision at the nearest unmethylated GATC site, either 5’ or 3’ to the mismatch, in the newly synthesized strand.



MutH shares sequence homology with the type II restriction endonuclease, Sau3AI. Both enzymes recognize and cleave GATC sequences. MutH does not bind or cleave fully methylated GATC sites, whereas Sau3AI cleaves fully, hemi and unmethylated GATC sites. Next in the process is that helicaseII (UvrD) unwinds the DNA and SSB binds to the resulting single strands. When the incision is 5’ to the mismatch, ExoVII or RecJ exonucleases hydrolyze the nicked strand in a 5’ 3’ direction. When the incision is 3’ to the mismatch, ExoI, ExoVII or ExoX exonucleases hydrolyze the nicked strand in a 3’ 5’ direction. DNA polymerase III holoenzyme fills the gap with new DNA. DNA ligase seals the remaining nick and deoxyadenosine methylase adds a methyl group to the GATC site.



All organisms that have a mismatch repair system have MutS and MutL homologs. However, MutH is present only in *E. coli* and some other gram-negative bacteria.

90. Mismatch Repair in Eukaryotes

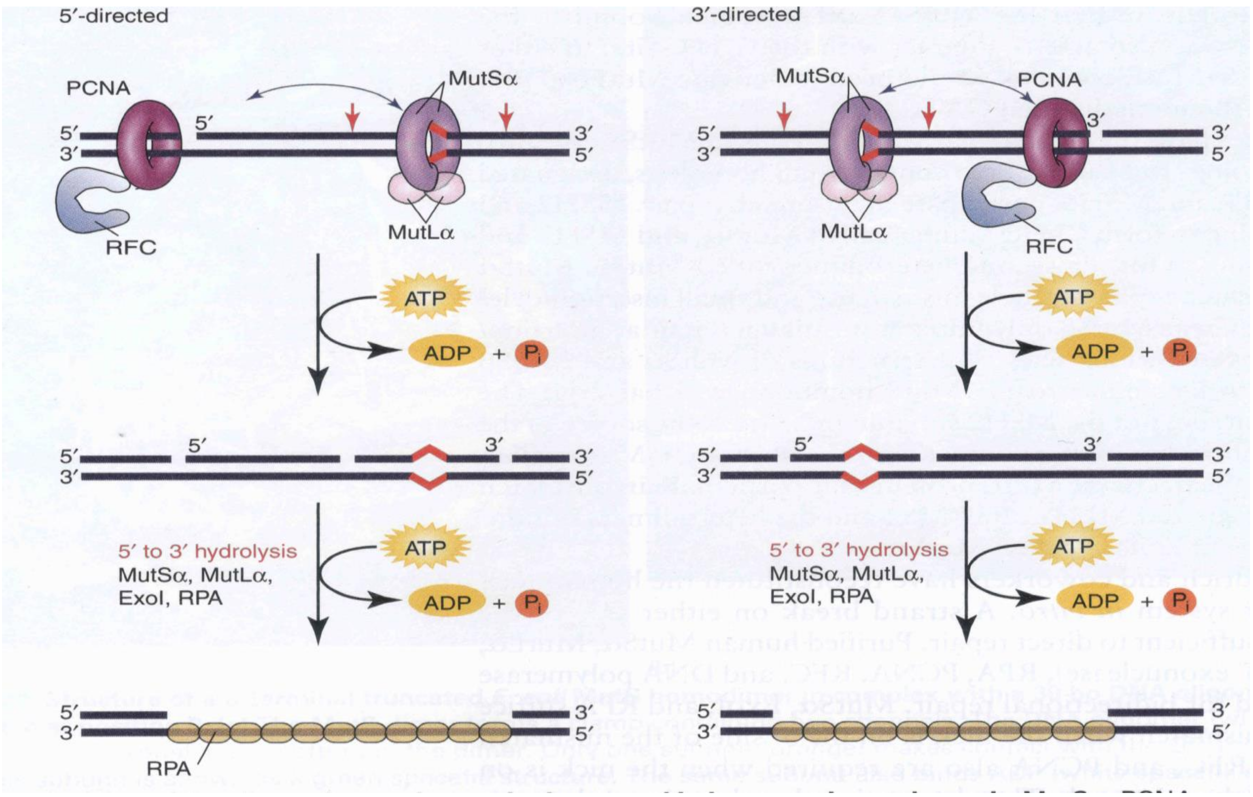
Eukaryotes have proteins that are homologous to MutS and MutL but lack homologs to MutH. There are three human MutS homologs designated as MSH2, MSH3, and MSH6 which participate in mismatch repair. MSH2 and MSH6 combine to form a hetrodimer called MutSα, and MSH2 and MSH3 combine to form a second heterodimer called MutSβ. The structures of MutSα and MutSβ are thought to be similar to the MutS homodimer in bacteria.

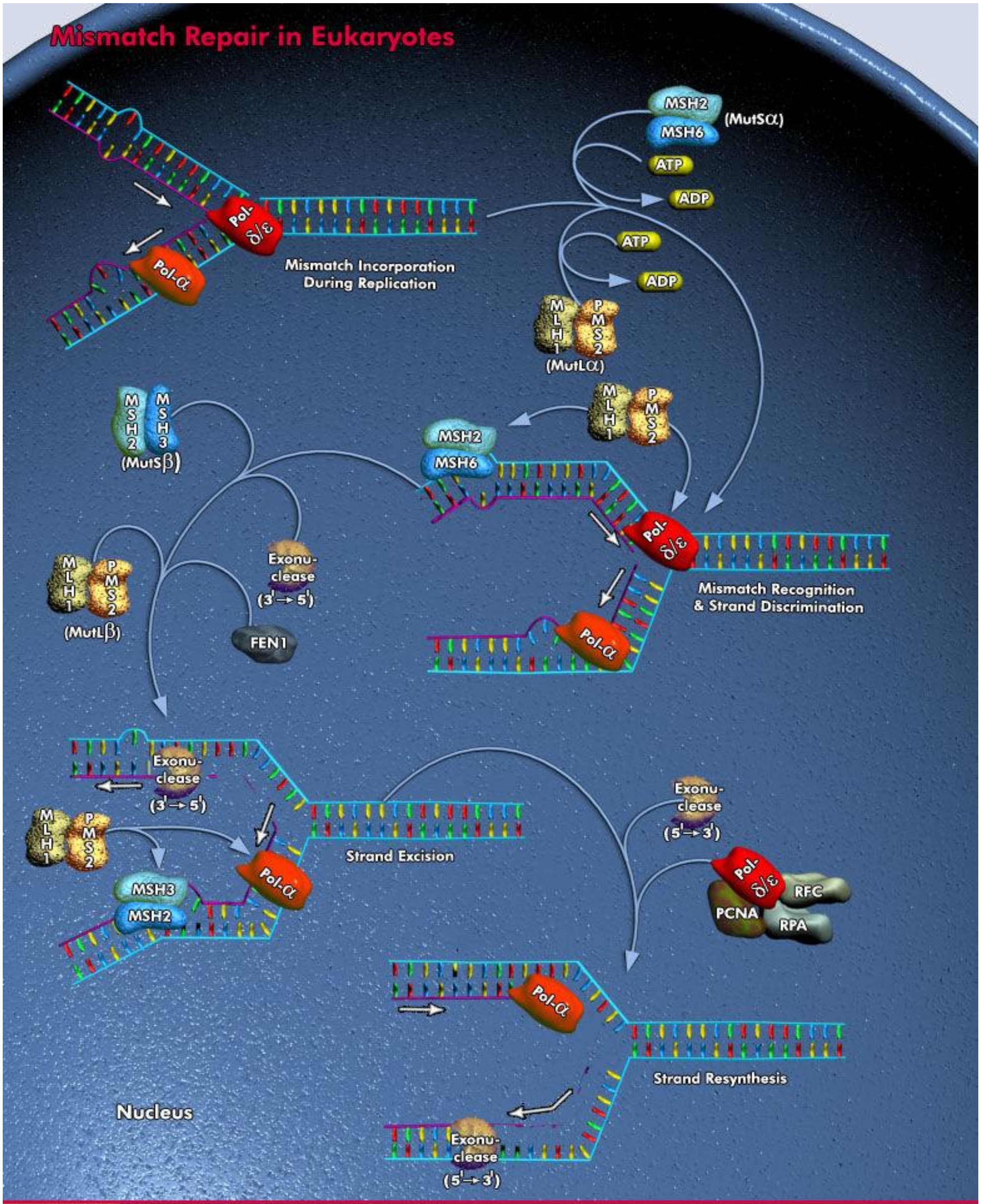
MutSα initiates mismatch repair at single mismatches and small insertions/deletion loops, whereas MutSβ only initiates mismatch repair at insertion/deletion loops of various sizes. Mammalian homologs of the bacterial MutL protein that participate in mismatch repair are designated as MLH1 and PMS2 and the heterodimer containing these two subunits is called MutLα. Paul Modrich and coworkers have reconstituted the human mismatch repair system *in vitro*.

A strand break on either side of the mismatch is sufficient to direct repair.

91. **Human Mismatch Repair System**

Paul Modrich and coworkers have reconstituted the human mismatch repair system *in vitro*. A strand break on either side of the mismatch is sufficient to direct repair. Purified human MutSα, MutLα, ExoI (a 5' to 3' exonuclease), RPA, PCNA, RFC, and DNA ploymerase□ are required for bidirectional repair. MutSα, ExoI, and RPA are adequate to excise a mismatch when the nick is on the 5' side of the mismatch but MutLα, PCNA, and RFC are also required when the nick is on the 3' side of the mismatch. The observation that the mismatch repair system can degrade newly synthesized strands with a nick on the 3'-side of the mismatch was very puzzling because ExoI degrades DNA in a 5' to 3' direction. Modrich and co-workers solved the puzzle by demonstrating that MutLα is a latent endonuclease that is activated in a mismatch. Once activated, MutLα preferentially makes incisions in the strand that already has a nick, that is, the discontinuous strand during replication. The endonuclease activity appears to require an amino acid motif present in the PMS2 but not the MLH1 subunit. In the human mismatch repair pathway, MutSα, PCNA, and RFC cooperate to activate the latent MutLα endonuclease. This MutLα endonuclease then nicks the discontinuous strand of a DNA duplex on both the 5' and 3' sides of the mismatch. When the original nick is on the 3' side of the mismatch, MutLα incisions produce a new 5' terminus on the far side of the mismatch that can serve as an entry site for MutSα-activated ExoI. This activated ExoI then removes the mismatch using its 5' to 3' exonuclease activity. RPA stimulates ExoI activity in the presence of MutSα as long as the mismatch is present. Once the mismatch has been removed, RPA inhibits the exonuclease, probably by displacing ExoI from the DNA.


5′-directed
3′-directed
PCNA
MutSα
MutLα
RFC
5′
3′
ATP
ADP + P_i
5′ to 3′ hydrolysis
MutSα, MutLα,
ExoI, RPA
RPA



Mismatch Repair in Eukaryotes

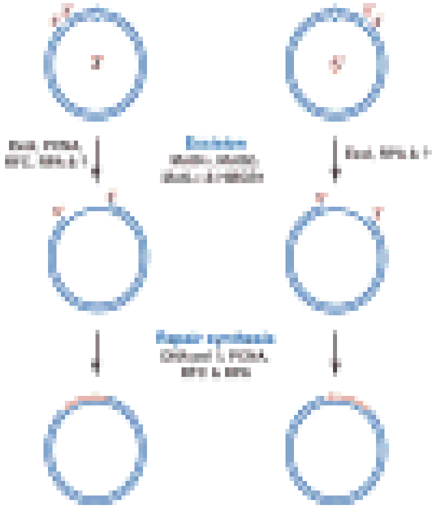
91. Human Mismatch Repair System

Inactivation of the human mismatch repair system confers a large increase in spontaneous mutability and a strong predisposition to tumor development. Mismatch repair provides several genetic stabilization functions: it corrects DNA biosynthetic errors, ensures the fidelity of genetic recombination, and participates in the earliest steps of checkpoint and apoptotic responses to several classes of DNA damage. Defects in this pathway are the cause of typical and atypical hereditary nonpolyposis colon cancer, but may also play a role in the development of 15 to 25% of sporadic tumors that occur in a number of tissues. The system is also of biomedical interest because mismatch repair-deficient tumor cells are resistant to certain cytotoxic chemotherapeutic drugs, a manifestation of its involvement in the DNA damage response. Of the several mutation avoidance functions of mismatch repair, the reaction responsible for replication error correction has been the most thoroughly studied, and the discussion that follows is restricted to this pathway.

Mismatch Repair in Eukaryotic Cell Extracts

Correction of DNA biosynthetic errors requires targeting of mismatch repair to the newly synthesized strand at the replication fork. In contrast to *E. coli*, where mismatch repair is directed by the transient absence of adenine methylation at d(GATC) sites within newly synthesized DNA, the strand signals that direct replication error correction in eukaryotes have not been identified. However, the function of the hemimethylated d(GATC) strand signal in *E. coli* mismatch repair is provision of a nick on the unmethylated strand, which serves as the actual signal that directs the reaction . Similarly, a strand-specific nick or gap is sufficient to direct mismatch repair in extracts of mammalian and *Drosophila* cells, as well as *Xenopus* egg extracts. These findings, coupled with the observation that mismatch repair is more efficient on the lagging strand at the replication fork, suggest that DNA termini that occur as natural intermediates during replication (3'-terminus on the leading strand; 3' and 5' termini on the lagging strand) may suffice as strand signals to direct the correction of DNA biosynthetic errors in eukaryotic cells.

Available information on the mechanism of eukaryotic mismatch repair is largely derived from analysis of the nick-directed repair of circular heteroduplexes in mammalian cell extracts. The strand break that directs repair may reside either 3' or 5' to the mispair as viewed along the shorter path linking the two sites in the circular substrate, and processing of such molecules in extracts is largely restricted to this region. Examination of intermediates produced in HeLa nuclear extracts when repair DNA synthesis is blocked has demonstrated that mismatch-provoked excision removes that portion of the incised strand spanning the shorter path between the nick and the mismatch, with excision tracts extending from the strand break to a number of sites within a region ≈ 90 to 170 nucleotides beyond the mispair. Radiolabeling of repair DNA synthesis tracts is also consistent with this view. The mammalian repair system thus displays a bidirectional capability in the sense that it responds to both 3'- and 5'-heteroduplex orientations, and functionality is retained at nick-mismatch separation distances as large as 1,000 base pairs (bp).



Substrates and requirements for *in vitro* mismatch repair

Mammalian MutS and MutL Activities

The activities responsible for initiation of *E. coli* mismatch repair are MutS and MutL, which function as homo-oligomers. MutS is responsible for mismatch recognition and MutL serves to interface mismatch recognition by MutS to activation of downstream activities. Mammalian cells possess two MutS activities that function as heterodimers and share MSH2 as a common subunit : MutSα (MSH2•MSH6 heterodimer) and MutSβ (MSH2•MSH3 heterodimer). MutSα, which represents 80 - 90% of the cellular MSH2, preferentially recognizes base-base mismatches and insertion/deletion (ID) mispairs in which one strand contains 1 or 2 unpaired nucleotides, but is also capable of recognition of larger ID heterologies with reduced affinity. MutSβ recognizes ID mismatches of 2 to about 10 nucleotides, weakly recognizes single-nucleotide ID mispairs, and is essentially inert on base-base mismatches. *MSH2*and *MSH6* defects have been implicated in tumor development, but the cancer predisposition conferred by*MSH6* inactivation is less severe. The association of *MSH3* defects with tumor development appears to be limited.

Three eukaryotic MutL activities have been identified, and like eukaryotic MutS activities function as heterodimeric complexes, with MLH1 serving as a common subunit. MutLα, a heterodimer of MLH1 and PMS2, is the primary MutL activity in human mitotic cells and supports repair initiated by either MutSα or MutSβ. MutLα accounts for ≈ 90% of the MLH1 in human cells, but two low abundance complexes involving MLH1 have also been identified. A human MLH1•PMS1 heterodimer (MutLβ) has been isolated, but involvement in mismatch repair has not been demonstrated. However, the MutLγ MLH1•MLH3 complex has been reported to support modest levels of base-base and single nucleotide ID mismatch repair *in vitro*, events that are presumably initiated by MutSα. Genetic inactivation of *MLH1*or *PMS2* confers cancer predisposition, but mutations in *PMS1* do not. Involvement of *MLH3* defects in tumor development is uncertain.

Chapter # 12 Central Dogma of Life

The ‘Central Dogma’ is the process by which the instructions in DNA are converted into a functional product. It was first proposed in 1958 by Francis Crick, discoverer of the structure of DNA.

The central dogma of molecular biology explains the flow of genetic information, from DNA to RNA, to make a functional product, a protein[?]

The central dogma suggests that DNA contains the information needed to make all of our proteins, and that RNA is a messenger that carries this information to the ribosomes[?].

The ribosomes serve as factories in the cell where the information is ‘translated’ from a code into the functional product.

The process by which the DNA instructions are converted into the functional product is called gene expression[?].

Gene expression has two key stages - transcription[?] and translation[?].

In transcription, the information in the DNA of every cell is converted into small, portable RNA messages.

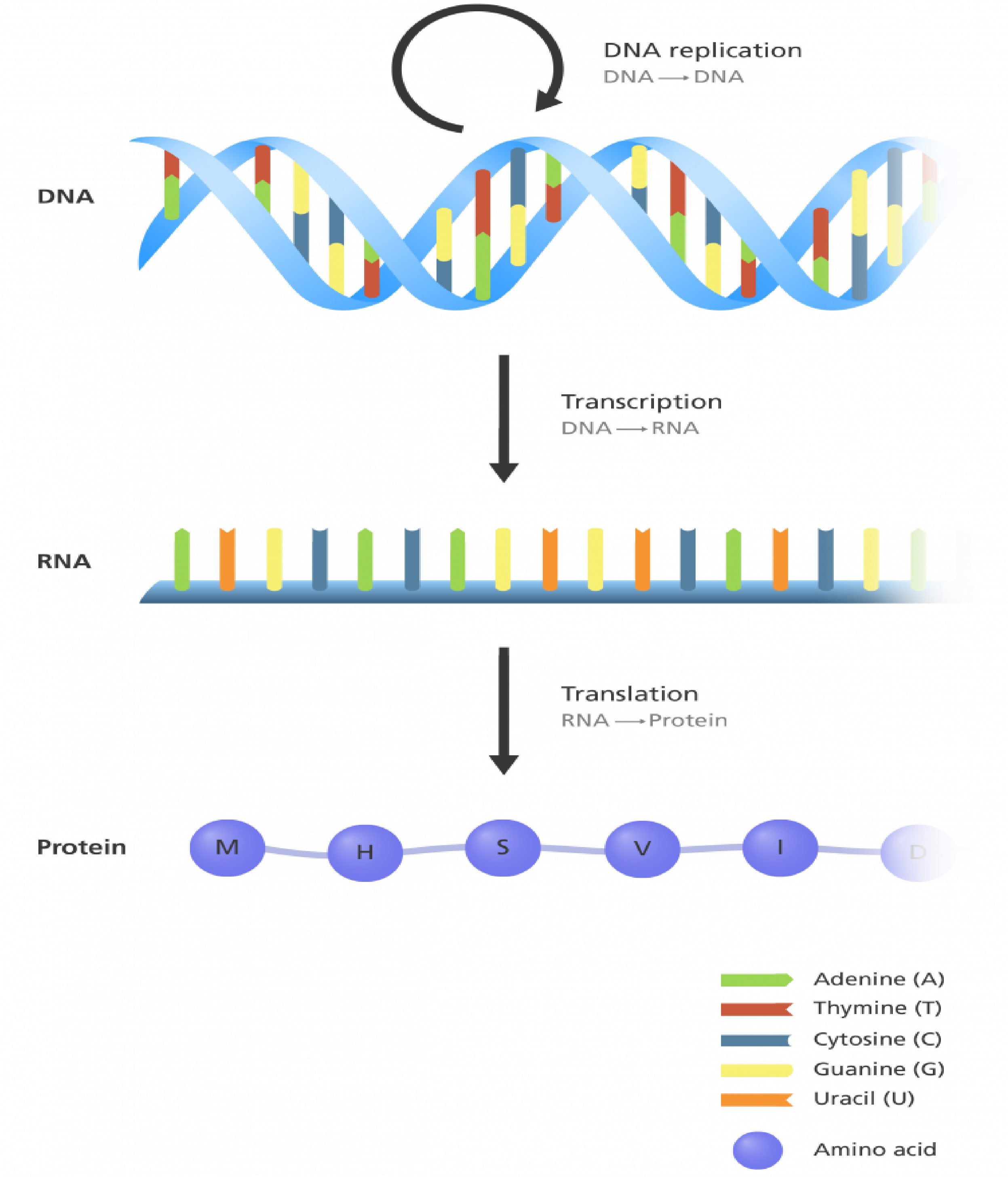
During translation, these messages travel from where the DNA is in the cell nucleus to the ribosomes where they are ‘read’ to make specific proteins.

The central dogma states that the pattern of information that occurs most frequently in our cells is:

From existing DNA to make new DNA (DNA replication[?])

From DNA to make new RNA (transcription)

From RNA to make new proteins (translation).



Reverse transcription is the transfer of information from RNA to make new DNA, this occurs in the case of retroviruses, such as HIV[?]. It is the process by which the genetic information from RNA is assembled into new DNA.

The central dogma has also been described as "DNA makes RNA and RNA makes protein,"[3] a positive statement which was originally termed the sequence hypothesis by Crick. However, this simplification does not make it clear that the central dogma as stated by Crick does not preclude the reverse flow of information from RNA to DNA, only ruling out the flow from protein to RNA or DNA. Crick's use of the word dogma was unconventional, and has been controversial.

The dogma is a framework for understanding the transfer of sequence information between information-carrying biopolymers, in the most common or general case, in living organisms. There are 3 major classes of such biopolymers: DNA and RNA (both nucleic acids), and protein. There are 3×3 = 9 conceivable direct transfers of information that can occur between these. The dogma classes these into 3 groups of 3: 3 general transfers (believed to occur normally in most cells), 3 special transfers (known to occur, but only under specific conditions in case of some viruses or in a laboratory), and 3 unknown transfers (believed never to occur). The general transfers describe the normal flow of biological information: DNA can be copied to DNA (DNA replication), DNA information can be copied into mRNA (transcription), and proteins can be synthesized using the information in mRNA as a template (translation).

DNA replications

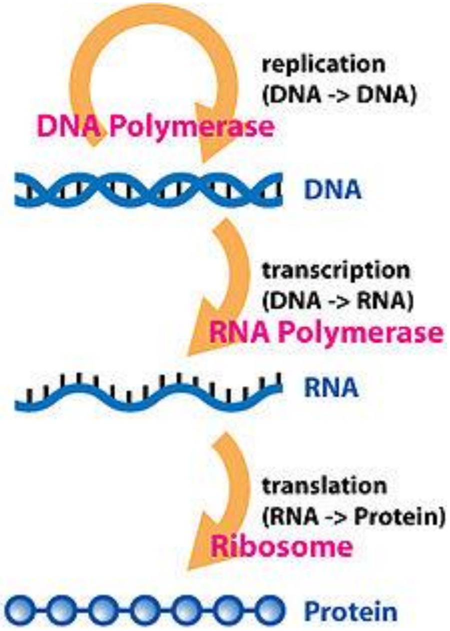
In the sense that DNA replication must occur if genetic material is to be provided for the progeny of any cell, whether somatic or reproductive, the copying from DNA to DNA arguably is the fundamental step in the central dogma. A complex group of proteins called the replisome performs the replication of the information from the parent strand to the complementary daughter strand.

The replisome comprises:

a helicase that unwinds the superhelix as well as the double-stranded DNA helix to create a replication fork SSB protein that binds open the double-stranded DNA to prevent it from reassociating RNA primase that adds a complementary RNA primer to each template strand as a starting point for replication DNA polymerase III that reads the existing template chain from its 3' end to its 5' end and adds new complementary nucleotides from the 5' end to the 3' end of the daughter chain DNA polymerase I that removes the RNA primers and replaces them with DNA.

DNA ligase that joins the two Okazaki fragments with phosphodiester bonds to produce a continuous chain. This process typically takes place during S phase of the cell cycle.

Transcription



Transcription is the process by which the information contained in a section of DNA is replicated in the form of a newly assembled piece of messenger RNA (mRNA). Enzymes facilitating the process include RNA polymerase and transcription factors. In eukaryotic cells the primary transcript is (pre-mRNA). Pre-mRNA must be processed for translation to proceed. Processing includes the addition of a 5' cap and a poly-A tail to the pre-mRNA chain, followed by splicing. Alternative splicing occurs when appropriate, increasing the diversity of the proteins that any single mRNA can produce. The product of the entire transcription process that began with the production of the pre-mRNA chain, is a mature mRNA chain.

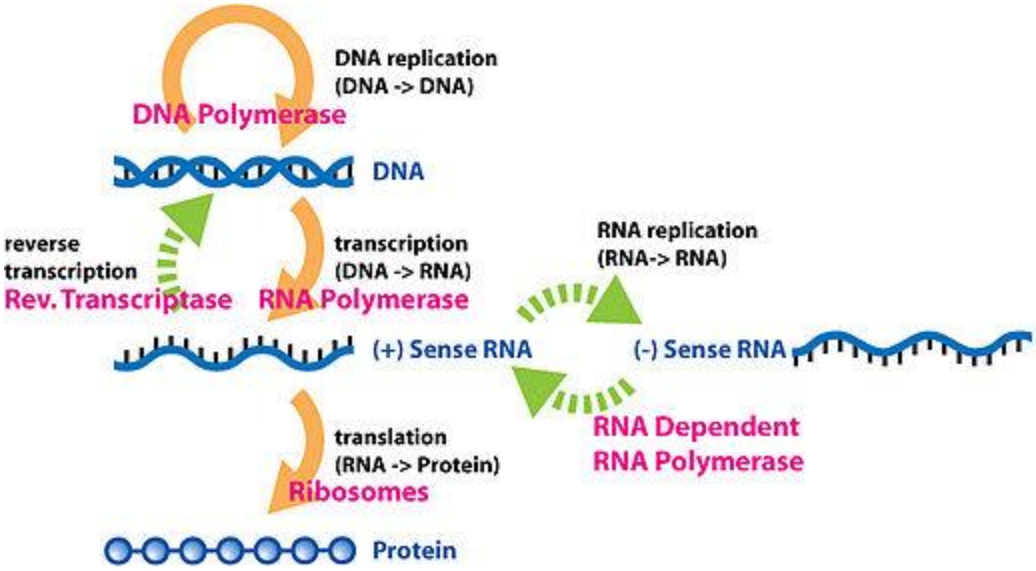
Translation

The mature mRNA finds its way to a ribosome, where it gets translated. In prokaryotic cells, which have no nuclear compartment, the processes of transcription and translation may be linked together without clear separation. In eukaryotic cells, the site of transcription (the cell nucleus) is usually separated from the site of translation (the cytoplasm), so the mRNA must be transported out of the nucleus into the cytoplasm, where it can be bound by ribosomes. The ribosome reads the mRNA triplet codons, usually beginning with an AUG (adenine−uracil−guanine), or initiator methionine codon downstream of the ribosome binding site. Complexes of initiation factors and elongation factors bring aminoacylated transfer RNAs (tRNAs) into the ribosome-mRNA complex, matching the codon in the mRNA to the anti-codon on the tRNA. Each tRNA bears the appropriate amino acid residue to add to the polypeptide chain being synthesised. As the amino acids get linked into the growing peptide chain, the chain begins folding into the correct conformation. Translation ends with a stop codon which may be a UAA, UGA, or UAG triplet.

The mRNA does not contain all the information for specifying the nature of the mature protein. The nascent polypeptide chain released from the ribosome commonly requires additional processing before the final product emerges. For one thing, the correct folding process is complex and vitally important. For most proteins it requires other chaperone proteins to control the form of the product. Some proteins then excise internal segments from their own peptide chains, splicing the free ends that border the gap; in such processes the inside "discarded" sections are called inteins. Other proteins must be split into multiple sections without splicing. Some polypeptide chains need to be cross-linked, and others must be attached to cofactors such as haem (heme) before they become functional.

Special transfers of biological sequential information

Reverse transcription



Unusual flow of information highlighted in green

Reverse transcription is the transfer of information from RNA to DNA (the reverse of normal transcription). This is known to occur in the case of retroviruses, such as HIV, as well as in eukaryotes, in the case of retrotransposons and telomere synthesis. It is the process by which genetic information from RNA gets transcribed into new DNA.

RNA replication

RNA replication is the copying of one RNA to another. Many viruses replicate this way. The enzymes that copy RNA to new RNA, called RNA-dependent RNA polymerases, are also found in many eukaryotes where they are involved in RNA silencing.

RNA editing, in which an RNA sequence is altered by a complex of proteins and a "guide RNA", could also be seen as an RNA-to-RNA transfer.

Direct translation from DNA to protein

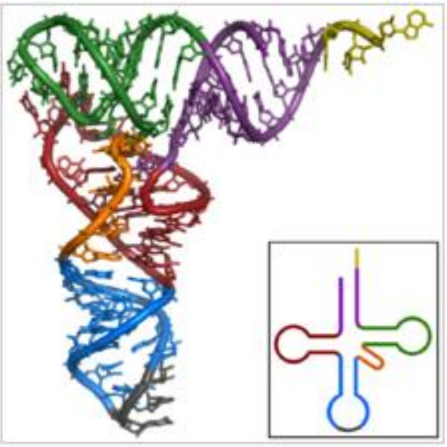
Direct translation from DNA to protein has been demonstrated in a cell-free system (i.e. in a test tube), using extracts from E. coli that contained ribosomes, but not intact cells. These cell fragments could synthesize proteins from single-stranded DNA templates isolated from other organisms (e,g., mouse or toad), and neomycin was found to enhance this effect. However, it was unclear whether this mechanism of translation corresponded specifically to the genetic code.

tRNA and genetic code:

Transfer RNA, or tRNA, is a member of a nucleic acid family called ribonucleic acids. RNA molecules are comprised of nucleotides, which are small building blocks for both RNA and DNA. tRNA has a very specific purpose: to bring protein subunits, known as amino acids, to the ribosome where proteins are constructed.

One of the discoverers of DNA, Francis Crick, first suggested the existence of tRNA. At the time, scientists knew that genetic information was kept in the nucleus as DNA and that DNA carried the instructions on how to make proteins. DNA doesn't leave the nucleus though, so our cells make a copy of the DNA called messenger RNA, or mRNA.

mRNA leaves the nucleus and is bound by ribosomes, the molecular machines that act as the factory that makes proteins. Scientists understood that while DNA and RNA have almost the same alphabet, proteins are very different. Francis Crick proposed that there must be a small molecule capable of translating mRNA into proteins. Other scientists proved his theory with the discovery of tRNA.



Tertiary structure of tRNA. *CCA tail* in yellow, *Acceptor stem* in purple, *Variable loop* in orange, *D arm* in red, *Anticodon arm* in blue with *Anticodon* in black, *T arm* in green.

The structure of tRNA

Function of tRNA

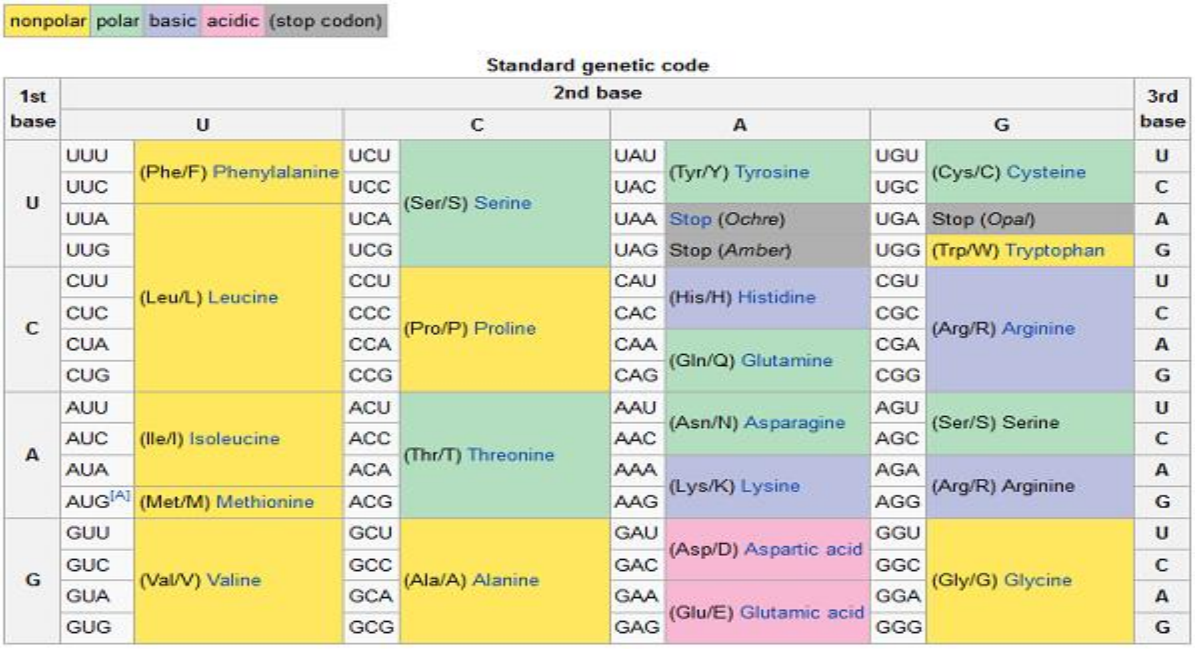
The job of tRNA is to read the message of nucleic acids, or nucleotides, and translate it into proteins, or amino acids. The process of making a protein from an mRNA template is called translation.

How does tRNA read the mRNA? It reads the mRNA in three-letter nucleotide sequences called codons. Each individual codon corresponds to an amino acid. There are four nucleotides in mRNA. There is one tRNA molecule for each and every codon.

Interestingly, there are only 21 amino acids. This brings up the idea that our genetic code is redundant. That is, we have 64 codons but only 21 amino acids. How do we resolve this? More than one codon can specify for an amino acid.

This table (Figure 2) shows all the combinations of nucleic acids, or codons, as well as which amino acid is specified by which codon. As you can see, not every amino acid has four codons. In fact, methionine only has one.

Notice, however, that each codon has only one corresponding amino acid. Thus we say that the genetic code is redundant, but not ambiguous. For example, the codons GUU, GUC, GUA, and GUG all code for Valine (redundancy), and none of them specify any other amino acid (no ambiguity).



nonpolar	polar	basic	acidic	(stop codon)

Standard genetic code

1st base	2nd base								3rd base
	U		C		A		G		
U	UUU	(Phe/F) Phenylalanine	UCU	(Ser/S) Serine	UAU	(Tyr/Y) Tyrosine	UGU	(Cys/C) Cysteine	U
	UUC		UCC		UAC		UGC		C
	UUA	(Leu/L) Leucine	UCA		UAA	Stop (*Ochre*)	UGA	Stop (*Opal*)	A
	UUG		UCG		UAG	Stop (*Amber*)	UGG	(Trp/W) Tryptophan	G
C	CUU		CCU	(Pro/P) Proline	CAU	(His/H) Histidine	CGU	(Arg/R) Arginine	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	(Gln/Q) Glutamine	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	(Ile/I) Isoleucine	ACU	(Thr/T) Threonine	AAU	(Asn/N) Asparagine	AGU	(Ser/S) Serine	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	(Lys/K) Lysine	AGA	(Arg/R) Arginine	A
	AUG[A]	(Met/M) Methionine	ACG		AAG		AGG		G
G	GUU	(Val/V) Valine	GCU	(Ala/A) Alanine	GAU	(Asp/D) Aspartic acid	GGU	(Gly/G) Glycine	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	(Glu/E) Glutamic acid	GGA		A
	GUG		GCG		GAG		GGG		G

The table of Amino Acids and Codons

So, we now know that the job of tRNA is to bring an amino acid to the ribosome. We also know that each codon has its own tRNA and that each tRNA has its own amino acid attached to it. Further, we know that the job of tRNA is to transport amino acids to the ribosome for protein production.

Chapter # 13 TRANSCRIPTION

RNA Polymerases:

RNA polymerase is an enzyme that is responsible for copying a DNA sequence into an RNA sequence, duyring the process of transcription. As complex molecule composed of protein subunits, RNA polymerase controls the process of transcription, during which the information stored in a molecule of DNA is copied into a new molecule of messenger RNA.

RNA polymerases have been found in all species, but the number and composition of these proteins vary across taxa. For instance, bacteria contain a single type of RNA polymerase, while eukaryotes (multicellular organisms and yeasts) contain three distinct types. In spite of these differences, there are striking similarities among transcriptional mechanisms. For example, all species require a mechanism by which transcription can be regulated in order to achieve spatial and temporal changes in gene expression.

BACTERIAL TRANSCRIPTION

Bacterial transcription is the process in which messenger RNA transcripts of genetic material in bacteria are produced, to be translated for the production of proteins. Bacterial transcription occurs in the cytoplasm alongside translation. Unlike in eukaryotes, bacterial 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 bacteria genetic material is not enclosed in a membrane-enclosed nucleus and has access to ribosomes in the cytoplasm.

Transcription is known to be controlled by a variety of regulators in bacteria. Many of these transcription factors are homodimers containing helix-turn-helix DNA-binding motifs

Initiation

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.

The -35 region and the -10 ("Pribnow box") region comprise the core 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. At this stage, the DNA is double-stranded ("closed"). This holoenzyme/wound-DNA structure is referred to as the *closed complex*.

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*.

The RNA polymerase transcribes the DNA (the beta subunit initiates the synthesis), 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.

The σ-factor eventually dissociates from the core enzyme, and elongation proceeds.

Elongation

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 disfavoured at the initiation site.

Termination

Two termination mechanisms are well known:

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.

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 utilisation 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 ρ-factor reaches the RNAP, it causes RNAP to dissociate from the DNA, terminating transcription

TRANSCRIPTION IN EUKARYOTES

Eukaryotic transcription is the elaborate process that eukaryotic cells use to copy genetic information stored in DNA into units of RNA replica. Gene transcription occurs in both eukaryotic and prokaryotic cells. Unlike prokaryotic RNA polymerase that initiates the transcription of all different types of RNA, RNA polymerase in eukaryotes (including humans) comes in three variations, each encoding a different type of gene. A eukaryotic cell has a nucleus that separates the processes of transcription and translation. Eukaryotic transcription occurs within the nucleus where DNA is packaged into nucleosomes and higher order chromatin structures. The complexity of the eukaryotic genome necessitates a great variety and complexity of gene expression control.

Overview

Transcription is the process of copying genetic information stored in a DNA strand into a transportable complementary strand of RNA. Eukaryotic transcription takes place in the nucleus of the cell and proceeds in three sequential stages: initiation, elongation, and termination. The transcriptional machinery that catalyzes this complex reaction has at its core three multi-subunit RNA polymerases. RNA polymerase I is responsible for transcribing RNA that codes for genes that become structural components of the ribosome.

Protein coding genes are transcribed into messenger RNAs (mRNAs) that carry the information from DNA to the site of protein synthesis. Although mRNAs possess great diversity, they are not the most abundant RNA species made in the cell. The so called non-coding RNAs account for the large majority of the transcriptional output of a cell. These non-coding RNAs perform a variety of important cellular functions.

RNA Polymerase

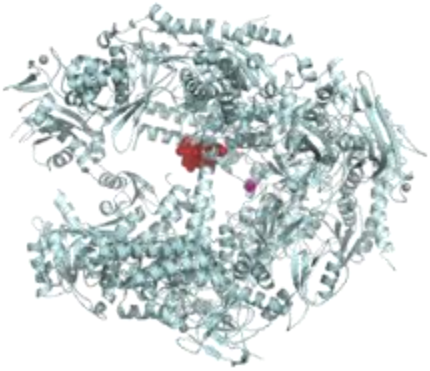
Eukaryotes have three nuclear RNA polymerases, each with distinct roles and properties

Name	**Location**	**Product**
RNA Polymerase I (Pol I, Pol A)	nucleolus	larger ribosomal RNA (rRNA) (28S, 18S, 5.8S)
RNA Polymerase II (Pol II, Pol B)	nucleus	messenger RNA (mRNA), most small nuclear RNAs (snRNAs), small interfering RNA (siRNAs) and micro RNA (miRNA).
RNA Polymerase III (Pol III, Pol C)	nucleus (and possibly the nucleolus-nucleoplasm interface)	transfer RNA (tRNA), other small RNAs (including the small 5S ribosomal RNA (5s rRNA), snRNA U6, signal recognition particle RNA (SRP RNA) and other stable short RNAs

RNA polymerase I (Pol I) catalyses the transcription of all rRNA genes except 5S. These rRNA genes are organised into a single transcriptional unit and are transcribed into a continuous transcript. This precursor is then processed into three rRNAs: 18S, 5.8S, and 28S. The transcription of rRNA genes takes place in a specialised structure of the nucleus called the nucleolus, where the transcribed rRNAs are combined with proteins to form ribosomes.

RNA polymerase II (Pol II) is responsible for the transcription of all mRNAs, some snRNAs, siRNAs, and all miRNAs. Many Pol II transcripts exist transiently as single strand precursor RNAs (pre-RNAs) that are further processed to generate mature RNAs. For example, precursor mRNAs (pre-mRNAs)are extensively processed before exiting into the cytoplasm through the nuclear pore for protein translation.

RNA polymerase III (Pol III) transcribes small non-coding RNAs, including tRNAs, 5S rRNA, U6 snRNA, SRP RNA, and other stable short RNAs such as ribonuclease P RNA.



Structure of eukaryotic RNA polymerase II (light blue) in complex with α-amanitin (red), a strong poison found in death cap mushrooms that targets this vital enzyme

RNA Polymerases I, II, and III contain 14, 12, and 17 subunits, respectively. All three eukaryotic polymerases have five core subunits that exhibit homology with the β, β', α^{I}, α^{II}, and ω subunits of E. coli RNA polymerase. An identical ω-like subunit (RBP6) is used by all three eukaryotic polymerases, while the same α-like subunits are used by Pol I and III. The three eukaryotic polymerases share four other common subunits among themselves. The remaining subunits are unique to each RNA polymerase. The additional subunits found in Pol I and Pol III relative to Pol II, are homologous to Pol II transcription factors.

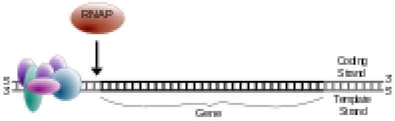
Crystal structures of RNA polymerases I and II provide an opportunity to understand the interactions among the subunits and the molecular mechanism of eukaryotic transcription in atomic detail.

The carboxyl terminal domain (CTD) of RPB1, the largest subunit of RNA polymerase II, plays an important role in bringing together the machinery necessary for the synthesis and processing of Pol II transcripts. Long and structurally disordered, the CTD contains multiple repeats of heptapeptide sequence YSPTSPS that are subject to phosphorylation and other posttranslational modifications during the transcription cycle. These modifications and their regulation constitute the operational code for the CTD to control transcription initiation, elongation and termination and to couple transcription and RNA processing.

Initiation

The initiation of gene transcription in eukaryotes occurs in specific steps. First, an RNA polymerase along with general transcription factors binds to the promoter region of the gene to form a closed complex called the preinitiation complex. The subsequent transition of the complex from the closed state to the open state results in the melting or separation of the two DNA strands and the positioning of the template strand to the active site of the RNA polymerase. Without the need of a primer, RNA polymerase can initiate the synthesis of a new RNA chain using the template DNA strand to guide ribonucleotide selection and polymerization chemistry.

However, many of the initiated syntheses are aborted before the transcripts reach a significant length (~10 nucleotides). During these abortive cycles, the polymerase keeps making and releasing short transcripts until it is able to produce a transcript that surpasses ten nucleotides in length. Once this threshold is attained, RNA polymerase escapes the promoter and transcription proceeds to the elongation phase.



Here is a diagram of the attachment of RNA polymerase II to the de-helicized DNA.

Eukaryotic promoters and general transcription factors

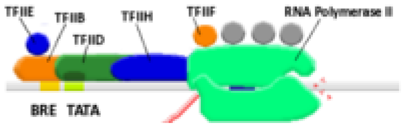
Pol II-transcribed genes contain a region in the immediate vicinity of the transcription start site (TSS) that binds and positions the preinitiation complex. This region is called the core promoter because of its essential role in transcription initiation. Different classes of sequence elements are found in the promoters. For example, the TATA box is the highly conserved DNA recognition sequence for the TATA box binding protein, TBP, whose binding initiates transcription complex assembly at many genes.

Eukaryotic genes also contain regulatory sequences beyond the core promoter. These cis-acting control elements bind transcriptional activators or repressors to increase or decrease transcription from the core promoter. Well-characterized regulatory elements include enhancers, silencers, and insulators. These regulatory sequences can be spread over a large genomic distance, sometimes located hundreds of kilobases from the core promoters.

General transcription factors are a group of proteins involved in transcription initiation and regulation. These factors typically have DNA-binding domains that bind specific sequence elements of the core promoter and help recruit RNA polymerase to the transcriptional start site. General transcription factors for RNA polymerase II include TFIID, TFIIA, TFIIB, TFIIF, TFIIE, and TFIIH.

Assembly of preinitiation complex

To prepare for transcription, a complete set of general transcription factors and RNA polymerase need to be assembled at the core promoter to form the ~2 million dalton preinitiation complex. For example, for promoters that contain a TATA box near the TSS, the recognition of TATA box by the TBP subunit of TFIID initiates the assembly of a transcription complex. The next proteins to enter are TFIIA and TFIIB, which stabilize the DNA-TFIID complex and recruit Pol II in association with TFIIF and additional transcription factors. TFIIF serves as the bridge between the TATA-bound TBP and polymerase. One of the last transcription factors to be recruited to the preinitiation complex is TFIIH, which plays an important role in promoter melting and escape.



The diagram describes the eukaryotic pre-initiation complex which includes the general transcription factors and RNA Polymerase II. Credit: ArneLH.

Promoter melting and open complex formation

For pol II-transcribed genes, and unlike bacterial RNA polymerase, promoter melting requires hydrolysis of ATP and is mediated by TFIIH. TFIIH is a ten-subunit protein, including both ATPase and protein kinase activities. While the upstream promoter DNA is held in a fixed position by TFIID, TFIIH pulls downstream double-stranded DNA into the cleft of the polymerase, driving the separation of DNA strands and the transition of the preinitiation complex from the closed to open state. TFIIB aids in open complex formation by binding the melted DNA and stabilizing the transcription bubble.

Abortive initiation

Once the initiation complex is open, the first ribonucleotide is brought into the active site to initiate the polymerization reaction in the absence of a primer. This generates a nascent RNA chain that forms a hetero-duplex with the template DNA strand. However, before entering the elongation phase, polymerase may terminate prematurely and release a short, truncated transcript. This process is called abortive initiation. Many cycles of abortive initiation may occur before the transcript grows to sufficient length to promote polymerase escape from the promoter. Throughout abortive initiation cycles, RNA polymerase remains bound to the promoter and pulls downstream DNA into its catalytic cleft in a scrunching-kind of motion.

Promoter escape

When a transcript attains the threshold length of ten nucleotides, it enters the RNA exit channel. The polymerase breaks its interactions with the promoter elements and any regulatory proteins associated with the initiation complex that it no longer needs. Promoter escape in eukaryotes requires ATP hydrolysis and, in the case of Pol II-phosphorylation of the CTD. Meanwhile, the transcription bubble collapses down to 12-14 nucleotides, providing kinetic energy required for the escape.

Elongation

After escaping the promoter and shedding most of the transcription factors for initiation, the polymerase acquires new factors for the next phase of transcription: elongation.[21][22] Transcription elongation is a processive process. Double stranded DNA that enters from the front of the enzyme is unzipped to avail the template strand for RNA synthesis. For every DNA base pair separated by the advancing polymerase, one hybrid RNA:DNA base pair is immediately formed. DNA strands and nascent RNA chain exit from separate channels; the two DNA strands reunite at the trailing end of the transcription bubble while the single strand RNA emerges alone.

Elongation factors

Among the proteins recruited to polymerase are elongation factors, thus called because they stimulate transcription elongation. There are different classes of elongation factors. Some factors can increase the overall rate of transcribing, some can help the polymerase through transient pausing sites, and some can assist the polymerase to transcribe through chromatin. One of the elongation factors, P-TEFb, is particularly important. P-TEFb phosphorylates the second residue (Ser-2) of the CTD repeats (YSPTSPS) of the bound Pol II. P-TEFb also phosphorylates and activates SPT5 and TAT-SF1. SPT5 is a universal transcription factor that helps recruit 5'-capping enzyme to Pol II with a CTD phosphorylated at Ser-5. TAF-SF1 recruits components of the RNA splicing machinery to the Ser-2 phosphorylated CTD. P-TEFb also helps suppress transient pausing of polymerase when it encounters certain sequences immediately following initiation.

Transcription fidelity

Transcription fidelity is achieved through multiple mechanisms. RNA polymerases select correct nucleoside triphosphate (NTP) substrate to prevent transcription errors. Only the NTP which correctly base pairs with the coding base in the DNA is admitted to the active center.

RNA polymerase performs two known proof reading functions to detect and remove misincorporated nucleotides: pyrophosphorylytic editing and hydrolytic editing. The former removes the incorrectly inserted ribonucleotide by a simple reversal of the polymerization reaction, while the latter involves backtracking of the polymerase and cleaving of a segment of error-containing RNA product. Elongation factor TFIIS stimulates an inherent ribonuclease activity in the polymerase, allowing the removal of misincorporated bases through limited local RNA degradation. Note that all reactions (phosphodiester bond synthesis, pyrophosphorolysis, phosphodiester bond hydrolysis) are performed by RNA polymerase by using a single active center.

Pausing, poising, and backtracking

Transcription elongation is not a smooth ride along the DNA railway. For proofreading, the polymerase is made to back-up, erase some of the RNA it has already made and have another go at transcription. In general, RNA polymerase does not transcribe through a gene at a constant pace. Rather it pauses periodically at certain sequences, sometimes for long periods of time before resuming transcription. In extreme cases, for example, when the polymerase encounters a damaged nucleotide, it comes to a complete halt. More often, an elongating polymerase is stalled near the promoter. Promoter-proximal pausing during early elongation is a commonly used mechanism for regulating genes poised to be expressed rapidly or in a coordinated fashion. Pausing is mediated by a complex called NELF (negative elongation factor) in collaboration with DSIF (DRB-sensitivity-inducing factor containing SPT4/SPT5). The blockage is released once the polymerase receives an activation signal, such as the phosphorylation of Ser-2 of CTD tail by P-TEFb. Other elongation factors such as ELL and TFIIS stimulate the rate of elongation by limiting the length of time that polymerase pauses.

RNA processing

Elongating polymerase is associated with a set of protein factors required for various types of RNA processing. mRNA is capped as soon as it emerges from the RNA-exit channel of the polymerase. After capping, dephosphorylation of Ser-5 within the CTD repeats may be responsible for dissociation of the capping machinery. Further phosphorylation of Ser-2 causes recruitment of the RNA splicing machinery that catalyzes the removal of non-coding introns to generate mature mRNA. Alternative splicing expands the protein complements in eukaryotes. Just as with 5’-capping and splicing, the CTD tail is involved in recruiting enzymes responsible for 3’-polyadenylation, the final RNA processing event that is coupled with the termination of transcription.

Termination

The last stage of transcription is termination, which leads to the dissociation of the complete transcript and the release of RNA polymerase from the template DNA.The process differs for each of the three RNA polymerases. The mechanism of termination is the least understood of the three transcription stages.

Factor-dependent termination

The termination of transcription of pre-rRNA genes by polymerase Pol I is performed by a system that needs a specific transcription termination factor. The mechanism used bears some resemblance to the rho-dependent termination in prokaryotes. Eukaryotic cells contain hundreds of ribosomal DNA repeats, sometimes distributed over multiple chromosomes. Termination of transcription occurs in the ribosomal intergenic spacer region that contains several transcription termination sites upstream of a Pol I pausing site. Through a yet unknown mechanism, the 3’-end of the transcript is cleaved, generating a large primary rRNA molecule that is further processed into the mature 18S, 5.8S and 28S rRNAs.

As Pol II reaches the end of a gene, two protein complexes carried by the CTD, CPSF (cleavage and polyadenylation specificity factor) and CSTF (cleavage stimulation factor), recognize the poly-A signal in the transcribed RNA. Poly-A-bound CPSF and CSTF recruit other proteins to carry out RNA cleavage and then polyadenylation. Poly-A polymerase adds approximately 200 adenines to the cleaved 3’ end of the RNA without a template. The long poly-A tail is unique to transcripts made by Pol II.

In the process of terminating transcription by Pol I and Pol II, the elongation complex does not dissolve immediately after the RNA is cleaved. The polymerase continues to move along the template, generating a second RNA molecule associated with the elongation complex. Two models have been proposed to explain how termination is achieved at last. The allosteric model states that when transcription proceeds through the termination sequence, it causes disassembly of elongation factors and/or an assembly of termination factors that cause conformational changes of the elongation complex. The torpedo model suggests that a 5' to 3' exonuclease degrades the second RNA as it emerges from the elongation complex. Polymerase is released as the highly processive exonuclease overtakes it. It is proposed that an emerging view will express a merge of these two models.

Factor-independent termination

RNA polymerase III can terminate transcription efficiently without involvement of additional factors. The Pol III termination signal consists of a stretch of thymines (on the nontemplate strand) located within 40bp downstream from the 3' end of mature RNAs. The poly-T termination signal pauses Pol III and causes it to back track to the nearest RNA hairpin to become a “dead-end” complex. Consistent with the allosteric mechanism of termination, the RNA hairpin allosterically opens Pol III and causes the elongation complex to disintegrate. The extensive structure embedded in the Pol III-transcript thus is responsible for the factor-independent release of Pol III at the end of a gene. RNA-duplex-dependent termination is an ancient mechanism that dates back to the last universal common ancestor.

Eukaryotic transcriptional control

The regulation of gene expression in eukaryotes is achieved through the interaction of several levels of control that acts both locally to turn on or off individual genes in response to a specific cellular need and globally to maintain a chromatin-wide gene expression pattern that shapes cell identity. Because eukaryotic genome is wrapped around histones to form nucelosomes and higher-order chromatin structures, the substrates for transcriptional machinery are in general partially concealed. Without regulatory proteins, many genes are expressed at low level or not expressed at all. Transcription requires displacement of the positioned nucleosomes to enable the transcriptional machinery to gain access of the DNA.

All steps in the transcription are subject to some degree of regulation. Transcription initiation in particular is the primary level at which gene expression is regulated. Targeting the rate-limiting initial step is the most efficient in terms of energy costs for the cell. Transcription initiation is regulated by cis-acting elements (enhancers, silencers, isolators) within the regulatory regions of the DNA, and sequence-specific trans-acting factors that act as activators or repressors. Gene transcription can also be regulated post-initiation by targeting the movement of the elongating polymerase.

Global control and epigenetic regulation

The eukaryotic genome is organized into a compact chromatin structure that allows only regulated access to DNA. The chromatin structure can be globally “open” and more transcriptionally permissive or globally “condensed” and transcriptionally inactive. The former (euchromatin) is lightly packed and rich in genes under active transcription. The latter (heterochromatin) includes gene-poor regions such as telomeres and centromeres but also regions with normal gene density but transcriptionally silenced. Transcription can be silenced by histone modification (deaceltylation and methylation), RNA interference, and/or DNA methylation.

The gene expression patterns that define cell identity have to be inherited through cell division. This process is called epigenetic regulation. DNA methylation is reliably inherited through the action of maintenance methylases that modify the nascent DNA strand generated by replication. In mammalian cells, DNA methylation is the primary marker of transcriptionally silenced regions. Specialized proteins can recognize the marker and recruit histone deacetylases and methylases to re-establish the silencing. Nucleosome histone modifications could also be inherited during cell division, however, it is not clear whether it can work independently without the direction by DNA methylation.

Gene-specific activation

The two main tasks of transcription initiation are to provide RNA polymerase with an access to the promoter and to assemble general transcription factors with polymerase into a transcription initiation complex. Diverse mechanisms of initiating transcription by overriding inhibitory signals at the gene promoter have been identified. Eukaryotic genes have acquired extensive regulatory sequences that encompass a large number of regulator-binding sites and spread overall kilobases (sometimes hundreds of kilobases) from the promoter–-both upstream and downstream. The regulator binding sites are often clustered together into units called enhancers. Enhancers can facilitate highly cooperative action of several transcription factors (which constitute enhanceosomes). Remote enhancers allow transcription regulation at a distance. Insulators situated between enhancers and promoters help define the genes that an enhancer can or cannot influence.

Eukaryotic transcriptional activators have separate DNA-binding and activating functions. Upon binding to its cis-element, an activator can recruit polymerase directly or recruit other factors needed by the transcriptional machinery. An activator can also recruit nucleosome modifiers that alter chromatin in the vicinity of the promoter and thereby help initiation. Multiple activators can work together, either by recruiting a common or two mutually dependent components of the transcriptional machinery, or by helping each other bind to their DNA sites. These interactions can synergize multiple signaling inputs and produce intricate transcriptional responses to address cellular needs.

Gene-specific repression

Eukaryotic transcription repressors share some of the mechanisms used by their prokaryotic counterparts. For example, by binding to a site on DNA that overlaps with the binding site of an activator, a repressor can inhibit binding of the activator. But more frequently, eukaryotic repressors inhibit the function of an activator by masking its activating domain, preventing its nuclear localization, promoting its degradation, or inactivating it through chemical modifications. Repressors can directly inhibit transcription initiation by binding to a site upstream of a promoter and interacting with the transcriptional machinery. Repressors can indirectly repress transcription by recruiting histone modifiers (deacetylases and methylases) or nucelosome remodeling enzymes that affect the accessibility of the DNA. Repressing histone and DNA modifications are also the basis of transcriptional silencing that can spread along the chromatin and switch off multiple genes.

Elongation and termination control

The elongation phase starts once assembly of the elongation complex has been completed, and progresses until a termination sequence is encountered. The post-initiation movement of RNA polymerase is the target of another class of important regulatory mechanisms. For example, the transcriptional activator Tat affects elongation rather than initiation during its regulation of HIV transcription. In fact, many eukaryotic genes are regulated by releasing a block to transcription elongation called promoter-proximal pausing. Pausing can influence chromatin structure at promoters to facilitate gene activity and lead to rapid or synchronous transcriptional responses when cells are exposed to an activation signal. Pausing is associated with the binding of two negative elongation factors, DSIF (SPT4/SPT5) and NELF, to the elongation complex. Other factors can also influence the stability and duration of the paused polymerase.[44] Pause release is triggered by the recruitment of the P-TEFb kinase.

Transcription termination has also emerged as an important area of transcriptional regulation. Termination is coupled with the efficient recycling of polymerase. The factors associated with transcription termination can also mediate gene looping and thereby determine the efficiency of re-initiation.

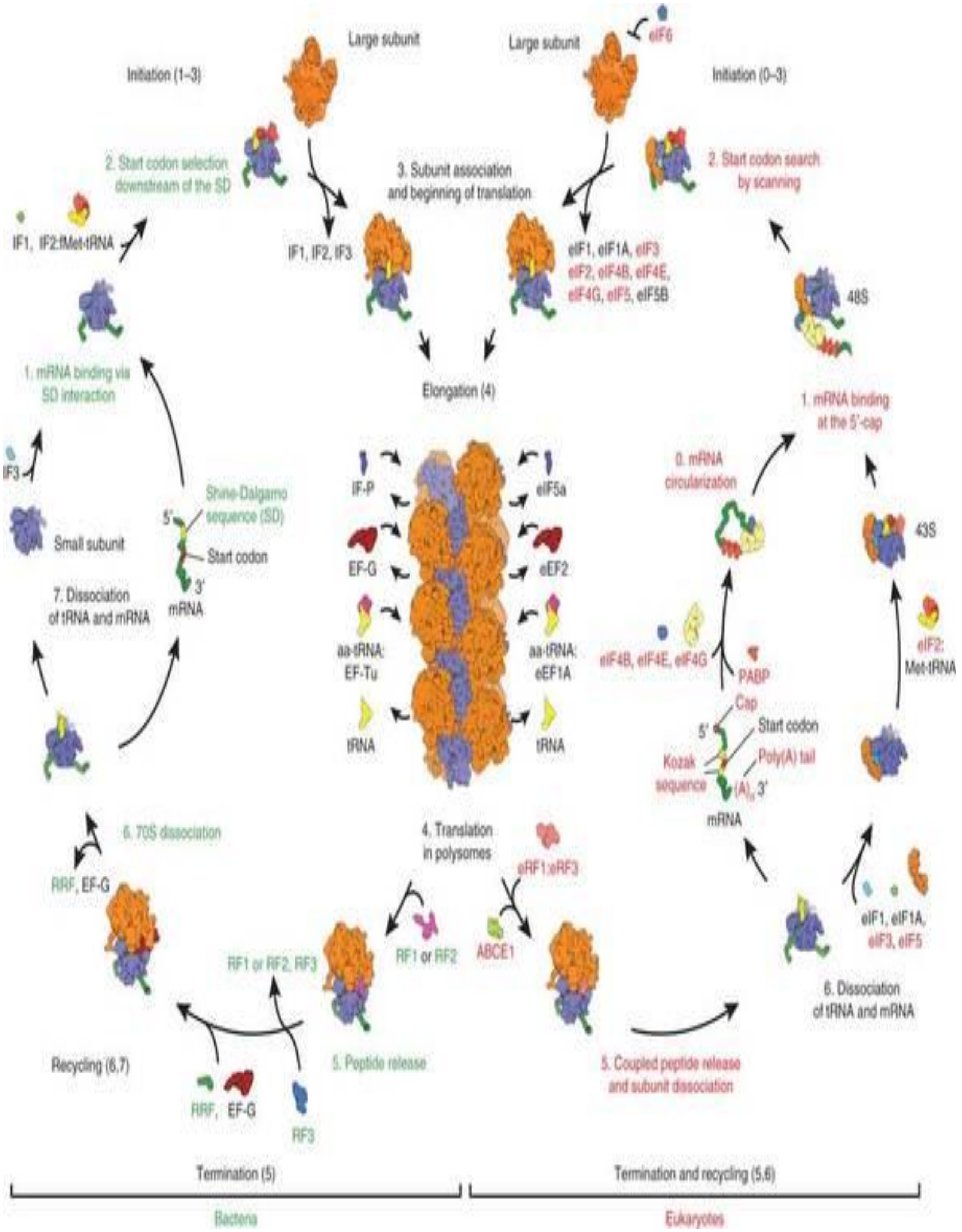
Transcription-coupled DNA repair

When transcription is arrested by the presence of a lesion in the transcribed strand of a gene, DNA repair proteins are recruited to the stalled RNA polymerase to initiate a process called transcription-coupled repair. Central to this process is the general transcription factor TFIIH that has ATPase activity. TFIIH causes a conformational change in the polymerase, to expose the transcription bubble trapped inside, in order for the DNA repair enzymes to gain access to the lesion. Thus, RNA polymerase serves as damage-sensing protein in the cell to target repair enzymes to genes that are being actively transcribed.

Comparisons between prokaryotic and eukaryotic transcription

Eukaryotic transcription is more complex than prokaryotic transcription. For instance, in eukaryotes the genetic material (DNA), and therefore transcription, is primarily localized to the nucleus, where it is separated from the cytoplasm (in which translation occurs) by the nuclear membrane. This allows for the temporal regulation of gene expression through the sequestration of the RNA in the nucleus, and allows for selective transport of mature RNAs to the cytoplasm. Bacteria do not have a distinct nucleus that separates DNA from ribosome and mRNA is translated into protein as soon as it is transcribed. The coupling between the two processes provides an important mechanism for prokaryotic gene regulation.

At the level of initiation, RNA polymerase in prokaryotes (bacteria in particular) binds strongly to the promoter region and initiates a high basal rate of transcription. No ATP hydrolysis is needed for the close-to-open transition, promoter melting is driven by binding reactions that favor the melted conformation. Chromatin greatly impedes transcription in eukaryotes. Assembly of large multi-protein preinitiation complex is required for promoter-specific initiation. Promoter melting in eukaryotes requires hydrolysis of ATP. As a result, eukaryotic RNA polymerases exhibit a low basal rate of transcription initiation.


Large subunit
Initiation (1–3)
2. Start codon selection downstream of the SD
IF1, IF2:fMet-tRNA
3. Subunit association and beginning of translation
IF1, IF2, IF3
Large subunit
eIF6
Initiation (0–3)
2. Start codon search by scanning
eIF1, eIF1A, eIF3
eIF2, eIF4B, eIF4E,
eIF4G, eIF5, eIF5B
48S
1. mRNA binding via SD interaction
Elongation (4)
1. mRNA binding at the 5'-cap
IF3
Shine-Dalgarno sequence (SD)
5'
Start codon
3'
mRNA
Small subunit
7. Dissociation of tRNA and mRNA
IF-P
EF-G
aa-tRNA: EF-Tu
tRNA
eIF5a
eEF2
aa-tRNA: eEF1A
tRNA
0. mRNA circularization
43S
eIF4B, eIF4E, eIF4G
PABP
eIF2: Met-tRNA
Cap
5'
Start codon
Kozak sequence
Poly(A) tail
(A)n 3'
mRNA
6. 70S dissociation
RRF, EF-G
4. Translation in polysomes
eRF1:eRF3
ABCE1
RF1 or RF2, RF3
RF1 or RF2
eIF1, eIF1A, eIF3, eIF5
6. Dissociation of tRNA and mRNA
Recycling (6,7)
5. Peptide release
RRF, EF-G
RF3
5. Coupled peptide release and subunit dissociation
Termination (5)
Termination and recycling (5,6)
Bacteria
Eukaryotes

CHAPTER # 14 RNA SPLICING

Pre-mRNA Splicing

Because eukaryotic pre-mRNAs are transcribed from intron containing genes, the sequences encoded by the intronic DNA must be removed from the primary transcript prior to the RNA's becoming biologically active. The process of intron removal is called RNA splicing, or pre-mRNA splicing**.** The intron-exon junctions (splice-sites) in the precursor mRNA (pre-mRNA) of eukaryotes are recognized by trans-acting factors (prokaryotes RNAs are mostly polycistronic). In pre-mRNA splicing the intronic sequences are excised and the exons are ligated to generate the spliced mRNA.

Group I introns occur in nuclear, mitochondrial and chloroplast rRNA genes, group II introns in mitochondrial and chloroplast mRNA genes.

Many of the group I and group II introns are self-splicing in that no additional protein factors are necessary for the intron to be efficiently and accurately excised and the strands reattached. "The nucleotide sequence of group II self-splicing introns is highly conserved, and hence these introns fold into an evolutionarily conserved three-dimensional structure, which can undergo a self-splicing reaction in the absence of any trans-acting factors.

In contrast, the nucleotide sequences and length of nuclear pre-mRNA introns is highly variable, except for the short conserved sequences at the 5´ and 3´ splice sites and the branch points. Therefore nuclear pre-mRNA splicing requires trans-acting factors, which interact with these short conserved sequences, and from which the catalytically active spliceosome is assembled.

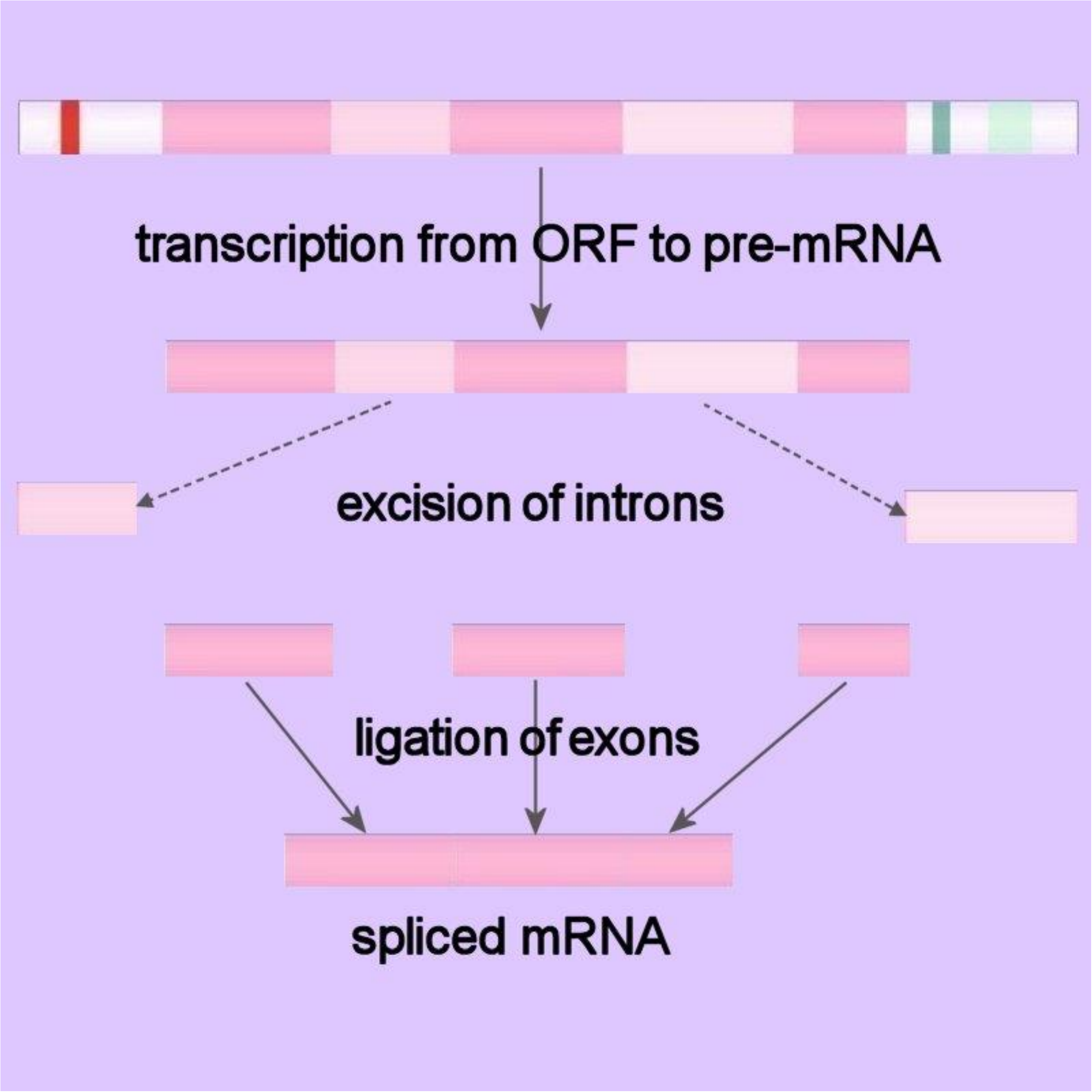
The conserved sequences are: 5' splice site = AGguragu; 3' splice site = yyyyyyy nagG (y= pyrimidine); branch site = ynyuray (r = purine, n = nucleotide)

Expressed differently, the highly conserved, consensus sequence for the **5'** donor splice site is (for RNA): (A or C)AG/GUAAGU. That is, most exons end with AG and introns begin with GU (GT for DNA). The highly conserved, consensus sequence for the **3'** acceptor splice site is (for RNA): (C/U)less than 10N(C/T)AG/G, where most introns end in AG after a long stretch of pyrimidines. The branch site within introns (area of lariat formation close to the acceptor site during splicing) has the consensus sequence UAUA**A**C. In most cases, U can be replaced by C and A can be replaced by G. However, the penultimate (bold) A residue is fully conserved (invariant).

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

Pre-mRNA processing takes place in the nucleus of eukaryotes, whereas lack of a nuclear membrane in prokaryotes permits initiation of translation while transcription is not yet complete.

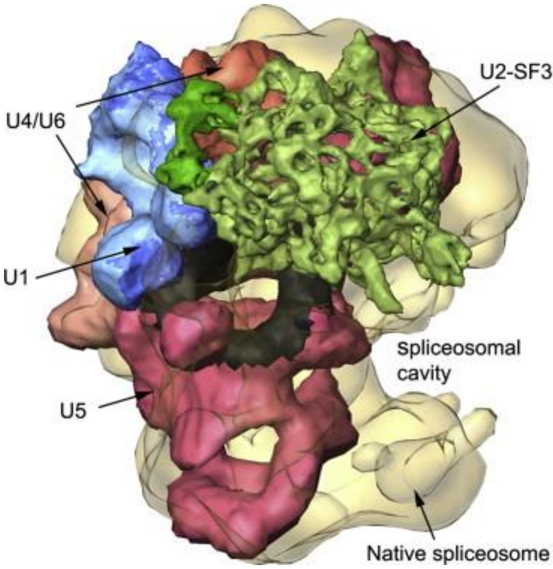
Pre-mRNA processing events include capping of the 5’ end on the pre-mRNA, pre-mRNA splicing to remove intronic sequences, and polyadenylation of the 3’ end of the pre-mRNA.



SPLICEOSOME MACHINERY

The spliceosome has been described as one of "the most complex macromolecular machines known," "composed of as many as 300 distinct proteins and five RNAs" (Nilsen, 2003). The animation above reveals this astonishing machine at work on the precursor mRNA, cutting out the non-coding introns and splicing together the protein-coding exons.

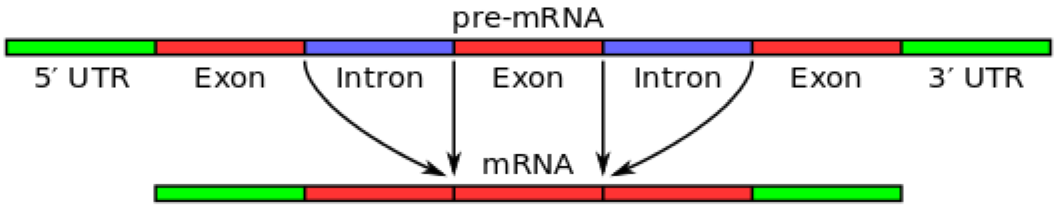
A spliceosome is a large and complex molecular machine found primarily within the splicing speckles of the cell nucleus of eukaryotic cells. The spliceosome is assembled from snRNAs and protein complexes. The spliceosome removes introns from a transcribed pre-mRNA, a kind of primary transcript. This process is generally referred to as splicing. Only eukaryotes have spliceosomes and metazoans have a second spliceosome, the minor spliceosome.



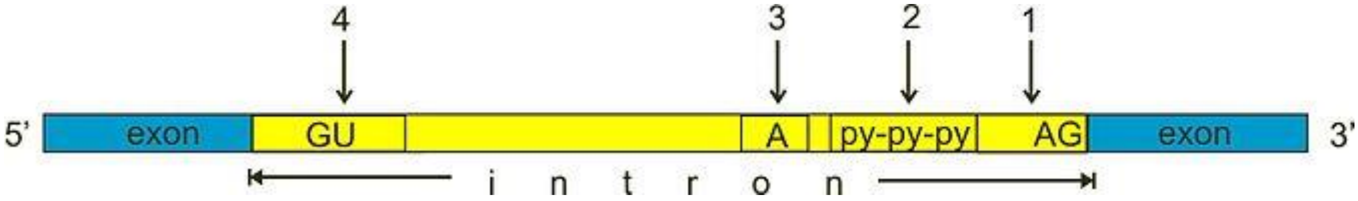
Introns (which, unlike exons, do not code for proteins) can be of considerable length in higher eukaryotes, even spanning many thousands of bases and sometimes comprising some 90% of the precursor mRNA. In contrast, lower eukaryotes such as yeast possess fewer and shorter introns, which are typically fewer than 300 bases in length. Since introns are the non-coding segments of genes, they are removed from the mRNA before it is translated into a protein. This is not to say, of course, that introns are without important function in the cell (as I discuss here).

Comprising the spliceosome, shown at right (excerpted from Frankenstein *et al*., 2012) are several small nuclear ribonucleoproteins (snRNPs) -- called U1, U2, U4, U5 and U6 -- each of which contains an RNA known as an snRNA (typically 100-300 nucleotides in length) -- and many other proteins that each contribute to the process of splicing by recognizing sequences in the mRNA or promoting rearrangements in spliceosome conformation. The spliceosome catalyzes a reaction that results in intron removal and the "gluing" together of the protein-coding exons.

RNA Splicing



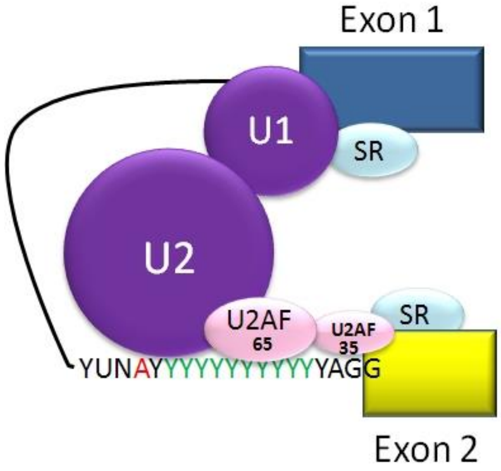
The first stage in RNA splicing is recognition by the spliceosome of splice sites between introns and exons. Key to this process are short sequence motifs. These include the 5' and 3' splice sites (typically a GU and AG sequence respectively); the branch point sequence (which contains a conserved adenosine important to intron removal); and the polypyrimidine tract (which is thought to recruit factors to the branch point sequence and 3' splice site). These sequence motifs are represented in the illustration below:



The U1 snRNP recognizes and binds to the 5' splice site. The branch point sequence is identified and bound by the branch-point-binding protein (BBP). The 3' splice site and polypyrimidine tract are recognized and bound by two specific components of a protein complex called U2 auxiliary factor (U2AF): U2AF35 and U2AF65 respectively.

Once these initial components have bound to their respective targets, the rest of the spliceosome assembles around them. Some of the previously bound components are displaced at this stage: For instance, the BBP is displaced by the U2 snRNP, and the U2AF complex is displaced by a complex of U4-U5-U6 snRNPs. The U1 and U4 snRNPs are also released. The first transesterification reaction then takes place, and a cut is made at the 5' splice site and the 5' end of the intron is subsequently connected to the conserved adenine found in the branch point sequence, forming the so-called "lariat" structure. This is followed by the second transesterification reaction which results in the splicing together of the two flanking exons. See this page for a helpful animation of the splicing process.

Other Important Protein Factors



Many other proteins play crucial roles in the RNA splicing process. One essential component is PRP8, a large protein that is located near the catalytic core of the spliceosome and that is involved in a number of critical molecular rearrangements that take place at the active site (for a review, see Grainger and Beggs, 2005). What is interesting is that this protein, though absolutely crucial to the RNA splicing machinery, bears no obvious homology to other known proteins.

The SR proteins, characterized by their serine/arginine dipeptide repeats and which are also essential, bind to the pre-mRNA and recruit other spliceosome components to the splice sites (Lin and Fu, 2007). SR proteins can be modified depending on the level of phosphorylation at their serine residues, and modulation of this phosphorylation helps to regulate their activity, and thus coordinate the splicing process (Saitoh *et al*., 2012; Plocinik *et al*., 2011; Zhong *et al*., 2009; Misteli *et al*., 1998). The illustration above (from here) shows the binding of SR proteins to splicing enhancer sites, which promotes the binding of U1 snRNP to the 5' splice site, and U2AF protein to the polypyrimidine tract and 3' splice site.

There are also ATPases that promote the structural rearrangements of snRNAs and release by the spliceosome of mRNA and the intron lariat. It is even thought that ATP-dependent RNA helicases play a significant role in "proofreading" of the chosen splice site, thus preventing the potentially catastrophic consequences of incorrect splicing (Yang *et al*., 2013; Semlow and Staley, 2012; Egecioglu and Chanfreau, 2011).

Composition

Each spliceosome is composed of five small nuclear RNAs (snRNA), and a range of associated protein factors. When these small RNA are combined with the protein factors, they make an RNA-protein complex called snRNP.

The snRNAs that make up the major spliceosome are named U1, U2, U4, U5, and U6, and participate in several RNA-RNA and RNA-protein interactions. The RNA component of the small nuclear ribonucleic protein or snRNP (pronounced "snurp") is rich in uridine (the nucleoside analog of the uracil nucleotide).

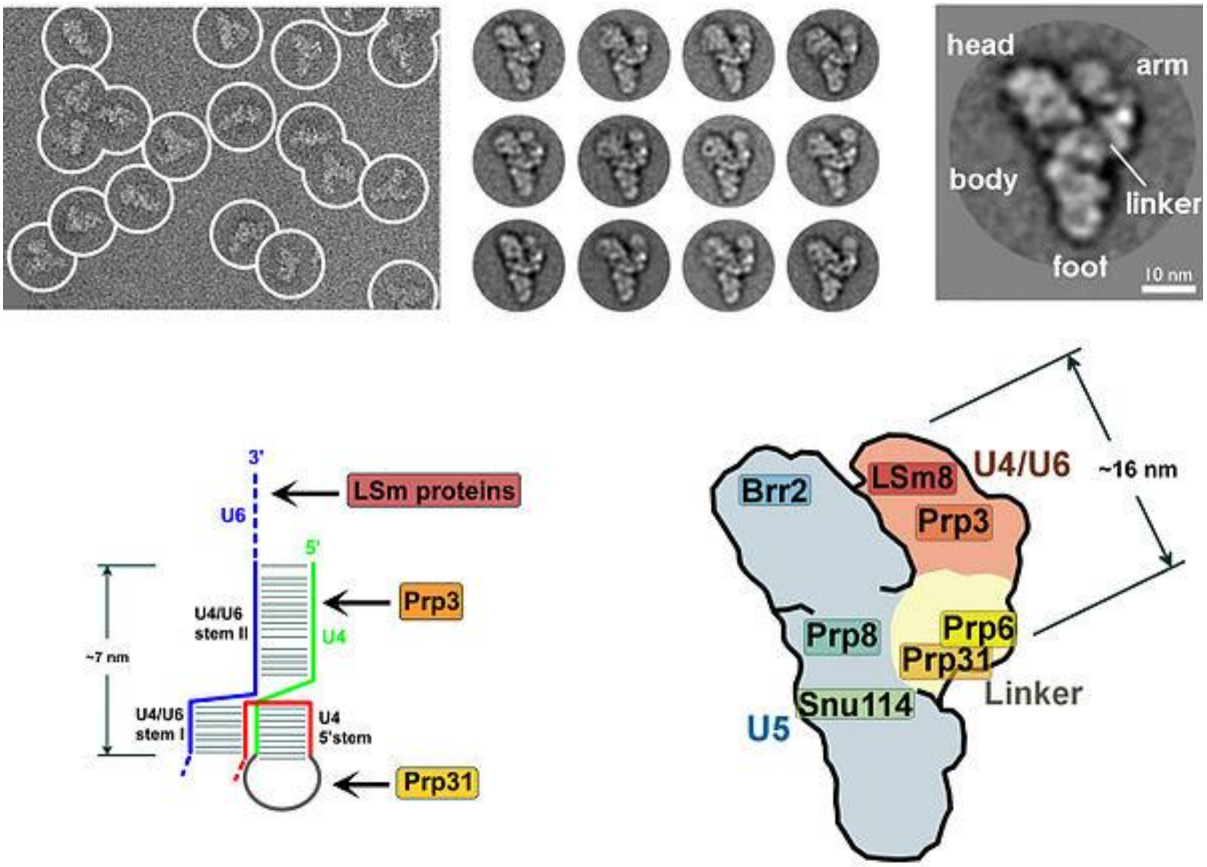
The canonical assembly of the spliceosome occurs anew on each hnRNA (pre-mRNA). The hnRNA contains specific sequence elements that are recognized and utilized during spliceosome assembly. These include the 5' end splice, the branch point sequence, the polypyrimidine tract, and the 3' end splice site. The spliceosome catalyzes the removal of introns, and the ligation of the flanking exons.

Introns typically have a GU nucleotide sequence at the 5' end splice site, and an AG at the 3' end splice site. The 3' splice site can be further defined by a variable length of polypyrimidines, called the polypyrimidine tract (PPT), which serves the dual function of recruiting factors to the 3' splice site and possibly recruiting factors to the branch point sequence (BPS). The BPS contains the conserved Adenosine required for the first step of splicing.

A group of less abundant snRNAs, U11, U12, U4atac, and U6atac, together with U5, are subunits of the so-called minor spliceosome that splices a rare class of pre-mRNA introns, denoted U12-type. The minor spliceosome is located in the nucleus like its major counterpart, though there are exceptions in some specialised cells including anucleate platelets and the dendroplasm of neuronal cells.

New evidence derived from the first crystal structure of a group II intron suggests that the spliceosome is actually a ribozyme, and that it uses a two–metal ion mechanism for catalysis.

In addition, many proteins exhibit a zinc-binding motif, which underscores the importance of zinc metal in the splicing mechanism.



Above are electron microscopy fields of negatively stained yeast (*Saccharomyces cerevisiae*) tri-snRNPs. Below left is a schematic illustration of the interaction of tri-snRNP proteins with the U4/U6 snRNA duplex. Below right is a cartoon model of the yeast tri-snRNP with shaded areas corresponding to U5 (gray), U4/U6 (orange) and the linker region (yellow).

Alternative splicing

Alternative splicing (the re-combination of different exons) is a major source of genetic diversity in eukaryotes. Splice variants have been used to account for the relatively small number of genes in the human genome. For years the estimate widely varied, with top estimates reaching 100,000 genes, but now, due to the Human Genome Project, the figure is believed to be closer to 20,000 genes. One particular Drosophila gene (Dscam, the

Drosophila homolog of the human Down syndrome cell adhesion molecule DSCAM) can be alternatively spliced into 38,000 different mRNA

The Exon Junction Complex

The exon junction complex (EJC) is a protein complex comprised of several protein components (RNPS1, Y14, SRm160, Aly/REF and Magoh) left behind near splice junctions by the splicing process (Hir and Andersen, 2008). Their function is to mark the transcript as processed, and thus ready for export from the nucleus to the cytoplasm, and translation at the ribosome. The EJC is typically found 20 to 24 nucleotides upstream of the splice junction.

The EJC also plays an important role in nonsense mediated decay, a surveillance system used in eukaryotes to destroy transcripts containing premature stop codons (Trinkle-Mulcahy *et al*., 2009; Chang *et al*., 2007; Gehring *et al*., 2005). Upon encountering an EJC during translation, the ribosome displaces the complex from the mRNA. The ribosome then continues until it reaches a stop codon. If, however, the mRNA contains a stop codon before the EJC, the nonsense mediated decay pathway is triggered. The EJC and its position thus contribute to transcript quality control.

The Evolution of the Spliceosome

A popular hypothesis regarding the origins of the spliceosome is that its predecessor was self-splicing RNA introns (e.g. Valadkhan, 2007). Such a hypothesis makes sense of several observations. For example, a simpler way to achieve splicing presumably would be to bring the splice sites together in one step to directly cleave and rejoin them. The proposed scenario, however, would explain the use of a lariat intermediate, since a lariat is generated by group II RNA intron sequences (Lambowitz1 and Zimmerly, 2011; Vogel and Borner, 2002).

The hypothesis also helps to clarify why RNA molecules play such an important part in the splicing process. Examples of self-splicing RNA introns still exist today (e.g., in the nuclear rRNA genes of the ciliate *Tetrahymena*) (Hagen and Cech, 1999; Price *et al*., 1995; Price and Cech, 1988; Kruger *et al*., 1982).

These observations may be taken as evidence as to the spliceosome's evolutionary predecessor, but they are hardly helpful in elucidating a plausible scenario for transitioning from one to the other. The spliceosome machinery is far more complex and sophisticated than autocatalytic ribozymes, involving not just five RNAs but hundreds of proteins.

CHAPTER # 15 RNA EDITING

RNA editing is a molecular process through which some cells can make discrete changes to specific nucleotide sequences within a RNA molecule after it has been generated by RNA polymerase. RNA editing is relatively rare, and common forms of RNA processing (e.g. splicing, 5'-capping and 3'-polyadenylation) are not usually included as editing. Editing events may include the insertion, deletion, and base substitution of nucleotides within the edited RNA molecule.

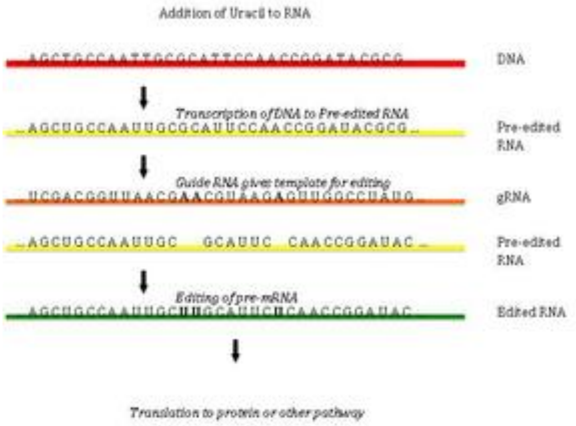
RNA editing has been observed in some tRNA, rRNA, mRNA or miRNA molecules of eukaryotes and their viruses, archaea and prokaryotes. RNA editing occurs in the cell nucleus and cytosol, as well as within mitochondria and plastids. In vertebrates, editing is rare and usually consists of a small number of changes to the sequence of affected molecules. In other organisms, extensive editing (*pan-editing*) can occur; in some cases the majority of nucleotides in a mRNA sequence may result from editing.

RNA-editing processes show great molecular diversity, and some appear to be evolutionarily recent acquisitions that arose independently. The diversity of RNA editing phenomena includes nucleobase modifications such as cytidine (C) to uridine (U) and adenosine (A) to inosine (I) deaminations, as well as non-templated nucleotide additions and insertions. RNA editing in mRNAs effectively alters the amino acid sequence of the encoded protein so that it differs from that predicted by the genomic DNA sequence.

Editing by insertion or deletion

RNA editing through the addition and deletion of uracil has been found in kinetoplasts from the mitochondria of Trypanosoma brucei[3] Because this may involve a large fraction of the sites in a gene, it is sometimes called "pan-editing" to distinguish it from topical editing of one or a few sites.

Pan-editing starts with the base-pairing of the unedited primary transcript with a guide RNA (gRNA), which contains complementary sequences to the regions around the insertion/deletion points. The newly formed double-stranded region is then enveloped by an editosome, a large multi-protein complex that catalyzes the editing. The editosome opens the transcript at the first mismatched nucleotide and starts inserting uridines. The inserted uridines will base-pair with the guide RNA, and insertion will continue as long as A or G is present in the guide RNA and will stop when a C or U is encountered. The inserted nucleotides cause a frameshift and result in a translated protein that differs from its gene.

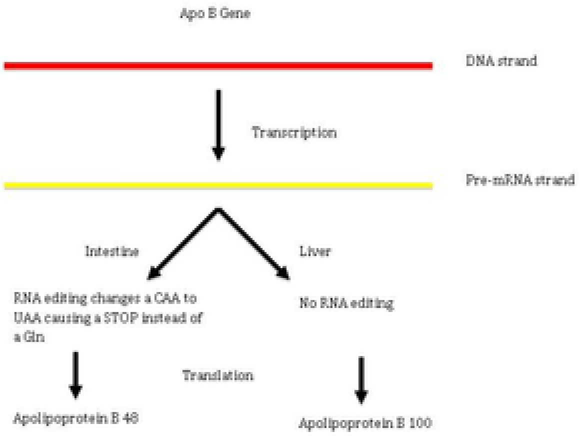

Addition of Uracil to RNA
AGCTGCCAATTGCGCATTCCAACCGGATACGCG
DNA
Transcription of DNA to Pre-edited RNA
...AGCUGCCAAUUGCGCAUUCCAACCGGAUACGCG...
Pre-edited RNA
Guide RNA gives template for editing
...UCGACGGUUAACGAACGUAAGAGUUGGCCUAUG...
gRNA
...AGCUGCCAAUUGC GCAUUC CAACCGGAUAC...
Pre-edited RNA
Editing of pre-mRNA
...AGCUGCCAAUUGCUUGCAUUCUCAACCGGAUAC...
Edited RNA
Translation to protein or other pathway

The Effect of Uracil Insertion in pre-mRNA transcripts

The mechanism of the editosome involves an endonucleolytic cut at the mismatch point between the guide RNA and the unedited transcript. The next step is catalyzed by one of the enzymes in the complex, a terminal U-transferase, which adds Us from UTP at the 3’ end of the mRNA. The opened ends are held in place by other proteins in the complex. Another enzyme, a U-specific exoribonuclease, removes the unpaired Us. After editing has made mRNA complementary to gRNA, an RNA ligase rejoins the ends of the edited mRNA transcript. As a consequence, the editosome can edit only in a 3’ to 5’ direction along the primary RNA transcript. The complex can act on only a single guide RNA at a time. Therefore, a RNA transcript requiring extensive editing will need more than one guide RNA and editosome complex.

Editing by deamination

C-to-U editing



The Effect of C-to-U RNA Editing on the Human ApoB gene

The editing involves cytidine deaminase that deaminates a cytidine base into a uridine base. An example of C-to-U editing is with the apolipoprotein B gene in humans. Apo B100 is expressed in the liver and apo B48 is expressed in the intestines. In the intestines, the mRNA has a CAA sequence edited to be UAA, a stop codon, thus producing the shorter B48 form.

C-to-U editing often occurs in the mitochondrial RNA of flowering plants. Different plants have different degrees of C-to-U editing; eight editing events occur in mitochondria of the moss Funaria hygrometrica , where as over 1700 editing events occur in the lycophytes Isoetes engelmanii. C-to-U editing is performed by members of the pentatricopeptide repeat (PPR) protein family. Angiosperms have large PPR families, acting as *trans* -factors for *cis* -elements lacking a consensus sequence; Arabidopsis has around 450 members in its PPR family. There have been a number of discoveries of PPR proteins in both plastids and mitochondria.

A-to-I editing

A-to-I editing is the main form of RNA editing in mammals and occurs in regions of double-stranded RNA (dsRNA). Adenosine deaminases acting on RNA (ADARs) are the RNA-editing enzymes involved in the hydrolytic deamination of Adenosine to Inosine (A-to-I editing). A-to-I editing can be specific (a single adenosine is edited within the stretch of dsRNA) or promiscuous (up to 50% of the adenosines are edited). Specific editing occurs within short duplexes (e.g., those formed in an mRNA where intronic sequence base pairs with a complementary exonic sequence), while promiscuous editing occurs within longer regions of duplex (e.g., pre- or pri-miRNAs, duplexes arising from transgene or viral expression, duplexes arising from paired repetitive elements). There are many effects of A-to-I editing, arising from the fact that I behaves as if it is G both in translation and when forming secondary structures. These effects include alteration of coding capacity, altered miRNA or siRNA target populations, heterochromatin formation, nuclear sequestration, cytoplasmic sequestration, endonucleolytic cleavage by Tudor-SN, inhibition of miRNA and siRNA processing, and altered splicing.

Alternative mRNA editing

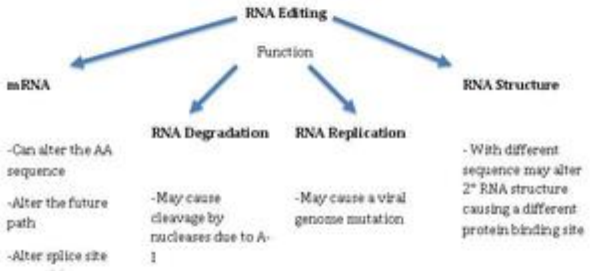
Alternative U-to-C mRNA editing was first reported in WT1 (Wilms Tumor-1) transcripts, and non-classic G-A mRNA changes were first observed in HNRNPK (heterogeneous nuclear ribonucleoprotein K) transcripts in both malignant and normal colorectal samples. The latter changes were also later seen alongside non-classic U-to-C alterations in brain cell TPH2 (tryptophan hydroxylase 2) transcripts. Although the reverse amination might be the simplest explanation for U-to-C changes, transamination and transglycosylation mechanisms have been proposed for plant U-to-C editing events in mitochondrial transcripts. A recent study reported novel G-to-A mRNA changes in WT1 transcripts at two hotspots, proposing the APOBEC3A (apolipoprotein B mRNA editing enzyme, catalytic polypeptide 3A) as the enzyme implicated in this class of alternative mRNA editing. It was also shown that alternative mRNA changes were associated with canonical WT1 splicing variants, indicating their functional significance.

RNA editing in plant mitochondria and plastids

It has been shown in previous studies that the only types of RNA editing seen in the plants' mitochondria and plastids are conversion of C-to-U and U-to-C (very rare). RNA-editing sites are found mainly in the coding regions of mRNA, introns, and other non-translated regions. In fact, RNA editing can restore the functionality of tRNA molecules. The editing sites are found primarily upstream of mitochondrial or plastid RNAs. While the specific positions for C to U RNA editing events have been fairly well studied in both the mitochondrion and plastid, the identity and organization of all proteins comprising the editosome have yet to be established. Members of the expansive PPR protein family have been shown to function as *trans*-acting factors for RNA sequence recognition. Specific members of the MORF (Multiple Organellar RNA editing Factor) family are also required for proper editing at several sites. As some of these MORF proteins have been shown to interact with members of the PPR family, it is possible MORF proteins are components of the editosome complex. An enzyme responsible for the trans- or deamination of the RNA transcript remains elusive, though it has been proposed that the PPR proteins may serve this function as well.

RNA editing is essential for the normal functioning of the plant’s translation and respiration activity. Editing can restore the essential base-pairing sequences of tRNAs, restoring functionality. It has also been linked to the production of RNA-edited proteins that are incorporated into the polypeptide complexes of the respiration pathway. Therefore, it is highly probable that polypeptides synthesized from unedited RNAs would not function properly and hinder the activity of both mitochondria and plastids.

C-to-U RNA editing can create start and stop codons, but it cannot destroy existing start and stop codons. A cryptic start codon is created when the codon ACG is edited to AUG.


RNA Editing
Function
mRNA
RNA Degradation
RNA Replication
RNA Structure
-Can alter the AA sequence
-Alter the future path
-Alter splice site
-May cause cleavage by nucleases due to A-I
-May cause a viral genome mutation
- With different sequence may alter 2° RNA structure causing a different protein binding site

Summary of the Various Functions of RNA Editing

RNA editing in viruses

RNA editing in viruses (i.e., measles, mumps, or parainfluenza) are used for stability and generation of protein variants. Viral RNAs are transcribed by a virus-encoded RNA-dependent RNA polymerase, which is prone to pausing and “stuttering” at certain nucleotide combinations. In addition, up to several hundred non-templated A's are added by the polymerase at the 3’ end of nascent mRNA. These As help stabilize the mRNA. Furthermore, the pausing and stuttering of the RNA polymerase allows the incorporation of one or two Gs or As upstream of the translational codon. The addition of the non-templated nucleotides shifts the reading frame, which generates a different protein.

Origin and evolution of RNA editing

The RNA-editing system seen in the animal may have evolved from mononucleotide deaminases, which have led to larger gene families that include the apobec-1 and adar genes. These genes share close identity with the bacterial deaminases involved in nucleotide metabolism. The adenosine deaminase of *E. coli* cannot deaminate a nucleoside in the RNA; the enzyme’s reaction pocket is too small to for the RNA strand to bind to. However, this active site is widened by amino acid changes in the corresponding human analog genes, APOBEC1 and ADAR, allowing deamination. The gRNA-mediated pan-editing in trypanosome mitochondria, involving templated insertion of U residues, is an entirely different biochemical reaction. The enzymes involved have been shown in other studies to be recruited and adapted from different sources. But, the specificity of nucleotide insertion via the interaction between the gRNA and mRNA are similar to the tRNA editing processes in the animal and Acanthamoeba mithochondria. Eukaryotic ribose methylation of rRNAs by guide RNA molecules is a similar form of modification.

Thus, RNA editing evolved more than once. Several adaptive rationales for editing have been suggested.[45] Editing is often described as a mechanism of correction or repair to compensate for defects in gene sequences. However, in the case of gRNA-mediated editing, this explanation does not seem possible because if a defect happens first, there is no way to generate an error-free gRNA-encoding region, which presumably arises by duplication of the original gene region. This thinking leads to an evolutionary proposal called "constructive neutral evolution" in which the order of steps is reversed, with the gratuitous capacity for editing preceding the "defect".

RNA editing may be involved in RNA degradation

A study looked at the involvement of RNA editing in RNA degradation. The researchers specifically looked at the interaction between ADAR and UPF1, an enzyme involved in the nonsense-mediated mRNA decay pathway (NMD). They found that ADAR and UPF1 are found within the suprasliceosome and they form a complex that leads to the down-regulation of specific genes. The exact mechanism or the exact pathways that these two are involved in are unknown at this time. The only fact that this research has shown is that they form a complex and down-regulate specific genes.

CHAPTER # 16 TRANSLATION

In molecular biology and genetics, translation is the process in which cellular ribosomes create proteins.

In translation, messenger RNA (mRNA)—produced by transcription from DNA—is decoded by a ribosome to produce a specific amino acid chain, or polypeptide. The polypeptide later folds into an active protein and performs its functions in the cell. The ribosome facilitates decoding by inducing the binding of complementary tRNA anticodon sequences to mRNA codons. The tRNAs carry specific amino acids that are chained together into a polypeptide as the mRNA passes through and is "read" by the ribosome. The entire process is a part of gene expression.

In brief, translation proceeds in four phases:

Initiation: The ribosome assembles around the target mRNA. The first tRNA is attached at the start codon.

Elongation: The tRNA transfers an amino acid to the tRNA corresponding to the next codon.

Translocation: The ribosome then moves (*translocates)* to the next mRNA codon to continue the process, creating an amino acid chain.

Termination: When a stop codon is reached, the ribosome releases the polypeptide.

In bacteria, translation occurs in the cell's cytoplasm, where the large and small subunits of the ribosome bind to the mRNA. In eukaryotes, translation occurs in the cytosol or across the membrane of the endoplasmic reticulum in a process called vectorial synthesis. In many instances, the entire ribosome/mRNA complex binds to the outer membrane of the rough endoplasmic reticulum (ER); the newly created polypeptide is stored inside the ER for later vesicle transport and secretion outside of the cell.

Many of transcribed RNA, such as transfer RNA, ribosomal RNA, and small nuclear RNA, do not undergo translation into proteins.

A number of antibiotics act by inhibiting translation. These include anisomycin, cycloheximide, chloramphenicol, tetracycline, streptomycin, erythromycin, and puromycin. Prokaryotic ribosomes have a different structure from that of eukaryotic ribosomes, and thus antibiotics can specifically target bacterial infections without any harm to a eukaryotic host's cells.

The basic process of protein production is addition of one amino acid at a time to the end of a protein. This operation is performed by a ribosome. The choice of amino acid type to add is determined by an mRNA molecule. Each amino acid added is matched to a three nucleotide subsequence of the mRNA. For each such triplet possible, the corresponding amino acid is accepted. The successive amino acids added to the chain are matched to successive nucleotide triplets in the mRNA. In this way the sequence of nucleotides in the template mRNA chain determines the sequence of amino acids in the generated amino acid chain. Addition of an amino acid occurs at the C-terminus of the peptide and thus translation is said to be amino-to-carboxyl directed.

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 molecules translate this code to a specific sequence of amino acids. The ribosome is a multisubunit 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 synthetases (enzymes) catalyze the bonding between specific tRNAs and the amino acids that their anticodon sequences call for. The product of this reaction is an aminoacyl-tRNA. This aminoacyl-tRNA is carried to the ribosome by EF-Tu, where mRNA codons are matched through complementary base pairing to specific tRNA anticodons.

Aminoacyl-tRNA synthetases that mispair tRNAs with the wrong amino acids can produce mischarged aminoacyl-tRNAs, which can result in inappropriate amino acids at the respective position in protein. This "mistranslation" of the genetic code naturally occurs at low levels in most organisms, but certain cellular environments cause an increase in permissive mRNA decoding, sometimes to the benefit of the cell.

The ribosome has three sites for tRNA to bind. They are the aminoacyl site (abbreviated A), the peptidyl site (abbreviated P) and the exit site (abbreviated E). With respect to the mRNA, the three sites are oriented 5’ to 3’ E-P-A, because ribosomes move toward the 3' end of mRNA. The A site binds the incoming tRNA with the complementary codon on the mRNA. The P site holds the tRNA with the growing polypeptide chain. The E site holds the tRNA without its amino acid. When an aminoacyl-tRNA initially binds to its corresponding codon on the mRNA, it is in the A site. Then, a peptide bond forms between the amino acid of the tRNA in the A site and the amino acid of the charged tRNA in the P site. The growing polypeptide chain is transferred to the tRNA in the A site. Translocation occurs, moving the tRNA in the P site, now without an amino acid, to the E site; the tRNA that was in the A site, now charged with the polypeptide chain, is moved to the P site. The tRNA in the E site leaves and another aminoacyl-tRNA enters the A site to repeat the process.

After the new amino acid is added to the chain, and after the mRNA is released out of the nucleus and into the ribosome's core, the energy provided by the hydrolysis of a GTP bound to the translocase EF-G (in prokaryotes) and eEF-2 (in eukaryotes) moves the ribosome down one codon towards the 3' end. 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 4*n*-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 6-9 amino acid residues per second).

In activation, the correct amino acid is covalently bonded to the correct transfer RNA (tRNA). 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 the 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). No tRNA can recognize or bind to this codon. Instead, the stop codon induces the binding of a release factor protein that prompts the disassembly of the entire ribosome/mRNA complex.

The process of translation is highly regulated in both eukaryotic and prokaryotic organisms. Regulation of translation can impact the global rate of protein synthesis which is closely coupled to the metabolic and proliferative state of a cell. In addition, recent work has revealed that genetic differences and their subsequent expression as mRNAs can also impact translation rate in an RNA-specific manner.

CHAPTER # 17 POST TRANSLATIONAL MODIFICATIONS

Protein post-translational modification (PTM) increases the functional diversity of the proteome by the covalent addition of functional groups or proteins, proteolytic cleavage of regulatory subunits or degradation of entire proteins. These modifications include phosphorylation, glycosylation, ubiquitination, nitrosylation, methylation, acetylation, lipidation and proteolysis and influence almost all aspects of normal cell biology and pathogenesis. Therefore, identifying and understanding PTMs is critical in the study of cell biology and disease treatment and prevention.

Introduction

Within the last few decades, scientists have discovered that the human proteome is vastly more complex than the human genome. While it is estimated that the human genome comprises between 20,000 and 25,000 genes, the total number of proteins in the human proteome is estimated at over 1 million. These estimations demonstrate that single genes encode multiple proteins. Genomic recombination, transcription initiation at alternative promoters, differential transcription termination, and alternative splicing of the transcript are mechanisms that generate different mRNA transcripts from a single gene .

The increase in complexity from the level of the genome to the proteome is further facilitated by protein post-translational modifications (PTMs). PTMs are chemical modifications that play a key role in functional proteomics, because they regulate activity, localization and interaction with other cellular molecules such as proteins, nucleic acids, lipids, and cofactors.

Post-translational modifications are key mechanisms to increase proteomic diversity. While the genome comprises 20-25,000 genes, the proteome is estimated to encompass over 1 million proteins. Changes at the transcriptional and mRNA levels increase the size of the transcriptome relative to the genome, and the myriad of different post-translational modifications exponentially increases the complexity of the proteome relative to both the transcriptome and genome.

Additionally, the human proteome is dynamic and changes in response to a legion of stimuli, and post-translational modifications are commonly employed to regulate cellular activity. PTMs occur at distinct amino acid side chains or peptide linkages and are most often mediated by enzymatic activity. Indeed, it is estimated that 5% of the proteome comprises enzymes that perform more than 200 types of post-translational modifications (4). These enzymes include kinases, phosphatases, transferases and ligases, which add or remove functional groups, proteins, lipids or sugars to or from amino acid side chains, and proteases, which cleave peptide bonds to remove specific sequences or regulatory subunits. Many proteins can also modify themselves using autocatalytic domains, such as autokinase and autoprotolytic domains.

Post-translational modification can occur at any step in the "life cycle" of a protein. For example, many proteins are modified shortly after translation is completed to mediate proper protein folding or stability or to direct the nascent protein to distinct cellular compartments (e.g., nucleus, membrane). Other modifications occur after folding and localization are completed to activate or inactivate catalytic activity or to otherwise influence the biological activity of the protein. Proteins are also covalently linked to tags that target a protein for degradation. Besides single modifications, proteins are often modified through a combination of post-translational cleavage and the addition of functional groups through a step-wise mechanism of protein maturation or activation.

Protein PTMs can also be reversible depending on the nature of the modification. For example, kinases phosphorylate proteins at specific amino acid side chains, which is a common method of catalytic activation or inactivation. Conversely, phosphatases hydrolyze the phosphate group to remove it from the protein and reverse the biological activity. Proteolytic cleavage of peptide bonds is a thermodynamically favorable reaction and therefore permanently removes peptide sequences or regulatory domains.

Consequently, the analysis of proteins and their post-translational modifications is particularly important for the study of heart disease, cancer, neurodegenerative diseases and diabetes. The characterization of PTMs, although challenging, provides invaluable insight into the cellular functions underlying etiological processes. Technically, the main challenges in studying post-translationally modified proteins are the development of specific detection and purification methods. Fortunately, these technical obstacles are being overcome with a variety of new and refined proteomics technologies.

Post-Translational Modifications

As noted above, the large number of different PTMs precludes a thorough review of all possible protein modifications. Therefore, this overview only touches on a small number of the most common types of PTMs studied in protein research today. Furthermore, greater focus is placed on phosphorylation, glycosylation and ubiquitination, and therefore these PTMs are described in greater detail on pages dedicated to the respective PTM.

Phosphorylation

Reversible protein phosphorylation, principally on serine, threonine or tyrosine residues, is one of the most important and well-studied post-translational modifications. Phosphorylation plays critical roles in the regulation of many cellular processes including cell cycle, growth, apoptosis and signal transduction pathways.

Glycosylation

Protein glycosylation is acknowledged as one of the major post-translational modifications, with significant effects on protein folding, conformation, distribution, stability and activity. Glycosylation encompasses a diverse selection of sugar-moiety additions to proteins that ranges from simple monosaccharide modifications of nuclear transcription factors to highly complex branched polysaccharide changes of cell surface receptors. Carbohydrates in the form of aspargine-linked (N-linked) or serine/threonine-linked (O-linked) oligosaccharides are major structural components of many cell surface and secreted proteins.

Ubiquitination

Ubiquitin is an 8-kDa polypeptide consisting of 76 amino acids that is appended to the Îµ-NH2 of lysine in target proteins via the C-terminal glycine of ubiquitin. Following an initial monoubiquitination event, the formation of a ubiquitin polymer may occur, and polyubiquitinated proteins are then recognized by the 26S proteasome that catalyzes the degradation of the ubiquitinated protein and the recycling of ubiquitin.

S-Nitrosylation

Nitric oxide (NO) is produced by three isoforms of nitric oxide synthase (NOS) and is a chemical messenger that reacts with free cysteine residues to form S-nitrothiols (SNOs). S-nitrosylation is a critical PTM used by cells to stabilize proteins, regulate gene expression and provide NO donors, and the generation, localization, activation and catabolism of SNOs are tightly regulated.

S-nitrosylation is a reversible reaction, and SNOs have a short half life in the cytoplasm because of the host of reducing enzymes, including glutathione (GSH) and thioredoxin, that denitrosylate proteins. Therefore, SNOs are often stored in membranes, vesicles, the interstitial space and lipophilic protein folds to protect them from denitrosylation. For example, caspases, which mediate apoptosis, are stored in the mitochondrial intermembrane space as SNOs. In response to extra- or intracellular cues, the caspases are released into the cytoplasm, and the highly reducing environment rapidly denitrosylates the proteins, resulting in caspase activation and the induction of apoptosis.

S-nitrosylation is not a random event, and only specific cysteine residues are S-nitrosylated. Because proteins may contain multiple cysteines and due to the labile nature of SNOs, S-nitrosylated cysteines can be difficult to detect and distinguish from non-S-nitrosylated amino acids. The biotin switch assay, developed by Jaffrey et al., is a common method of detecting SNOs, and the steps of the assay are listed below:

All free cysteines are blocked.

All remaining cysteines (presumably only those that are denitrosylated) are denitrosylated.

The now-free thiol groups are then biotinylated.

Biotinylated proteins are detected by SDS-PAGE and Western blot analysis or mass spectrometry.

Methylation

The transfer of one-carbon methyl groups to nitrogen or oxygen (N- and O-methylation, respectively) to amino acid side chains increases the hydrophobicity of the protein and can neutralize a negative amino acid charge when bound to carboxylic acids. Methylation is mediated by methyltransferases, and S-adenosyl methionine (SAM) is the primary methyl group donor.

Methylation occurs so often that SAM has been suggested to be the most-used substrate in enzymatic reactions after ATP. Additionally, while N-methylation is irreversible, O-methylation is potentially reversible. Methylation is a well-known mechanism of epigenetic regulation, as histone methylation and demethylation influences the availability of DNA for transcription. Amino acid residues can be conjugated to a single methyl group or multiple methyl groups to increase the effects of modification.

N-Acetylation

N-acetylation, or the transfer of an acetyl group to nitrogen, occurs in almost all eukaryotic proteins through both irreversible and reversible mechanisms. N-terminal acetylation requires the cleavage of the N-terminal methionine by methionine aminopeptidase (MAP) before replacing the amino acid with an acetyl group from acetyl-CoA by N-acetyltransferase (NAT) enzymes. This type of acetylation is co-translational, in that N-terminus is acetylated on growing polypeptide chains that are still attached to the ribosome. While 80-90% of eukaryotic proteins are acetylated in this manner, the exact biological significance is still unclear.

Acetylation at the ε-NH2 of lysine (termed lysine acetylation) on histone N-termini is a common method of regulating gene transcription. Histone acetylation is a reversible event that reduces chromosomal condensation to promote transcription, and the acetylation of these lysine residues is regulated by transcription factors that contain histone acetyletransferase (HAT) activity. While transcription factors with HAT activity act as transcription co-activators, histone deacetylase (HDAC) enzymes are co-repressors that reverse the effects of acetylation by reducing the level of lysine acetylation and increasing chromosomal condensation.

Sirtuins (silent information regulator) are a group of NAD-dependent deacetylases that target histones. As their name implies, they maintain gene silencing by hypoacetylating histones and have been reported to aid in maintaining genomic stability.

While acetylation was first detected in histones, cytoplasmic proteins have been reported to also be acetylated, and therefore acetylation seems to play a greater role in cell biology than simply transcriptional regulation. Furthermore, crosstalk between acetylation and other post-translational modifications, including phosphorylation, ubiquitination and methylation, can modify the biological function of the acetylated protein.

Protein acetylation can be detected by chromosome immunoprecipitation (ChIP) using acetyllysine-specific antibodies or by mass spectrometry, where an increase in histone by 42 mass units represents a single acetylation.

Lipidation

Lipidation is a method to target proteins to membranes in organelles (endoplasmic reticulum [ER], Golgi apparatus, mitochondria), vesicles (endosomes, lysosomes) and the plasma membrane. The four types of lipidation are:

C-terminal glycosyl phosphatidylinositol (GPI) anchor

N-terminal myristoylation

S-myristoylation

S-prenylation

Each type of modification gives proteins distinct membrane affinities, although all types of lipidation increase the hydrophobicity of a protein and thus its affinity for membranes. The different types of lipidation are also not mutually exclusive, in that two or more lipids can be attached to a given protein.

GPI anchors tether cell surface proteins to the plasma membrane. These hydrophobic moieties are prepared in the ER, where they are then added to the nascent protein en bloc. GPI-anchored proteins are often localized to cholesterol- and sphingolipid-rich lipid rafts, which act as signaling platforms on the plasma membrane. This type of modification is reversible, as the GPI anchor can be released from the protein by phosphoinositol-specific phospholipase C. Indeed, this lipase is used in the detection of GPI-anchored proteins to release GPI-anchored proteins from membranes for gel separation and analysis by mass spectrometry.

N-myristoylation is a method to give proteins a hydrophobic handle for membrane localization. The myristoyl group is a 14-carbon saturated fatty acid (C14), which gives the protein sufficient hydrophobicity and affinity for membranes, but not enough to permanently anchor the protein in the membrane. N-myristoylation can therefore act as a conformational localization switch, in which protein conformational changes influence the availability of the handle for membrane attachment. Because of this conditional localization, signal proteins that selectively localize to membrane, such as Src-family kinases, are N-myristoylated.

N-myristoylation is facilitated specifically by N-myristoyltransferase (NMT) and uses myristoyl-CoA as the substrate to attach the myristoyl group to the N-terminal glycine. Because methionine is the N-terminal amino acid of all eukaryotic proteins, this PTM requires methionine cleavage by the above-mentioned MAP prior to addition of the myristoyl group; this represents one example of multiple PTMs on a single protein.

S-palmitoylation adds a C16 palmitoyl group from palmitoyl-CoA to the thiolate side chain of cysteine residues via palmitoyl acyl transferases (PATs). Because of the longer hydrophobic group, this anchor can permanently anchor the protein to the membrane. This localization can be reversed, though, by thioesterases that break the link between the protein and the anchor; thus, S-palmitoylation is used as an on/off switch to regulate membrane localization. S-palmitoylation is often used to strengthen other types of lipidation, such as myristoylation or farnesylation (see below). S-palmitoylated proteins also selectively concentrate at lipid rafts.

S-prenylation covalently adds a farnesyl (C15) or geranylgeranyl (C20) group to specific cysteine residues within 5 amino acids from the C-terminus via farnesyl transferase (FT) or geranylgeranyl transferases (GGT I and II). Unlike S-palmitoylation, S-prenylation is hydrolytically stable. Approximately 2% of all proteins are prenylated, including all members of the Ras superfamily. This group of molecular switches is farnesylated, geranylgeranylated or a combination of both. Additionally, these proteins have specific 4-amino acid motifs at the C-terminus that determine the type of prenylation at single or dual cysteines. Prenylation occurs in the ER and is often part of a stepwise process of PTMs that is followed by proteolytic cleavage by Rce1 and methylation by isoprenyl cysteine methyltransferase (ICMT).

Proteolysis

Peptide bonds are indefinitely stable under physiological conditions, and therefore cells require some mechanism to break these bonds. Proteases comprise a family of enzymes that cleave the peptide bonds of proteins and are critical in antigen processing, apoptosis, surface protein shedding and cell signaling.

The family of over 11,000 proteases varies in substrate specificity, mechanism of peptide cleavage, location in the cell and the length of activity. While this variation suggests a wide array of functionalities, proteases can generally be separated into groups based on the type of proteolysis. Degradative proteolysis is critical to remove unassembled protein subunits and misfolded proteins and to maintain protein concentrations at homeostatic concentrations by reducing a given protein to the level of small peptides and single amino acids. Proteases also play a biosynthetic role in cell biology that includes cleaving signal peptides from nascent proteins and activating zymogens, which are inactive enzyme precursors that require cleavage at specific sites for enzyme function. In this respect, proteases act as molecular switches to regulate enzyme activity.

Proteolysis is a thermodynamically favorable and irreversible reaction. Therefore, protease activity is tightly regulated to avoid uncontrolled proteolysis through temporal and/or spatial control mechanisms including regulation by cleavage in cis or trans and compartmentalization (e.g., proteasomes, lysosomes).

The diverse family of proteases can be classified by the site of action, such as aminopeptidases and carboxypeptidase, which cleave at the amino or carboxy terminus of a protein, respectively. Another type of classification is based on the active site groups of a given protease that are involved in proteolysis. Based on this classification strategy, greater than 90% of known proteases fall into one of four categories as follows:

Serine proteases

Cysteine proteases

Aspartic acid proteases

Zinc metalloproteases

CHAPTER # 18 REGULATORY ELEMENTS

RNA molecules that act as regulators were known in bacteria for years before the first microRNAs (miRNAs) and short interfering RNAs (siRNAs) were discovered in eukaryotes. In 1981, the ~108 nucleotide RNA I was found to block ColE1 plasmid replication by base pairing with the RNA that is cleaved to produce the replication primer. This work was followed by the 1983 discovery of a ~70 nucleotide RNA which is transcribed from the pOUT promoter of the Tn10 transposon and represses transposition by preventing translation of the transposase mRNA. The first chromosomally-encoded small RNA regulator, reported in 1984, was the 174 nucleotide *Escherichia coli* MicF RNA, which inhibits translation of the mRNA encoding the major outer membrane porin OmpF. These first small RNA regulators, and a handful of others, were identified by gel analysis due to their abundance, by multicopy phenotypes, or by serendipity.

While a few bacterial RNA regulators were identified early on, their prevalence and their contributions to numerous physiological responses were not initially appreciated. In 2001–2002, four groups reported the identification of many new small RNAs through systematic computational searches for conservation and orphan promoter and terminator sequences in the intergenic regions of *E. coli*. Additional RNAs were discovered by direct detection using cloning-based techniques or microarrays with probes in intergenic regions. Variations of these approaches, aided by the availability of many new bacterial genome sequences, have led to the identification of regulatory RNAs in an ever-increasing number of bacteria. Enabled by recent technical advances, including multilayered computational searches deep sequencing and tiled microarrays with full genome coverage, hundreds of candidate regulatory RNA genes in various bacteria have now been predicted. In *E. coli* alone, ~80 small transcripts have been verified, increasing the total number of genes identified for this organism by 2%.

In this review, we will focus our discussion on bacterial small RNAs that act as regulators. A limited number of small RNAs carry out specific housekeeping functions, namely the 4.5S RNA component of the signal recognition particle, the RNase P RNA responsible for processing of tRNAs and other RNAs, and tmRNA, which acts as both a tRNA and mRNA to tag incompletely translated proteins for degradation and to release stalled ribosomes. We will not discuss these RNAs further, although their actions, as well as those of some tRNAs, can have regulatory consequences.

In addition, a few defining features are worthy of mention at the outset. Riboswitches are part of the mRNA they regulate, usually found within the 5' untranslated region (5'-UTR), and hence act in *cis*. Most of the regulatory RNAs that act in *trans* by base pairing with other RNAs are synthesized as discrete transcripts with dedicated promoter and terminator sequences. Given that the longest of these RNAs, RNAIII of *Staphylococcus aureus*, is still only 514 nucleotides, the RNAs are commonly referred to as small RNAs (sRNAs). We prefer this term to “noncoding RNA”, the term frequently used in eukaryotes, since a number of the sRNAs, including RNAIII, also encode proteins. In contrast to the base pairing sRNAs, some sRNAs that modulate protein activity as well as the CRISPR RNAs are processed out of longer transcripts.

Regulatory Functions of Bacterial RNAs

Regulatory RNAs can modulate transcription, translation, mRNA stability, and DNA maintenance or silencing. They achieve these diverse outcomes through a variety of mechanisms, including changes in RNA conformation, protein binding, base pairing with other RNAs, and interactions with DNA.

Riboswitches

Perhaps the simplest bacterial RNA regulatory elements are sequences at the 5' end of mRNAs, or less frequently at the 3' end, that can adopt different conformations in response to environmental signals, including stalled ribosomes, uncharged tRNAs, elevated temperatures, or small molecule ligands. These elements were first described decades ago in elegant studies characterizing transcription attenuation. In this process, stalled ribosomes lead to changes in mRNA structure, affecting transcription elongation through the formation of terminator or antiterminator structures in the mRNA. Later studies showed that sequences found in transcripts encoding tRNA synthetases, termed “Tboxes”, bind the corresponding uncharged tRNAs, and that other leader sequences, known as “RNA thermometers”, fold in a manner that is sensitive to temperature. In both of these cases, the alternate structures lead to changes in the expression of the downstream gene.

More recently, it was found that leader sequences could bind small molecules and adopt different conformations in the presence or absence of metabolites. These metabolite sensors, denoted “riboswitches”, directly regulate the genes involved in the uptake and use of the metabolite. In fact, in some cases, the presence of a riboswitch upstream of an uncharacterized or mis-annotated gene has helped to clarify the physiological role of the gene product. An ever-increasing number and variety of riboswitches are being identified in bacteria, as well as in some eukaryotes. For example, as many as 2% of all *Bacillus subtilis* genes are regulated by riboswitches which bind metabolites ranging from flavin mononucleotide (FMN) and thiamin pyrophosphate to S-adenosylmethionine, lysine and guanine.

Riboswitches generally consist of two parts: the aptamer region, which binds the ligand, and the so-called expression platform, which regulates gene expression through alternative RNA structures that affect transcription or translation. Upon binding of the ligand, the riboswitch changes conformation. These changes usually involve alternative hairpin structures which form or disrupt transcriptional terminators or antiterminators, or which occlude or expose ribosome binding sites. In general, most riboswitches repress transcription or translation in the presence of the metabolite ligand; only a few riboswitches that activate gene expression have been characterized.

Gene Arrangement and Regulatory Functions of Ligand- and Protein-binding Regulatory RNAs

Due to the modular nature of riboswitches, the same aptamer domain can mediate different regulatory outcomes or operate through distinct mechanisms in different contexts. For example, the cobalamin riboswitch, which binds the coenzyme form of vitamin B_{12}, operates by transcription termination for the *btuB* genes in Gram-positive bacteria but modulates translation initiation for the *cob* operons of Gram-negative bacteria. Some transcripts carry tandem riboswitches, which can integrate distinct physiological signals, and one notable riboswitch, the *glmS* leader sequence, even acts as a ribozyme to catalyze self-cleavage. Upon binding of its cofactor glucosamine-6-phosphate, the *glmS* riboswitch cleaves itself and inactivates the mRNA encoding the enzyme that generates glucosamine-6-phosphate, thus effecting a negative feedback loop for metabolite levels. In principle, riboswitches could be used in conjunction with any reaction associated with RNA, not just transcription, translation and RNA processing, but also RNA modification, localization or splicing.

Generally, the riboswitches in Gram-positive bacteria affect transcriptional attenuation, while the riboswitches in Gram-negative bacteria more frequently inhibit translation. Possibly the preferential use of transcriptional termination in Gram-positive organisms is linked to the fact that genes are clustered together in larger biosynthetic operons where more resources would be wasted if the full-length transcript is synthesized. Gram-positive organisms also appear to rely more on *cis*-acting riboswitches than Gram-negative organisms, for which more *trans*-acting sRNA regulators are known. Research directions pursued in studies of the different organisms, however, may bias these generalizations.

sRNAs That Modulate Protein Activity

Three protein-binding sRNAs have intrinsic activity (RNase P) or contribute essential functions to a ribonucleoprotein particle (4.5S and tmRNA). In contrast, three other protein-binding sRNAs (CsrB, 6S, and GlmY) act in a regulatory fashion to antagonize the activities of their cognate proteins by mimicking the structures of other nucleic acids.

The CsrB and CsrC RNAs of *E. coli* modulate the activity of CsrA, an RNA-binding protein that regulates carbon usage and bacterial motility upon entry into stationary phase and other nutrient-poor conditions. CsrA dimers bind to GGA motifs in the 5’ UTR of target mRNAs, thereby affecting the stability and/or translation of the mRNA. The CsrB and CsrC RNAs each contain multiple GGA binding sites, 22 and 13 respectively, for CsrA. Thus, when CsrB and CsrC levels increase, the sRNAs effectively sequester the CsrA protein away from mRNA leaders. Transcription of the *csrB* and *csrC* genes is induced by the BarA-UvrB two-component regulators when cells encounter nutrient poor growth conditions, though the signal for this induction is not known. The CsrB and CsrC RNAs also are regulated at the level of stability through the CsrD protein, a cyclic di-GMP binding protein, which recruits RNase E to degrade the sRNAs. CsrB and CsrC homologs (such as RsmY and RsmZ) have been found to antagonize the activities of CsrA homologs in a range of bacteria including *Salmonella*, *Erwinia*, *Pseudomonas*, and *Vibrio* where they impact secondary metabolism, quorum sensing and epithelial cell invasion.

The *E. coli* 6S RNA mimics an open promoter to bind to and sequester the σ^{70}-containing RNA polymerase. When 6S is abundant, especially in stationary phase, it is able to complex with much of the σ^{70}-bound, housekeeping form of RNA polymerase, but is not associated with the σ^{S}-bound, stationary phase form of RNA polymerase. The interaction between 6S and σ^{70}-holoenzyme inhibits transcription from certain σ^{70} promoters and increases transcription from some σ^{S} regulated promoters, in part by altering the competition between σ^{70}-and σ^{S}-holoenzyme binding to promoters. Interestingly, the 6S RNA can serve as a template for the transcription of 14–20 nucleotide product RNAs (pRNAs) by RNA polymerase, especially during outgrowth from stationary phase. In fact, it is thought that transcription from 6S when NTP concentrations increase may be a way to release σ^{70}-RNA polymerase. It is not known whether the pRNAs themselves have a function. The 6S RNA is processed out of a longer transcript and accumulates during stationary phase, but the details of this regulation have not been elucidated. There are multiple 6S homologs in a number of organisms, including two in *B. subtilis*. The roles of these homologs again are not known, but it is tempting to speculate that they inhibit the activities of alternative σ factor forms of RNA polymerase.

One additional sRNA, GlmY, has recently been proposed to have a protein-binding mode of action and is thought to function by titrating an RNA processing factor away from a homologous sRNA, GlmZ. Both GlmZ and GlmY promote accumulation of the GlmS glucosamine-6-phosphate synthase, however they do so by distinct mechanisms. The full-length GlmZ RNA base pairs with and activates translation of the *glmS* mRNA. Although the GlmY RNA is highly homologous to GlmZ in sequence and predicted secondary structure, GlmY lacks the region that is complementary to the *glmS* mRNA target and does not directly activate *glmS* translation. Instead, GlmY expression inhibits a GlmZ processing event that renders GlmZ unable to activate *glmS* translation. Although not yet conclusively shown, GlmY most likely stabilizes the full-length GlmZ by competing with GlmZ for binding to the YhbJ protein that targets GlmZ for processing. The GlmY RNA is also processed and its levels are negatively regulated by poly-adenylation.

CsrB RNA simulates an mRNA element, 6S imitates a DNA structure, and GlmY mimics another sRNA, raising the question as to what other molecules, nucleic acid or otherwise, might yet uncharacterized sRNAs mimic?

***Cis*-encoded Base Pairing sRNAs**

In contrast to the few known protein-binding sRNAs, most characterized sRNAs regulate gene expression by base pairing with mRNAs and fall into two broad classes: those having extensive potential for base pairing with their target RNA and those with more limited complementarity. We will first focus on sRNAs that are encoded in *cis* on the DNA strand opposite the target RNA and share extended regions of complete complementarity with their target, often 75 nucleotides or more. While the two transcripts are encoded in the same region of DNA, they are transcribed from opposite strands as discrete RNA species and function in *trans* as diffusible molecules. For the few cases where it has been examined, the initial interaction between the sRNA and target RNA involves only limited pairing, though the duplex can subsequently be extended. The most well-studied examples of *cis*-encoded antisense sRNAs reside on plasmids or other mobile genetic elements, however chromosomal versions of these sRNAs increasingly are being found.

Gene Arrangement and Regulatory Functions of Base Pairing Regulatory RNAs

Most of the *cis*-encoded antisense sRNAs expressed from bacteriophage, plasmids and transposons function to maintain the appropriate copy number of the mobile element. They achieve this through a variety of mechanisms, including inhibition of replication primer formation and transposase translation, as mentioned for plasmid ColE1 RNA I and Tn10 pOUT RNA, respectively. Another common group act as antitoxins to repress the translation of toxic proteins that kill cells from which the mobile element has been lost.

In general, the physiological roles of the *cis*-encoded antisense sRNAs expressed from bacterial chromosomes are less well understood. A subset promote degradation and/or repress translation of mRNAs encoding proteins that are toxic at high levels. In *E. coli,* there are also two sRNAs, IstR and OhsC, that are encoded directly adjacent to genes encoding potentially toxic proteins. Although these sRNAs are not true antisense RNAs, they do contain extended regions of perfect complementarity (19 and 23 nucleotides) with the toxin mRNAs. Interestingly, most of these sRNAs appear to be expressed constitutively. Some of the chromosomal antitoxin sRNAs are homologous to plasmid antitoxin sRNAs (for example, the Hok/Sok loci present in the *E. coli* chromosome) or are located in regions acquired from mobile elements (for example, the RatA RNA of *B. subtilis* found in a remnant of a cryptic prophage). These observations indicate that the antitoxin sRNA and corresponding toxin genes might have been acquired by horizontal transfer. The chromosomal versions may simply be non-functional remnants. However, some *cis*-encoded antisense antitoxin sRNAs do not have known homologs on mobile elements. In addition, given that bacteria have multiple copies of several loci, all of which are expressed in the cases examined, it is tempting to speculate that the antitoxin sRNAs-toxin proteins encoded on the chromosome provide beneficial functions. Although high levels of the toxins kill cells, more moderate levels produced from single-copy loci under inducing conditions may only slow growth. Thus one model proposes that chromosomal toxin-antitoxin modules induce slow growth or stasis under conditions of stress to allow cells time to repair damage or otherwise adjust to their environment. Another possibility is that certain modules may be retained in bacterial chromosomes as a defense against plasmids bearing homologous modules, assuming that the chromosomal antisense sRNA can repress the expression of the plasmid-encoded toxin.

Another group of *cis*-encoded antisense sRNAs modulates the expression of genes in an operon. Some of these sRNAs are encoded in regions complementary to intervening sequence between ORFs. For example, in *E. coli*, base pairing between the stationary phase-induced GadY antisense sRNA and the *gadXW* mRNA leads to cleavage of the duplex between the *gadX* and *gadW* genes and increased levels of a *gadX* transcript. For the virulence plasmid pJM1 of *Vibrio anguillarum*, the interaction between the RNAβ antisense sRNA and the *fatDCBAangRT* mRNA leads to transcription termination after the *fatA* gene, thus reducing expression of the downstream *angRT* genes. In *Synechocystis*, the iron-stress repressed IsrR antisense sRNA base pairs with sequences within *isiA* coding region of the *isiAB* transcript and leads to decreased levels of an *isiA* transcript. In this case, it is not known whether *isiB* expression is also affected.

The list of *cis*-encoded antisense sRNAs is far from complete, especially for chromosomal versions, and other mechanisms of action are sure to be found.

***Trans*-encoded Base Pairing sRNAs**

Another class of base pairing sRNAs is the *trans*-encoded sRNAs, which, in contrast to the *cis*-encoded antisense sRNAs, share only limited complementarity with their target mRNAs. These sRNAs regulate the translation and/or stability of target mRNAs and are, in many respects, functionally analogous to eukaryotic miRNAs.

The majority of the regulation by the known *trans*-encoded sRNAs is negative. Base pairing between the sRNA and its target mRNA usually leads to repression of protein levels through translational inhibition, mRNA degradation, or both. The bacterial sRNAs characterized to date primarily bind to the 5' UTR of mRNAs and most often occlude the ribosome binding site, though some sRNAs such as GcvB and RyhB inhibit translation through base pairing far upstream of the AUG of the repressed gene. The sRNA-mRNA duplex is then frequently subject to degradation by RNase E. For the few characterized sRNA-mRNA interactions, the inhibition of ribosome binding is the main contributor to reduced protein levels, while the subsequent degradation of the sRNA-mRNA duplex is thought to increase the robustness of the repression and make the regulation irreversible. However, sRNAs can also activate expression of their target mRNAs through an anti-antisense mechanism whereby base pairing of the sRNA disrupts an inhibitory secondary structure which sequesters the ribosome binding site. Theoretically, base pairing between a *trans*-encoded sRNA and its target could promote transcription termination or antitermination, as has been found for some *cis*-encoded sRNAs, or alter mRNA stability through changes in poly-adenylation.

For *trans*-encoded sRNAs, there is little correlation between the chromosomal location of the sRNA gene and the target mRNA gene. In fact, each *trans*-encoded sRNA typically base pairs with multiple mRNAs. The capacity for multiple base pairing interactions results from the fact that *trans*-encoded sRNAs make more limited contacts with their target mRNAs in discontinuous patches, rather than extended stretches of perfect complementarity, as for *cis*-encoded antisense sRNAs. The region of potential base pairing between *trans*-encoded sRNAs and target mRNAs typically encompasses ~10–25 nucleotides, but in all cases where it has been examined only a core of the nucleotides seem to be critical for regulation. For example, although the SgrS sRNA has the potential to form 23 base pairs with the *ptsG* mRNA across a stretch of 32 nucleotides, only four single mutations in SgrS significantly affected downregulation of *ptsG*.

In many cases, the RNA chaperone Hfq is required for *trans*-encoded sRNA-mediated regulation, presumably to facilitate RNA-RNA interactions due to limited complementarity between the sRNA and target mRNA. The hexameric Hfq ring, which is homologous to Sm and Sm-like proteins involved in splicing and mRNA decay in eukaryotes, may actively remodel the RNAs to melt inhibitory secondary structures. Hfq also may serve passively as a platform to allow sRNAs and mRNAs to sample potential complementarity, effectively increasing the local concentrations of sRNAs and mRNAs. It should be noted that when the *E. coli* SgrS RNA is pre-annealed with the *ptsG* mRNA in vitro, the Hfq protein is no longer required. However, in vivo in *E. coli*, sRNAs no longer regulate their target mRNAs in *hfq* mutant strains, and all *trans*-encoded base pairing sRNAs examined to date co-immunoprecipitate with Hfq. In fact, enrichment of sRNAs by co-immunoprecipitation with Hfq proved to be a fruitful approach to identify and validate novel sRNAs in *E. coli* and has been extended to other bacteria, such as *S. typhimurium*.

Beyond facilitating base pairing, Hfq contributes to sRNA regulation through modulating sRNA levels. Somewhat counterintuitively, most *E. coli* sRNAs are less stable in the absence of Hfq, presumably because Hfq protects sRNAs from degradation in the absence of base pairing with mRNAs. Once base paired with target mRNAs, many of the known sRNA-mRNA pairs are subject to degradation by RNase E, and Hfq may also serve to recruit RNA degradation machinery through its interactions with RNase E and other components of the degradosome. In addition, competition between sRNAs for binding to Hfq may be a factor controlling sRNA activity in vivo.

Although all characterized *E. coli trans*-encoded sRNAs require Hfq for regulation of their targets, the need for an RNA chaperone may not be universal. For example, VrrA RNA repression of OmpA protein expression in *V. cholerae* is not eliminated in *hfq* mutant cells, though the extent of repression is higher in cells expressing Hfq. In general, longer stretches of base pairing, as is the case for the *cis*-encoded antisense sRNAs that usually do not require Hfq for function, and/or high concentrations of the sRNA may obviate a chaperone requirement.

In contrast to *cis*-encoded sRNAs, several of which are expressed constitutively, most of the *trans*-encoded sRNAs are synthesized under very specific growth conditions. In *E. coli* for example, these regulatory RNAs are induced by low iron (Fur-repressed RyhB), oxidative stress (OxyR-activated OxyS), outer membrane stress (σ^E-induced MicA and RybB), elevated glycine (GcvA-induced GcvB), changes in glucose concentration (CRP-repressed Spot42 and CRP-activated CyaR), and elevated glucose-phosphate levels (SgrR-activated SgrS). In fact, it is possible that every major transcription factor in *E. coli* may control the expression of one or more sRNA regulators. It is also noteworthy that a number of the sRNAs are encoded adjacent to the gene encoding their transcription regulator, including *E. coli* OxyR-OxyS, GcvA-GcvB, and SgrR-SgrS.

The fact that a given base pairing sRNA often regulates multiple targets means that a single sRNA can globally modulate a particular physiological response, in much the same manner as a transcription factor, but at the post-transcriptional level. Well-characterized regulatory effects of these sRNAs include the down regulation of iron-sulfur cluster containing enzymes under conditions of low iron (*E. coli* RyhB), repression of outer membrane porin proteins under conditions of membrane stress (*E. coli* MicA and RybB), and repression of quorum sensing at low cell density (*Vibrio* Qrr). The fact that direct or indirect negative feedback regulation is observed for a number of sRNAs emphasizes that sRNAs are integrated into regulatory circuits. In *E. coli* for example, *ryhB* is repressed when iron is released after RyhB down-regulates iron-sulfur enzymes, and *micA* and *rybB* are repressed when membrane stress is relieved upon their down-regulation of outer membrane porins. As another example, the Qrr sRNAs in *Vibrio* base pair with and inhibit expression of the mRNAs encoding the transcription factors responsible for the activation of the *qrr* genes.

CRISPR RNAs

A unique class of recently discovered regulatory RNAs is the CRISPR RNAs, which provide resistance to bacteriophage and prevent plasmid conjugation. CRISPR systems share certain similarities with eukaryotic siRNA-driven gene silencing, although they exhibit distinct features as well, and present an exciting new arena of RNA research. The CRISPR sequences have been found in ~40% of bacteria and ~90% of archaea sequenced to date, emphasizing their wide-ranging importance.

CRISPR sequences (**C**lustered **R**egularly **I**nterspaced **S**hort **P**alindromic **R**epeats) are highly variable DNA regions which consist of a ~550 bp leader sequence followed by a series of repeat-spacer units. The repeated DNA can vary from 24 to 47 base pairs, but the same repeat sequence usually appears in each unit in a given CRISPR array, and is repeated two to 249 times. The repeat sequences diverge significantly between bacteria, but can be grouped into 12 major types and often contain a short 5–7 base pair palindrome. Unlike other repeated sequences in bacterial chromosomes, the CRISPR repeats are regularly interspersed with unique spacers of 26 to 72 base pairs; these spacers are not typically repeated in a given CRISPR array. Although the repeats can be similar between species, the spacers between the repeats are not conserved at all, often varying even between strains, and are most often found to be homologous to DNA from phages and plasmids, an observation that was initially perplexing.

Gene Arrangement and Regulatory Functions of CRISPR RNAs

Adjacent to the CRISPR DNA array are several CRISPR-associated (CAS) genes. Two to six core CAS genes seem to be associated with most CRISPR systems, but different CRISPR subtypes also have specific CAS genes encoded in the flanking region. Other CAS genes, that are never present in strains lacking the repeats, may be found in genomic locations distant from the CRISPR region(s). The molecular functions of the CAS proteins are still mostly obscure, but they often contain RNA- or DNA-binding domains, helicase motifs, and endo- or exonuclease domains.

After the initial report of CRISPR sequences in 1989, several different hypotheses were advanced as to possible functions of these repeats. The proposal that CRISPRs confer resistance to phages came in 2005 with findings that the spacers often contain homology to phage or plasmids. Another major advance was the discovery that the CRISPR DNA arrays are transcribed in bacteria and archaea. The full-length CRISPR RNA initially extends the length of the entire array, but is subsequently processed into shorter fragments the size of a single repeat-spacer unit. Recently, it was shown that the *E. coli* K12 CasA-E proteins associate to form a complex termed Cascade, for **C**RISPR-**As**sociated **C**omplex for **A**ntiviral **De**fense. The CasE protein within the Cascade complex is responsible for processing of the full-length CRISPR RNA transcript.

Importantly, it was demonstrated that new spacers corresponding to phage sequences are integrated into existing CRISPR arrays during phage infection and that these new spacers confer resistance to subsequent infections with the cognate phage, or other phage bearing the same sequence. The new spacers are inserted at the beginning of the array, such that the 5' end of the CRISPR region is hypervariable between strains and conveys information about the most recent phage infections, while the 3' end spacers are consequences of more ancient infections. Single nucleotide point mutations in the bacterial spacers or the phage genome abolish phage resistance and, further, introduction of novel phage sequences as spacers in engineered CRISPR arrays provides de novo immunity to bacteria that have never encountered this phage. Similar observations were recently made for spacers found to correspond to sequences present on conjugative plasmids.

These findings, together with the observation that some CAS genes encode proteins with functions potentially analagous to eukaryotic RNAi enzymes, have led to a model for CRISPR RNA function. The CRISPR DNA array is transcribed into a long RNA, which is processed by the Cascade complex of CAS proteins into a single repeat-spacer unit known as a crRNA. The crRNAs, which are single-stranded unlike double-stranded siRNAs, are retained in the Cascade complex. By analogy with eukaryotic RNAi systems, Cascade or other CAS effector proteins may then direct base pairing of the crRNA spacer sequence with phage or plasmid nucleic acid targets. Until recently, it was not known whether the crRNAs would target DNA or RNA, but CRISPR spacers generated from both strands of phage genes can effectively confer phage resistance. In addition, the insertion of an intron into the target gene DNA in a conjugative plasmid abolishes interference by crRNAs, even though the uninterrupted target sequence is regenerated in the spliced mRNA. These results all point to DNA as the direct target, but how the crRNAs interact with the DNA and what occurs subsequently are still unknown. Further studies addressing the details of the molecular mechanism behind CRISPR RNA-mediated "silencing" of foreign DNA and how new spacers are selected and then acquired are eagerly anticipated and will provide further insight into the similarities and differences with the eukaryotic RNAi machinery.

The CRISPR system has broad evolutionary implications. The extreme variability of CRISPR arrays between organisms and even strains of the same species provides useful tools for researchers to genotype strains and to study horizontal gene transfer and micro-evolution. The CRISPR loci record the history of recent phage infection and allow differentiation between strains of the same species. This property can be used to identify pathogenic bacterial strains and track disease progression world-wide, as well as to monitor the population dynamics of non-pathogenic bacteria. Additionally, the presence of phage sequences within the CRISPR arrays that confer resistance against infection provide a strong selective pressure for the mutation of phage genomes and may partially underlie the rapid phage mutation rate.

Dual Function RNAs

The distinctions between some of the categories of RNA regulators discussed above as well as between the RNA regulators and other RNAs can be blurry. For example, a few of the *trans*-encoded base pairing sRNAs encode proteins in addition to base pairing with target mRNAs. The *S. aureus* RNAIII has been shown to base pair with mRNAs encoding virulence factors and a transcription factor, but also encodes a 26 amino acid δ-hemolysin peptide. Similarly, the *E. coli* SgrS RNA, which blocks translation of the *ptsG* mRNA encoding a sugar-phosphate transporter, is translated to produce the 43 amino acid SgrT protein. In this case, the SgrT protein is thought to reinforce the regulation exerted by SgrS by independently down-regulating glucose uptake through direct or indirect inhibition of the PtsG protein. We predict that other regulatory sRNAs will be found to encode small proteins and that conversely some mRNAs encoding small proteins will be found to have additional roles as sRNA regulators. It also deserves mention that some of the *cis*-encoded antisense sRNAs, in addition to regulating their cognate sense mRNA, may base pair with other mRNAs via limited complementarity or, in independent roles, bind proteins to affect other functions. Similarly, while riboswitches are synthesized as part of an mRNA, the small transcripts that are generated by transcription attenuation or autocleavage potentially could go on to perform other functions as their own entities.

Factors Influencing Regulation by RNAs

While there has been a great explosion in the discovery and characterization of RNA regulators in the past ten years, a number of critical questions about their regulatory mechanisms remain to be answered.

RNA Structures, Levels, and Localization

What are the structures of the RNAs and how do they impact ligand, protein and mRNA binding? Three-dimensional structures for several riboswitches, both in the presence and absence of their respective ligands, have been solved in recent years. These studies have shown that some riboswitches have a single, localized ligand-binding pocket. In these cases, the conformational changes induced by ligand binding are confined to a small region. In other riboswitches, the ligand-binding site is comprised of at least two distinct sites, such that ligand binding results in more substantial changes in the global tertiary structure. In contrast, no three-dimensional structures have been solved for bacterial sRNAs. In fact, the secondary structures for only a limited number of sRNAs have been probed experimentally. Another generally unknown quantity, which has important implications for how an RNA interacts with other molecules, is the concentration of the RNA. After induction, the OxyS RNA has been estimated to be present at 4,500 molecules per cell, but it is not known whether this is typical for other sRNAs and whether all of the sRNA molecules are active. Do nucleotide modifications or metabolite binding alter the abundance or activities of any of the sRNAs? It is also intriguing to ask whether any of the regulatory RNAs show specific subcellular localization or are even secreted. In eukaryotes, localization of regulatory RNAs to specific subcellular structures, such as P bodies and Cajal bodies, is connected to their functions. It is plausible that subcellular localization similarly impacts regulatory RNA function in bacteria. In support of this idea, RNase E has been found to bind membranes in vitro, and membrane targeting of the *ptsG* mRNA-encoded protein is required for efficient SgrS sRNA repression of this transcript. Another attractive, but untested, hypothesis is that bacterial RNAs might be secreted into a host cell where they could modulate eukaryotic cell functions.

Proteins Involved

What proteins are associated with regulatory RNAs and how do the proteins impact the actions of the RNAs? So far much of the attention has been focused on the RNA chaperone Hfq. Even so, the details of how this protein binds to sRNAs and impacts their functions are murky. For example, structural and mutational studies indicate that both faces of the donut-like Hfq hexamer can make contacts with RNA, but it is not clear whether the sRNA and mRNA bind both faces simultaneously, whether the sRNA and mRNA bind particular faces, and whether base pairing is facilitated by changes in RNA structure or proximity between the two RNAs or both. The Hfq protein has been shown to copurify with the ribosomal protein S1, components of the RNase E degradosome, and polynucleotide phosphorylase, among others, but these are all abundant RNA-binding proteins and the in vivo relevance of these interactions is poorly understood. In addition, only half of all sequenced Gram-negative and Gram-positive species and one archaeon have Hfq homologs. Do other proteins substitute for Hfq in the organisms that do not have homologs, or does base pairing between sRNAs and their target mRNAs not require an RNA chaperone in these cases?

It is likely still other proteins that act on or in conjuction with the regulatory RNAs remain to be discovered. The RNase E and RNase III endonucleases are known to cleave base pairing sRNAs and their targets, but these may not be the only ribonucleases to degrade the RNAs. Pull-down experiments with tagged sRNAs indicate that other proteins, such as RNA polymerase, also bind the RNA regulators, but again the physiological relevance of this interaction is not known. In addition, genetic studies hint at the involvement of proteins such as YhbJ, which antagonizes GlmY and GlmZ activity, though the activity of this protein is still mysterious.

Requirements for Productive Base Pairing

What are the rules for productive base pairing? *Trans*-encoded sRNAs bind to their target mRNAs using discontiguous and imperfect base pairing, of which often only a core set of interactions is essential, stimulating questions as to how specificity between sRNAs and mRNAs is imparted and how such limited pairing can cause translation inhibition or RNA degradation. Several algorithms for the predictions of base pairing targets for *trans*-encoded sRNAs have been developed and reviewed in. However, the accuracy of these predictions has been varied. For some sRNAs, such as RyhB and GcvB, there are distinct conserved single-stranded regions, which appear to be required for base pairing with most targets and are associated with more accurate predictions. For other sRNAs such as OmrA and OmrB, few known targets were predicted in initial searches. Mutational studies to define the base pairing interactions with known OmrA and OmrB targets highlight possible impediments to computational predictions. These can include the lack of knowledge about the sRNA domains required for base pairing, limited base pairing interactions, and base pairing to mRNA regions outside the immediate vicinity of the ribosome binding site. Recent systematic analysis indicates sRNAs can block translation by pairing with sequences in the coding region, as far downstream as the fifth codon. Other factors such as the position of Hfq binding and the secondary structures of both the mRNA and sRNA are also likely to impact base pairing in ways that have not been formalized. In vitro studies exploring the role of Hfq in facilitating the pairing between the RprA and DsrA RNAs and the *rpoS* mRNA show that binding between Hfq, the mRNA and the sRNAs is clearly influenced by what portion of the *rpoS* 5' leader is assayed. With an increasing number of validated targets that can serve as training sets, the ability to accurately predict targets should significantly improve.

As with eukaryotic miRNAs and siRNAs, there may be mechanistic differences between the *trans*- and *cis*-encoded base pairing sRNAs based on their different properties. *Trans*-encoded sRNAs, which have imperfect base pairing with their targets like miRNAs, often interact with Hfq. In contrast, *cis*-encoded sRNAs, which have complete complementarity with targets like siRNAs, do not appear to require Hfq, but tend to be more structured and may use other factors to aid in base pairing. These differences may have broader implications for the types of targets regulated, the nature of the proteins required, as well as the mechanistic details of base pairing.

New Mechanisms of Action

What novel mechanisms of action remain to be uncovered? Most sRNAs characterized to date base pair in the 5’ UTR of target mRNAs near the ribosome binding site, however other locations for base pairing and consequent mechanisms of regulation are possible. Only a few bacterial ribozymes have been described. Will other sRNAs or riboswitches be found to have enzymatic activity? As already alluded to, the mechanism of crRNA action in targeting and interfering with DNA is not understood. Completely novel mechanisms may be revealed by further studies of the CRISPR sequences. Finally, nearly a third of the *E. coli* sRNAs identified to date, and the vast majority of those in other organisms, have yet to be characterized in significant detail. These too may have unanticipated roles and modes of action.

Physiological Roles of Regulatory RNAs

In addition to further exploring the mechanisms by which riboswitches, sRNAs and crRNAs act, it is worth reflecting on what is known, as well as what is not understood, about the physiological roles of these regulators.

Association with Specific Responses

A number of themes are emerging with respect to the physiological roles of riboswitches and sRNAs. In general terms, riboswitches, protein binding sRNAs, *trans*-encoded base pairing sRNAs and some *cis*-encoding base pairing sRNAs mediate responses to changing environmental conditions by modulating metabolic pathways or stress responses. Riboswitches and T-boxes tend to regulate biosynthetic genes, as these elements directly sense the concentrations of various metabolites, while some RNA thermometers, such as the

5'-UTR of the mRNA encoding the heat shock sigma factor, control transcriptional regulators. The CsrB and 6S families of sRNAs also control the expression of large numbers of genes in response to decreases in nutrient availability by repressing the activities of global regulators. The *trans*-encoded base pairing sRNAs mostly contribute to the ability to survive various environmental insults by modulating the translation of regulators or repressing the synthesis of unneeded proteins. In particular, it is intriguing that a disproportionate number of *trans*-encoded sRNAs regulate outer membrane proteins (MicA, MicC, MicF, RybB, CyaR, OmrA and OmrB) or transporters (SgrS, RydC, GcvB). Other pervasive themes include RNA-mediated regulation of iron metabolism, not only in bacteria but also in eukaryotes, as well as RNA regulators of quorum sensing.

Pathogenesis presents a set of behaviors one might expect to be regulated by sRNAs since bacterial infections involve multiple rounds of rapid and coordinated responses to changing conditions. The central role of sRNAs in modulating the levels of outer membrane proteins, which are key targets for the immune system, as well as other responses important for survival under conditions found in host cells, such as altered iron levels, also implicates these RNA regulators in bacterial survival in host cells. Indeed, although these studies are still at the early stages, several sRNAs have been shown to alter infection. These include members of the CsrB family of sRNAs in *Salmonella*, *Erwinia*, *Yersinia*, *Vibrio* and *Pseudomonads* which bind to and antagonize CsrA family proteins that are global regulators of virulence genes; RyhB of *Shigella* which represses a transcriptional activator of virulence genes; RNAIII of *Staphylococcus* which both base pairs with mRNAs encoding virulence factors and encodes the δ-hemolysin peptide; and the Qrr sRNAs of *Vibrio* which regulate quorum and *hfq* mutants of a wide range of bacteria also show reduced virulence. Some sRNAs, such as a number of sRNAs encoded in *Salmonella* and *Staphylococcus* pathogenicity islands, show differential expression under pathogenic conditions. Other sRNAs, such as five in *Listeria monocytogenes*, are specific to pathogenic strains. Finally, thermosensors and riboswitches can have roles in as regulators of pathogenesis, upregulating virulence genes upon increased temperature encountered in host cells or upon binding signals such as the "second messenger" cyclic di-GMP. Further studies of these and other pathogenesis-associated regulatory RNAs could lead to opportunities for interfering with disease.

A subset of the *cis*-encoded antisense sRNAs expressed from bacterial chromosomes act as antitoxins but their physiological roles are not clear. They may also be involved in altering cell metabolism in response to various stresses enabling survival. Alternatively, they may play a role in protecting against foreign DNA. This is clearly the function of CRISPR RNAs, which have been demonstrated to repress bacteriophage and plasmid entry into the cell, and in principle could be used to silence genes from other mobile elements.

Physiological Roles of Multiple Copies

Some sRNAs including OmrA/OmrB, Prr1/Prr2, Qrr1–5, 6S homologs, CsrB homologs, GlmY/GlmZ, and several toxin-antitoxin modules are present in multiple copies in a given bacterium. Although the physiological advantages of the repeated sRNA genes are only understood in a subset of cases, multiple copies can have several different roles.

Possible Roles of Duplicated RNA Genes

Firstly, homologous RNAs can act redundantly, serving as back ups in critical pathways or to increase the sensitivity of a response. In *V. cholerae*, any single Qrr RNA is sufficient to repress quorum sensing by down regulating the HapR transcription factor, and the deletion of all four *qrr* genes is required to constitutively activate the quorum sensing behaviors. Since the effectiveness of sRNA regulation is directly related to their abundance relative to mRNA targets, this redundancy has been proposed to permit an ultrasensitive, switch-like response for quorum sensing in *V. cholerae* and may help amplify a small input signal to achieve a large output.

Secondly, repeated RNAs can act additively, as in the case of the *V. harveyi* Qrr sRNAs. In this case, the five *qrr* genes have divergent promoter regions and are differentially expressed, suggesting each Qrr sRNA may respond to different metabolic indicators to integrate various environmental signals. Deletion of individual Qrr genes affects the extent of quorum sensing behaviors, indicating they do not act redundantly. Rather, the total amount of Qrr sRNAs in *V. harveyi* produces distinct levels of regulated genes, such that altering the abundance of any given Qrr sRNA changes the extent of the response. This additive regulation is thought to allow fine-tuning of *luxR* levels across a gradient of expression, leading to precise, tailored amounts of gene expression. It is surprising that within the same quorum sensing system in two related species of *Vibrio*, the multiple Qrr sRNAs operate according to two distinct mechanisms. While the reason for this is not clear, the difference illustrates the evolvability of RNA regulators and the regulatory nuances that can be provided by having multiple copies.

A third possibility is that the duplicated RNAs can act independently of each other. This could occur in several ways. For base pairing sRNAs, each sRNA could regulate a different set of genes, most likely in a somewhat overlapping manner. For protein-binding sRNAs, different homologs could interact with distinct proteins, giving rise to variations in the core complexes. As mentioned above, various *B. subtilis* 6S isoforms could repress RNA polymerase bound to different σ factors. Homologous RNA species also can employ very different mechanisms of action, as observed for the *E. coli* GlmY and GlmZ RNAs. GlmZ functions by base pairing, while GlmY likely acts as a mimic to titrate away YhbJ and other factors that inactive GlmZ.

In some cases it is still perplexing why multiple copies are maintained. One example is the toxin-antitoxin modules, which are not only encoded by multiple genes in *E. coli* chromosomes, but which can vary in gene number even within the same species. Redundant RNAs may simply indicate a recent evolutionary event, which has not yet undergone variation to select new functions. Alternatively, additional genes may be selected by the pressure to maintain at least one copy across a population. Complete answers to the question of why various regulatory RNA genes are duplicated await more characterization of each set of RNAs.

Advantages of Regulatory RNAs

RNA regulators have several advantages over protein regulators. They are less costly to the cell and can be faster to produce, since they are shorter than most mRNAs (~100–200 nucleotides compared to 1,000 nucleotides for the average ~350 amino acid *E. coli* protein) and do not require the extra step of translation.

The effects of the RNA regulators themselves also can be very fast. For *cis*-acting riboswitches, the coupling of a sensor directly to an mRNA allows a cell to respond to the signal in an extremely rapid and sensitive manner. Similarly, since sRNAs are faster to produce than proteins and act post-transcriptionally, it was anticipated that, in the short term, they could shut off or turn on expression more rapidly than protein-based transcription factors. Indeed this expectation is supported by some dynamic simulations. Other unique aspects of sRNA regulation revealed by recent modeling studies are related to the threshold linear response provided by sRNAs, in contrast to the straight linear response provided by transcription factors. Most sRNAs characterized thus far act stoichiometrically through the noncatalytic mechanisms of mRNA degradation or competitive inhibition of translation, reactions in which the relative concentrations of the sRNA and mRNA are critical. Thus for negatively-acting sRNAs, when [sRNA] ≫ [mRNA], gene expression is tightly shut off, but when [mRNA] ≫ [sRNA], the sRNA has little effect on expression. This threshold property of sRNA repression suggests that sRNAs are not generally as effective as proteins at transducing small or transient input signals. In contrast, when input signals are large and persistent, sRNAs are hypothesized to be better than transcription factors at strongly and reliably repressing proteins levels, as well as at filtering noise. Moreover, sRNA-based regulation is thought to be ultra-sensitive to changes in sRNA and mRNA levels around the critical threshold, especially in the case of multiple, redundant sRNAs as in the *V. cholerae* Qrr quorum sensing system, which is proposed to lead to switch-like “all or nothing” behavior.

Additional features of different subsets of the RNA regulators provide other advantages. Some riboswitches lead to transcription termination or self-cleavage and some base pairing sRNAs direct the cleavage of their targets, rendering their regulatory effects irreversible. For the *cis*-encoded antisense sRNAs and the CRISPR RNAs, the extensive complementarity with the target nucleic acids imparts extremely high specificity. In contrast, the ability of *trans*-encoded sRNAs to regulate many different genes allows these sRNAs to control entire physiological networks with varying degrees of stringency and outcomes. The extent and quality of base pairing with sRNAs can prioritize target mRNAs for differential regulation and could be used by cells to integrate different states into gene expression programs. In addition, when multiple target mRNAs of a given sRNA are expressed in a cell, their relative abundance and binding affinities can strongly influence expression of each other through cross-talk. Conversely, competition between different sRNAs for Hfq or a specific mRNA is likely to alter dynamics within a regulatory network. Finally, base pairing flexibility presumably also allows rapid evolution of sRNAs and mRNA targets.

Moreover, while not an advantage *per se*, RNA regulators usually act at a level complementary to protein regulators, most often functioning at the post-transcriptional level as opposed to transcription factors that act before sRNAs or enzymes such as kinases or proteases that act after sRNAs. Different combinations of these protein and RNA regulators can provide a variety of regulatory outcomes, such as extremely tight repression, an expansion in the genes regulated in response to a single signal or conversely an increase in the number of signals sensed by a given gene (Shimoni et al., 2007).

Evolution of Regulatory RNAs

We do not yet know whether all bacteria contain regulatory RNAs or whether we are coming close to having identified all sRNAs and riboswitches in well-studied bacteria. Given the redundancy in the sRNAs being found, the searches for certain classes of sRNAs, in particular sRNAs encoded in intergenic regions and expressed under typical laboratory conditions, appear near saturation in *E. coli*. However, other types of sRNAs, such as *cis*-encoded antisense sRNAs and sRNAs whose expression is tightly regulated, may still be missing from the lists of identified RNA regulators.

Are RNA regulators remnants of the RNA world or are the genes recent additions to bacterial genomes? We propose that the answer to this question is both. Some of the regulators such as riboswitches and CRISPR systems, which are very broadly conserved, are likely to have ancient evolutionary origins. In contrast, while regulation by base pairing may long have been in existence, individual antisense regulators, both *cis*- and *trans*-encoded sRNAs may be recently acquired and rapidly evolving. This is exemplified by the poor conservation of sRNA sequences across bacteria. For example, the Prr RNAs of *Pseudomonas* bear almost no resemblance to the equivalent RyhB sRNA of *E. coli* although both are repressed by Fur and act on similar targets (Wilderman et al., 2004). One might imagine that the expression of a spurious transcript, either antisense or with limited complementarity to a bona fide mRNA, which provides some selective advantage could easily be fixed in a population.

It is intriguing to note that distinct RNA regulators have been used to solve specific regulatory problems, emphasizing the pervasiveness and adaptability of RNA-mediated regulation. For example, in *B. subtilis*, the *glmS* mRNA is inactivated by the self-cleavage of the glucosamine-6-phosphate-responsive *cis*-acting riboswitch (Collins et al., 2007), whereas in *E. coli*, the *glmS* mRNA is positively regulated by the two *trans*-acting sRNAs GlmY and GlmZ. As another example, RyhB-like *trans*-encoded sRNAs repress the expression of iron-containing enzymes during iron starvation in various bacteria, while the *cis*-encoded IsiR sRNA of *Synechocystis* represses expression of the IsiA protein, a light harvesting antenna, under iron replete conditions .

Applications of Regulatory RNAs

The central roles played by RNA regulators in cellular physiology make them attractive for use as tools to serve as biosensors or to control bacterial growth either positively or negatively. Endogenous RNAs could serve as signals of the environmental status of the cell. For example, the levels of the RyhB and OxyS sRNAs, respectively, are powerful indicators of the iron status and hydrogen peroxide concentration in a cell. CRISPR sequences provide insights into the history of the extracellular DNA encountered by the bacteria and have been used to genotype strains during infectious disease outbreaks. Regarding the control of bacterial cell growth, one can imagine how riboswitches might be exploited as drug targets given their potential to bind a wide variety of compounds. Similarly, since interference with the functions of some of the sRNAs is detrimental to growth and several sRNAs contribute to virulence, these regulators and their interacting proteins also could be targeted by antibacterial therapies. Alternatively, ectopic expression of specific regulatory RNAs might be used to increase stress resistance and facilitate bacterial survival in various industrial or ecological settings.

RNA also presents a powerful system for rational design as it is modular, easily synthesized and manipulated, and can attain an enormous diversity of sequence, structure, and function. Although less developed than in eukaryotes, the application of synthetic RNAs is being explored in bacteria. For example, riboswitch elements have been engineered to use novel ligands, and sRNAs have been designed to base pair with novel transcripts. Engineered CRISPR repeats present an obvious mechanism by which to repress uptake of specific DNA sequences. Limitations to these approaches include incomplete repression observed for the synthetic riboswitches and base pairing sRNAs thus far, off target effects, as well as problems in delivering the RNA regulators into cells where they might be of greatest utility. Nevertheless, synthetic RNAs have potential to provide a variety of useful tools and therapeutics in the future.

CHAPTER # 19 REGULATION OF GENE EXPRESSION

Regulation of gene expression includes a wide range of mechanisms that are used by cells to increase or decrease the production of specific gene products (protein or RNA), and is informally termed *gene regulation*. Sophisticated programs of gene expression are widely observed in biology, for example to trigger developmental pathways, respond to environmental stimuli, or adapt to new food sources. Virtually any step of gene expression can be modulated, from transcriptional initiation, to RNA processing, and to the post-translational modification of a protein.

Gene regulation is essential for viruses, prokaryotes and eukaryotes as it increases the versatility and adaptability of an organism by allowing the cell to express protein when needed. Although as early as 1951, Barbara McClintock showed interaction between two genetic loci, Activator (*Ac*) and Dissociator (*Ds*), in the color formation of maize seeds, the first discovery of a gene regulation system is widely considered to be the identification in 1961 of the *lac* operon, discovered by Jacques Monod, in which some enzymes involved in lactose metabolism are expressed by *E. coli* only in the presence of lactose and absence of glucose.

Furthermore, in multicellular organisms, gene regulation drives the processes of cellular differentiation and morphogenesis, leading to the creation of different cell types that possess different gene expression profiles, and hence produce different proteins/have different ultrastructures that suit them to their functions (though they all possess the genotype, which follows the same genome sequence).

The initiating event leading to a change in gene expression include activation or deactivation of receptors. Also, there is evidence that changes in a cell's choice of catabolism leads to altered gene expressions

Regulated stages of gene expression

Any step of gene expression may be modulated, from the DNA-RNA transcription step to post-translational modification of a protein. The following is a list of stages where gene expression is regulated, the most extensively utilised point is Transcription Initiation:

- Chromatin domains
- Transcription
- Post-transcriptional modification
- RNA transport
- Translation
- mRNA degradation

Modification of DNA

In eukaryotes, the accessibility of large regions of DNA can depend on its chromatin structure, which can be altered as a result of histone modifications directed by DNA methylation, ncRNA, or DNA-binding protein. Hence these modifications may up or down regulate the expression of a gene. Some of these modifications that regulate gene expression are inheritable and are referred to as epigenetic regulation.

Structural

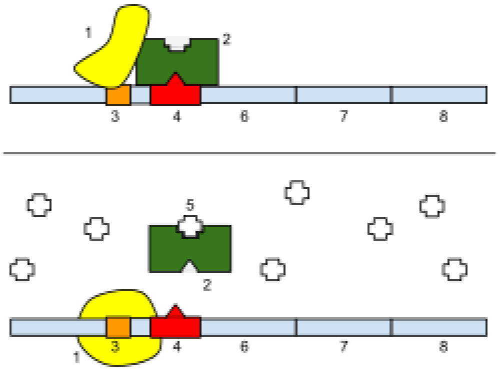
Transcription of DNA is dictated by its structure. In general, the density of its packing is indicative of the frequency of transcription. Octameric protein complexes called nucleosomes are responsible for the amount of supercoiling of DNA, and these complexes can be temporarily modified by processes such as phosphorylation or more permanently modified by processes such as methylation. Such modifications are considered to be responsible for more or less permanent changes in gene expression levels.

Chemical

Methylation of DNA is a common method of gene silencing. DNA is typically methylated by methyltransferase enzymes on cytosine nucleotides in a CpG dinucleotide sequence (also called "CpG islands" when densely clustered). Analysis of the pattern of methylation in a given region of DNA (which can be a promoter) can be achieved through a method called bisulfite mapping. Methylated cytosine residues are unchanged by the treatment, whereas unmethylated ones are changed to uracil. The differences are analyzed by DNA sequencing or by methods developed to quantify SNPs, such as Pyrosequencing (Biotage) or MassArray (Sequenom), measuring the relative amounts of C/T at the CG dinucleotide. Abnormal methylation patterns are thought to be involved in oncogenesis.

Histone acetylation is also an important process in transcription. Histone acetyltransferase enzymes (HATs) such as CREB-binding protein also dissociate the DNA from the histone complex, allowing transcription to proceed. Often, DNA methylation and histone deacetylation work together in gene silencing. The combination of the two seems to be a signal for DNA to be packed more densely, lowering gene expression.

Regulation of transcription



***1*: RNA Polymerase, *2*: Repressor, *3*: Promoter, *4*: Operator, *5*: Lactose, *6*: lacZ, *7*: lacY, *8*: lacA. Top**: The gene is essentially turned off. There is no lactose to inhibit the repressor, so the repressor binds to the operator, which obstructs the RNA polymerase from binding to the promoter and making lactase. **Bottom**: The gene is turned on. Lactose is inhibiting the repressor, allowing the RNA polymerase to bind with the promoter, and express the genes, which synthesize lactase. Eventually, the lactase will digest all of the lactose, until there is none to bind to the repressor. The repressor will then bind to the operator, stopping the manufacture of lactase.

Regulation of transcription thus controls when transcription occurs and how much RNA is created. Transcription of a gene by RNA polymerase can be regulated by at least five mechanisms:

- **Specificity factors** alter the specificity of RNA polymerase for a given promoter or set of promoters, making it more or less likely to bind to them (i.e., sigma factors used in prokaryotic transcription).

- **Repressors** bind to the **Operator**, coding sequences on the DNA strand that are close to or overlapping the promoter region, impeding RNA polymerase's progress along the strand, thus impeding the expression of the gene.The image to the right demonstrates regulation by a repressor in the lac operon.
- **General transcription factors** position RNA polymerase at the start of a protein-coding sequence and then release the polymerase to transcribe the mRNA.
- **Activators** enhance the interaction between RNA polymerase and a particular promoter, encouraging the expression of the gene. Activators do this by increasing the attraction of RNA polymerase for the promoter, through interactions with subunits of the RNA polymerase or indirectly by changing the structure of the DNA.
- **Enhancers** are sites on the DNA helix that are bound by activators in order to loop the DNA bringing a specific promoter to the initiation complex. Enhancers are much more common in eukaryotes than prokaryotes, where only a few examples exist (to date).
- **Silencers** are regions of DNA sequences that, when bound by particular transcription factors, can silence expression of the gene.

Post-transcriptional regulation

After the DNA is transcribed and mRNA is formed, there must be some sort of regulation on how much the mRNA is translated into proteins. Cells do this by modulating the capping, splicing, addition of a Poly(A) Tail, the sequence-specific nuclear export rates, and, in several contexts, sequestration of the RNA transcript. These processes occur in eukaryotes but not in prokaryotes. This modulation is a result of a protein or transcript that, in turn, is regulated and may have an affinity for certain sequences.

Three prime untranslated regions and microRNAs

Three prime untranslated regions (3'-UTRs) of messenger RNAs (mRNAs) often contain regulatory sequences that post-transcriptionally influence gene expression. Such 3'-UTRs often contain both binding sites for microRNAs (miRNAs) as well as for regulatory proteins. By binding to specific sites within the 3'-UTR, miRNAs can decrease gene expression of various mRNAs by either inhibiting translation or directly causing degradation of the transcript. The 3'-

UTR also may have silencer regions that bind repressor proteins that inhibit the expression of a mRNA.

The 3'-UTR often contains miRNA response elements (MREs). MREs are sequences to which miRNAs bind. These are prevalent motifs within 3'-UTRs. Among all regulatory motifs within the 3'-UTRs (e.g. including silencer regions), MREs make up about half of the motifs.

As of 2014, the miRBase web site, an archive of miRNA sequences and annotations, listed 28,645 entries in 233 biologic species. Of these, 1,881 miRNAs were in annotated human miRNA loci. miRNAs were predicted to have an average of about four hundred target mRNAs (affecting expression of several hundred genes). Freidman et al. estimate that >45,000 miRNA target sites within human mRNA 3'-UTRs are conserved above background levels, and >60% of human protein-coding genes have been under selective pressure to maintain pairing to miRNAs.

Direct experiments show that a single miRNA can reduce the stability of hundreds of unique mRNAs.[7] Other experiments show that a single miRNA may repress the production of hundreds of proteins, but that this repression often is relatively mild (less than 2-fold).

The effects of miRNA dysregulation of gene expression seem to be important in cancer. For instance, in gastrointestinal cancers, a 2015 paper identified nine miRNAs as epigenetically altered and effective in down-regulating DNA repair enzymes.

The effects of miRNA dysregulation of gene expression also seem to be important in neuropsychiatric disorders, such as schizophrenia, bipolar disorder, major depressive disorder, Parkinson's disease, Alzheimer's disease and autism spectrum disorders.[12][13][14]

Regulation of translation

The translation of mRNA can also be controlled by a number of mechanisms, mostly at the level of initiation. Recruitment of the small ribosomal subunit can indeed be modulated by mRNA secondary structure, antisense RNA binding, or protein binding. In both prokaryotes and eukaryotes, a large number of RNA binding proteins exist, which often are directed to their target sequence by the secondary structure of the transcript, which may change depending on certain conditions, such as temperature or presence of a ligand (aptamer). Some transcripts act as ribozymes and self-regulate their expression.

Examples of gene regulation

- Enzyme induction is a process in which a molecule (e.g., a drug) induces (i.e., initiates or enhances) the expression of an enzyme.
- The induction of heat shock proteins in the fruit fly *Drosophila melanogaster*.
- The Lac operon is an interesting example of how gene expression can be regulated.
- Viruses, despite having only a few genes, possess mechanisms to regulate their gene expression, typically into an early and late phase, using collinear systems regulated by anti-terminators (lambda phage) or splicing modulators (HIV).
- GAL4 is a transcriptional activator that controls the expression of GAL1, GAL7, and GAL10 (all of which code for the metabolic of galactose in yeast). The GAL4/UAS system has been used in a variety of organisms across various phyla to study gene expression.

Developmental biology

Main article: morphogen

A large number of studied regulatory systems come from developmental biology. Examples include:

- The colinearity of the Hox gene cluster with their nested antero-posterior patterning
- It has been speculated that pattern generation of the hand (digits - interdigits) The gradient of Sonic hedgehog (secreted inducing factor) from the zone of polarizing activity in the limb, which creates a gradient of active Gli3, which activates Gremlin, which inhibits BMPs also secreted in the limb, resulting in the formation of an alternating pattern of activity as a result of this reaction-diffusion system.
- Somitogenesis is the creation of segments (somites) from a uniform tissue (Pre-somitic Mesoderm, PSM). They are formed sequentially from anterior to posterior. This is achieved in amniotes possibly by means of two opposing gradients, Retinoic acid in the

anterior (wavefront) and Wnt and Fgf in the posterior, coupled to an oscillating pattern (segmentation clock) composed of FGF + Notch and Wnt in antiphase.

- Sex determination in the soma of a Drosophila requires the sensing of the ratio of autosomal genes to sex chromosome-encoded genes, which results in the production of sexless splicing factor in females, resulting in the female isoform of doublesex.

Circuitry

Up-regulation and down-regulation

Up-regulation is a process that occurs within a cell triggered by a signal (originating internal or external to the cell), which results in increased expression of one or more genes and as a result the protein(s) encoded by those genes. On the converse, **down-regulation** is a process resulting in decreased gene and corresponding protein expression.

- Up-regulation occurs, for example, when a cell is deficient in some kind of receptor. In this case, more receptor protein is synthesized and transported to the membrane of the cell and, thus, the sensitivity of the cell is brought back to normal, reestablishing homeostasis.

- Down-regulation occurs, for example, when a cell is overstimulated by a neurotransmitter, hormone, or drug for a prolonged period of time, and the expression of the receptor protein is decreased in order to protect the cell (see also tachyphylaxis).

Inducible vs. repressible systems

Gene Regulation can be summarized by the response of the respective system:

- Inducible systems - An inducible system is off unless there is the presence of some molecule (called an inducer) that allows for gene expression. The molecule is said to "induce expression". The manner by which this happens is dependent on the control mechanisms as well as differences between prokaryotic and eukaryotic cells.

- Repressible systems - A repressible system is on except in the presence of some molecule (called a corepressor) that suppresses gene expression. The molecule is said to "repress expression". The manner by which this happens is dependent on the control mechanisms as well as differences between prokaryotic and eukaryotic cells.

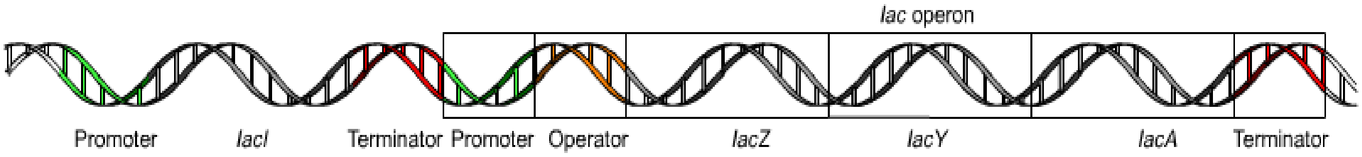
The GAL4/UAS system is an example of both an inducible and repressible system. GAL4 binds an upstream activation sequence (UAS) to activate the transcription of the GAL1/GAL7/GAL10 cassette. On the other hand, a MIG1 response to the presence of glucose can inhibit GAL4 and therefore stop the expression of the GAL1/GAL7/GAL10 cassette.

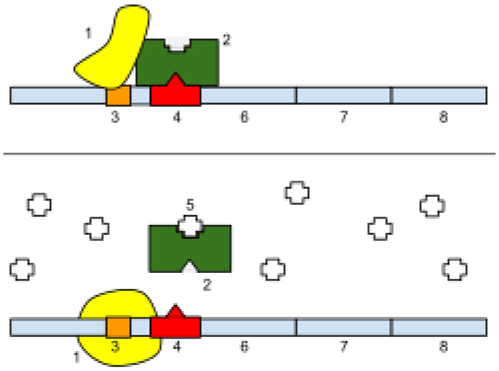
lac **operon**

From Wikipedia, the free encyclopedia

lac **operon** (lactose operon) is an operon required for the transport and metabolism of lactose in *Escherichia coli* and many other enteric bacteria. Although glucose is the preferred carbon source for most bacteria, the *lac* operon allows for the effective digestion of lactose when glucose is not available. Gene regulation of the *lac* operon was the first genetic regulatory mechanism to be understood clearly, so it has become a foremost example of prokaryotic gene regulation. It is often discussed in introductory molecular and cellular biology classes at universities for this reason.

Bacterial operons are polycistronic transcripts that are able to produce multiple proteins from one mRNA transcript. In this case, when lactose is required as a sugar source for the bacterium, the three genes of the lac operon can be expressed and their subsequent proteins translated: *lacZ*, *lacY*, and *lacA*. The gene product of *lacZ* is β-galactosidase which cleaves lactose, a disaccharide, into glucose and galactose. LacY encodes lactose permease, a protein which becomes embedded in the cytoplasmic membrane to enable transport of lactose into the cell. Finally, *lacA* encodes galactoside O-acetyltransferase.

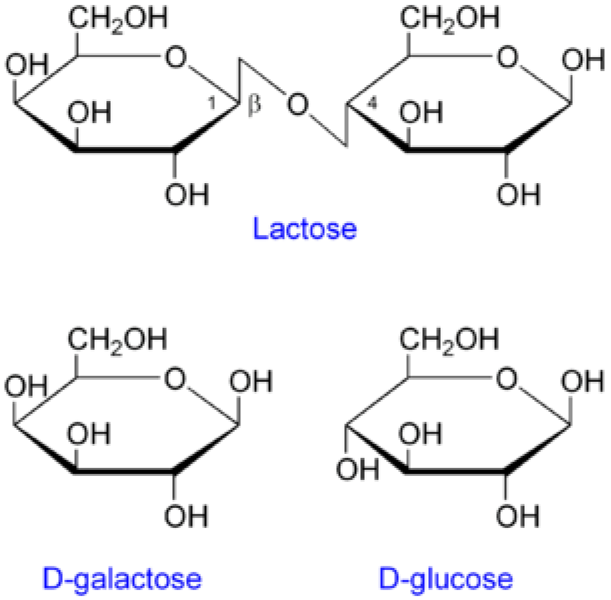




The lac operon. Top:Repressed, Bottom:Active. ***1*****: RNA Polymerase,** ***2*****: Repressor,** ***3*****: Promoter,** ***4*****: Operator,** ***5*****: Lactose,** ***6*****: lacZ,** ***7*****: lacY,** ***8*****: lacA.**

It would be wasteful to produce the enzymes when there is no lactose available or if there is a more preferable energy source available, such as glucose. The *lac* operon uses a two-part control mechanism to ensure that the cell expends energy producing the enzymes encoded by the *lac* operon only when necessary. In the absence of lactose, the *lac* repressor halts production of the enzymes encoded by the *lac* operon. In the presence of glucose, the catabolite activator protein (CAP), required for production of the enzymes, remains inactive, and EIIA^{Glc} shuts down lactose permease to prevent transport of lactose into the cell. This dual control mechanism causes the sequential utilization of glucose and lactose in two distinct growth phases, known as diauxie.

Structure of the *lac* operon



Structure of lactose and the products of its cleavage.

- The *lac* operon consists of three structural genes, and a promoter, a terminator, regulator, and an operator. The three structural genes are: *lacZ, lacY*, and *lacA*.
    - *lacZ* encodes β-galactosidase (LacZ), an intracellular enzyme that cleaves the disaccharide lactose into glucose and galactose.
    - *lacY* encodes lactose permease (LacY), a transmembrane symporter that pumps β-galactosides into the cell using a proton gradient in the same direction.
    - *lacA* encodes galactoside O-acetyltransferase (LacA), an enzyme that transfers an acetyl group from acetyl-CoA to β-galactosides.

Only *lacZ* and *lacY* appear to be necessary for lactose catabolism.

Genetic nomenclature

Three-letter abbreviations are used to describe phenotypes in bacteria including *E. coli*.

Examples include:

- Lac (the ability to use lactose),
- His (the ability to synthesize the amino acid histidine)

- Mot (swimming motility)
- Sm^R (resistance to the antibiotic streptomycin)

In the case of Lac, wild type cells are Lac$^+$ and are able to use lactose as a carbon and energy source, while Lac$^-$ mutant derivatives cannot use lactose. The same three letters are typically used (lower-case, italicized) to label the genes involved in a particular phenotype, where each different gene is additionally distinguished by an extra letter. The *lac* genes encoding enzymes are *lacZ*, *lacY,* and *lacA*. The fourth *lac* gene is *lacI*, encoding the lactose repressor—"I" stands for *inducibility*.

One may distinguish between *structural* genes encoding enzymes, and regulatory genes encoding proteins that affect gene expression. Current usage expands the phenotypic nomenclature to apply to proteins: thus, LacZ is the protein product of the *lacZ* gene, β-galactosidase. Various short sequences that are not genes also affect gene expression, including the *lac* promoter, *lac p*, and the *lac* operator, *lac o*. Although it is not strictly standard usage, mutations affecting *lac o* are referred to as *lac o*^{c}, for historical reasons.

Regulation

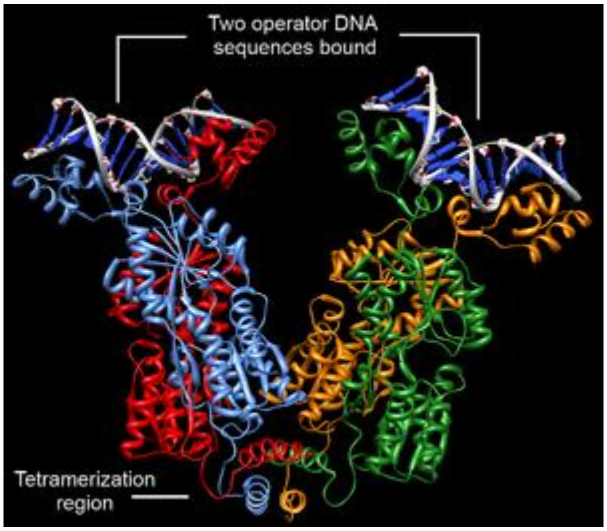
Specific control of the *lac* genes depends on the availability of the substrate lactose to the bacterium. The proteins are not produced by the bacterium when lactose is unavailable as a carbon source. The *lac* genes are organized into an operon; that is, they are oriented in the same direction immediately adjacent on the chromosome and are co-transcribed into a single polycistronic mRNA molecule. Transcription of all genes starts with the binding of the enzyme RNA polymerase (RNAP), a DNA-binding protein, which binds to a specific DNA binding site, the promoter, immediately upstream of the genes. Binding of RNA polymerase to the promoter is aided by the cAMP-bound catabolite activator protein (CAP, also known as the cAMP receptor protein). However, the *lacI* gene (regulatory gene for *lac* operon) produces a protein that blocks RNAP from binding to the promoter of the operon. This protein can only be removed when allolactose binds to it, and inactivates it. The protein that is formed by the *lacI* gene is known as the lac repressor. The type of regulation that the *lac* operon undergoes is referred to as negative inducible, meaning that the gene is turned off by the regulatory factor (*lac* repressor) unless some molecule (lactose) is added. Because of the presence of the *lac* repressor protein, genetic engineers who replace the *lacZ* gene with another gene will have to grow the experimental bacteria on agar with lactose available on it. If they do not, the gene they are trying to express will not be expressed as the repressor protein is still blocking RNAP from binding to the promoter and transcribing the gene. Once the repressor is removed, RNAP then proceeds to transcribe all three genes (*lacZYA*) into mRNA. Each of the three genes on the mRNA strand has its own Shine-Dalgarno sequence, so the genes are independently translated. The DNA sequence of the *E. coli* lac operon, the lacZYA mRNA, and the *lacI* genes are available from GenBank.

The first control mechanism is the regulatory response to lactose, which uses an intracellular *regulatory protein* called the *lactose repressor* to hinder production of β-galactosidase in the absence of lactose. The *lacI* gene coding for the repressor lies nearby the *lac* operon and is always expressed (*constitutive*). If lactose is missing from the growth medium, the repressor binds very tightly to a short DNA sequence just downstream of the promoter near the beginning of *lacZ* called the *lac operator*. The repressor binding to the operator interferes with binding of RNAP to the promoter, and therefore mRNA encoding LacZ and LacY is only made at very low levels. When cells are grown in the presence of lactose, however, a lactose metabolite called allolactose, made from lactose by the product of the lacZ gene, binds to the repressor, causing an allosteric shift. Thus altered, the repressor is unable to bind to the operator, allowing RNAP to transcribe the *lac* genes and thereby leading to higher levels of the encoded proteins.

The second control mechanism is a response to glucose, which uses the catabolite activator protein (CAP) homodimer to greatly increase production of β-galactosidase in the absence of glucose. Cyclic adenosine monophosphate (cAMP) is a signal molecule whose prevalence is inversely proportional to that of glucose. It binds to the CAP, which in turn allows the CAP to bind to the CAP binding site (a 16 bp DNA sequence upstream of the promoter on the left in the diagram below, about 60 bp upstream of the transcription start site), which assists the RNAP in binding to the DNA. In the absence of glucose, the cAMP concentration is high and binding of CAP-cAMP to the DNA significantly increases the production of β-galactosidase, enabling the cell to hydrolyse lactose and release galactose and glucose.

More recently inducer exclusion was shown to block expression of the *lac* operon when glucose is present. Glucose is transported into the cell by the PEP-dependent phosphotransferase system. The phosphate group of phosphoenolpyruvate is transferred via a phosphorylation cascade consisting of the general PTS (phosphotransferase system) proteins HPr and EIA and the glucose-specific PTS proteins EIIA^{Glc} and EIIB^{Glc}, the cytoplasmic domain of the EII glucose transporter. Transport of glucose is accompanied by its phosphorylation by EIIB^{Glc}, draining the phosphate group from the other PTS proteins, including EIIA^{Glc}. The unphosphorylated form of EIIA^{Glc} binds to the *lac* permease and prevents it from bringing lactose into the cell. Therefore, if both glucose and lactose are present, the transport of glucose blocks the transport of the inducer of the *lac* operon.

Multimeric nature of the repressor



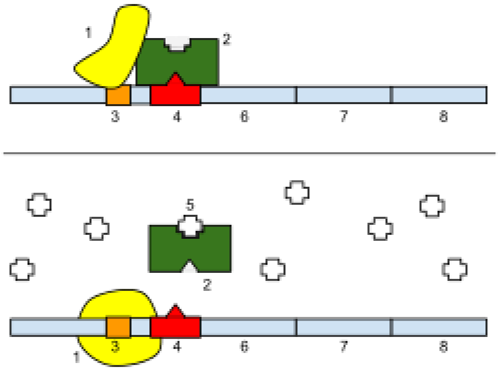
Tetrameric LacI binds two operator sequences and induces DNA looping. Two dimeric *LacI* functional subunits (red+blue and green+orange) each bind a DNA operator sequence (labeled). These two functional subunits are coupled at the tetramerization region (labeled); thus, tetrameric *LacI* binds two operator sequences. This allows tetrameric *LacI* to induce DNA looping.

The lac repressor is a tetramer of identical subunits. Each subunit contains a helix-turn-helix (HTH) motif capable of binding to DNA. The operator site where repressor binds is a DNA sequence with inverted repeat symmetry. The two DNA half-sites of the operator together bind to two of the subunits of the tetrameric repressor. Although the other two subunits of repressor are not doing anything in this model, this property was not understood for many years.

Eventually it was discovered that two additional operators are involved in *lac* regulation. One (O_3) lies about -90 bp upstream of O_1 in the end of the *lacI* gene, and the other (O_2) is about +410 bp downstream of O_1 in the early part of *lacZ*. These two sites were not found in the early work because they have redundant functions and individual mutations do not affect repression very much. Single mutations to either O_2 or O_3 have only 2 to 3-fold effects. However, their importance is demonstrated by the fact that a double mutant defective in both O_2 and O_3 is dramatically de-repressed (by about 70-fold).

In the current model, *lac* repressor is bound simultaneously to both the main operator O_1 and to either O_2 or O_3. The intervening DNA loops out from the complex. The redundant nature of the two minor operators suggests that it is not a specific looped complex that is important. One idea is that the system works through tethering; if bound repressor releases from O_1 momentarily, binding to a minor operator keeps it in the vicinity, so that it may rebind quickly. This would increase the affinity of repressor for O_1.

Mechanism of induction



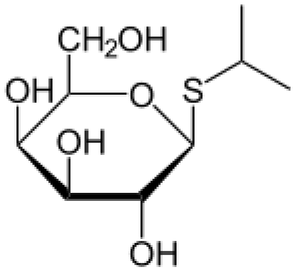
***1*: RNA Polymerase, *2*: Repressor, *3*: Promoter, *4*: Operator, *5*: Lactose, *6*: lacZ, *7*: lacY, *8*: lacA. Top**: The gene is essentially turned off. There is no allolactose to inhibit the *lac* repressor, so the repressor binds tightly to the operator, which obstructs the RNA polymerase from binding to the promoter, resulting in no *laczya* mRNA transcripts. **Bottom**: The gene is turned on. Allolactose inhibits the repressor, allowing the RNA polymerase to bind to the promoter and express the genes, resulting in production of LacZYA. Eventually, the enzymes will digest all of the lactose, until there is no allolactose that can bind to the repressor. The repressor will then bind to the operator, stopping the transcription of the LacZYA genes.

The repressor is an allosteric protein, i.e. it can assume either one of two slightly different shapes, which are in equilibrium with each other. In one form the repressor will bind to the operator DNA with high specificity, and in the other form it has lost its specificity. According to the classical model of induction, binding of the inducer, either allolactose or IPTG, to the repressor affects the distribution of repressor between the two shapes. Thus, repressor with inducer bound is stabilized in the non-DNA-binding conformation. However, this simple model cannot be the whole story, because repressor is bound quite stably to DNA, yet it is released rapidly by addition of inducer. Therefore it seems clear that inducer can also bind to the repressor when the repressor is already bound to DNA. It is still not entirely known what the exact mechanism of binding is.

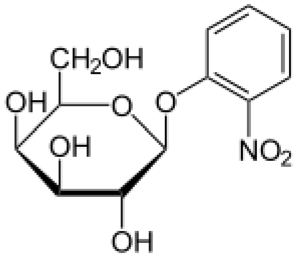
Role of non-specific binding

Non-specific binding of the repressor to DNA plays a crucial role in the repression and induction of the Lac-operon. The specific binding site for the Lac-repressor protein is the operator. The non-specific interaction is mediated mainly by charge-charge interactions while binding to the operator is reinforced by hydrophobic interactions. Additionally, there is an abundance of non-specific DNA sequences to which the repressor can bind. Essentially, any sequence that is not the operator, is considered non-specific. Studies have shown, that without the presence of non-specific binding, induction (or unrepression) of the Lac-operon could not occur even with saturated levels of inducer. It had been demonstrated that, without non-specific binding, the basal level of induction is ten thousand times smaller than observed normally. This is because the non-specific DNA acts as sort of a "sink" for the repressor proteins, distracting them from the operator. The non-specific sequences decrease the amount of available repressor in the cell. This in turn reduces the amount of inducer required to unrepress the system.

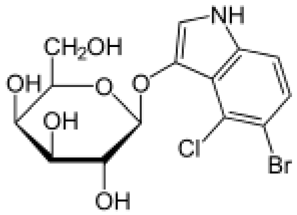
Lactose analogs



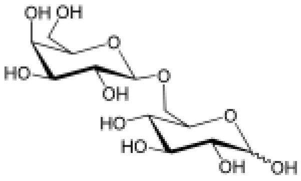
IPTG



ONPG



X-gal



allolactose

A number of lactose derivatives or analogs have been described that are useful for work with the lac operon. These compounds are mainly substituted galactosides, where the glucose moiety of lactose is replaced by another chemical group.

- Isopropyl-β-D-thiogalactoside (IPTG) is frequently used as an inducer of the *lac* operon for physiological work.[1] IPTG binds to repressor and inactivates it, but is not a substrate for β-galactosidase. One advantage of IPTG for *in vivo* studies is that since it cannot be metabolized by *E. coli* its concentration remains constant and the rate of expression of *lac p/o*-controlled genes, is not a variable in the experiment. IPTG intake is dependent on the action of lactose permease in *P. fluorescens*, but not in *E. coli*.

- Phenyl-β-D-galactose (phenyl-Gal) is a substrate for β-galactosidase, but does not inactivate repressor and so is not an inducer. Since wild type cells produce very little β-galactosidase, they cannot grow on phenyl-Gal as a carbon and energy source. Mutants lacking repressor are able to grow on phenyl-Gal. Thus, minimal medium containing only phenyl-Gal as a source of carbon and energy is selective for repressor mutants and operator mutants. If 10^8 cells of a wild type strain are plated on agar plates containing phenyl-Gal, the rare colonies which grow are mainly spontaneous mutants affecting the repressor. The relative distribution of repressor and operator mutants is affected by the target size. Since the *lacI gene* encoding repressor is about 50 times larger than the operator, repressor mutants predominate in the selection.

- Other compounds serve as colorful indicators of β-galactosidase activity.
  - ONPG is cleaved to produce the intensely yellow compound, orthonitrophenol, and is commonly used as a substrate for assay of β-galactosidase *in vitro*.
  - Colonies that produce β-galactosidase are turned blue by X-gal (5-bromo-4-chloro-3-indolyl-β-D-galactoside).

- Allolactose is an isomer of lactose and is the inducer of the lac operon. Lactose is galactose-(β1->4)-glucose, whereas allolactose is galactose-(β1->6)-glucose. Lactose is converted to allolactose by β-galactosidase in an alternative reaction to the hydrolytic one. A physiological experiment which demonstrates the role of LacZ in production of the "true" inducer in *E. coli* cells is the observation that a null mutant of *lacZ* can still produce LacY permease when grown with IPTG but not when grown with lactose. The explanation is that processing of lactose to allolactose (catalyzed by β-galactosidase) is needed to produce the inducer inside the cell.

Development of the classic model

The experimental microorganism used by François Jacob and Jacques Monod was the common laboratory bacterium, *E. coli*, but many of the basic regulatory concepts that were discovered by Jacob and Monod are fundamental to cellular regulation in all organisms. The key idea is that proteins are not synthesized when they are not needed--- *E. coli* conserves cellular resources and energy by not making the three Lac proteins when there is no need to metabolize lactose, such as when other sugars like glucose are available. The following section discusses how *E. coli* controls certain genes in response to metabolic needs.

During World War II, Monod was testing the effects of combinations of sugars as nutrient sources for *E. coli* and *B. subtilis*. Monod was following up on similar studies that had been conducted by other scientists with bacteria and yeast. He found that bacteria grown with two different sugars often displayed two phases of growth. For example, if glucose and lactose were both provided, glucose was metabolized first (growth phase I, see Figure 2) and then lactose (growth phase II). Lactose was not metabolized during the first part of the diauxic growth curve because β-galactosidase was not made when both glucose and lactose were present in the medium. Monod named this phenomenon diauxie.

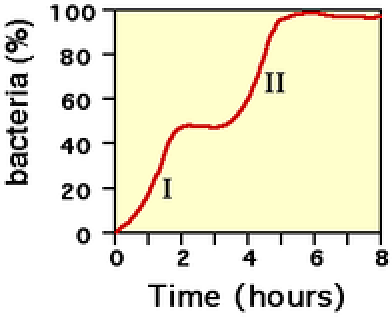


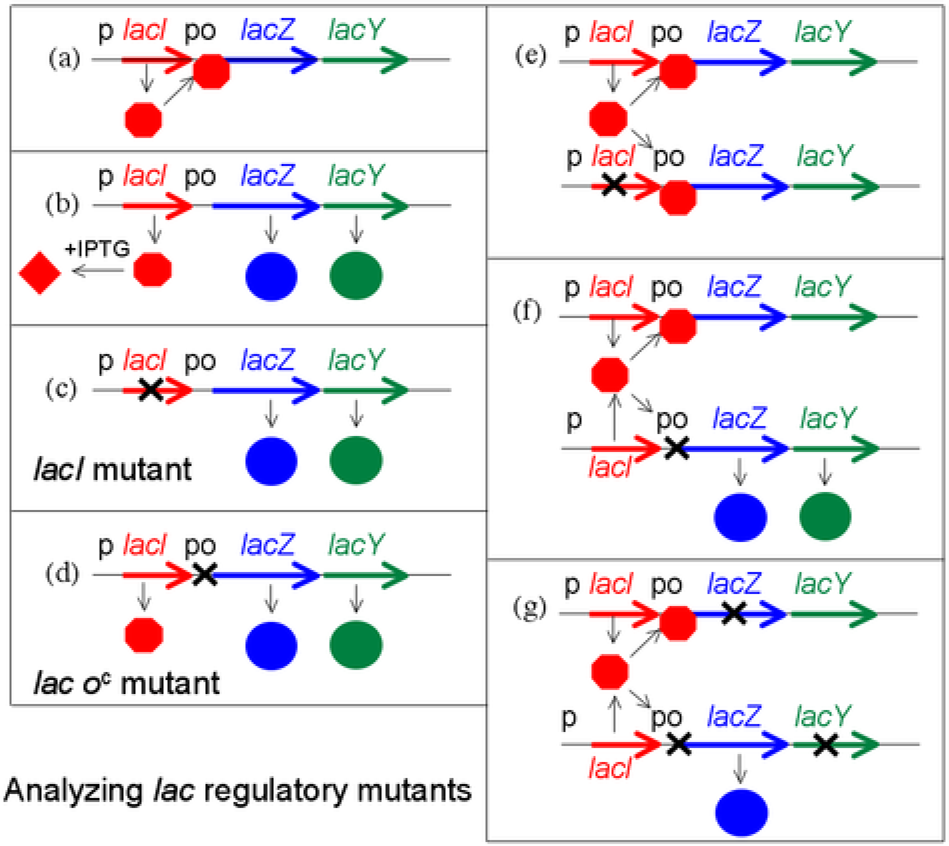
Figure 2: Monod's "bi-phasic" growth curve

Monod then focused his attention on the induction of β-galactosidase formation that occurred when lactose was the sole sugar in the culture medium.

Classification of regulatory mutants

A conceptual breakthrough of Jacob and Monod was to recognize the distinction between regulatory substances and sites where they act to change gene expression. A former soldier, Jacob used the analogy of a bomber that would release its lethal cargo upon receipt of a special radio transmission or signal. A working system requires both a ground transmitter and a receiver in the airplane. Now, suppose that the usual transmitter is broken. This system can be made to work by introduction of a second, functional transmitter. In contrast, he said, consider a bomber with a defective receiver. The behavior of *this* bomber cannot be changed by introduction of a second, functional aeroplane.

To analyze regulatory mutants of the *lac* operon, Jacob developed a system by which a second copy of the *lac* genes (*lacI* with its promoter, and *lacZYA* with promoter and operator) could be introduced into a single cell. A culture of such bacteria, which are diploid for the *lac* genes but otherwise normal, is then tested for the regulatory phenotype. In particular, it is determined whether LacZ and LacY are made even in the absence of IPTG (due to the *lactose repressor* produced by the mutant gene being non-functional). This experiment, in which genes or gene clusters are tested pairwise, is called a *complementation test*.



Analyzing *lac* regulatory mutants

This test is illustrated in the figure (*lacA* is omitted for simplicity). First, certain haploid states are shown (i.e. the cell carries only a single copy of the *lac* genes). Panel (a) shows repression, (b) shows induction by IPTG, and (c) and (d) show the effect of a mutation to the *lacI* gene or to the operator, respectively. In panel (e) the complementation test for repressor is shown. If one copy of the *lac* genes carries a mutation in *lacI*, but the second copy is wild type for *lacI*, the resulting phenotype is normal---but lacZ is expressed when exposed to inducer IPTG. Mutations affecting repressor are said to be *recessive* to wild type (and that wild type is *dominant*), and this is explained by the fact that repressor is a small protein which can diffuse in the cell. The copy of the *lac* operon adjacent to the defective *lacI* gene is effectively shut off by protein produced from the second copy of *lacI*.

If the same experiment is carried out using an operator mutation, a different result is obtained (panel (f)). The phenotype of a cell carrying one mutant and one wild type operator site is that LacZ and LacY are produced even in the absence of the inducer IPTG; because the damaged operator site, does not permit binding of the repressor to inhibit transcription of the structural genes. The operator mutation is dominant. When the operator site where repressor must bind is damaged by mutation, the presence of a second functional site in the same cell makes no difference to expression of genes controlled by the mutant site.

A more sophisticated version of this experiment uses *marked* operons to distinguish between the two copies of the *lac* genes and show that the unregulated structural gene(s) is(are) the one(s) next to the mutant operator (panel (g). For example, suppose that one copy is marked by a mutation inactivating *lacZ* so that it can only produce the LacY protein, while the second copy carries a mutation affecting *lacY* and can only produce LacZ. In this version, only the copy of the *lac* operon that is adjacent to the mutant operator is expressed without IPTG. We say that the operator mutation is *cis-dominant*, it is dominant to wild type but affects only the copy of the operon which is immediately adjacent to it.

This explanation is misleading in an important sense, because it proceeds from a description of the experiment and then explains the results in terms of a model. But in fact, it is often true that the model comes first, and an experiment is fashioned specifically to test the model. Jacob and Monod first imagined that there must be a *site* in DNA with the properties of the operator, and then designed their complementation tests to show this.

The dominance of operator mutants also suggests a procedure to select them specifically. If regulatory mutants are selected from a culture of wild type using phenyl-Gal, as described above, operator mutations are rare compared to repressor mutants because the target-size is so small. But if instead we start with a strain which carries two copies of the whole *lac* region (that is diploid for *lac*), the repressor mutations (which still occur) are not recovered because complementation by the second, wild type *lacI* gene confers a wild type phenotype. In contrast, mutation of one copy of the operator confers a mutant phenotype because it is dominant to the second, wild type copy.

Regulation by cyclic AMP

Explanation of diauxie depended on the characterization of additional mutations affecting the *lac* genes other than those explained by the classical model. Two other genes, *cya* and *crp*, subsequently were identified that mapped far from *lac*, and that, when mutated, result in a decreased level of expression in the *presence* of IPTG and even in strains of the bacterium lacking the repressor or operator. The discovery of cAMP in *E. coli* led to the demonstration that mutants defective the *cya* gene but not the *crp* gene could be restored to full activity by the addition of cAMP to the medium.

The *cya* gene encodes adenylate cyclase, which produces cAMP. In a *cya* mutant, the absence of cAMP makes the expression of the *lacZYA* genes more than ten times lower than normal. Addition of cAMP corrects the low Lac expression characteristic of *cya* mutants. The second gene, *crp*, encodes a protein called catabolite activator protein (CAP) or cAMP receptor protein (CRP).

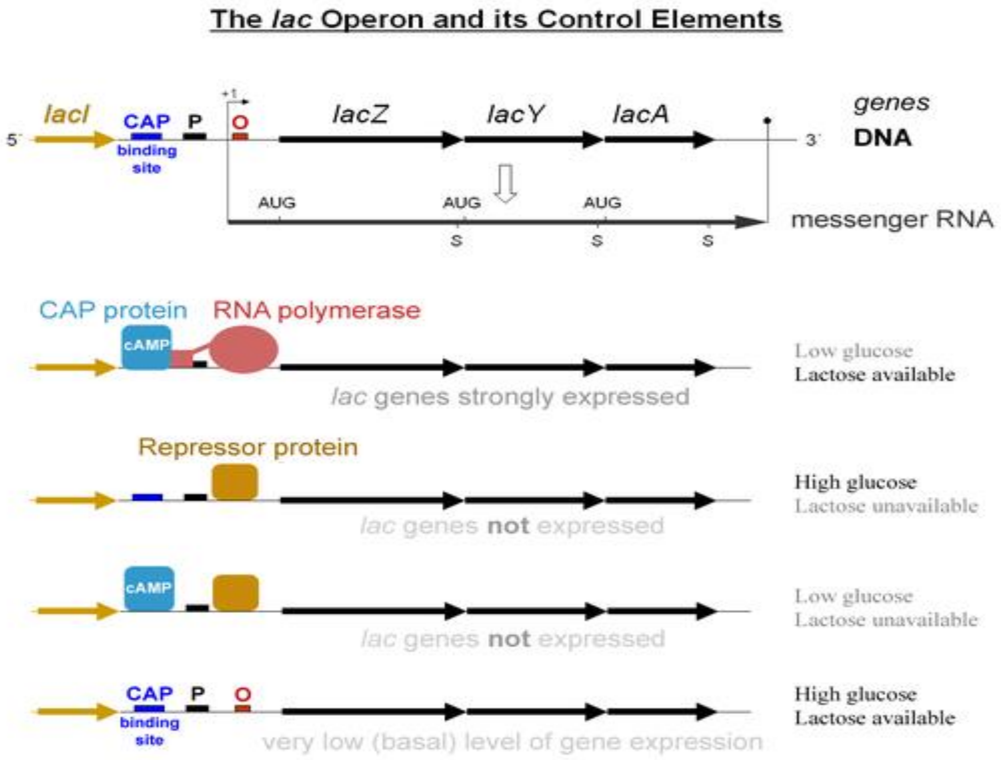
However the lactose metabolism enzymes are made in small quantities in the presence of both glucose and lactose (sometimes called leaky expression) due to the fact that the LacI repressor rapidly associates/dissociates from the DNA rather than tightly binding to it, which can allow time for RNAP to bind and transcribe mRNAs of *lacZYA*. Leaky expression is necessary in order to allow for metabolism of some lactose after the glucose source is expended, but before *lac* expression is fully activated.

In summary:

- When lactose is absent then there is very little Lac enzyme production (the operator has Lac repressor bound to it).
- When lactose is present but a preferred carbon source (like glucose) is also present then a small amount of enzyme is produced (Lac repressor is not bound to the operator).
- When glucose is absent, CAP-cAMP binds to a specific DNA site upstream of the promoter and makes a direct protein-protein interaction with RNAP that facilitates the binding of RNAP to the promoter.

The delay between growth phases reflects the time needed to produce sufficient quantities of lactose-metabolizing enzymes. First, the CAP regulatory protein has to assemble on the *lac* promoter, resulting in an increase in the production of *lac* mRNA. More available copies of the *lac* mRNA results in the production (see translation) of significantly more copies of LacZ (β-galactosidase, for lactose metabolism) and LacY (lactose permease to transport lactose into the cell). After a delay needed to increase the level of the lactose metabolizing enzymes, the bacteria enter into a new rapid phase of cell growth.

The diagram below summarizes these statements.



lac operon in detail

Two puzzles of catabolite repression relate to how cAMP levels are coupled to the presence of glucose, and secondly, why the cells should even bother. After lactose is cleaved it actually forms glucose and galactose (easily converted to glucose). In metabolic terms, lactose is just as *good* a carbon and energy source as glucose. The cAMP level is related not to intracellular glucose concentration but to the rate of glucose transport, which influences the activity of adenylate cyclase. (In addition, glucose transport also leads to direct inhibition of the lactose permease.) As to why *E. coli* works this way, one can only speculate. All enteric bacteria ferment glucose, which suggests they encounter it frequently. It is possible that a small difference in efficiency of transport or metabolism of glucose v. lactose makes it advantageous for cells to regulate the *lac* operon in this way.

Use in molecular biology

The *lac* gene and its derivatives are amenable to use as a reporter gene in a number of bacterial-based selection techniques such as two hybrid analysis, in which the successful binding of a transcriptional activator to a specific promoter sequence must be determined. In LB plates containing X-gal, the colour change from white colonies to a shade of blue corresponds to about 20-100 β-galactosidase units, while tetrazolium lactose and MacConkey lactose media have a range of 100-1000 units, being most sensitive in the high and low parts of this range respectively. Since MacConkey lactose and tetrazolium lactose media both rely on the products of lactose breakdown, they require the presence of both *lacZ* and *lacY* genes. The many *lac* fusion techniques which include only the *lacZ* gene are thus suited to the X-gal plates or ONPG liquid broths

Catabolite repression

Carbon catabolite repression, or simply **catabolite repression**, is an important part of global control system of various bacteria and other micro-organisms. Catabolite repression allows micro-organisms to adapt quickly to a preferred (rapidly metabolisable) carbon and energy source first. This is usually achieved through inhibition of synthesis of enzymes involved in catabolism of carbon sources other than the preferred one. The catabolite repression was first shown to be initiated by glucose and therefore sometimes referred to as the **glucose effect**. However, the term "glucose effect" is actually a misnomer since other carbon sources are known to induce catabolite repression.

Escherichia coli

Catabolite repression was extensively studied in *Escherichia coli*. *E. coli* grows faster on glucose than on any other carbon source. For example, if *E. coli* is placed on an agar plate containing only glucose and lactose, the bacteria will use glucose first and lactose second. When glucose is available in the environment, the synthesis of β-galactosidase is under repression due to the effect of catabolite repression caused by glucose. The catabolite repression in this case is achieved through the utilization of phosphotransferase system.

An important enzyme from the phosphotranferase system called Enzyme II A (**EIIA**) plays a central role in this mechanism. There are different catabolite-specific **EIIA** in a single cell, even though different bacterial groups have specificities to different sets of catabolites. In enteric bacteria one of the **EIIA** enzymes in their set is specific for glucose transport only. When glucose levels are high inside the bacteria, **EIIA** mostly exists in its unphosphorylated form. This leads to inhibition of adenylyl cyclase and lactose permease, therefore cAMP levels are low and lactose can not be transported inside the bacteria. After some time, the glucose is all used up and the second preferred carbon source (i.e. lactose) has to be used by bacteria. Absence of glucose will "turn off" catabolite repression.

Furthermore, when glucose levels are low the phosphorylated form of **EIIA** accumulates and consequently activates the enzyme adenylyl cyclase, which will produce high levels of cAMP. cAMP binds to catabolite activator protein (CAP) and together they will bind to a promoter sequence on the lac operon. However, this is not enough for the lactose genes to be transcribed. Lactose must be present inside the cell to remove the lactose repressor from the operator sequence (transcriptional regulation). When these two conditions are satisfied, it means for the bacteria that glucose is absent and lactose is available. Next, bacteria start to transcribe lactose gene and produce β-galactosidase enzymes for lactose metabolism. The example above is a simplification of a complex process. Catabolite repression is considered to be a part of global control system and therefore it affects more genes rather than just lactose gene transcription.

Bacillus subtilis

Gram positive bacteria such as *Bacillus subtilis* have a cAMP-independent catabolite repression mechanism controlled by catabolite control protein A (CcpA). In this alternative pathway CcpA negatively represses other sugar operons so they are off in the presence of glucose. It works by the fact that Hpr is phosphorylated by a specific mechanism, when glucose enters through the cell membrane protein EIIC, and when Hpr is phosphoralated it can then allow CcpA to block transcription of the alternative sugar pathway operons at their respective cre sequence binding sites. Note that *E. coli* has a similar cAMP-independent catabolite repression mechanism that utilizes a protein called catabolite repressor activator (Cra).

gal operon

The ***gal* operon** is a prokaryotic operon, which encodes enzymes necessary for galactose metabolism. The operon contains two operators, O_E (for external) and O_I. The former is just before the promoter, and the latter is just after the *galE* gene (the first gene in the operon).

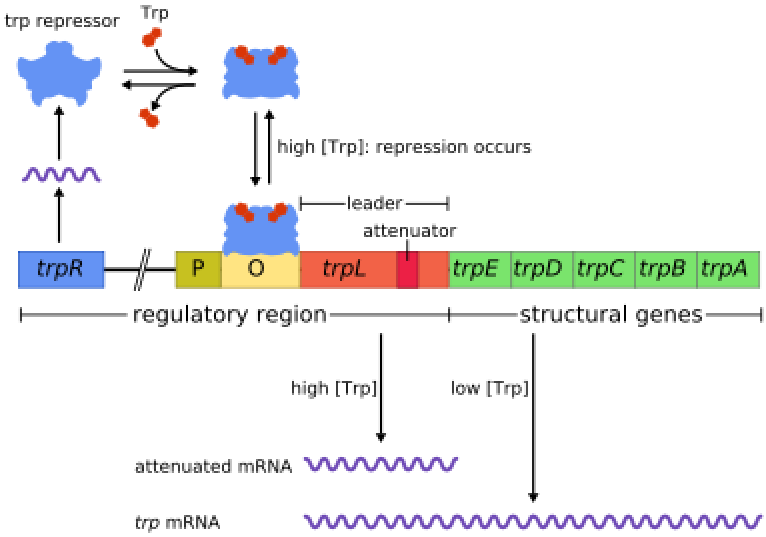
Repression of gene expression works via binding of repressor molecules to the two operators. These repressors dimerize, creating a loop in the DNA. The loop as well as hindrance from the external operator prevent RNA polymerase from binding to the promoter, and thus prevent transcription.

The *gal* operon of *E. coli* consists of 3 structural genes: *galE* (epimerase), *galT* (galactose transferase), and *galK* (galactokinase), which are transcribed from two overlapping promoters PG1 and PG2 upstream from *galE*. Regulation of the operon is complex since the GalE product, an epimerase that converts UDP-glucose into UDP-galactose, is required for the formation of UDP-galactose for cell wall biosynthesis, in particular the cell wall component lipopolysaccharide, even when cells are not using galactose as a carbon/energy source.

The gal operon is controlled by CRP-cAMP as for the lac operon. CRP-cAMP binds to -35 region promoting transcription from PG1 but inhibiting transcription from PG2. When cells are grown in glucose, basal level transcription occurs from PG2. The unlinked *galR* gene encodes the repressor for this system. A tetrameric GalR repressor binds to 2 operators, one located at +55 and one located at -60 relative to the PG1 start site. Looping of the DNA blocks the access of RNA polymerase to promoters and/or inhibits formation of the open complex. When GalR binds as a dimer to the -60 site only, promoter PG2 is activated, not repressed, allowing basal levels of GalE to be produced. In this state promoter PG1 is inactivated through interactions with the alpha subunit of RNA polymerase

trp operon

From Wikipedia, the free encyclopedia



Structure of the trp operon.

The *trp* operon is an operon — a group of genes that are used, or transcribed, together — that codes for the components for production of tryptophan. The *trp* operon is present in many bacteria, but was first characterized in *Escherichia coli*. The operon is regulated so that when tryptophan is present in the environment, the genes for tryptophan synthesis are not expressed. It was an important experimental system for learning about gene regulation, and is commonly used to teach gene regulation.

Discovered in 1953 by Jacques Monod and colleagues, the *trp* operon in *E. coli* was the first repressible operon to be discovered. While the *lac* operon can be activated by a chemical (allolactose), the tryptophan (Trp) operon is inhibited by a chemical (tryptophan). This operon contains five structural genes: trp E, trp D, trp C, trp B, and trp A, which encode tryptophan synthetase. It also contains a repressive regulator gene called trp R. Trp R has a promoter where RNA polymerase binds and synthesizes mRNA for a regulatory protein. The protein that is synthesized by trp R then binds to the operator which then causes the transcription to be blocked. In the *lac* operon, allolactose binds to the repressor protein, allowing gene transcription, while in the *trp* operon, tryptophan binds to the repressor protein effectively blocking gene transcription. In both situations, repression is that of RNA polymerase transcribing the genes in the operon. Also unlike the *lac* operon, the *trp* operon contains a leader peptide and an attenuator sequence which allows for graded regulation.

It is an example of repressible negative regulation of gene expression. Within the operon's regulatory sequence, the operator is blocked by the repressor protein in the presence of tryptophan (thereby preventing transcription) and is liberated in tryptophan's absence (thereby allowing transcription). The process of attenuation (explained below) complements this regulatory action.

Repression

The operon operates by a negative repressible feedback mechanism. The repressor for the trp operon is produced upstream by the trpR gene, which is constitutively expressed at a low level. Synthesized TrpR monomers associate into tetramers. These tetramers are inactive and are dissolved in the nucleoplasm. When tryptophan is present, these tryptophan repressor tetramers bind to tryptophan, causing a change in the repressor conformation, allowing the repressor to bind to the operator. This prevents RNA polymerase from binding to and transcribing the operon, so tryptophan is not produced from its precursor. When tryptophan is not present, the repressor is in its inactive conformation and cannot bind the operator region, so transcription is not inhibited by the repressor.

Attenuation

Mechanism of transcriptional attenuation of the *trp* operon.

Attenuation is a second mechanism of negative feedback in the *trp* operon. The repression system targets the intracellular trp concentration whereas the attenuation responds to the concentration of charged tRNA^{trp}. Thus, the trpR repressor decreases gene expression by altering the initiation of transcription, while attenuation does so by altering the process of transcription that's already in progress. While the TrpR repressor decreases transcription by a factor of 70, attenuation can further decrease it by a factor of 10, thus allowing accumulated repression of about 700-fold. Attenuation is made possible by the fact that in prokaryotes (which have no nucleus), the ribosomes begin translating the mRNA while RNA polymerase is still transcribing the DNA sequence. This allows the process of translation to affect transcription of the operon directly.

At the beginning of the transcribed genes of the *trp* operon is a sequence of at least 130 nucleotides termed the leader transcript (trpL). Lee and Yanofsky (1977) found that the attenuation efficiency is correlated with the stability of a secondary structure embedded in trpL, and the 2 constituent hairpins of the terminator structure were later elucidated by Oxender *et al.* (1979). This transcript includes four short sequences designated 1-4, each of which is partially complementary to the next one. Thus, three distinct secondary structures (hairpins) can form: 1-2, 2-3 or 3-4. The hybridization of sequences 1 and 2 to form the 1-2 structure is rare because the RNA Polymerase waits for a ribosome to attach before continuing transcription past sequence 1, however if the 1-2 hairpin were to form it would prevent the formation of the 2-3 structure (but not 3-4). The formation of a hairpin loop between sequences 2-3 prevents the formation of hairpin loops between both 1-2 and 3-4. The 3-4 structure is a transcription termination sequence (abundant in G/C and immediately followed by several uracil residues), once it forms RNA polymerase will disassociate from the DNA and transcription of the structural genes of the operon can not occur (see below for a more detailed explanation). The functional importance of the 2nd hairpin for the transcriptional termination is illustrated by the reduced transcription termination frequency observed in experiments destabilizing the central G+C pairing of this hairpin.

Part of the leader transcript codes for a short polypeptide of 14 amino acids, termed the leader peptide. This peptide contains two adjacent tryptophan residues, which is unusual, since tryptophan is a fairly uncommon amino acid (about one in a hundred residues in a typical *E. coli* protein is tryptophan). The strand 1 in trpL encompasses the region encoding the trailing residues of the leader peptide: Trp, Trp, Arg, Thr, Ser; conservation is observed in these 5 codons whereas mutating the upstream codons do not alter the operon expression. If the ribosome attempts to translate this peptide while tryptophan levels in the cell are low, it will stall at either of the two trp codons. While it is stalled, the ribosome physically shields sequence 1 of the transcript, preventing the formation of the 1-2 secondary structure. Sequence 2 is then free to hybridize with sequence 3 to form the 2-3 structure, which then prevents the formation of the 3-4 termination hairpin, which is why the 2-3 structure is called an anti-termination hairpin. In the presence of the 2-3 structure, RNA polymerase is free to continue transcribing the operon. Mutational analysis and studies involving complementary oligonucleotides demonstrate that the stability of the 2-3 structure corresponds to the operon expression level. If tryptophan levels in the cell are high, the ribosome will translate the entire leader peptide without interruption and will only stall during translation termination at the stop codon. At this point the ribosome physically shields both sequences 1 and 2. Sequences 3 and 4 are thus free to form the 3-4 structure which terminates transcription. This terminator structure forms when no ribosome stalls in the vicinity of the Trp tandem (i.e. Trp or Arg codon): either the leader peptide is not translated or the translation proceeds smoothly along the strand 1 with abundant charged tRNAtrp. More over, the ribosome is proposed to only block about 10 nts downstream, thus ribosome stalling in either the upstream Gly or further downstream Thr do not seem to affect the formation of the termination hairpin. The end result is that the operon will be transcribed only when tryptophan is unavailable for the ribosome, while the trpL transcript is constitutively expressed.

This attenuation mechanism is experimentally supported. First, the translation of the leader peptide and ribosomal stalling are directly evidenced to be necessary for inhibiting the transcription termination. Moreover, mutational analysis destabilizing or disrupting the base-pairing of the antiterminator hairpin results in increased termination of several folds; consistent with the attenuation model, this mutation fails to relieve attenuation even with starved Trp. In contrast, complementary oligonucleotides targeting strand 1 increases the operon expression by promoting the antiterminator formation. Furthermore, in histidine operon, compensatory mutation shows that the pairing ability of strands 2-3 matters more than their primary sequence in inhibiting attenuation.

In attenuation, where the translating ribosome is stalled determines whether the termination hairpin will be formed. In order for the transcribing polymerase to concomitantly capture the alternative structure, the time scale of the structural modulation must be comparable to that of the transcription. To ensure that the ribosome binds and begins translation of the leader transcript immediately following its synthesis, a pause site exists in the trpL sequence. Upon reaching this site, RNA polymerase pauses transcription and apparently waits for translation to begin. This mechanism allows for synchronization of transcription and translation, a key element in attenuation.

A similar attenuation mechanism regulates the synthesis of histidine, phenylalanine and threonine.

The quintessential example which still stands as the paradigm of transcriptional control is the Lac Operon, first developed by Jacob and Monod and verified each year faithfully by second year science students.

Background:

E. coli (we are back with these pesky little coliforms) have the ability to grow in media which contain lactose as their sole carbon source. Lactose is a sugar found in milk (a disaccharide).

To metabolise this sugar the bugs must produce two enzymes, beta galactosidase and lac permease. The lac permease allows the lactose to enter the cell. The beta galactosidase cleaves the bond joining the two monosaccharides (known as a beta galactoside bond, a type of glycosidic bond). Once the sugar has been cleaved the two monosaccharides can be utilized by the cell's glycolytic "house keeping" enzymes. If the E. coli are grown up on media with other carbon sources there is very little activity of these two enzymes.

Is this modulation of enzyme activity a transcriptional event or simply an activation/deactivation of pre-existing enzyme activity? The answer is it is a transcriptional event. To test for this a protein synthesis inhibitor is included in the incubation. The induction does not occur.

The term induction means that the activity of an enzyme (beta gal in our example) increases after the addition of a compound, in this case lactose. How does the presence of lactose, as the sole carbon source, control the level of transcription of the enzymes that catalyse its utilization?

The experiments by Jacob and Monod, working with bacteria containing an extra copy of the genes(extra chromosomal) for the lac operon, found

that the control of this gene expression had two elements; a cis acting factor and a trans acting factor. They isolated many mutants of E. coli where the lesion was either on the genomic DNA or on the extra chromosomal copy. From the analysis of these mutants the lac operon model was developed.

Mutants of E. coliwhich have lost the ability to control beta gal gene expression tend to fall into two categories: constitutive high producers and no producers. The high producers have high levels of beta galactosidase activity, whether lactose is present or not. The no producers have no activity whether lactose is present or not.

Cis acting factors could only affect gene expression on the same piece of DNA, while trans acting factors could influence gene expression on other copies of the gene located on separate pieces of DNA (extra chromosomal) in the cell.

The model proposed contains lac I, a promoter region, an operator region, the transcription unit which contains lac Z, lac Y and lac A. The gene product of lac I is a protein, known as the lac repressor. This protein is a tetramer with a high affinity for the operator region of the lac operon. There are only a few copies in the cell.

The promoter is the region where RNA polymerase binds (at the -10 and -35 regions). The repressor binds at the operator (-10 to 0).

Lac Z is the structural gene coding for beta galactosidase, lac Y for lac permease and lac A for a transacetylase of unknown function. These three genes are all transcribed in one long mRNA, known as a polycistronic mRNA.

The lac operon is under two forms of control; positive and negative control.

A protein is said to exert negative control when its binding prevents an event. The repressor is an example of negative control. When the repressor is bound to the operator the RNA polymerase cannot bind to the promoter. No RNA polymerase, no transcription no enzyme activity. Once lactose enters the cell a small amount of it is converted to allolactose by the few copies of beta gal present in the induced cell. The allolactose binds to the repressor and results in its dissociation from the operator.

In the lab we use a compound IPTG which is an analogue of the allolactose and is thus described as an inducer. A protein exerts positive control when its binding results in an event. The catabolite activator protein (CAP) or as it is sometimes known, the cAMP receptor protein (CRP) is an example of positive control. This protein binds to an activation site within the promoter region only when it is complexed to cAMP. cAMP is a derivative of ATP, synthesized by adenylyl cyclase and is a second messenger in both eukaryotic and prokaryotic cells. In this case the significance of cAMP is that its concentration is low when intracellular [glucose] is high and vice versa. The other important fact is that cAMP reversibly binds to CAP in a concentration dependent manner.

In summary: When intracellular glucose levels are high cAMP is
low CAP is not associated with cAMP and is not bound to the DNA.

When intracellular [glucose] is low cAMP is high cAMP-CAP is complexed and associated with the DNA at the activation site. This complex will increase the frequency of initiation of transcription by RNA polymerase at the lac promoter. This protein complex acts on a number of operons around the genome and is described as global regulation.

The best way to understand the lac operon is to take a couple of scenarios.

1.Lactose (+) and glucose (-)

2.Lactose (+) and glucose (+)

3.Lactose (-) and glucose (+)

4.Lactose (-) and glucose (-)

In scenario 1 some of the Lactose entering the cell via the few lac permease transporters available has been converted to allolactose and has resulted in the removal of the repressor from the operator. The

promoter is now unmasked and RNA polymerase can now bind and initiate transcription. However it won’t do this very frequently without the help of the cAMP-CAP bound to the activation site. This protein complex binding puts a 90o kink in the DNA and interacts with the alpha subunit of RNA polymerase. Without the cAMP CAP the lac promoter is a weak promoter varying significantly from the consensus sequence at -10 and -35. The combination of the two controls means beta gal and lac

permease are transcribed at high levels.

In scenario 2the repressor is off the operator but the CAP (without cAMP) is not bound to the DNA so initiation only occurs at a low rate

little transcription.

In scenario 3the repressor is bound to the operator and the CAP (without cAMP) is not bound to the DNA. Very little transcription of the lac operon genes is happening now.

In scenario 4the cell is starving! The repressor is on the operator but the cAMP CAP is on the DNA. If the repressor is bound there is no transcription RNA polymerase has no access. The regulation of the lac operon by the repressor is described as specific regulation of gene

expression; the control is only exerted over the three structural genes following the operator. The CAP-cAMP complex is an example of global regulation; it exerts its influence over a number of genes scattered throughout the genome. Genes which encode other catabolic enzymes involved in carbohydrate metabolism e.g. the arabinose operon (arabinose is another alternative sugar to glucose which E. coli can survive on if they have nothing else) also come under the control of the CAP-cAMP complex. Again, if glucose is present at a sufficient concentration then it is in the organism’s interest to down-regulate the synthesis of genes which catalyse the catabolism of arabinose. The idea is save the energy and use glucose. The enzymes which catalyse glucose catabolism are not inducible i.e glycolysis is too fundamental to survival to be switched on and off.

Cis acting factors: the operator sequence, the promoter sequence, the activation sequence.

Trans acting factors: the repressor, the sigma factor

The trp operon.

The trp operon is another example of specific control of gene expression operating in our friend E. coli. One of the aspects of this example that makes it different is the genes here are biosynthetic rather than catabolic.

Background.

E. coli have the ability to synthesise tryptophan (an amino acid) from the compound chorismate. It requires 5 enzymes to do this; the gene products of genes trpE to trpA.

The trp operon also has a leader sequence which will attenuate trpE – trpA expression at intermediate [trp]. The bacteria do not want to make the enzymes for tryptophan biosynthesis when there is plenty of tryptophan around. To control this an operon is in place with a repressor that binds to the trp operator when trp is bound (the opposite of the lac operon).

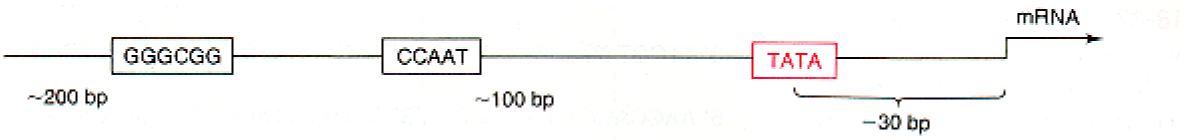
The association of the trp with the repressor protein is non-covalent and concentration dependent. Once the intracellular [trp] concentration falls the trp comes off the repressor and the repressor dissociates itself from the DNA leaving the promoter region free to bind RNA polymerase.

The repressor in this case works in the opposite way to the lac repressor. If the trp is present, the repressor binds, whereas in the lac operon, if the lactose is present the repressor dissociates. This behaviour is explained by the purpose of both enzymes. Beta galactosidase is a catalytic enzyme. It is needed when the lactose is present to metabolise the lactose. If glucose is around the bugs can use that instead and there is no need to make beta gal. In the case of the trp operon the gene products are biosynthetic enzymes. They do not need to be synthesized if there is enough of the product they are synthesizing. The trp repressor and the lac repressor are great examples of proteins that bind to DNA in a base sequence specific manner. Their affinity for DNA is altered several thousand fold by the binding of a metabolite; lactose or tryptophan. The difference is that the trp repressor has greater affinity for the operator when the trp is bound, while the lac repressor has greater affinity when the allolactose is NOT bound.

CHAPTER # 21 CONTROL OF GENE EXPRESSION IN EUKARYOTES

Eukaryotic Gene Control

Eukaryotic control sites include promoter consensus sequences similar to those in bacteria.



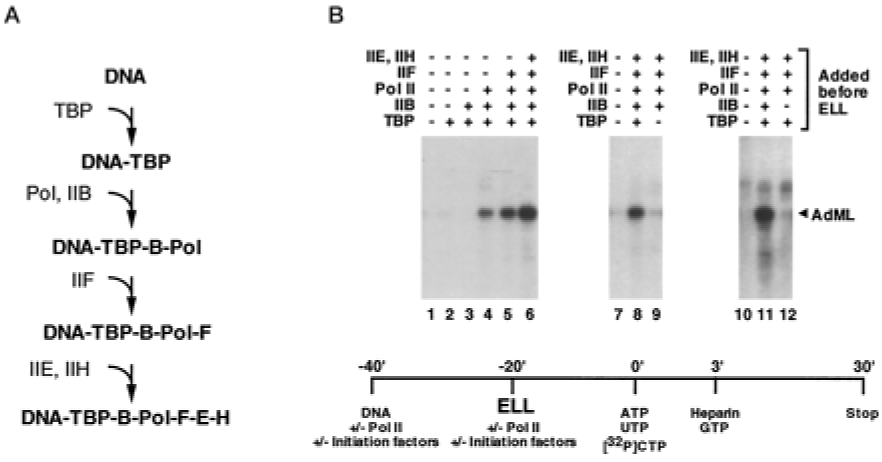
However, there can be many control sequences, called enhancers and silencers, responsive to many different signals. Enhancers were defined by cis/trans complementation experiments, in which their activation only occurred when they were present on the same DNA helix with the gene under their control. Thus they were originally called cis-acting elements; this terminology is still used in experiments defining new regulatory sites.

Three RNA Polymerases in Eukaryotes

Review from before break: Eukaryotes have three different RNA polymerases, which transcribe three different classes of genes. RNA pol II transcribes hnRNA (precursor to mRNA). RNA pol I and III transcribe functional RNAs such as rRNAs and tRNAs.

Initiation of RNA pol II transcription requires multiple basal transcription factors. Most of these were identified initially through biochemical approaches, i.e. fractionation of nuclear extracts (by chromatography or density gradient centrifugation) and reconsitution of transcription in vitro.

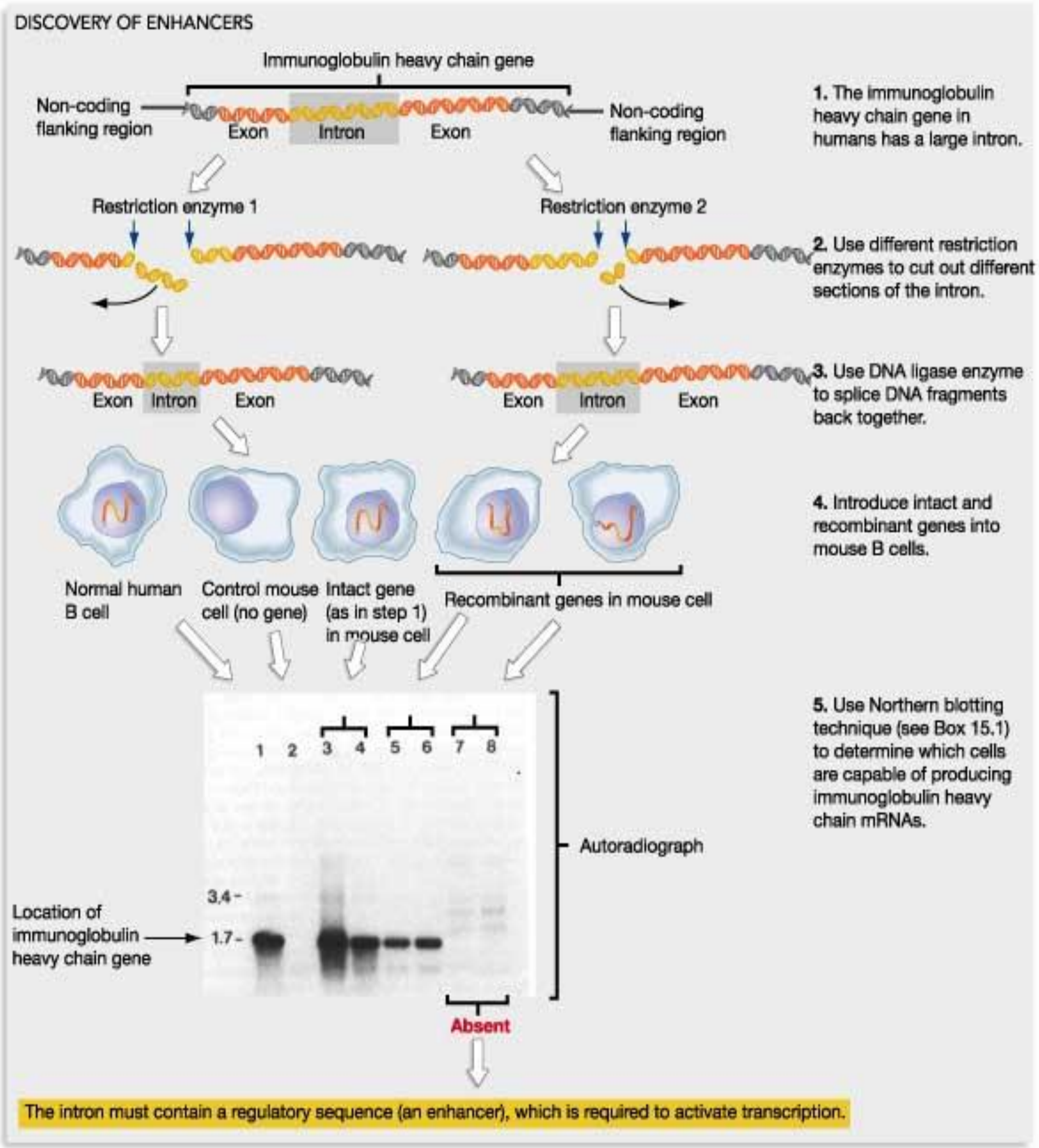
For example, in this experiment, different purified basal transcription factors (TBP, TFIIB, IIF, IIE, IIH) and RNA polymerase II were mixed and matched to see which would support transcription from the adenovirus major late promoter.



(Shilatifard A, Haque D, Conaway RC, Conaway JW. J Biol Chem 1997 Aug 29;272(35):22355-63)

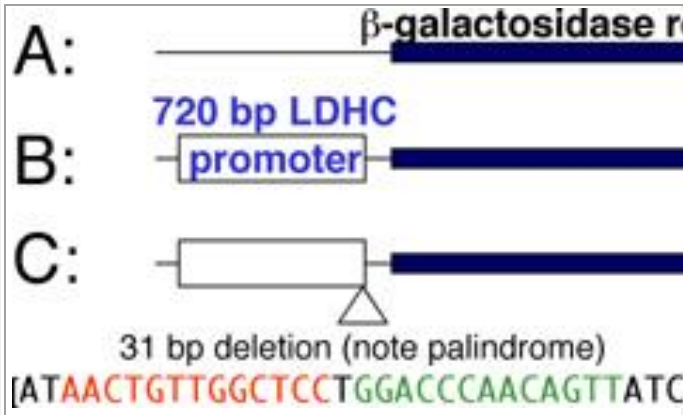
Discovery of Enhancers: Using recombinant DNA transfected into *cultured cells*.

Susumu Tonegawa: Transcription of the human antibody heavy chain gene is under control of enhancer elements in Intron 1.

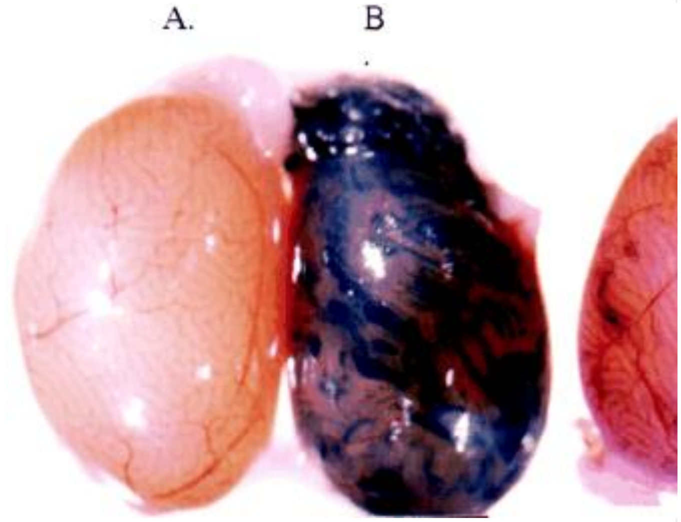

DISCOVERY OF ENHANCERS
Immunoglobulin heavy chain gene
Non-coding flanking region
Exon
Intron
Exon
Non-coding flanking region
1. The immunoglobulin heavy chain gene in humans has a large intron.
Restriction enzyme 1
Restriction enzyme 2
2. Use different restriction enzymes to cut out different sections of the intron.
Exon
Intron
Exon
Exon
Intron
Exon
3. Use DNA ligase enzyme to splice DNA fragments back together.
4. Introduce intact and recombinant genes into mouse B cells.
Normal human B cell
Control mouse cell (no gene)
Intact gene (as in step 1) in mouse cell
Recombinant genes in mouse cell
1
2
3
4
5
6
7
8
3.4
1.7
Location of immunoglobulin heavy chain gene
Autoradiograph
5. Use Northern blotting technique (see Box 15.1) to determine which cells are capable of producing immunoglobulin heavy chain mRNAs.
Absent
The intron must contain a regulatory sequence (an enhancer), which is required to activate transcription.

[Figure from Freeman, S. (2002) *Biological Science*]

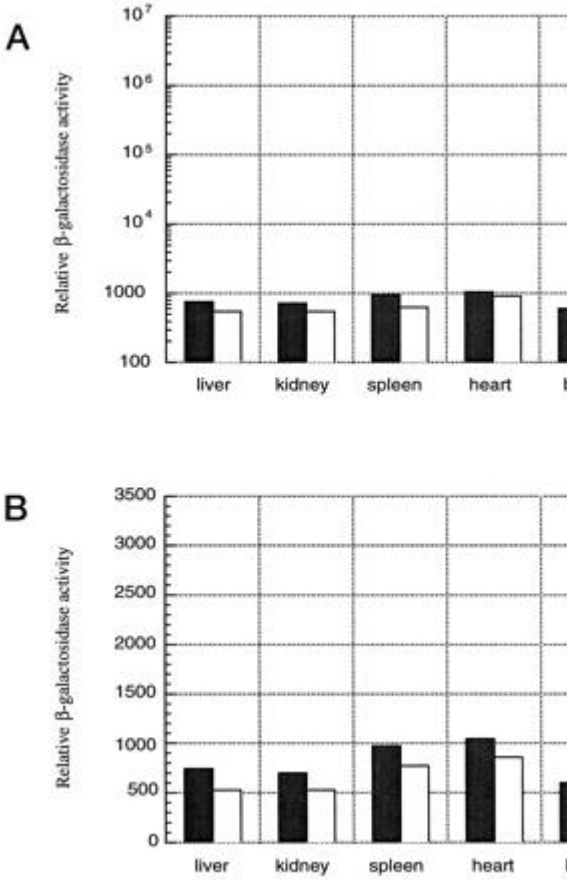
Assessment of enhancer elements using recombinant reporter genes in *transgenic mice*: Testis-specific Lactate Dehydrogenase C promoter.


A:
β-galactosidase r
720 bp LDHC
B:
promoter
C:
31 bp deletion (note palindrome)
[ATAACTGTTGGCTCCTGGACCCAACAGTTATC

Three different constructs contained different portions of the LDHC promoter coupled with beta galactosidase reporter gene.



X-gal staining indicates beta-galactosidase activity driven by DNA regulatory elements in the indicated construct. Which portion of the promoter supported the greatest testis-specific expression?

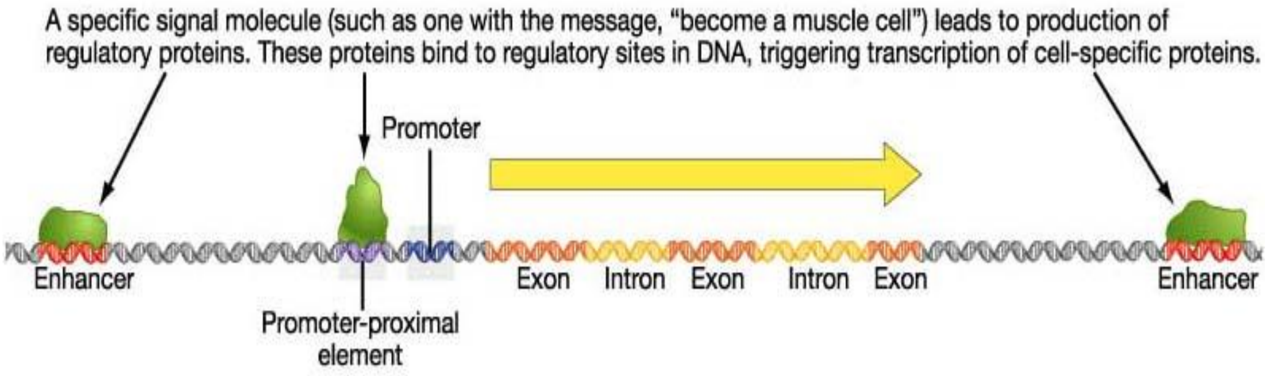


Tissue-specific function of regulatory elements in the LDHC promoter: Panel A: Compares beta-galactosidase activity in construct A (black bars) and construct B (white bars). Panel B: Compares beta-galactosidase activity in construct A (black bars) and construct C (white bars)

Note the difference in the scales on Y-axes in the two graphs.

From: Li, S, W. Zhou, L. Doglio, and E. Goldberg (1998) Transgenic Mice Demonstrate a Testis-specific Promoter for Lactate Dehydrogenase, LDHC *J. Biol. Chem.* 273:31191-31194.	

How do enhancer elements work to regulate transcription of specific genes in specific times and places? By serving as binding sites for transcription factors--proteins that regulate transcription.

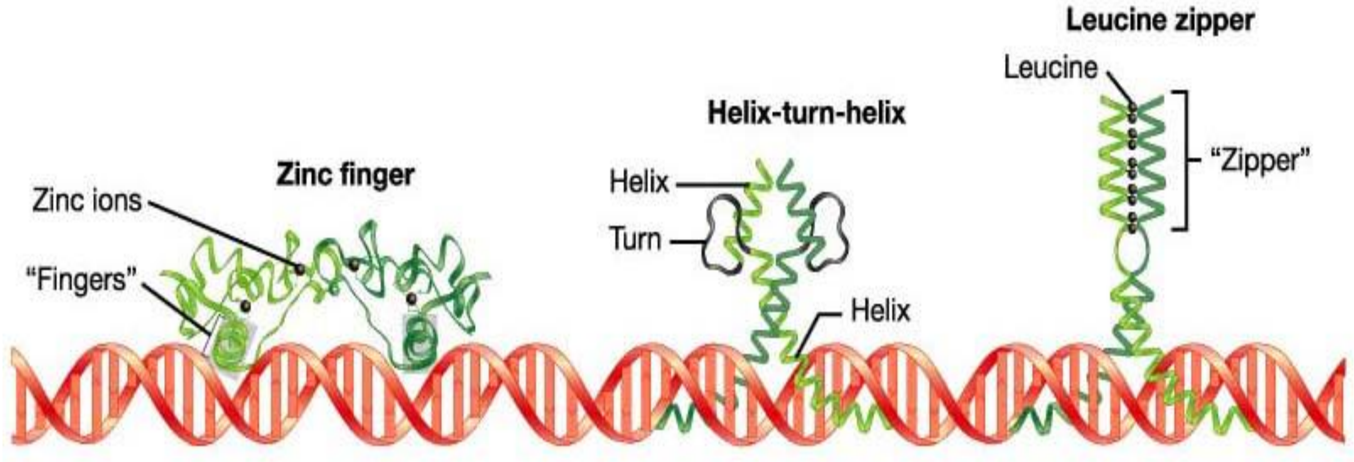


A specific signal molecule (such as one with the message, "become a muscle cell") leads to production of regulatory proteins. These proteins bind to regulatory sites in DNA, triggering transcription of cell-specific proteins.

[Figure from Freeman, S. (2002) *Biological Science*]

DNA:Protein and Protein:Protein interactions are important for transcription factor function. Note modular structure of transcription factors: one part of the protein is responsible for DNA binding, another for dimer formation, another for transcriptional activation (i.e. interaction with basal transcription machinery).

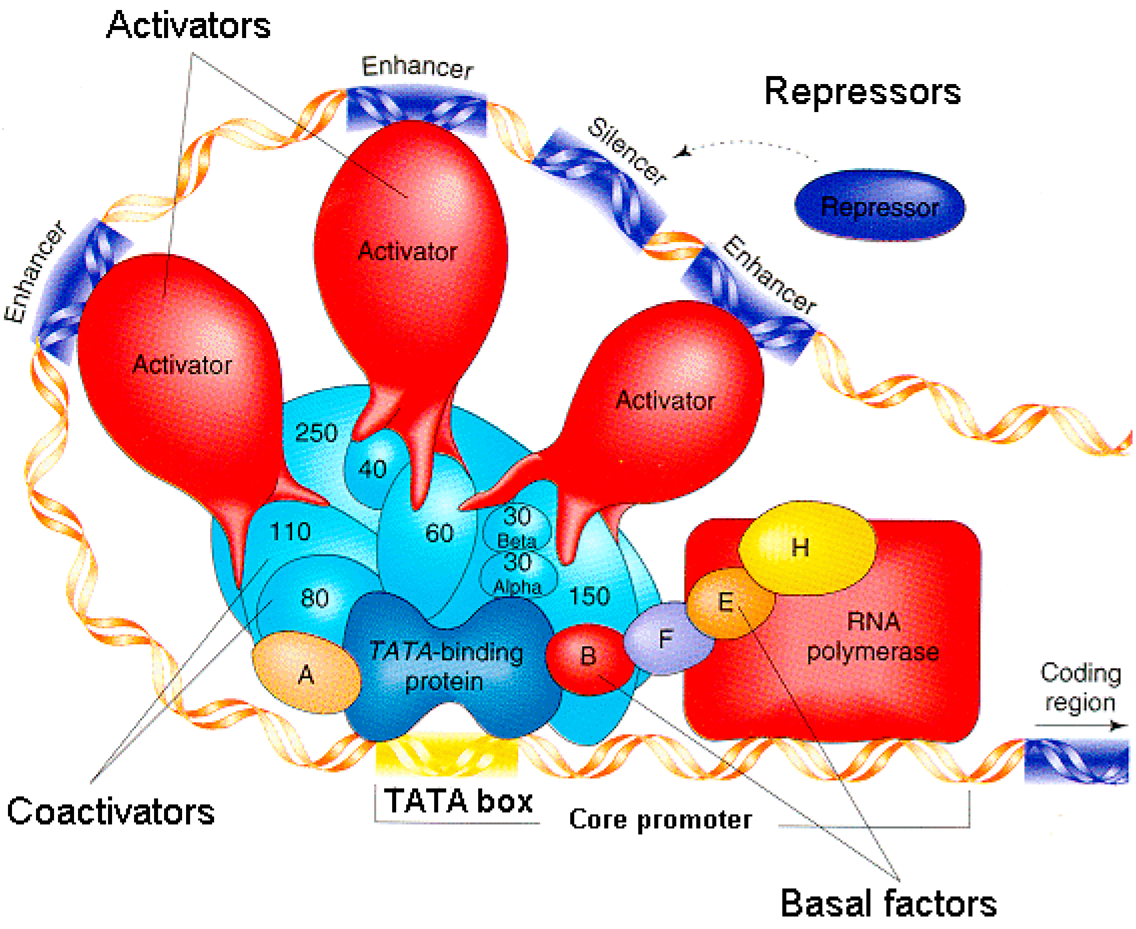
Dimer formation adds an extra element of complexity and versatility. Mixing and matching of proteins into different heterodimers and homodimers means that three distinct complexes can be formed from two proteins.



[Figure from Freeman, S. (2002) *Biological Science*]

A COMPREHENSIVE MODEL OF REGULATION OF RNA POLYMERASE II TRANSCRIPTION:

Although they are cis-acting, the enhancers and silencers can be strung out across 10-20 kilobases (thousands of base pairs) of DNA upstream. Some signals can even be downstream of the coding gene, or even found within introns (!) How can this be possible? Long regions of the DNA can loop over to enable the regulatory connections.



Based on Robert Tjian, "Molecular Machines that Control Genes," *Scientific American*.

- Activators bind to enhancer sites, controlled by hormones or other signals. They increase transcription of the regulated gene.
- Repressors bind to silencer sites, controlled by hormones or other signals. They decrease transcription of the regulated gene, possibly by interfering with activators.
- Coactivators bind to activators and/or repressors (at one end) and to basal factors (at the other end). The coactivators somehow communicate the signal from activators and/or repressors to the RNA polymerase.
- Basal factors act similarly to bacterial sigma factors. They enable RNA polymerase to initiate transcription. However, they require interaction with coactivators.

How were all these control proteins figured out? Robert Tjian explains some experiments:

- The researchers tested human cell extracts for a sigma-like protein: one that (1) bound to DNA and (2) stimulated RNA transcription in the test tube. They tested many, many proteins, and found one: SP1.
- SP1 only increased transcription when the DNA contained "GC box" sequence (an enhancer). Without GC box, only basal (low-level) transcription occurred.
- The "zinc finger" domain of SP1 was essential for binding to GC box. The "glutamine-rich domain" was not needed for DNA binding, but was needed to increase transcription. The researchers guessed that the glutamine-rich domain bound to basal factors needed for low-level transcription, and converted it to high-level transcription.
- Basal "Factor D" (known to bind TATA box) was suspected to be the target of SP1. To Tjian's surprise, however, when Factor D was better purified, SP1 failed to increase transcription. Therefore he guessed that Factor D included the TATA Binding Protein plus some other factor. The other factor(s) turned out to be eight coactivators.

Splicing of hnRNA to make mRNA

The first transcript of RNA from a eukaryotic gene is not yet ready for transcription. It is called hnRNA, for high-molecular-weight nuclear RNA. In order for the RNA to exit the nucleus, and for proteins to be translated by ribosomes in the cytoplasm, the following processing steps must first occur:

- Capping of the 5' sequence with 5' methyl-7-guanidine (the "m-7-G cap")
- Addition of a run of adenine nucleotides to the 3' OH end (the "poly-A tail")
- Splicing out of the intron sequences

Interestingly, retroviruses such as HIV which use an RNA genome have a "cap" and "tail," enabling them to mimic harmless messenger RNA.

Post-transcriptional control

Degradation of mRNA. Certain hormones can stimulate (or retard) the rate of degradation of mRNA, thereby decreasing (or increasing) its availability fortranslation to protein. Translational repression.Translation of mRNA can be repressed. For example, when iron is low, in human blood, a translational repressor protein binds to the mRNA encoding the iron carrier protein ferritin, and prevents translation of the iron carrier.

Post-translational control

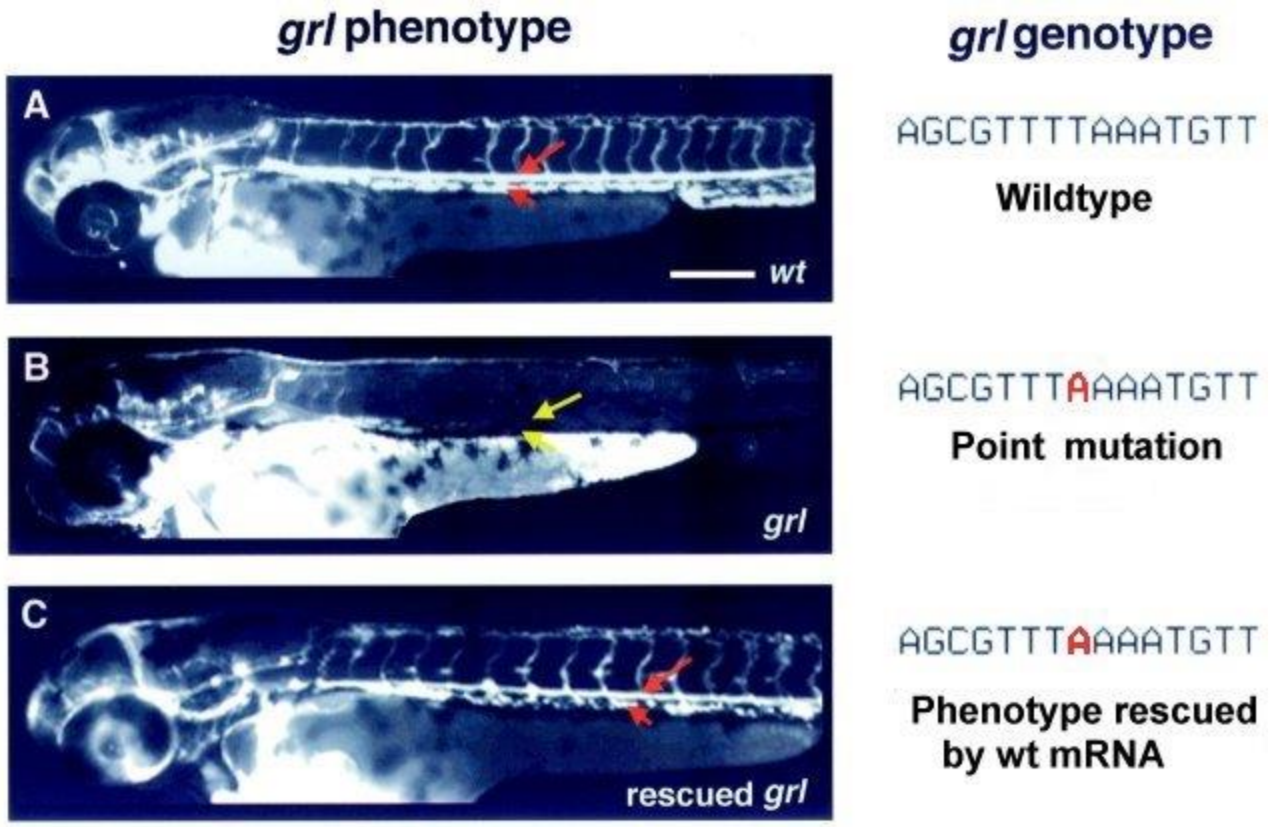
Protein cleavage and/or splicing. The initial polypeptide can be cut into different functional pieces, with different patterns of cleavage occurring in different tissues. In some cases, different pieces may be spliced together.
Chemical modification. Protein function can be modified by addition of methyl, phosphoryl, or glycosyl groups.
Signal sequences direct packaging and secretion. Some proteins have "signal sequences" which direct their packaging in the Golgi and movement through the endoplasmic reticulum (ER) to be secreted. The signal sequences usually end up cleaved off.

Zebrafish is a major model system for vertebrate development.
The "gridlock" gene *grl* was discovered as a major developmental signal distinguishing between arteries and veins in the early vertebrate embryo.

Genotype and phenotype of *grl.*

- **Chemical mutagenesis of a large population of zebrafish yielded some deformed embryos.**
- **One deformed embryo lacked circulation to the back and tail, due to a blocked arterial junction--"gridlock".**
- **By positional cloning (see below) the mutation was mapped and sequenced to a gene named *grl.***
- **A point substitution resulted in partial loss of function of the gene product.**

When *grl* mRNA was injected, the mutant embryo developed normal arterial circulation!

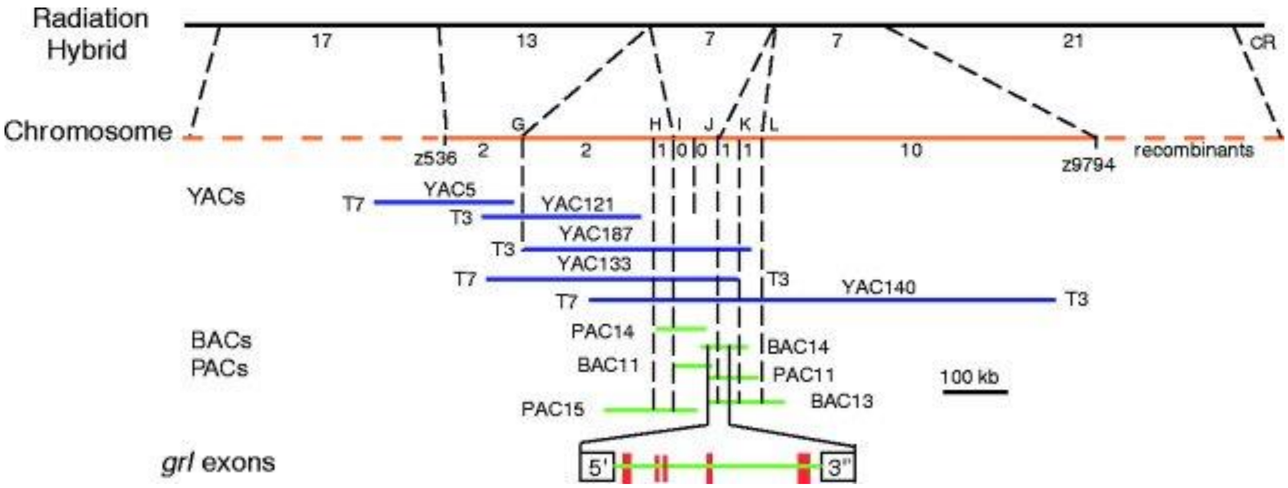


Zhong et al, 2000, Science 287:1820

How was *grl* mapped, located, and sequenced?
It is no small task to find one gene in a vertebrate genome of perhaps 50,000 genes, buried within 20X as much non-coding DNA. Friday afternoon's Advanced Topic session will present the details.

- **Classical recombinant mapping (meiotic crossover analysis) between hybrid *grl* carriers and fish with various genetic markers. These markers are sequence polymorphisms detectable by SSLP PCR.**
- **The position of the mutation was narrowed down to ever smaller chunks of DNA by radiation hybrid cloning, yeast artificial chromosomes (YACs), bacterial artificial chromosomes (BACs), and P1 phage clones (PACs).**

- **Bioinformatic analysis revealed exons and introns. BAC and PAC clones were used to screen embryonic cDNA libraries for genes expressed during early development.**
- **One gene was found which:**
    - **Was expressed ONLY in the embryonic aorta**
    - **Contains a point substitution of lysine instead of a stop codon--thus the protein extends 44 extra amino acids**



Zhong et al, 2000, Science 287:1820

What is the function of grl?
Bioinformatic analysis of the grl sequence identifies it as a transcription factor of the Helix-Loop-Helix family. This protein motif (short conserved protein sequence) is found in many *Drosophila* homeotic developmental genes (more later.) To find out about protein families and domains, see Procite. Click for Chimeview--Helix-Loop-Helix Model

The terminal ends of the protein form a clamp that binds DNA. Major or minor groove? Check the Chime model!

The protein encoded by grl appears to be a transcriptional repressor that distinguishes certain populations of aortic angioblasts (precursor cells of arterial structures).

CHAPTER # 22 INTRODUCTION TO PLASMIDS

Plasmids are commonly used to multiply (make many copies of) or express particular genes.

Types of Plasmids

Plasmids used in genetic engineering are called vectors . Plasmids serve as important tools in genetics and biotechnology labs, where they are commonly used to multiply (make many copies of) or express particular genes. Many plasmids are commercially available for such uses. The gene to be replicated is inserted into copies of a plasmid containing genes that make cells resistant to particular antibiotics. The gene is also inserted into a multiple cloning site (MCS, or polylinker), which is a short region containing several commonly used restriction sites allowing the easy insertion of DNA fragments.

A Plasmid Map of pUC19

pUC19 is one of a series of plasmid cloning vectors created by Messing and co-workers in the University of California. The p in its name stands for plasmid and UC represents the University in which it was created. It is a circular double stranded DNA and has 2686 base pairs. pUC19 is one of the most widely used vector molecules as the recombinants, or the cells into which foreign DNA has been introduced, can be easily distinguished from the non-recombinants based on color differences of colonies on growth media. pUC18 is similar to pUC19, but the multiple cloning site region is reversed.

Next, the plasmids are inserted into bacteria by a process called transformation. Then, the bacteria are exposed to the particular antibiotics. Only bacteria that take up copies of the plasmid survive, since the plasmid makes them resistant. In particular, the protecting genes are expressed (used to make a protein) and the expressed protein breaks down the antibiotics. In this way, the antibiotics act as a filter, selecting only the modified bacteria. Finally, these bacteria can be grown in large amounts, harvested, and lysed (often using the alkaline lysis method) to isolate the plasmid of interest.

Another major use of plasmids is to make large amounts of proteins. In this case, researchers grow bacteria containing a plasmid harboring the gene of interest. Just as the bacterium produces proteins to confer its antibiotic resistance, it can also be induced to produce large amounts of proteins from the inserted gene. This is a cheap and easy way of mass-producing a gene or the protein it then codes for; for example, insulin or even antibiotics.

One way of grouping plasmids is by their ability to transfer to other bacteria. Conjugative plasmids contain tra genes, which perform the complex process of conjugation, the transfer of plasmids to another bacterium. Non-conjugative plasmids are incapable of initiating conjugation, hence they can be transferred only with the assistance of conjugative plasmids. An intermediate class of plasmids are mobilizable, and carry only a subset of the genes required for transfer. They can parasitize a conjugative plasmid, transferring at high frequency only in its presence. Plasmids are now being used to manipulate DNA, and may possibly be a tool for curing many diseases.

It is possible for plasmids of different types to coexist in a single cell. Several different plasmids have been found in E. coli. However, related plasmids are often incompatible, in the sense that only one of them survives in the cell line, due to the regulation of vital plasmid functions. Thus, plasmids can be assigned into incompatibility groups.

Another way to classify plasmids is by function. There are five main classes:

- Fertility F-plasmids, which contain tra genes. They are capable of conjugation and result in the expression of sex pilli.
- Resistance plasmids, which contain genes that provide resistance against antibiotics or poisons. They were historically known as R-factors, before the nature of plasmids was understood.
- Col plasmids, which contain genes that code for bacteriocins, proteins that can kill other bacteria.
- Degradative plasmids, which enable the digestion of unusual substances, e.g. toluene and salicylic acid.
- Virulence plasmids, which turn the bacterium into a pathogen.

CHAPTER # 23 INTRODUCTION TO VECTORS

Vector (molecular biology)

In molecular cloning, a **vector** is a DNA molecule used as a vehicle to artificially carry foreign genetic material into another cell, where it can be replicated and/or expressed. A vector containing foreign DNA is termed recombinant DNA. The four major types of vectors are plasmids, viral vectors, cosmids, and artificial chromosomes. Of these, the most commonly used vectors are plasmids. Common to all engineered vectors are an origin of replication, a multicloning site, and a selectable marker.

The vector itself is generally a DNA sequence that consists of an insert (transgene) and a larger sequence that serves as the "backbone" of the vector. The purpose of a vector which transfers genetic information to another cell is typically to isolate, multiply, or express the insert in the target cell. Vectors called expression vectors (expression constructs) specifically are for the expression of the transgene in the target cell, and generally have a promoter sequence that drives expression of the transgene. Simpler vectors called transcription vectors are only capable of being transcribed but not translated: they can be replicated in a target cell but not expressed, unlike expression vectors. Transcription vectors are used to amplify their insert.

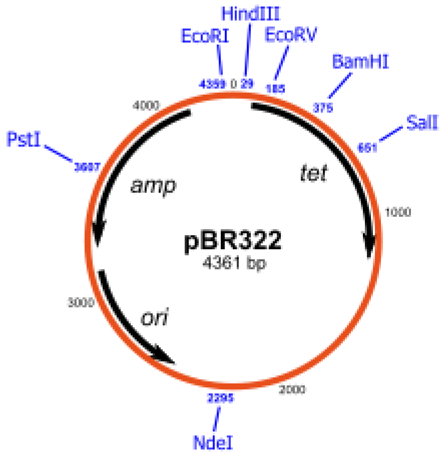
Insertion of a vector into the target cell is usually called transformation for bacterial cells, transfection for eukaryotic cells, although insertion of a viral vector is often called transduction.

Characteristics

Plasmids

Plasmids are double-stranded and generally circular DNA sequences that are capable of automatically replicating in a host cell. Plasmid vectors minimalistically consist of an origin of replication that allows for semi-independent replication of the plasmid in the host. Plasmids are found widely in many bacteria, for example in *Escherichia coli*, but may also be found in a few eukaryotes, for example in yeast such as *Saccharomyces cerevisiae*.[1] Bacterial plasmids may be conjugative/transmissible and non-conjugative:

- conjugative: mediate DNA transfer through conjugation and therefore spread rapidly among the bacterial cells of a population; e.g., F plasmid, many R and some col plasmids.
- nonconjugative- do not mediate DNA through conjugation, e.g., many R and col plasmids.



The pBR322 plasmid is one of the first plasmids widely used as a cloning vector.

Plasmids with specially-constructed features are commonly used in laboratory for cloning purposes. These plasmid are generally non-conjugative but may have many more features, notably a "multiple cloning site" where multiple restriction enzyme cleavage sites allow for the insertion of a transgene insert. The bacteria containing the plasmids can generate millions of copies of the vector within the bacteria in hours, and the amplified vectors can be extracted from the bacteria for further manipulation. Plasmids may be used specifically as transcription vectors and such plasmids may lack crucial sequences for protein expression. Plasmids used for protein expression, called expression vectors, would include elements for translation of protein, such as a ribosome binding site, start and stop codons.

Viral vectors

Viral vectors are generally genetically engineered viruses carrying modified viral DNA or RNA that has been rendered noninfectious, but still contain viral promoters and also the transgene, thus allowing for translation of the transgene through a viral promoter. However, because viral vectors frequently are lacking infectious sequences, they require helper viruses or packaging lines for large-scale transfection. Viral vectors are often designed for permanent incorporation of the insert into the host genome, and thus leave distinct genetic markers in the host genome after incorporating the transgene. For example, retroviruses leave a characteristic retroviral integration pattern after insertion that is detectable and indicates that the viral vector has incorporated into the host genome.

Transcription

Transcription is a necessary component in all vectors: the premise of a vector is to multiply the insert (although expression vectors later also drive the translation of the multiplied insert). Thus, even stable expression is determined by stable transcription, which generally depends on promoters in the vector. However, expression vectors have a variety of expression patterns: constitutive (consistent expression) or inducible (expression only under certain conditions or chemicals). This expression is based on different promoter activities, not post-transcriptional activities. Thus, these two different types of expression vectors depend on different types of promoters.

Viral promoters are often used for constitutive expression in plasmids and in viral vectors because they normally force constant transcription in many cell lines and types reliably.

Inducible expression depends on promoters that respond to the induction conditions: for example, the murine mammary tumor virus promoter only initiates transcription after dexamethasone application and the *Drosophilia* heat shock promoter only initiates after high temperatures.

Expression

Expression vectors produce proteins through the transcription of the vector's insert followed by translation of the mRNA produced, they therefore require more components than the simpler transcription-only vectors. Expression in different host organism would require different elements, although they share similar requirements, for example a promoter for initiation of transcription, a ribosomal binding site for translation initiation, and termination signals.

Prokaryotes expression vector

- Promoter - commonly used inducible promoters are promoters derived from *lac* operon and the T7 promoter. Other strong promoters used include Trp promoter and Tac Promoter, which a hybrid of both the Trp and Lac Operon promoters.
- Ribosome binding site (RBS) Follows the promoter, and promotes efficient translation of the protein of interest.
- Translation initiation site - Shine-Dalgarno sequence enclosed in the RBS, 8 base-pairs upstream of the AUG start codon.

Eukaryotes expression vector

Eukaryote expression vectors require sequences that encode for:

- Polyadenylation tail: Creates a polyadenylation tail at the end of the transcribed pre-mRNA that protects the mRNA from exonucleases and ensures transcriptional and translational termination: stabilizes mRNA production.
- Minimal UTR length: UTRs contain specific characteristics that may impede transcription or translation, and thus the shortest UTRs or none at all are encoded for in optimal expression vectors.
- Kozak sequence: Vectors should encode for a Kozak sequence in the mRNA, which assembles the ribosome for translation of the mRNA.

Features

Modern artificially-constructed vectors contain essential components as well as other additional features:

- Origin of replication: Necessary for the replication and maintenance of the vector in the host cell.
- Promoter: Promoters are used to drive the transcription of the vector's transgene as well as the other genes in the vector such as the antibiotic resistance gene. Some cloning vectors need not have a promoter for the cloned insert but it is an essential component of expression vectors so that the cloned product may be expressed.

- Cloning site: This may be a multiple cloning site or other features that allow for the insertion of foreign DNA into the vector through ligation.
- Genetic markers: Genetic markers for viral vectors allow for confirmation that the vector has integrated with the host genomic DNA.
- Antibiotic resistance: Vectors with antibiotic-resistance open reading frames allow for survival of cells that have taken up the vector in growth media containing antibiotics through antibiotic selection.
- Epitope: Vector contains a sequence for a specific epitope that is incorporated into the expressed protein. Allows for antibody identification of cells expressing the target protein.
- Reporter genes: Some vectors may contain a reporter gene that allows for identification of plasmid that contains inserted DNA sequence. An example is *lacZ-α* which codes for the N-terminus fragment of β-galactosidase, an enzyme that digests galactose. A multiple cloning site is located within *lacZ-α*, and an insert successfully ligated into the vector will disrupt the gene sequence, resulting in an inactive β-galactosidase. Cells containing vector with an insert may be identified using blue/white selection by growing cells in media containing an analogue of galactose (X-gal). Cells expressing β-galactosidase (therefore doesn't contain an insert) appear as blue colonies. White colonies would be selected as those that may contain an insert. Other commonly used reporters include green fluorescent protein and luciferase.
- Targeting sequence: Expression vectors may include encoding for a targeting sequence in the finished protein that directs the expressed protein to a specific organelle in the cell or specific location such as the periplasmic space of bacteria.
- Protein purification tags: Some expression vectors include proteins or peptide sequences that allows for easier purification of the expressed protein. Examples include polyhistidine-tag, glutathione-S-transferase, and maltose binding protein. Some of these tags may also allow for increased solubility of the target protein. The target protein is fused to the protein tag, but a protease cleavage site positioned in the polypeptide linker region between the protein and the tag allows the tag to be removed later.